Molecular evolution of tropinone-reductase-like and tau GST genes duplicated in tandem in Brassicaceae

Aura Rocio Navarro-Quezada

München 2007

Gedruckt mit Unterstützung des Deutschen Akademischen Austauschdienstes (DAAD)

Molecular evolution of tropinone-reductase-like and tau GST genes duplicated in tandem in Brassicaceae

Aura Rocio Navarro-Quezada

Dissertation
an der Fakultät Biologie II
der Ludwig–Maximilians–Universität
München

vorgelegt von
Aura Rocio Navarro-Quezada
aus Mexiko Stadt

München, den 17 Juli 2007

Erstgutachter: Prof. John Parsch

Zweitgutachter: Prof. Wolfgang Stephan

Tag der mündlichen Prüfung: 24 September 2007

Erklärung

Diese Dissertation wurde im Sinne von §13 Abs. 3 bzw. 4 der Promotionsordnung von Prof. Dr. John Parsch und Dr. Karl Schmid betreut.

Ehrenwörtliche Versicherung
Diese Dissertation wurde selbstständig, ohne unerlaubte Hilfe erarbeitet.

Hiermit erkläre ich, dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe.

München, der 17 Juli 2007

Note

In this thesis I present the results from my doctoral research, which I have done between December 2003 and July 2007. Most of the work was done under the supervision of Karl Schmid at the Max Planck Institute for Chemical Ecology in Jena. Chapters 1, 2 and 3 of this thesis are closely related to each other and the result of a collaboration between Karl Schmid and myself. Karl Schmid decided to study fast duplicating plant gene families. I did all of the experimental work, sequencing, assembly and analysis and wrote the manuscript as Chapter 1. Part of the analyses in Chapter 2 are included in Theresia Eichners Masters Thesis, that she submitted in January 2006 in the Bioinformatics Department of the Friedrich-Schiller University in Jena. I worked in collaboration with her for some of the analyses and Karl for others. I did all the experiments for the analyses of expression and wrote the manuscript. For Chapter 3 I did the analyses and wrote them up.

The work in this thesis was supported by a stipendium of the Deutsches Akademisches Austauschdienst (DAAD) for my Ph.D. work. The Max Planck Society and an Human Frontiers in Science Project funded the research part of the project.

Dedicatoria

A mi familia y amigos

’Un día, los hombres descubrirán un alfabeto en los ojos de las calcedonias, en los pardos terciopelos de la falena, y entonces se sabrá con asombro que cada caracol manchado era, desde siempre, un poema’-
Alejo Carpentier en ’Los pasos perdidos’

Contents

List of Figures

List of Tables

Abstract

The presence of multiple genes with high degree of similarity forming gene families is a universal phenomenon in living organisms. Gene duplication is an opportunity for evolving new functions from the newer gene, but also has a disadvantage due to local gene-rearrangement effects and, if duplications are numerous, through alterations of genome size. Therefore, selection is playing a central role in determining the fate of a duplicate gene. Plants are known to harbor numerous gene families, and are thus an ideal system to test the fate of gene duplicates. This thesis tackles the tropinone-reductase like enzymes (further TRL) and the tau GSTs located upstream from this gene family. TRL enzymes are short-chain dehydrogenases that are involved in a reduction step downstream in the synthesis of tropane alkaloids in Solanaceae, important defense compounds of plants. The function of TRLs in Brassicaceae is not clear, since most of the plants in this family do not produce tropane alkaloids, but some have been associated with the oxidative-stress response. This gene family contains 80% of its members duplicated in tandem in *Arabidopsis thaliana*. We profited from this fact to isolate 12 TRL (+ pseudogenes) from this species, further six species of Brassicaceae (*A. thaliana, A. lyrata, A. cebennensis, Capsella rubella, Boechera divaricarpa* and *Brassica rapa*), and one species from a closely related plant family, *Cleome spinosa*. We tested the role that selection plays in maintaining large numbers of this gene family. We used phylogenetic methods to analyze non-coding sequence evolution and identified regulatory motifs. We analyzed non-coding sequence evolution. Microarray expression data from *A. thaliana* and qPCR for *A. thaliana* and *A. lyrata* were analyzed to detect divergence in the expression patterns of orthologs and paralogs. TRL genes follow a gene birth and death dynamics. More probable, they originated from non-equal recombination of tandem duplicated genes. Positive selection at the origin of the duplicated genes allowed these to acquire differential expression patterns, leading to the preservation of numerous TRLs. The analysis of coding and non-coding sequences shows them to display correlated evolution, particularly in species recently separated by speciation. We further tested for selection on the tau glutathione-S-transferases (GST) enzymes, adjacent 3' in the genome to TRLs. Tau GSTs are unique to plants and are involved in detoxification. Multiple copies of these enzymes will allow flexibility in substrate specificity, which is important for the detoxification function. We detected positive selection among paralogs of tau GSTs supporting their potential of functional diversity, but we also detected negative selection among paralogs and groups of orthologs, indicating that more often their functions are conserved.

Introduction

0.1 Genomic approaches in the study of plant evolution

The availability of complete genome information for different organisms is providing the scientific community with large amounts of data that can be used to learn about complex genetic and evolutionary events occurring at large (i.e. genome-wide and duplicated gene block) and also very small scales (i.e. SNPs), and not only at chromosomal level, as was the case in classic genetic experiments.

One of the first organisms to have a genome sequenced was *Arabidopsis thaliana*, commonly known as the thale-cress, which belongs to the plant family Brassicaceae (previously Cruciferae) (1). This plant has a small genome size and a short generation time, which has made it useful as a laboratory model organism. Further sequencing projects are going on for two commercially important plants belonging to the same family as thale cress, *Brassica rapa* and *Brassica oleracea* (2). Other two wild relatives of *A. thaliana*, *Arabidopsis lyrata* and *Capsella rubella*, are being sequenced in a project of the JGI (http://www.jgi.doe.gov/sequencing/why/CSP2006/AlyrataCrubella.html) in an effort to enlarge the understanding of these plants genomes. *Arabidopsis thaliana* has often been used as the reference genome in genomic comparisons with *Arabidopsis* relatives, as for instance *Capsella rubella* (3) and *Brassica oleracea* (4). *A. lyrata* is a wild relative that has previously been compared to *A. thaliana*, because their contrasting mating systems allow evolutionary hypotheses of recombination and colonization to be tested (5). A recent review of the importance of the Brassicaceae in genetic studies has been published by **(author?)** (6).

Genomic studies can benefit from the comparative method to learn how genomes are built according to the biology of particular organisms. Neutral and selective hypotheses can be tested to find explanations for genome composition, as for instance, in a recent study on codon-bias in *A. thaliana* (7) that found a role of selection on both low and highly expressed genes. In this thesis, we used the information available from *A. thaliana* to isolate a genomic region containing a gene family clustered in tandem in seven related Brassicaceae species. For this purpose, we used bacterial artificial chromosomes (BACs) in order to isolate all TRL family members in the syntenic region. By comparing members of a gene family in closely related species, we could distinguish between the hypothesis of conversion vs. gene birth and death and infer the evolutionary history of the gene duplicates. In Chapter 1, from the sequenced BACs we compare 12 TRL genes, equivalent to 80% of those present from *A. thaliana*. In Chapter 3 we analyze of 7 GSTs, equivalent to 14% of GSTs, but to 25% of the tau GSTs present in *A. thaliana*.

0.2 Why study gene duplication?

Gene duplication has genetic and phenotypic important consequences and therefore might also have evolutionary consequences in a relatively short time scale. It is said to belong to the fast modes of evolution, since on the one hand it can rearrange and change loci locations quickly and on the other hand it provides raw material for fixing mutations to create novel genes. Researchers have been trying to estimate overall rates of gene duplication and loss. The rates and biological roles of duplication might vary in different organisms, depending on population structure and mating system (8). Rates of duplication in plants have been estimated as 0.0029/locus per MY in wheat (9) and 0.001-0.03 genes per MY for *Arabidopsis thaliana* (10).

By comparing duplicate genes across species, we can elucidate the different mechanisms responsible for keeping gene duplicates. In the evolution of gene families the following outcomes are possible: divergent evolution, concerted evolution and gene birth and death (for a recent review see (11)). The first scenario is what we would expect if duplicated genes evolve at different rates in each species, and will not be distinguished as duplicates after speciation. The second scenario is a product of the gene conversion mechanism. Gene conversion is due to recombination through non-homologous pairing occurs among duplicate gene loci, which homogenizes their gene sequences, so that these genes appear younger than the separation of the species. This last mechanism is more common in genes that interact in their function as, for instance, pathogen- and immune defense-related genes such as LRR (12) and MHCs (13). Gene birth and death is the most common fate of duplicate genes (14). This type of evolution predicts that genes duplicate in an ancestral species and diversify soon, in order for the gene duplicate not to be eliminated from the genome. Diversification takes place by the processes known as neo-functionalization (acquisition of a new function) or sub-functionalization (splitting of function among paralogs) (15). After speciation, when reconstructing the phylogeny from genes that diversified in the ancestors, the phylogeny will reflect the history of the gene, and not that of the species, unless the genes acquire species- specific functions. Some of the duplicates might lose their function, in a process known as non-functionalization, because of different selection pressures on the populations and species. The latter genes will be found as pseudogenes in the present time and eventually be eliminated from the genome (14).

Until now, most studies comparing regulatory and functional evolution in gene families are based on one organism, or evolutionary distant organisms, as human and mouse (16), making it difficult to compare evolution of protein and promoters. One exception are studies in *Drosophila* where multiple closely related species have been compared. (17). Our study addresses seven closely related species being separated by less than 20 MY.

0.3 Duplication of genes of secondary metabolism

In general, we expect genes in secondary metabolism to be flexible and if present in multiple copies, to diversify in function through sub- or neo-functionalization (18). Gene duplication has been proved to be at the base of evolutionary novelty in defense-related genes such as MAM (19), LRR (12), and for genes involved in flower pigmentation like chalcone- synthases (CHS) (20). Tropinone-reductase like genes (TRLs), the gene family on which the first two chapters are centered, appear to respond to environmental stimuli (21). A list of putative functions obtained from the literature and databases can be found in Table 1.1. These enzymes are probably more involved in secondary than in primary metabolism. We expected to find more divergence in TRLs *a priori* than in genes involved in primary metabolism and/or in genes involved in development, since secondary metabolism genes might not be essential in an organism's life. Multiple copies of a gene indicate that they have either diversified in function, by the process known as neo-functionalization, or that partial genes have complemented each other, by sub-functionalization. Another possibility would be redundancy, but this has not yet been demonstrated for genes controlling secondary metabolism. Until now, development and primary metabolism genes have been shown to display gene redundancy (22; 23). It has recently been proposed that genes essential in an organisms' life might be subject to stronger negative selection than 'non-essential' genes, and this might be reflected in lower duplication rates of these last genes (24). Although we cannot discuss whether TRL genes are essential for Brassicaceae, we can infer from reports in the literature that only some of them have been discovered as intervening in signalling pathways (see table 1.1). Furthermore, microarray and massive parallel signature sequence data (MPSS) from *Arabidopsis thaliana* show that not all the TRL genes appear to have a significant differential expression, as we indicate in the section *in silico* expression in this introduction. Expression of TRLs is discussed briefly in Chapter 1 (Figure 3), and described more extensively in Chapter 2.

0.4 The biological role of tropinone reductases in Solanaceae

Alkaloids are important plant defense compounds, as they are toxic for many herbivores, including humans. Therefore, these compounds have received special attention. Many have economic importance for human societies, as for instance, nicotine. Tropinone-reductases are involved in a parallel pathway to nicotine production, acting upstream from *N*-methyl-putrescine, a precursor of both nicotine and tropane alkaloids. Tropinone reductases are at the branch point of tropane alkaloid synthesis, preceding the synthesis of hyosciamine, scopolamine and calystegines in Solanaceous plants, but also of cocaine in the genus *Erythroxylum*. The pathway of synthesis is shown in Figure 1.

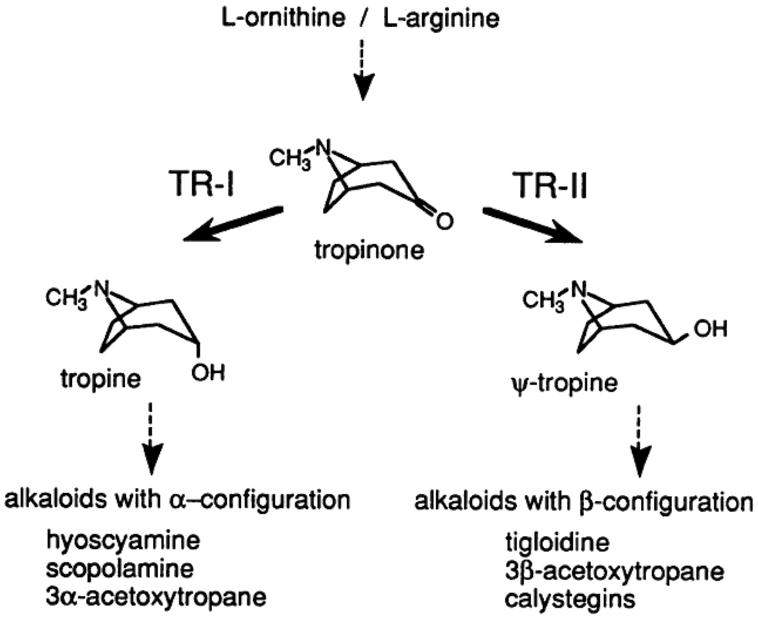


Figure 0: **Formation of tropane alkaloids catalyzed by TRs.** Modified from (25)

In Solanaceae, where tropinone reductase enzymes were originally described, the different enzymes are responsible for sterification of tropinone in different stereospecificities. TR-I produces tropine and TR-II produces pseudo-tropine, as shown in Figure 0. Tropine is the precursor of hyosciamine and scopolamine and pseudo-tropine is the precursor of calystegines. Two research groups have been studying both tropinone reductases, and therefore I refer to their work for further information on cristallization and biochemistry of the protein (26; 27; 25). One of this research groups recently found that some Brassicaceae do produce calystegines (28). A simplified representation of tropane alkaloids detected in sister species of some of the species studied in this thesis is shown in Figure 2.

Bioinformatic studies have proved that the sequence of *Arabidopsis* has multiple copies of genes it does not use in the predicted way from sequence similarity. A study by Allen (30) found that around 10% of the genes from other plants appear to have been lost in this model plant. Therefore the functions that these genes perform must either not be present or must be substituted by other proteins. In this same study it is mentioned that one of the genes missing is an upstream enzyme necessary for converting ornithine to putrescine, ornithine decarboxylase (ODC) (see Figure 1). The other enzyme that also precedes the conversion of arginine to putrescine, arginine decarboxylase (ADC), is present in *Arabidopsis*. Nevertheless, the abscence of the other pathway might limit the availability of putrescine, which might not be diverted into tropinone production, especially since putrescine is a buffer for abiotic stress used in other pathways.

0.5 Tropinone reductases in Brassicaceae

The study that found calystegines in Brassicaceae, did not detect these compounds in *Arabidopsis thaliana*. Furthermore, more than 50% of the Brassicaceae plants they tested, 24 out of 43, did not contain calystegines (28). This is surprising, since genes for tropane alkaloids are found in the complete or partial genomes already sequenced from *Arabidopsis thaliana* and *Brassica* species. In *Arabidopsis* multiple genes were found highly similar to TRLs. The proteins encoded by this genes must have an alternative function for this plant, since it does not contain any type of complex alkaloids. The question then rises, why does *Arabidopsis thaliana*

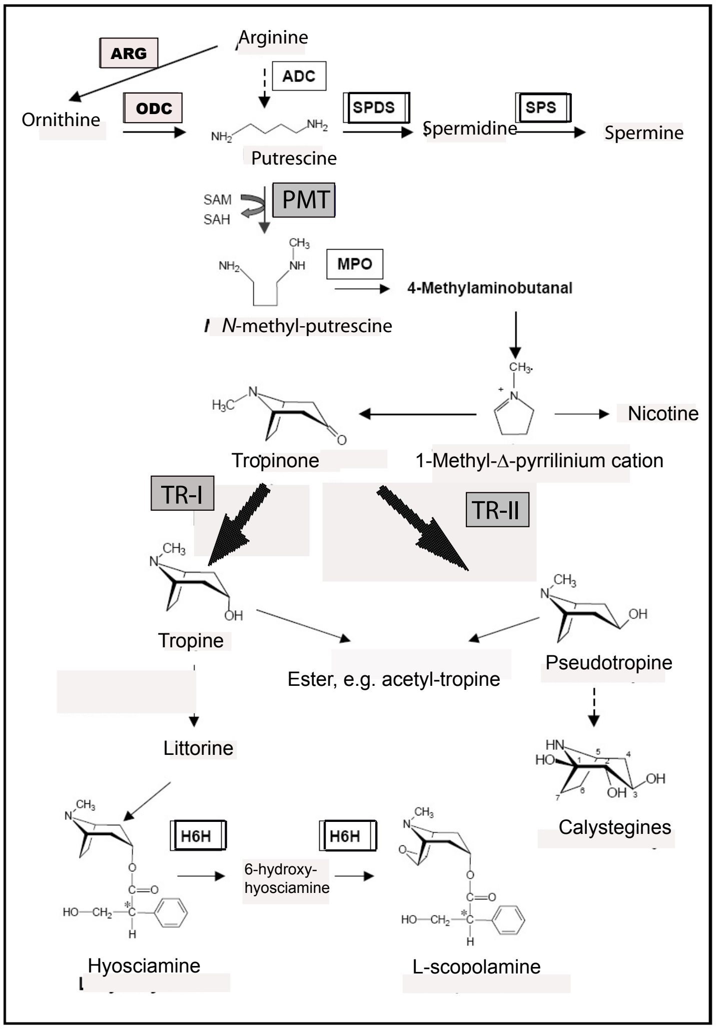


Figure 1: Formation of tropane and nicotine alkaloids in Solanaceae. Tropane alkaloid synthesis is catalyzed by TRs. Tropinone-reductase I and II are in grey squares respectively. Modified from (29)

keep multiple copies of TRL genes? (see (31)).

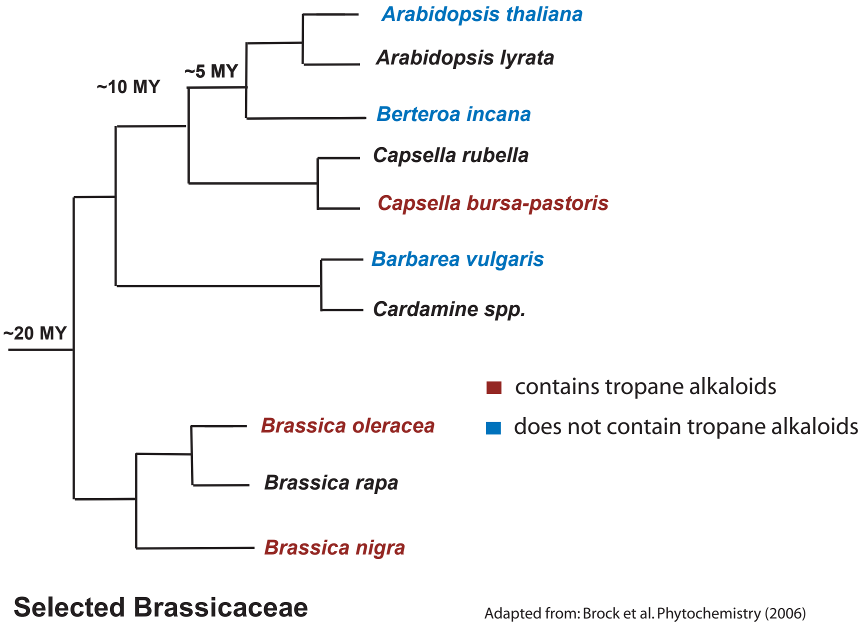


Figure 2: Presence of tropane alkaloids in *Arabidopsis* and related Brassicaceae. Tree was modified from (Brock et al. 28). Names in blue indicate absence, names in red indicate presence of tropane alkaloids and names in black indicate sister species not yet tested for these compounds (28).

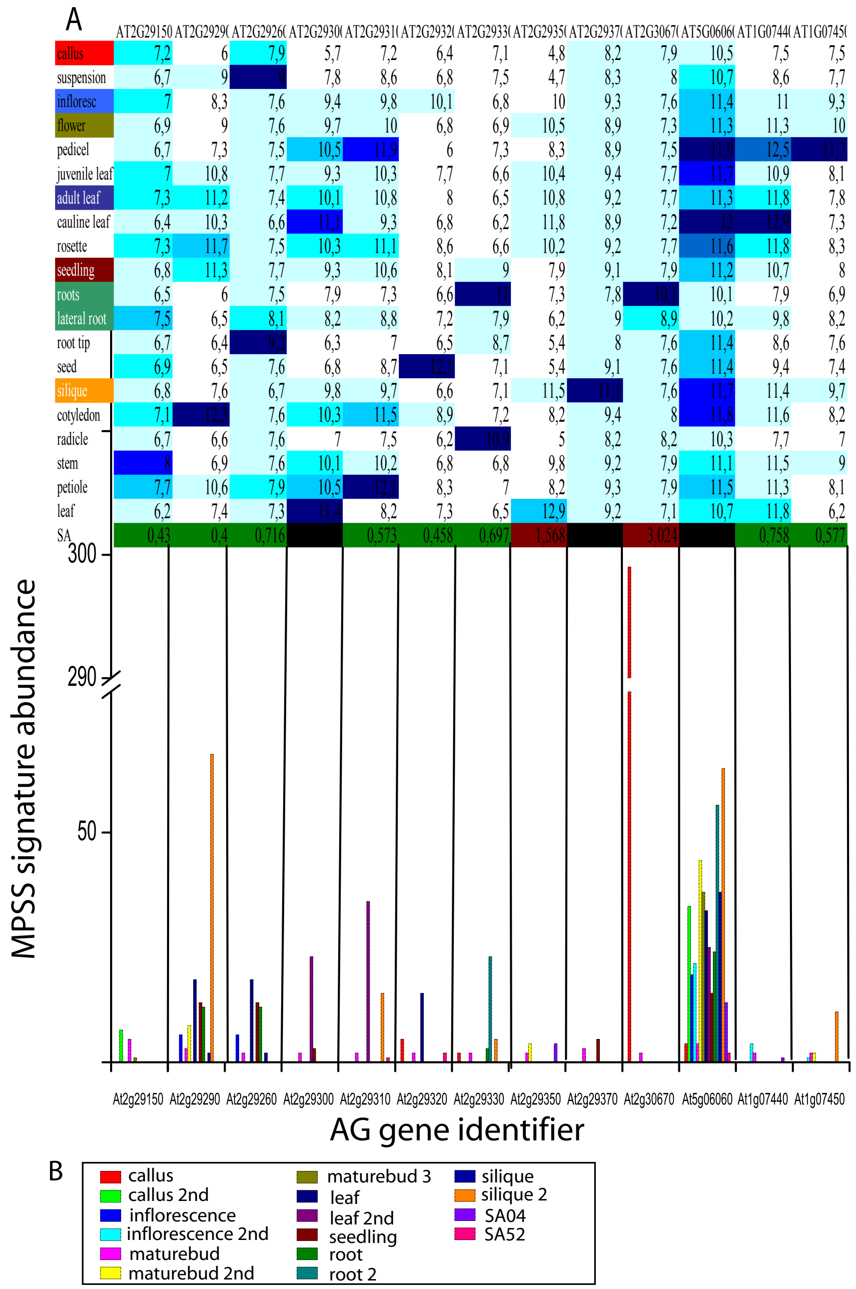
The function of one TRL enzyme from *Arabidopsis* has been reported in the literature. Physiological studies on senescence have shown that one of the enzymes encoded by the TRL genes, SAG13 (AGI: at2g29350), is associated with other enzymes in the senescence response (32; 33; 34). The other TRL genes have been found associated with other functions, most of them respond to hormone signals, as shown in Table 1.1. Differential expression patterns have been previously observed in response to stress (35).

0.6 Comparison of TRL expression data from available web resources

Publicly available databases for expression data from *A. thaliana* (36; 37) can give insight into the TRL gene function. A comparison between microarray and MPSS data at different developmental stages are shown in Figure 3. It appears that genes that are found in the 'ancient' clade from *A. thaliana* (clade A, Figure 1.8), At5g06060 and At2g29260, are over-expressed throughout development, although at different magnitudes in different tissues. At5g06060 especially, appears to be expressed in all organs and at all developmental stages. Specialization of function appears to occur after the first duplication, which is confirmed by the differential expression observed in the other TRLs, including those on chromosome 1. The gene At1g07440 showed slight differential expression in microarray data, but when expression is analyzed with MPSS, the gene is barely detected in inflorescence and due to salicylic acid induction (INF and S04 in Figure 3). This might indicate that this gene has lost its function. A more extensive analysis of possible factors inducing or repressing TRL expression can be found in Chapter 2.

0.7 Glutathione-S-transferases in plants

Glutathione-S-transferases are considered as one of the large and variable secondary defense enzymes. Despite their variability, they all have in common that they recognize and transport reactive electrophilic compounds, which are toxic to the cell (38). Typically, GSTs share only between 25-35% sequence similarity, although within a subfamily the similarity is usually more than 40% (39). One of the main functions is to catalyze the conjugation of a glutathione molecule to a variety of chemical compounds. Glutathione is highly polar and renders these compounds more soluble for excretion from the cell through the apoplast or vacuole through glutathione pumps. GSTs can detoxify breakdown products of lipid peroxidation or oxidative DNA degradation, and they can also function as peroxi-



	AT2G29150	AT2G2929(	AT2G2926(	AT2G2930(	AT2G2931(	AT2G2932(	AT2G2933(	AT2G2935(	AT2G2937(	AT2G3067(	AT5G0606(	AT1G0744(	AT1G0745(
callus	7,2	6	7,9	5,7	7,2	6,4	7,1	4,8	8,2	7,9	10,5	7,5	7,5
suspension	6,7	9	9	7,8	8,6	6,8	7,5	4,7	8,3	8	10,7	8,6	7,7
infloresc	7	8,3	7,6	9,4	9,8	10,1	6,8	10	9,3	7,6	11,4	11	9,3
flower	6,9	9	7,6	9,7	10	6,8	6,9	10,5	8,9	7,3	11,3	11,3	10
pedicel	6,7	7,3	7,5	10,5	11,9	6	7,3	8,3	8,9	7,5	11,9	12,5	11,7
juvenile leaf	7	10,8	7,7	9,3	10,3	7,7	6,6	10,4	9,4	7,7	11,7	10,9	8,1
adult leaf	7,3	11,2	7,4	10,1	10,8	8	6,5	10,8	9,2	7,7	11,3	11,8	7,8
cauline leaf	6,4	10,3	6,6	11,1	9,3	6,8	6,2	11,8	8,9	7,2	12	12,9	7,3
rosette	7,3	11,7	7,5	10,3	11,1	8,6	6,6	10,2	9,2	7,7	11,6	11,8	8,3
seedling	6,8	11,3	7,7	9,3	10,6	8,1	9	7,9	9,1	7,9	11,2	10,7	8
roots	6,5	6	7,5	7,9	7,3	6,6	11	7,3	7,8	10,1	10,1	7,9	6,9
lateral root	7,5	6,5	8,1	8,2	8,8	7,2	7,9	6,2	9	8,9	10,2	9,8	8,2
root tip	6,7	6,4	9,2	6,3	7	6,5	8,7	5,4	8	7,6	11,4	8,6	7,6
seed	6,9	6,5	7,6	6,8	8,7	12,1	7,1	5,4	9,1	7,6	11,4	9,4	7,4
silique	6,8	7,6	6,7	9,8	9,7	6,6	7,1	11,5	11,1	7,6	11,7	11,4	9,7
cotyledon	7,1	12,3	7,6	10,3	11,5	8,9	7,2	8,2	9,4	8	11,8	11,6	8,2
radicle	6,7	6,6	7,6	7	7,5	6,2	10,9	5	8,2	8,2	10,3	7,7	7
stem	8	6,9	7,6	10,1	10,2	6,8	6,8	9,8	9,2	7,9	11,1	11,5	9
petiole	7,7	10,6	7,9	10,5	12,1	8,3	7	8,2	9,3	7,9	11,5	11,3	8,1
leaf	6,2	7,4	7,3	11,4	8,2	7,3	6,5	12,9	9,2	7,1	10,7	11,8	6,2
SA	0,43	0,4	0,716		0,573	0,458	0,697	1,568		3,024		0,758	0,577

Figure 3: Expression of TRL genes along different plant development stages according to A) microarrays and B) MPSS. Values for experiments represented in both databases were kept, which are also those for which expression patterns differentiate between the tissues. Lines unite same genes measured with the two methods. A) adapted from Genevestigator Meta-analyzer tool, values are log_2 expression. Colored squares represent expression intensities going from 0 % (white square) to 100 % (dark blue square). B) MPSS signatures 17bp and 20bp, signatures were pooled or averaged. Specific signatures are found only for 13 out of 16 *A. thaliana* TRL genes, which are shown.

dases and scavenge radicals. GST conjugations also participate in the synthesis of secondary metabolites, and are induced by reactive oxygen species (ROS), ozone, wounding, ethylene, heavy metals, and pathogen attack. Tau GSTs are our subject of study in Chapter 3.

The best studied GST subfamilies from plants are the phi and tau GSTs, which seem to be involved in responses to different environmental stresses including drought, cold, heat, chemical compounds such as H2O2, SA, DTT, CuSO4, and also in herbicide metabolism (40).

0.8 Detecting selection using the Phylogenetic Analysis of Maximum Likelihood (PAML) Method

In Chapters one and three, I searched for positive selection using phylogenetic analysis of maximum likelihood (PAML). PAML estimates different parameters of sequences in a phylogeny, and performs a simulation using a maximum-likelihood algorithm, to obtain the 'most' probable evolutionary path. Parameters estimated by PAML are: κ (the transition over transversion ratio), ω (*dN/dS* ratio) and lnL (the natural logarithm of the likelihood), the last parameter is necessary to evaluate the simulation. This program allows different hypotheses of evolutionary rates to be tested (different values of ω) and positive selection to be identified. In order to detect different rates of evolution among orthologs, we first performed a pairwise comparison of *dN/dS* among orthologs from each of the orthologous clades (runmode=-2) and compared these rates. We performed the so-called site tests explained in Table 0. For exploratory reasons we performed M0, M1, M2, M3 and M7 vs M8. These tests calculate a prior expectation of rates of change estimated for different number of sites. Table 0 displays the parameters estimated and the number of sites each model allows. We performed M8 vs M8a (41) and we fixed $\omega < 1$ as a further comparison. We used the branch-site models MA vs. MB to detect selection in particular parts of the tree (42). These tests allow us to contrast different rates of evolution in different parts of the phylogeny of TRL genes. We were interested in those branches at the base of duplication events, splitting 'novel set of genes', since positive selection might be detected in this part of the tree. MA and MB have been used previously to detect differential evolution of gene families (43); MA estimated two ratios, one for labeled branches (foreground), and another for the unlabeled branches (background) whereas MB estimated three ratios, one for foreground and two for background branches (see Table 1.7). In Chapter 1 we tested selection using clades B and C separately, and constructed a 'short' tree with one sequence representing each clade (see Figure 1.8). The 'short' tree we call a paralog tree, although it contains ancient duplicates that have diverged in different species. Log-likelihoods obtained for each model tested can be compared using a likelihood ratio test (LRT, in (44)). We tested for positive selection in the internal branches of the tree, after the initial separation from the proteins in the 'ancestral' clade (A), using the branch-site models. For Chapter 3 we used the complete tree to test for positive selection, as the number of sequences was not as large.

Table 0: Site and branch-site models of variable ω ratios among sites from PAML used in this thesis.

Model code	p	Parameters	Notes
M0 (one ratio)	1	ω	One ω ratio for all sites
M1 (neutral)	1	p_0	$p_1 = 1 - p_0$, $\omega_0 = 0$, $\omega_1 = 1$
M2 (selection)	3	p_0, p_1, ω_2	$p_2 = 1 - p_0 - p_1$, $\omega_0 = 0$, $\omega_1 = 1$
M3 (discrete)	2K-1	p_0, p_1,...,p_{K-2}	$p_{K-1} = 1 - p_0 - p_1 - ...p_{K-2}$
M7 (beta)	2	p, q	From $B(p,q)$
M8 (beta & ω)	4	p_0, p, q, ω	p_0 from $B\ p,q\ and\ 1 - p_0\ with\ \omega$
MA	3	p_0, p_1, ω_2	$\omega_0, \omega_1 \leq 1$ are fixed
MB	5	p_0, p_1, ω_0, ω_1, ω_2	

Adapted from (45), p=proportion of sites.

List of Abbreviations

TRL Tropinone-reductase-like

GST Glutathione-S-transferases

dN Number of non-synonymous substitutions, i.e. those causing an aminoacid replacement

dS Number of synonymous substitutions, i.e. silent mutations not causing aminoacid replacement

PAML Phylogenetic analysis of maximum likelihood

MA M stand for Model, A for the designation of the model

BLAST Best local alignment search tool

NJ Neighbor joining

ω equals to d_N/d_S, the nonsynonymous over synonymous substitution ratio

ℓ likelihood

p probability

σ standard deviation

BAC Bacterial artificial chromosome

bp Base pair

PCC Pearson correlation coefficient

r Pearson correlation

CCOs Conserved clusters of orthologs

PCR Polymerase Chain Reaction

qPCR Quantitative polymerase chain reaction

qRT-PCR Quantitative RT real time polymerase chain reaction

RT Reverse transcriptase

SAG Senescence associated gene

LRR Leucine rich repeats

Chapter 1

Different fates in a (gene) family: birth and death evolution in tropinone-reductase-like genes from Brassicaceae.

A. Navarro-Quezada, S. Gebauer-Jung, K.J. Schmid

17th October 2007

1.1 Abstract Chapter 1

The role of selection in the maintenance of multiple tandemly duplicated gene families is not yet clear. To gain insight into this question, we studied the tropinone-reductase-like (TRL) gene family (short-chain dehydrogenases) duplicated in tandem on chromosome 2 from *Arabidopsis thaliana*. We sequenced homologous TRL genes from BAC and cosmid libraries from *Arabidopsis lyrata*, *Arabidopsis cebennensis*, *Capsella rubella*, *Boechera divaricarpa* and from *Cleome spinosa* (Cleomaceae). The data were complemented with previously published gene data from *Brassica rapa*. We found TRLs to be located in a gene-duplicate-rich region where most of the genes are associated with response to stress. TRLs have undergone gene birth and death within the Brassicaceae, increasing from 4 TRL gene copies in *B. rapa* to 16 copies in *C. rubella*. Most of the duplications occurred once in the Brassicaceae ancestor (around 10 ±5 MY ago). TRL genes have been transposed and inverted. Tests for selection using phylogenetic analysis of maximum likelihood found evidence for positive selection early in the TRL history. This indicates neo-functionalization occurred at the origin of the major duplications. After the initial duplication, most duplicate loci have been retained under selective constraint. Mapping positively selected sites onto an available three-dimensional TRL structure from *A. thaliana* using maximum likelihood and a Markov Chain Monte Carlo approach shows the few changes that have occurred in TRLs map close to substrate binding sites. One group (out of 9) of TRL orthologs shows evidence for recent positive selection. We found that the closely related species *Cleome spinosa* contains 11 TRL copies that have evolved independently from the Brassicaceae TRLs. This shows that genes in secondary metabolism can undergo parallel duplication with independent subsequent diversification. Tropinone reductase enzymes might have been recruited for different functions, among them tropane alkaloid synthesis, as has been shown for some Brassicaceous plants.

1.2 Introduction Chapter 1

Gene duplications are known to be a major source of evolutionary novelties and phenotypic variation (46). Duplicated genes can originate from whole genome-, segmental- or tandem duplications. Theoretical considerations suggest it is highly unlikely that duplicated genes remain completely redundant in their functions over extended evolutionary periods. More probably they will assume different evolutionary trajectories, characterized as non-functionalization (loss of function), neo-functionalization (acquisition of a new function) or sub-functionalization (splitting of function among paralogs) (15). The last two processes allow gene duplicates to be preserved, as they provide functional gene copies upon which selection can act. Gene duplications can also be considered as 'reserves' facilitating the adaptation of organisms to a changing environment (11). At the same time, duplications may also show deleterious effects on genome structure since they can induce chromosome rearrangements by nonhomologous recombination and other mechanisms (47). Since the duplication of genes is an ongoing process in all species whose genomes have been studied so far, it is important to understand how duplicated genes evolve and why clusters of gene duplicates exist.

Plant genomes contain a high proportion of duplicated genes. Polyploidization is frequent among plant species and an important component of plant genome evolution (48). For example, in *A. thaliana* one recent large-scale duplication occurred 24-40 MY ago, and a second one in the more distant past (49). Additional large-scale duplications may have occurred independently in different lineages of the order Brassicales (50). After large-scale duplications, numerous genes are deleted rapidly (51), suggesting that gene numbers do not increase by large jumps. Furthermore, the proportion of genes that is retained differs significantly among functional classes (52), indicating differential selective pressures and gene-dosage effects. Plants also have a higher proportion of tandemly duplicated genes than do model species from other kingdoms (1; 53), suggesting gene duplication by non-homologous recombination is an important mechanism for generating phenotypic diversity, for example in response to pathogens or herbivores. Differences in expression of gene duplicates has been detected shortly after duplication, particularly for genes duplicated in tandem (54).

Although the patterns of gene duplication are well described, relatively little is known about the interplay of large-scale and tandem duplications, and the evolutionary dynamics of this process. Comparisons between the complete genome sequences of poplar and *A. thaliana* identified variation in gene copy number of numerous gene families between the two evolutionary lineages during the last 80 MY that likely represent adaptations to the different life histories and ecologies of the two species (55). The comparative sequencing of selected tandemly repeated gene families among accessions of *A. thaliana* (56), and in closely related species (57; 58; 19) demonstrated a rapid evolution of gene tandems with respect to sequence divergence and copy number that appears to be partially driven by positive Darwinian selection.

To further understand the molecular and genome evolution of tandemly repeated genes, we decided to characterize one of the largest gene tandems in the *A. thaliana* genome, the tropinone- reductase-like (TRL) gene family, which consists of sixteen complete genes and three pseudogenes (1). Twelve paralogs are located in tandem on chromosome 2 together with three pseudogenes, another single paralog is located further downstream on the same chromosome. Two members are duplicated in tandem on chromosome 1, and a single TRL gene is located on chromosome 5. The enzymes encoded by TRL genes are highly conserved at the amino acid level (64 - 99.9% identity among paralogs) and they are oxidoreductases of the short-chain dehydrogenase protein superfamily to which alcohol dehydrogenase (*Adh*) also belongs. The TRLs from *A. thaliana* are closely related to tropinone reductases from the Solanaceae plant family, which catalyze a reduction step in the synthesis of tropane alkaloids (27). Since *A. thaliana* and other Brassicaceae apparently do not produce tropane-like alkaloids (31) (but see (28)), the presence of TRLs suggests other functions in this group of plants. For example, Lohman (32) showed that one TRL gene, SAG13 (AGI identifier At2g29350), is expressed during leaf senescence and this gene has often been used as a cellular marker for this stage (33). Studies of expressed sequence tags (ESTs, 59), massively parallel signature sequencing (MPSS, 36) and microarray data (35; 37) showed that some TRLs are induced by biotic (pathogen infection) and abiotic (cold, salt stress) stresses, and also plant hormones such as jasmonate and salycilic acid (21). A list of putative functions obtained from the literature and databases can be found in Table 1.1.

Table 1.1: Putative function of tropinone-reductase-like enzymes (TRLs)

TRL (AGI)	Elicitor	Experimental condition	Citation
At2g29350*	senescence	early response	(33; 32)
At2g29350	salicylic acid/pathogen	upregulated by infection	(60)
At2g29350	salicylic acid/methyl jasmonate	downregulated	(21)
At2g29350	salicylic acid	upregulated after 4h	(36)
At2g29340	cold	downregulated after 24h	(61)
At2g29330	salicylic acid	slight upregulation	(21)
At2g29330	brassinosteroid response	upregulation, late stage	(62)
At2g30670	brassinosteroid response	upregulation, late stage	(62)
At2g29320	salicylic Acid	upregulation 4h,52h	(36)
At2g29340	salicylic Acid	upregulation 4h,52h	(36)
At5g06060	ethylene	upregulation	(21)

*SAG13,Senescence Associated Gene

In this Chapter, we present a comparative genomic and molecular evolutionary analysis of the large TRL gene cluster on chromosome 2 in the *A. thaliana* genome with 12 (of a total of 16 genes in the genome) and three pseudogenes. The homologous region was obtained from other closely related plant species and completely sequenced. We conducted a phylogenetic analysis of TRLs across seven species to gain insight into the duplication dynamics and to test hypotheses on the role of natural selection in TRL evolution. We expected to detect positive selection at some point of the duplication history indicative of functional diversification between paralogs, since TRLs likely acquired new functions other than alkaloid synthesis despite the high overall level of sequence conservation.

The phylogeny and genomic position of TRLs in seven closely related species indicates that non-homologous recombination is responsible for the TRL gene duplication. From the phylogeny, we also learned that they are evolving according to gene birth and death. This mode of evolution was also supported by the presence of non-functional TRLs (pseudogenes) in all the analyzed species. From this observation, we predicted different rates of change for protein sequences along the phylogeny, which we tested using PAML. Different rates of evolution do not appear evident comparing whithin orthologous groups, as they show that TRLs are recently subject to purifying selection. We detected positive selection when comparing ancestral branches to those branches after the duplications occurred, possibly diversifying the function of the TRL enzymes. We propose that TRLs were kept in numerous copies in the Brassicaceae genome, as they proved to have significant functions, which is evidenced by negative selection in the present orthologous groups. Markov Chain Monte Carlo Analysis of significant aminoacid changes detect few changes that could alter the functionality of new TRLs from the ancestral ones. Six sites (Type I in the DIVERGE analysis) were detected that allowed the split into two large groups of TRLs. These changes were mapped onto the three-dimensional structure of an available TRL, and were found to be close to substrate binding sites, which supports the hypothesis that these enzymes diversified in function due to few changes in the proteins.

1.3 Materials and Methods Chapter 1

1.3.1 Sequencing the TRL gene cluster

BAC libraries from five species closely related to *A. thaliana* were screened for TRL genes. They include *Arabidopsis lyrata*, *Arabidopsis cebennensis*, *Boechera divaricarpa*, *Capsella rubella*, and a more distantly related species, *Cleome spinosa* (Capparales). A phylogeny of the species is shown in Figure 1.1 and information on the BAC libraries is available in Table 1.2. To ensure that only BAC clones containing the homeologous regions containing the TRLs on chromosome 2 were identified, the BAC libraries were hybridized with conserved genes flanking the TRL cluster. They include six flanking genes and five central genes (see Fig. 1.2): At2g29120 (ion-channel protein), At2g29130 (putative laccase), At2g29140, At2g29190 and At2g29200 (pumilio RNA-binding protein), At2g29170 (tropinone-reductase-like), At2g29220 (kinase), At2g29370 (tropinone-reductase-like), At2g29380 (protein phosphatase type 2C), At2g29390 and At2g29400 (glutathione-S-transferases). Probes for these genes were obtained by designing primers with the PRIMER3 program (63) based on the sequence of the Col-0 accession of *A. thaliana*. Products from standard PCR reactions were run on an agarose gel to confirm the presence of a single band. 100 ng of the PCR products were labeled using the ECL Direct Nucleic Acid labeling and detection system (Amersham Biosciences) and used as probes to hybridize nylon filters with the BAC clone DNA. At least ten positive BAC clones from every species were further characterized. BAC-DNA was isolated from 200 ml bacterial culture with the NucleoBond BAC100 Kit (Macherey-Nagel, Dueren, Germany). Dot blots of BAC-DNA were made and separately hybridized with the flanking genes to identify those clones that contain the complete TRL cluster, indicated by successful hybridization with flanking genes from both sides. The BACs were subjected to a restriction digest with EcoRI and XhoI enzymes and separated on an agarose gel to identify distinct BAC clones. A single clone from each species was chosen for complete sequencing.

Table 1.2: Origin of BAC libraries

Species	BAC Library created by
Arabidopsis lyrata spp.lyrata	Dr. June Nasrallah, Cornell U
Arabidopsis cebennensis	Keygene, Netherlands
Boechera divaricarpa	LION Biosciences, Heidelberg, Germany
Capsella rubella	Keygene, Netherlands
Cleome spinosa	Keygene, Netherlands

The *Brassica rapa* TRL genes have been published before (64), and their sequence was kindly supplied by Dr. K. Murase (Nara Institute of Science and Technology, Japan) published. Shotgun sequencing of the BAC clones containing the TRL gene cluster was partly outsourced to a commercial company (Windsor Pond Associates, Chicago, Illinois, USA) and was partially done in-house. We used enzyme restriction (Sau3A and Tsp590, New England Biolabs) to construct 2-9 Kb fragment subclone libraries. Ligations were performed using *BamHI* or *EcoRI* restricted and dephosporylated pUC19. Subclones were sequenced using the M13 primers, in an ABI3700 capillary sequencer.

1.3.2 Sequence assembly and annotation

Sequences were edited and assembled into contigs using the programs PHRED and PHRAP (65). Finishing was carried out with CONSED (66) and the AUTOFINISH and AUTOPCRAMPLIFY programs (67). Gaps were closed by primer walking and the correct assembly of repetitive regions was checked by comparing the observed and expected lengths of PCR products spanning these regions. Annotation was carried out using a software pipeline (written by Steffi Gebauer-Jung), which combines BLAST searches (68) and GENEWISE (69) with *ab-initio* programs such as GENEMARK (70) and GENSCAN (71). The pipeline creates a XML file for editing with the APOLLO annotation editor (72). A pseudogene was identified as a gene that either lost one or more exons or displayed a clear frameshift. The identity of all ORFs identified by the pipeline and considered to be true genes was determined by BLAST comparisons with *A. thaliana* genes. Transposable elements were detected by BLAST searches against the Plant Transposable Element Database (AT-TED; `http://www.biology.mcgill.ca/faculty/bureau`). In addition, repetitive elements were detected with REPEATMASKER (`www.repeatmasker.org`).

1.3.3 Phylogenetic analysis

Alignment of TRL genes was done initially with CLUSTALW and checked for misalignments by eye. Phylogenetic trees were constructed from both coding DNA and amino acid sequences using SEQBOOT to produce 100 bootstrapped sets and DNAML and PROML programs from the PHYLIP package, using default options for all 100 trees (73). We also used the program NEIGHBOR to construct a neighbor-joining (NJ) tree of amino acid sequences. Among all genes, we defined groups of orthologous genes as those that grouped together in a clade whose supporting branches have a bootstrap value higher than 50%. In parallel, a Markov Chain Monte Carlo simulation was done using MrBayes (74) for phylogenetic reconstruction of the amino acid sequences. The parameters used were Ngen (number of generations)= 10,000; Datatype=protein; Aamodel= mixed (sets fixed rates for variable amino acids); number of states= 20 (sets the frequencies of the amino acid states from the mixed models) and covarion= no. After 10,000 generations, 75,000 trees were obtained and a consensus was built from these. This tree was compared with the neighbor-joining tree, which confirmed the robustness of this tree. As the Bayesian tree had better confidence at nodes and branches, we used this tree for further analyses for positive selection (described below).

1.3.4 Tests of positive selection

Tests of positive selection in protein coding genes were conducted on codon-based nucleotide alignments with the PAML package (75). These tests are based on a comparison of the ratio of nonsynonymous to synonymous divergence by calculating the ratio $\omega = d_N/d_S$. A value of $\omega = 1$ is expected for a protein evolving under no selective constraint; $\omega < 1$ indicates purifying selection and $\omega > 1$ positive selection. This test for positive selection is conservative when applied to the whole coding sequence. However, because most amino acid sites are expected to be under some level of selective constraint, numerous modifications have been developed to allow tests of positive selection on particular codons or evolutionary lineages. PAML implements a maximum likelihood (ML) method that jointly estimates model parameters (ω values of different codon classes and/or lineages; the transition/transversion ratio κ) and the model likelihood. The likelihoods are used to compare different evolutionary models with and without positive selection (i.e., a separate class of codons with $\omega > 1$) in a likelihood ratio test (LRT). A test statistic is calculated as twice the difference of the individual likelihoods: $2\Delta = 2 \times (l_1 - l_2)$. This statistic is approximately χ^2-distributed and corresponding tables are used to look up the significance threshold; the degrees of freedom of the test are calculated as the differences among the model parameters.

To detect different rates of evolution within and among orthologous clades, we first performed a pairwise comparison (runmode= -2 in PAML). We obtained ω for pairs of genes within each orthologous clade, calculated averages and standard deviations for all pairwise comparisons, and performed a Z-test to test for statistical significant differences from the mean. To test for positive selection on individual

codons, we compared the site models M7 vs. M8 (44). Both models assume that ω ratios are distributed among sites according to a beta distribution with $f(x; p, q)$ where the distribution can take any shape depending on p and q. M8 is an extension of M7 with an independent class of sites; its ω value is estimated from the data. Positive selection is detected when this ω value is > 1 and the LRT is significant. We also compared M8 vs M8a (as in (41)) and performed two additional tests for the large clades, as M8 was significant for most of these. The difference between the two distributions is that M8a estimates the beta distribution centered on $\omega = 1$. This is achieved by fixing $\omega_s = 1$ .We fixed $\omega > 1$ as a further comparison, also known as M8b, and performed another test centering the beta distribution at $\omega < 1$, which we call M8c, for comparison.
We used the branch-site models MA vs. MB to detect positive selection on codons in preselected branches of the phylogeny (the 'foreground' branch; 42) in which positive selection was allowed to occur. Model A (MA) estimates two ratios, one for labeled branches (foreground), and another for the unlabeled branches (background), where the background sites are evolving neutrally. Model B (MB) estimated three ratios, one for foreground and two for background branches (see Table 1.7). In both models, positive selection is allowed in foreground branches only. Since ML analysis is computationally expensive for large numbers of sequences, tests of selection were carried out separately for the three major clusters (A, B and C from the tree in Figure 1.3) and for a reduced set with one sequence for each paralogous clade, also called the 'short' tree (Figure 1.8). A schematic representation of the 'short tree' showing the branches after which positive selection is allowed, is shown in Figure 1.7. PAML also performs a Bayesian test for detecting positively selected sites. This test estimates a prior probability of ω from site frequencies from the data, and calculates a posterior probability, estimating the probability of sites, including those with $\omega > 1$. This test is used to identify sites that are positively selected specially in M7 and M8, when the LRT is not significant, although it is conservative (76). One of the assumptions of the PAML program is that no recombination has occurred between the analyzed loci. To test for recombination or gene conversion between paralogs, we used the GENECONV program (77).

1.3.5 Mapping selected sites to a TRL protein structure to detect functional important changes

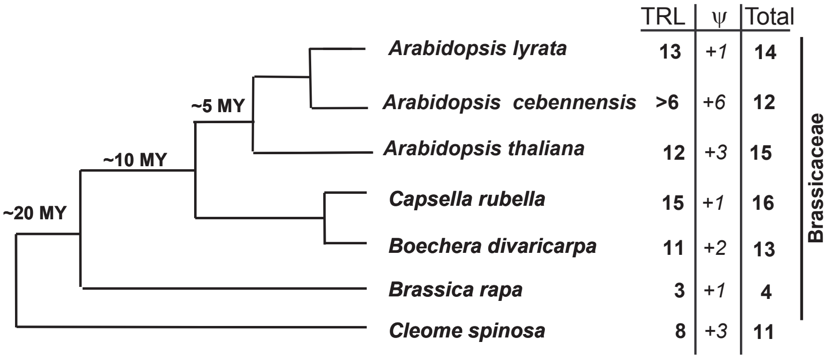
The crystal structure of one TRL from *A. thaliana* (At1g07740) was recently determined in a structural genomics project (http://www.uniprot.org; PDB code Q9ASC2, (78)). The location of positively selected amino acid residues was visualized with PYMOL (79). Functional properties of positively selected sites were investigated with the program DIVERGE (80), which implements a Markov Chain Monte Carlo method (MCMC) to estimate probabilities of amino acid variation and changes among paralogs to infer evolutionary and functionally important sites from amino acid sequences (81). The MCMC method constructs a transition probability matrix for given time periods, and this is estimated for each site. The program DIVERGE uses a maximum likelihood approximation. It uses information from the protein phylogeny and divides the phylogeny into subclusters, analyzing the probability that each of these sites will change according to prior probabilities and observed changes. It further uses information from the protein structure to assign functional importance to changes in the properties of amino acids at one site. Changes are mapped according to posterior probabilities detecting those sites that are significantly different and produce functional changes among groups of paralogs (in this case, clades B and C). This functionally important variation is named Type I divergence. The method can also detect those sites that do not cause functional changes, as they are conserved among paralogs but are divergent among species. These are called Type II divergent sites, and can result in changes in charge and hydrophobicity. We used this method, since we were interested in mapping those sites that are split among the different duplicated gene clusters, which might be of functional importance. If these sites also showed evidence for positive selection, this might explain the divergence of the orthologous groups that allowed numerous copies of TRLs to be preserved in the genomes of Brassicaceae.

1.4 Results Chapter 1

1.4.1 Comparative genomics of the TRL region

The TRL regions of seven species were compared for this study. We had to sequence three BAC clones from *B. divaricarpa* to cover the whole genomic TRL region (Figure 1.2). The BAC clone from *A. cebennensis* does not cover the complete TRL region, and we were not able to identify another clone that would cover the missing part. The sequence from *B. rapa* was the shortest, 10 kb, and contains only the TRL genes and one neighboring gene on each side. The insert size of the sequenced BAC clones ranged from 63 kb in one of the three clones from *Boechera divaricarpa*, to 198 kb for the *Arabidopsis lyrata* clone.

At least four TRL genes were found in all species included (Figures 1.1 and 1.2). The *Brassica rapa* sequence had the lowest number of TRLs and three complete and one truncated TRL ORFs. *Capsella rubella* has the largest number of paralogs with 15 complete ORFs and one pseudogene. The most distantly related species, *Cleome spinosa*, contains eight complete TRL copies plus three pseudogenes. The TRL genes are surrounded by other genes likely involved in the stress response, such as glutathione-S-transferases (82) and small heat shock proteins (of Type II). As in the TRL genes, there is copy number variation within the neighboring gene families; copy number variation seems to be a general feature of plant gene families and is not unique for TRLs. However, since we do not have the complete sequence information on these clusters from the other species, we did not analyze them further.



	TRL	ψ	Total
Arabidopsis lyrata	13	*+1*	14
Arabidopsis cebennensis	>6	*+6*	12
Arabidopsis thaliana	12	*+3*	15
Capsella rubella	15	*+1*	16
Boechera divaricarpa	11	*+2*	13
Brassica rapa	3	*+1*	4
Cleome spinosa	8	*+3*	11

Figure 1.1: Phylogenetic relationships of the studied species displaying TRL gene copy number and pseudogenes in italics. The total number of TRLs found in tandem in the syntenic region to chromosome 2 from *A. thaliana* is displayed in the last column. TRL genes seem to expand along the phylogeny, but there are differences in gene function loss. Adapted from O'Kane and Al-Shehbaz (2003).

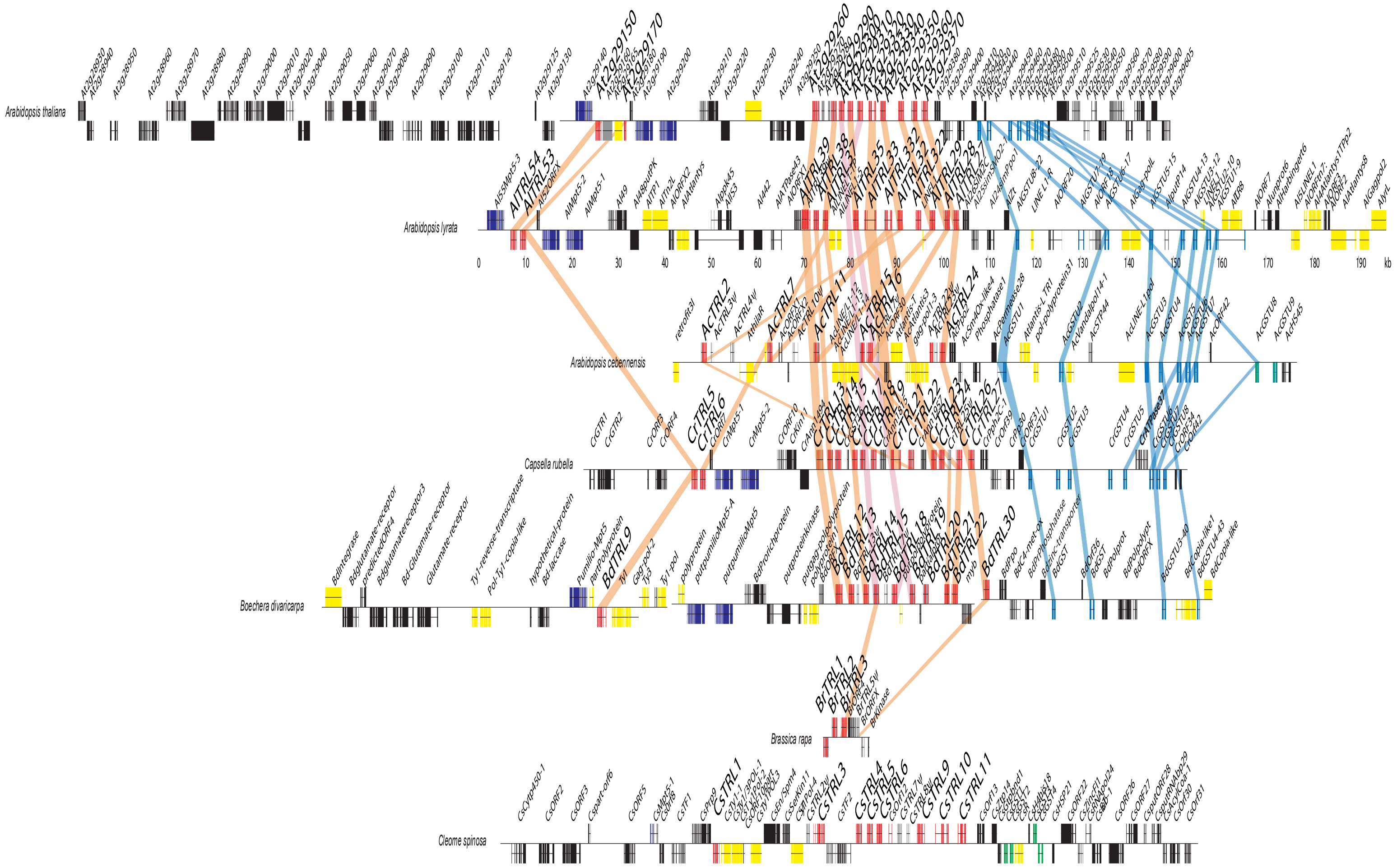


Figure 1.2: Composition and alignment of homologous regions to *A. thaliana* sequenced from *A. lyrata*, *A. cebennensis*, *Boechera divaricarpa*, *Capsella rubella*, *Brassica rapa* and *Cleome spinosa*. Red lines unite orthologous TRL genes (red squares), blue lines unite orthologous glutathione-S-transferases (blue squares). (These last genes are discussed in chapter 3.)

Transposable elements (TEs) identified by BLAST searches against the Plant Transposable Element Database and by REPEATMASKER were counted separately for each BAC clone (Supplementary Table 1.3). The highest proportion (expressed as percent BAC sequence consisting of TEs) occurs in *A. lyrata* (18.9%), followed by *A. cebennensis* (16.8%) and *B. divaricarpa* (15%). At the lower end are *A. thaliana* (4.5%) and *C. rubella* (0.45%). *Cleome spinosa* is intermediate with 9.8% . There is no significant correlation of TE density (Pearson $r = -0.25$) and TE number (Pearson $r = -0.22$) with genome size (data from 83). It should be noted that such a correlation is not necessarily expected, because the BAC clones cover different sections of the TRL regions in the species. Nevertheless, the differences observed between *A. lyrata* and *A. thaliana* correspond to the results of a comparison of randomly sequence genomic regions from both species, suggesting there is a relationship with the genome size (Oyama et al., submitted).

1.4.2 Phylogenetic history of TRL genes

After constructing phylogenetic trees, we identified several major clades (Figure 1.3). They consist of an ancient clade with homologs from several dicot and monocot species, including the tropinone reductase enzymes from the Solanaceae family (clade A), and two 'modern' clades that contain only homologs from the Brassicales (clades B and C). Clade A contains at least one gene of each of the studied species, except for *Brassica rapa* and *A. cebennensis*, from which we might not have obtained all TRL paralogs (see below). One poplar and two rice genes obtained from Genbank, and an additional TRL gene from *A. thaliana* located on another chromosome (At5g06060, Figure 1.3) also group within this clade. Because of the similarity with rice and poplar TRL genes, clade A is most probably ancestral to all other Brassicaceae and *Cleome spinosa* TRLs. Duplicate genes for this last species are all paralogous, i.e. they group together and form independent clades. One cluster of *C. spinosa* paralogs is nested within clade B (Cleo-par1; Figure 1.3); two additional clusters of *C. spinosa* paralogs are basal to clade C (Cleo-par2 and Cleo-par3). No representative of this species is found in the 'ancient' clade A and no *C.*

Species	Ty1-copia	Ty3-Gypsy	LINE	Other	Total bp transposon	% of total BAC length (bp)
A. lyrata	2	2	5	Atlantys(3x), Harbinger, IS3	37622	18.94
A. cebennensis	1 (retrofit)	0	3	Vandal,Atlantys(3x),non-LTR,Athila	22663	16.84
A. thaliana	1	0	0	MuDR	4958	4.75
C. rubella	0	0	0	unknown	640	0.50
B. divaricarpa	9	1	0	Evelknievel (2x), integrase	28705	14.99
*B. rapa**	0	0	0	retroelement, Ac-like	1800	3.42
C. spinosa	4	0	0	En/Spm, non-LTR	14630	9.81

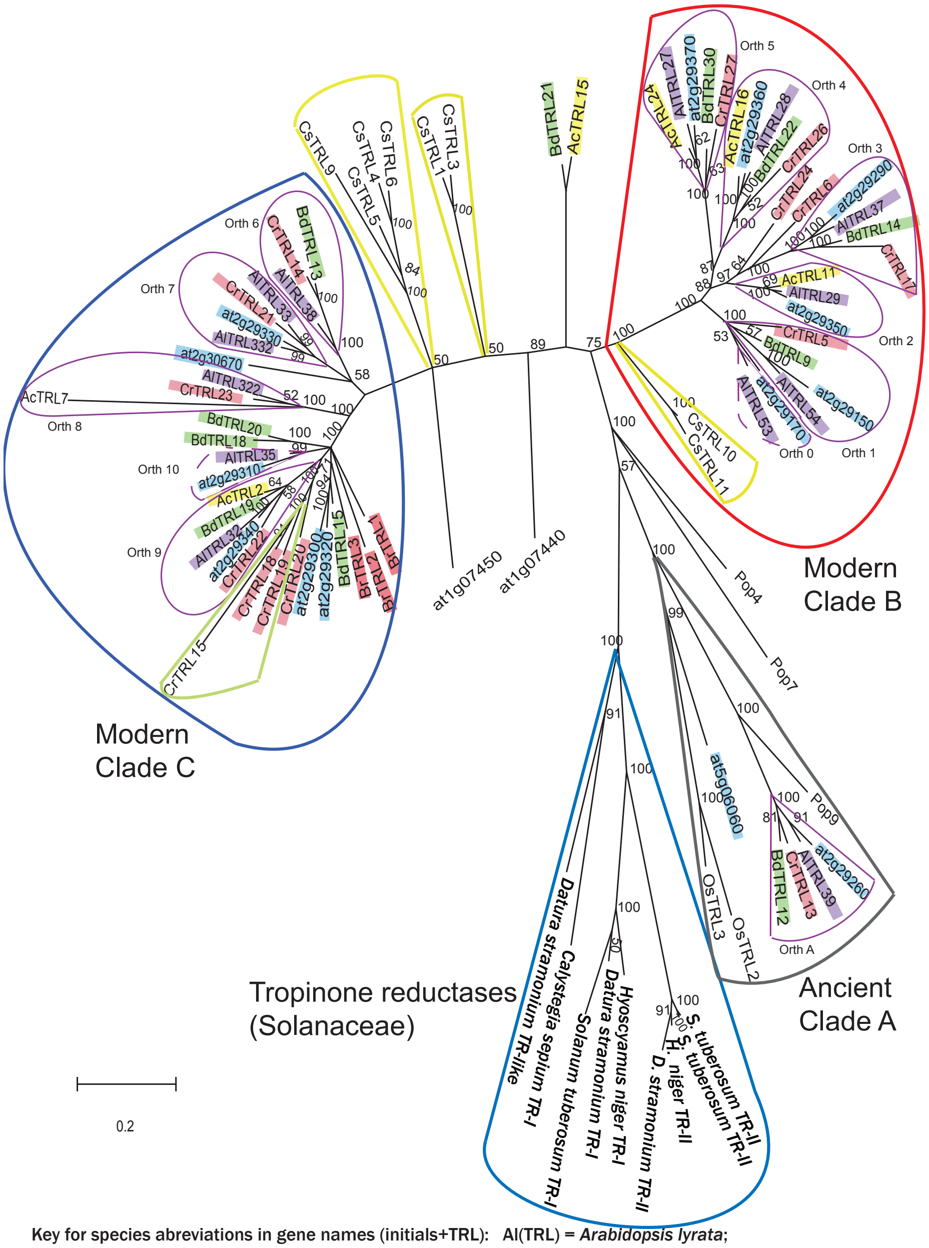
Table 1.3: Numbers and types of transposable elements from the BAC sequences. Data for *B. rapa* are obtained from (64).

spinosa paralog is found mixed together with Brassicaceae in the 'modern' clades, B and C.

Brassica rapa represents the most basal lineage in our sample of Brassicaceae species (Figure 1.1). The sequence contig available for this species contains only three TRLs and one TRL pseudogene. The three functional paralogs are basal to orthologous groups in clade C, although with a bootstrap support <50%. The pseudogene falls within clade B (Figure 1.4) and may consist of a non-functional descendant of the common ancestor of ortholog group 3 in Figure 1.1.

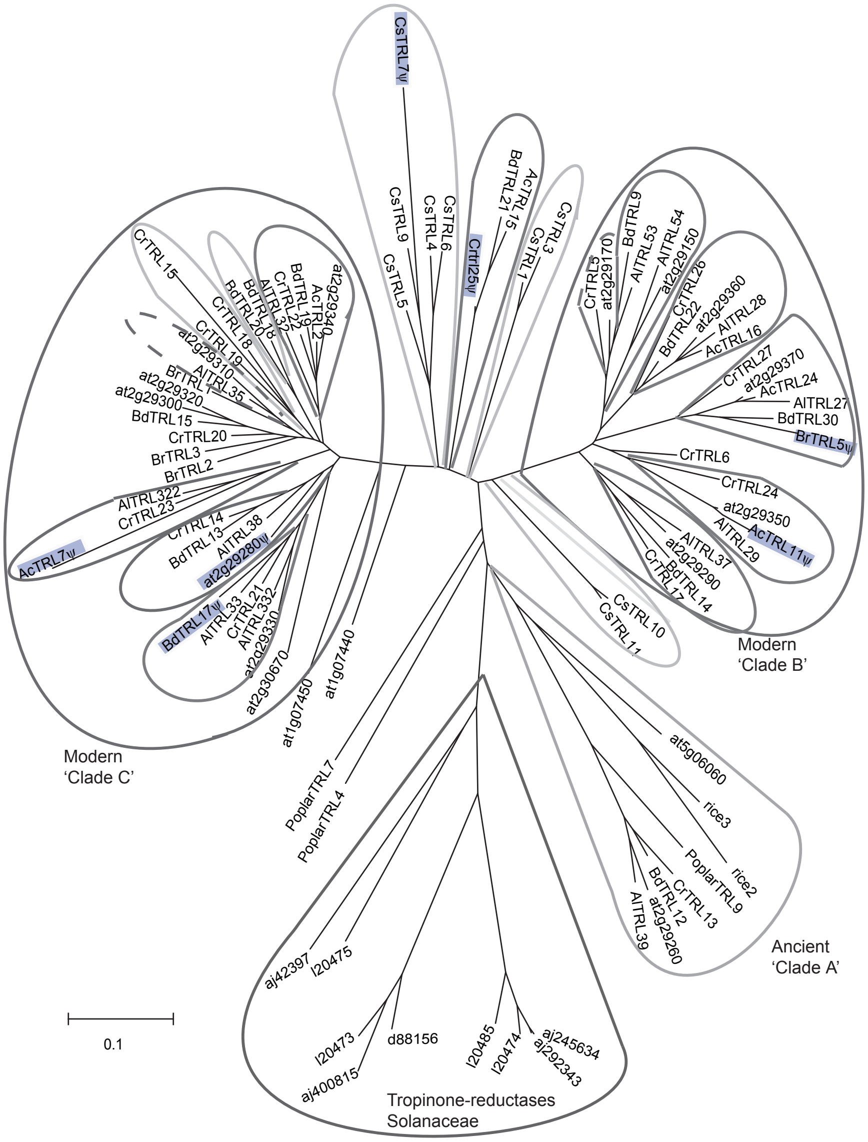
The presence of four TRL genes in *B. rapa* could indicate that there was an expansion of TRL genes after the split from the common ancestor with the rest of the studied Brassicaceae, or that substantial gene loss has occurred in this lineage. This latter hypothesis is difficult to test with the available data, because the available *B. rapa* genomic region is only around one-tenth the size of the genomic regions sequenced for the other species. Since the *B. rapa* BAC was obtained from another group as a partial sequence, some of the TRLs in the gene cluster might be missing. One way to test if *B. rapa* contains only four TRLs is to perform a reciprocal BLAST search of the TRL coding sequences from the *Arabidopsis spp.* from our study, in EST sequence databases from *Brassica spp.* species. We first compared the complete *A. thaliana* contig to the *Brassica oleraceaea* database at the Plant Genetics and Genomics Centre, DPI, Victoria, Australia (http://hornbill.cspp.latrobe.edu.au/brassica.html) using the BLAST program. We found only four EST clones containing TRLs. When compared to *Brassica napus* databases (John Innes Center, Norwich, UK), we found six complete TRL ESTs, and four partial TRLs, or pseudogenes (data not shown). Considering that this species is a tetraploid, we might expect the base number of TRLs to be three to five, considering that one might have become a pseudogene. Therefore our *Brassica rapa* clone might contain all or at least most of the TRLs present in this plant.

More recent clades within the Brassicaceae (Figure 1.1) include the *Boechera divaricarpa* and *Capsella rubella* TRLs. They consist of eleven complete TRL paralogs and two pseudogenes in the former species, and fifteen complete TRLs and one pseudogene in the latter. The most recent clade comprises the three *Arabidopsis*



Key for species abreviations in gene names (initials+TRL): Al(TRL) = *Arabidopsis lyrata*; Ac(TRL)= *Arabidopsis cebennensis*; Bd(TRL)= *Boechera divaricarpa;* Cr(TRL)= *Capsella rubella*; Br(TRL)= *Brassica rapa*; Cs(TRL)= *Cleome spinosa*. AGI abbreviations are used for *Arabidopsis thaliana*.

Figure 1.3: Bayesian consensus tree of amino acid sequences for all sequenced Brassicaceae TRLs including the outgroups Cleomaceae and Solanaceae obtained using MrBayes. Light yellow grouped genes are paralogs from *Cleome spinosa* indicating independent duplication of TRLs in this species. Genes circled by fine purple lines are orthologous groups of Brassicaceae TRLs, further analyzed in PAML. Solanaceae, poplar and rice genes were obtained from Genbank. Consensus was obtained from the MrBayes output from $n = 75,000$ trees after 10,000 generations.



Key for species abreviations in gene names (initials+TRL): Al(TRL) = *Arabidopsis lyrata*; Ac(TRL)= *A. cebennensis;* Bd(TRL)= *Boechera divaricarpa;* Br(TRL)= *Brassica rapa*; Cr(TRL)= *Capsella rubella;* Cs(TRL)= *Cleome spinosa.* AGI abbreviations are used for *Arabidopsis thaliana* and Solanaceae.

Figure 1.4: Phylogenetic tree of all TRL-genes including pseudogenes that were easy to align. The tree is one neighbor-joining tree constructed using amino acid sequences after 1000 bootstrap runs. Pseudogenes are shadowed in blue and labelled with ψ after the gene's name. Note that branches leading to pseudogenes are much longer, indicating less constrained evolution.

species. The incompletely sequenced *A. cebennensis* TRL region harbors six complete TRLs and six pseudogenes. *A. lyrata* contains 13 complete TRLs plus one pseudogene, and *A. thaliana* contains 12 copies of TRLs and three pseudogenes.

There are three TRLs from *A. thaliana* that are found on other chromosomes. Two copies are duplicated in tandem on chromosome 1 and one copy is found on chromosome 5. Both genes on chromosome one (At1g07440 and At1g07450) are at the base of clade C and appear to be duplications that occured independently from the rest of the chromosome 2 TRLs. The duplications on chromosome 1 appear to have occurred relatively recently from the *A. thaliana* paralogons website (http://wolfe.gen.tcd.ie/athal/dup). They are duplicates of the ancestral TRL on chromosome 2 (at2g29260) localized in clade A, which itself is an ancient duplicate from the gene on chromosome 5, at5g06060. The age of the duplications is confirmed by the silent divergence (dS) values, $dS = 1.59, 1.62$, between the chromosome 1 and all chromosome 2 genes, but the average dS value is $dS = 15.34$ among chromosome 5 and all the other TRLs. The paralogons in this website also indicate that all TRLs on chromosome 2 are derived from at2g29260, as they all 'collapse' into (i.e. are more similar to) this gene.

1.4.3 Dynamics of in tandem TRL genes

The phylogenetic trees of TRLs with complete ORFs (Figure 1.3) together with pseudogenes (Figure 1.4) were used to arrange Brassicaceae TRLs according to orthology (Figures 1.2 and 1.5). Since TRLs from *Cleome spinosa* have undergone independent evolution and are not orthologous, they could not be aligned accordingly, and only gene order is depicted (Figure 1.2). This representation allows us to infer evolutionary dynamics and helps to distinguish inversions and gene losses at particular loci, schematized in Figure 1.5. Also, in some cases, it allowed us to estimate the time of gene loss and pseudogenization with respect to other orthologs (Supplementary Table 1.8). After separation from the Cleome- Brassicaceae ancestor, ten gains by duplication occurred in the genome segement of *C. spinosa*. In Brassicaceae, we recognized 14 gains by duplication, consisting of seven recent

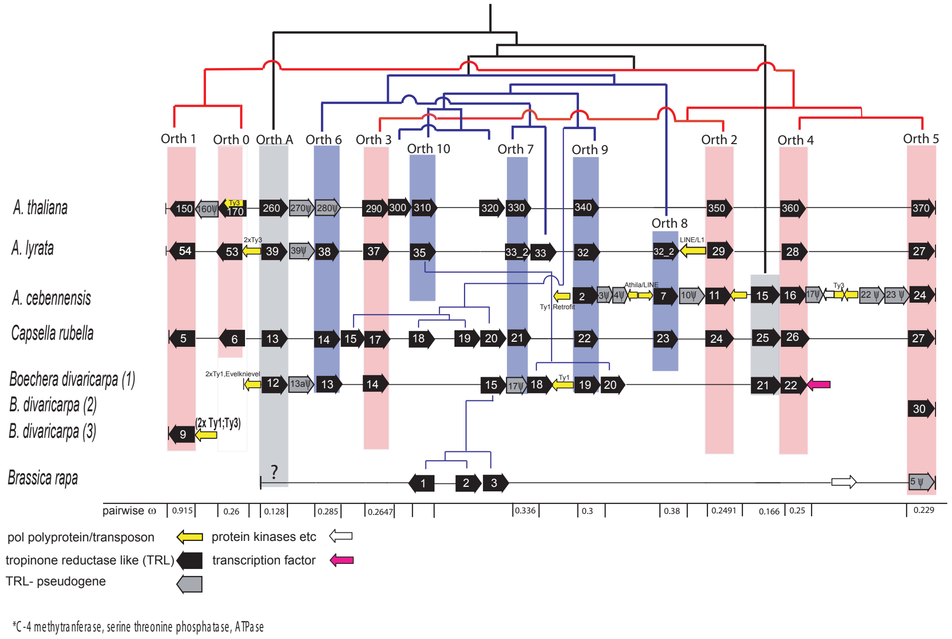


Figure 1.5: Phylogeny of tropinone-reductase-like genes *in tandem* from Brassicaceae. They are aligned according to orthology. The phylogeny does not reflect the succesion on the chromosome and the alignment of the genes does not reflect ages of duplication. The *A. thaliana* genes are used as a reference. Red lines indicate clade B and blue lines clade C. Black lines lead to the ancestral clade. TRL genes are depicted by black, the pseudogenes by gray arrows. Shadowed regions indicate homology, the dark shadowing indicates the ancestral genes. Fine lines conect paralogs. The table in the bottom displays average pairwise ω values for sequences in each orthologous clade.

duplications, which are found in single species, and seven ’gains’ of genes or loci, which are more ancient duplication events as they are shared between two or more species. Genes were lost from a particular position four times (see Table 1.8) and at least three times genes lost their function and became pseudogenes as a result of exon loss or frameshifts.

This type of duplication-retention-loss dynamic is known as the gene birth and death process (14). In the TRL region, genes are more often retained than lost (Supplementary Table 1.8) suggesting the functional importance of this gene family. Another variable feature of TRLs is their intron exon structure (see Table 1.4). Most of the TRLs have five exons (and four introns) but 10 out of 71 TRLs (14%) display a different structure, containing at least four exons and at most six.

A schematic alignment of TRL genes is shown in Figure 1.6. This alignment of paralogous genes (where each orthologous clade is represented by a single paralog) displays conserved and variable sites. The binding site and proton acceptor site are conserved, but neighboring sites are variable (Figure 1.6) (27).

1.4.4 Maximum likelihood analysis of positive selection

The phylogenetic tree did not look as if conversion had occurred, as each of the orthologous clades contained the species tree. Exceptions are the paralogs of *Cleome spinosa*, indicated in the tree as par1-3, and some duplicates within *Capsella rubella* and *A. thaliana*. We tested for gene conversion using GENECONV on these species, but it was not significant (results not shown; for a discussion see Appendix 1). Also, when reconstructing a phylogeny with the splitstree method (see Appendix 1 for an explanation), we find that branches are well defined and no recombination appears to be happening.

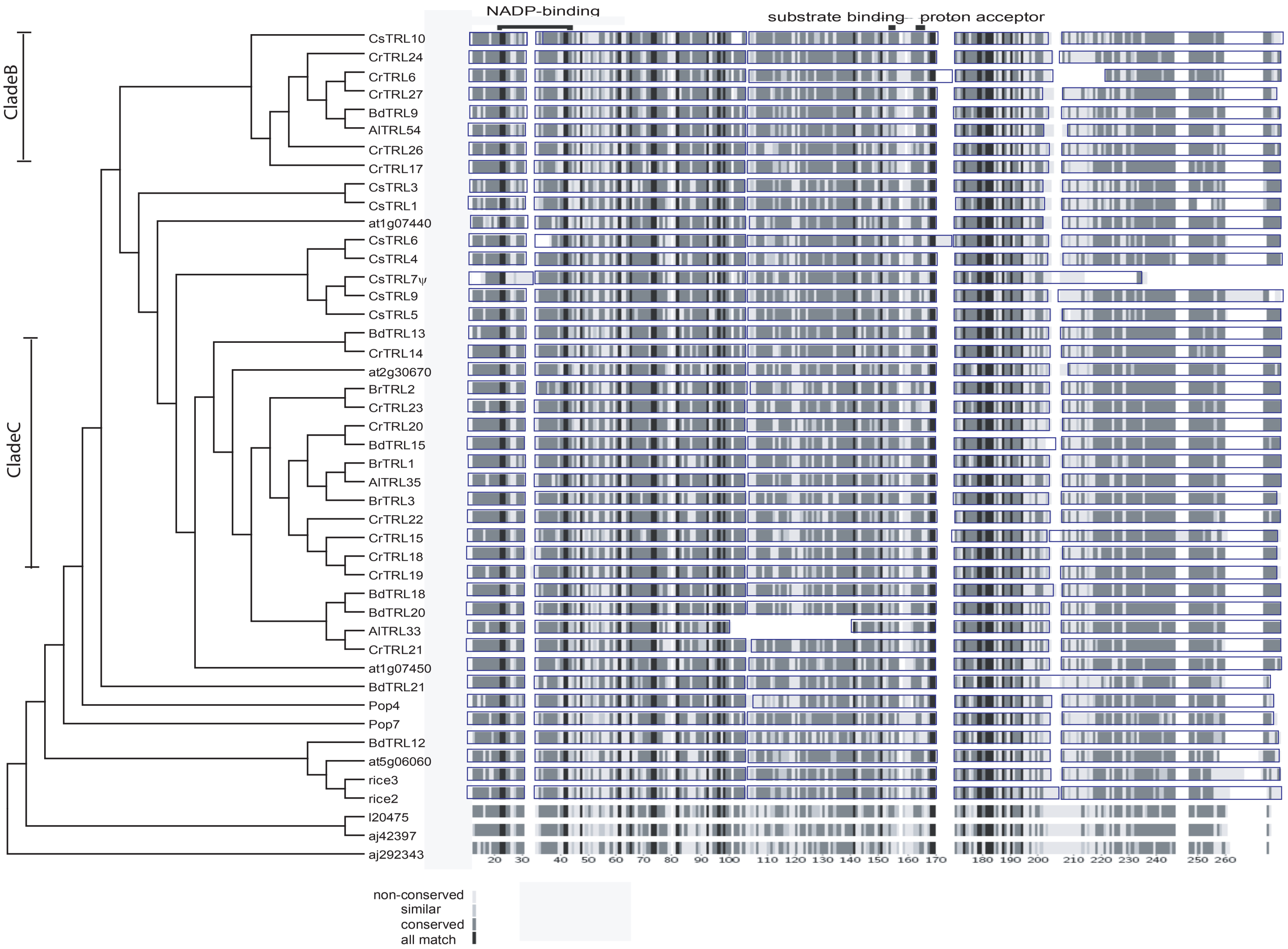


Figure 1.6: Comparison of amino acid sequences of ancient paralogs used in the analysis 'short tree'. Displayed are conserved, 'similar' and variable sites. Rectangles around the fingerprint indicate exons, white spaces are gaps, but when short also demarcate exon start-end.

Locus (from *A. thaliana*	*A. thaliana*	*A. lyrata*	*A. cebennensis*	*Boechera divaricarpa*	*Brassica rapa*	*Capsella rubella*	*Cleome spinosa*
1 (At2g29150)	5	5 (AlTRL54)		6 (BdTRL9)		5 (CrTRL5)	5 (CsTRL1)
2 (At2g29170)	5	5 (AlTRL53)				5 (CrTRL6)	
3 (At2g29260)	5	5 (AlTRL39)		5 (BdTRL12)		5 (CrTRL13)	3 (CsTRL2ψ)
4 (none)		2 ($AlTRL39\psi$)		3 (BdTRL13aψ)		6 (CrTRL15)	
5 (At2g29280)	3(ψ)	5 (AlTRL38)		5 (BdTRL13)		5 (CrTRL14)	
6 (At2g29290)	5	5 (AlTRL37)		5 (BdTRL14)		5 (CrTRL17)	6 (CsTRL3)
7 (At2g29300)	5					5 (CrTRL18)	5 (CsTRL4)
8 (At2g29310)	5	5 (AlTRL35)			5 (BrTRL1)	5 (CrTRL19)	5 (CsTRL5)
9 (At2g29320)	5			5 (BdTRL15)	5 (BrTRL2)	5 (CrTRL20)	5 (CsTRL6)
10 (none)				3 ($BdTRL17\psi$)			4 ($CsTRL7\psi$)
11 (At2g29330)	5	5 (AlTRL33)		5 (BdTRL18)	5 (BrTRL3)	5 (CrTRL21)	
12 (none)		5 (AlTRL33_2)					3 (CsTRL8j)
13 (At2g29340)	6	5 (AlTRL32)	5 (AcTRL2)	5 (BdTRL19)		5 (CrTRL22)	6 (CsTRL9)
14 (none)			2; 3(ψ) (AcTRL3;AcTRL4)	5 (BdTRL20)			
15 (none)		5 (AlTRL32_2)	5 (AcTRL7)			5 (CrTRL23)	
16 (At2g29350)	5	5 (AlTRL29)	5 (AcTRL10ψ+AcTRL11ψ)			5 (CrTRL24)	
17 (none)			4 (AcTRL15)	4 (BdTRL21)		3 (CrTRL25ψ)	5 (CsTRL10)
18 (At2g29360)	5	5 (AlTRL28)	5 (AcTRL16)	5 (BdTRL22)		5 (CrTRL26)	5 (CsTRL11)
19 (none)			4 (AcTRL22); 3 ($AcTRL23\psi$)				
20 (At2g29370)	5	6 (AlTRL27)	5 (AcTRL24)	5 (BdTRL30)	3 ($BrTRL5\psi$)	5 (CrTRL27)	
Total							

Table 1.4: Exon counts of TRL loci. The reference genome is *A. thaliana*

Table 1.5: Pairwise estimates of ω. Z-values test if ω is significantly different from neutrality ($\omega = 1$).

Within group	n	κ	dN	dS	ω	σ (ω)	Z-value
Orthologous group 1	4	1.7263	5.3285	7.8097	0.9147	0.6546	0.1580
Orthologous group 2	3	1.5548	0.0188	0.0767	0.2491	0.03044	3.082**
Orthologous group 3	3	1.6648	0.0757	0.3008	0.2647	0.06155	2.5034**
Orthologous group 4	5	2.0977	0.0556	0.2219	0.2496	0.0535	4.315**
Orthologous group 5	5	1.6648	0.0575	0.2578	0.2286	0.04360	3.074**
Orthologous group 6	3	2.3760	0.0480	0.1677	0.2854	0.0550	2.088**
Orthologous group 7	3	1.4956	0.5029	1.2493	0.3356	0.1886	2.641**
Orthologous group 8	4	1.5893	0.1751	0.4434	0.3836	0.2112	2.809**
Orthologous group 9	5	2.7773	0.0536	0.1889	0.3000	0.0866	2.6108**
Average					0.3639	0.1623	

The rapid and independent amplification of the TRL family in different lineages suggests a selection-driven diversification of this gene family. To test this hypothesis, we conducted tests of positive selection in the protein-coding regions of TRLs. These tests are based on the ratio of non-synonymous (d_N) to synonymous substitutions (d_S), where $\omega = d_N/d_S$.

We first tested the hypothesis that selection pressures are not variable within orthologous clusters, which we can expect to evolve under purifying selection. A comparison of the site models M7 versus M8 (Table 1.6) reveals that all but one of the orthologous groups are evolving under purifying selection. M8 has a significantly higher likelihood in all of the cases, and according to its estimates some sites evolve with $\omega > 1$, but their frequency is low and not significant. The only exception is orthologous group 1, which shows 3% of its sites evolving with $\omega > 1$ and a positively selected site identified with BEB with a p-value of 0.982, which is significant. This orthologous group contains the genes orthologous to At2g29150 from *A. thaliana*, AlTRL54, CrTRL5 and BdTRL9. The ancestor of this group of orthologous genes originated before *Arabidopsis spp.* and *Capsella-Boechera* and underwent an inversion (see Ortholgous group 1 in Figure 1.5). It then duplicated generating Orthlogous group 0 in Figure 1.5 in the *Capsella-Arabidopsis* ancestor.

We found that within each orthologous group, all estimates of ω are significantly below 1 (Z-value, Table 1.5), except one estimate of ω that does not differ significantly from neutrality ($\omega = 1$). $\omega \geq 1$ values indicate that evolution occurs more rapidly, $\omega = 1$ if the genes evolve neutrally or $\omega > 1$ if they evolve faster than the other clades. Among orthologous groups, ω values range from 0.072 ± 0.043 to 0.3744 ± 0.21 indicating purifying selection and functional conservation within orthologous groups.

In contrast to evolution within ortholog clusters, it might be expected that positive selection for functional divergence occurred between paralogs soon after gene duplication. Thus, we tested for positive selection in the internal 'duplication' branches of clades A, B and C and also on the 'short tree' of ancient paralogs, using branch-site models (Figures 1.7, 1.8). These tests have been used previously to detect differential evolution in gene families (43).

For all of the tested groups, clade A, clade B, clade C and the 'short' tree of ancient paralogs, MB had a better likelihood than MA, but the likelihood was better with M8ωfree in all cases (Table 1.7). In the 'short' tree, we detected 8.5% of the sites evolving with $\omega = 1.29$ using M8.

Using branch-site models, MB was more likely, with background branches evolving under strong constraint, with $\omega \approx 0$, and a second type of sites evolving under weaker constraints $\omega = 0.55$. For the foreground branches, which allow positive selection, 3.2% of the sites are evolving with $\omega = 2.71$. Clade B displayed only 1.05% of positively selected sites with M8ω, where $\omega = 1.4$. The MB test was still significant compared to MA ($p = 4.36E - 27 * *^2$), but when compared to M8, this last test had a better likelihood. Positively selected sites detected with this test were different from the ones detected previously with the M0-M8 models. Clade C had the best likelihood with positive selection using M8, which showed that 3.7% of sites evolve with $\omega = 1.29$ (bottom of Figure 1.8). MB showed that in foreground branches, 4.87% of the sites are evolving with $\omega = 2.44$.

Further tests for positive selection in the phylogeny using M8a and M8b as in (41), and M8c did not outperform M8 (Table 1.7). M8 was more likely for both clades B and C (Table 1.7), although for the 'short' ancient paralog tree, M8B ($\omega > 1$) had a better likelihood, although not significant.

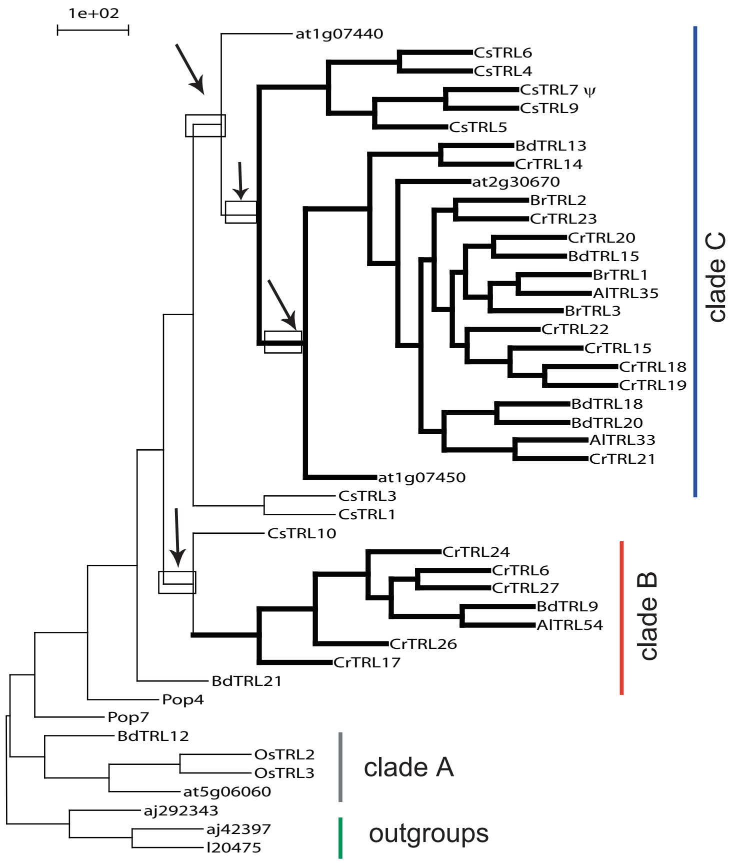


Figure 1.7: Tree of paralogs where positive selection was tested using the PAML branch-site analysis MA and MB. Those branches were tested, as they separate clades from the rest of the sequences, for instance, clade B vs. rest of the tree. Arrows point at the branches that split background from foreground branches, also denoted by squares. Positive selection is allowed in foreground branches only. Thicker lines indicate those branches where positive selection is more likely to have occurred.

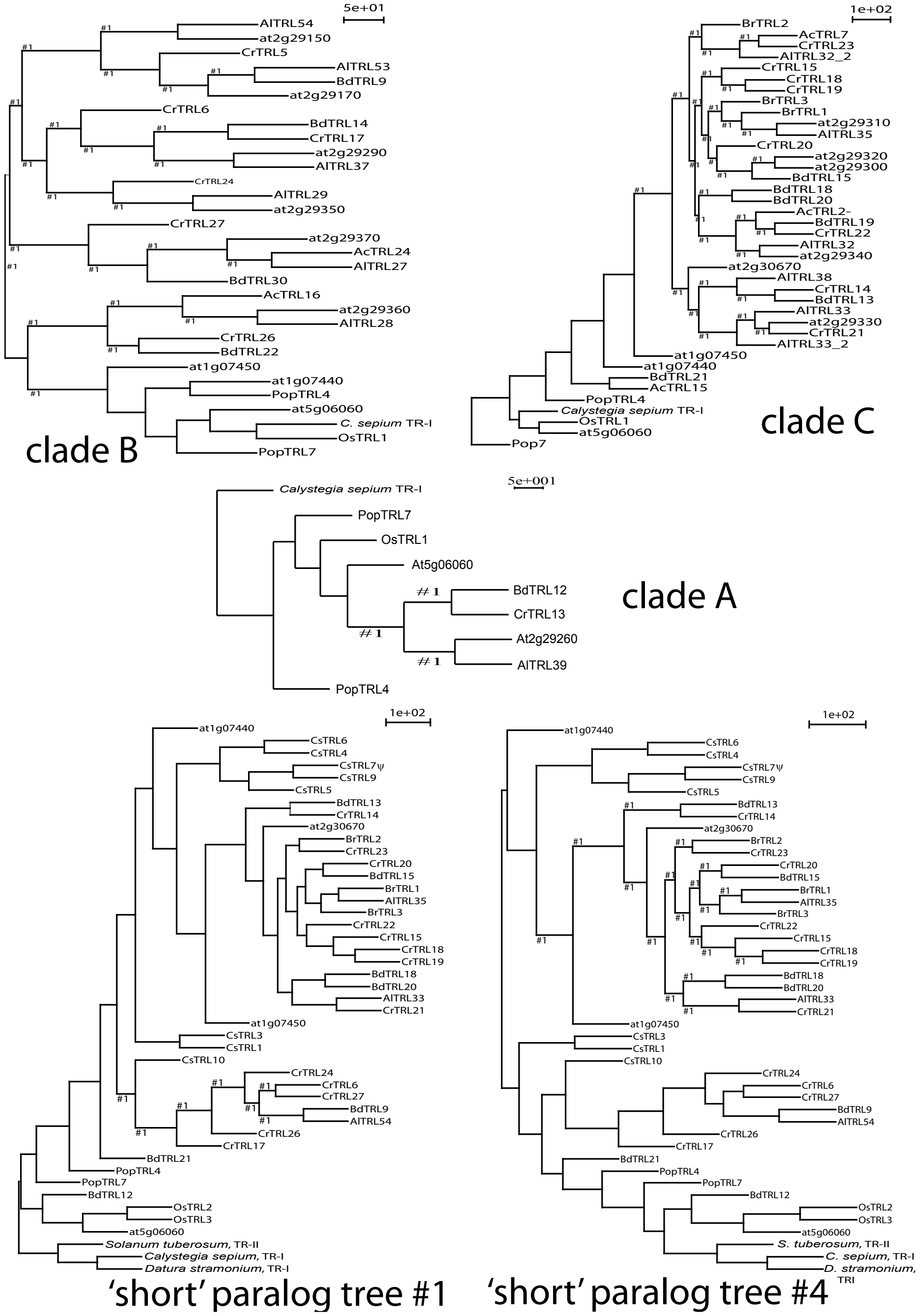


Figure 1.8: **Trees for clade B, clade C and ancient paralogs where positive selection was detected using the PAML branch-site analysis.** Branches labeled with #1 are those where positive selection was applied in the branch-site tests MA and MB that had the best likelihoods. Other trees were tested with fewer and other labelled branches, but they had a lower likelihoods and are therefore not shown.

Table 1.6: Parameter estimates and tests of selection using branch site models for single ortholog clusters and clades. Test for selection is M7 vs. M8.

Model	Parameter estimates	ℓ	$2\Delta l$	P
Ortholog cluster 1, n=4				
M7	$\omega_{1..n} = 0 - 1$	-3409.789	72.57	$1.75E - 16 * *^{1}$
	$p = 0.14853$ $q = 0.34143$			
M8	$\omega_{1..n} = 0 - 1.67$ $\omega_s = 1.67$	-3373.505		
	$p_0 = 0.57$ $p = 12.237$ $q = 99.0$			
Ortholog cluster 2, n=3				
M7	$\omega_{1..n} = 0.1827 - 0.3074$	-1314.556	0	0.9999
	$p = 31.57$ $q = 99.0$			
M8	$\omega_{1..n} = 0.1827 - 1$ $\omega_s = 1$	-1314.556		
	$p_0 = 1$ $p = 31.58$ $q = 99.0$			
Ortholog cluster 3, n=3				
M7	$\omega_{1..n} = 0 - 0.9957$	-1857.9565	3	0.2354
	$p = 0.1286$ $q = 0.3264$			
M8	$\omega_{1..n} = 0.042 - 3.41$ $\omega_s = 3.41$	-1856.5101		
	$p_0 = 0.9696$ $p = 0.4087$ $q = 1.4447$			
Ortholog cluster 4, n=5				
M7	$\omega_{1..n} = 0 - 0.98$	-1963.7443	0	0.9999
	$p = 0.1378$ $q = 0.3934$			
M8	$\omega_{1..n} = 0 - 1$ $\omega_s = 1$	-1963.7443		
	$p_0 = 1$ $p = 0.1378$ $q = 0.3934$			
Ortholog cluster 5, n=5				
M7	$\omega_{1..n} = 0 - 0.9243$	-1995.3465	0.003	0.9987
	$p = 0.2885$ $q = 0.7534$			
M8	$\omega_{1..n} = 0 - 1.0$ $\omega_s = 1$	-1995.3465		
	$p_0 = 1$ $p = 0.2885$ $q = 0.7534$			
Ortholog cluster 6, n=3				
M7	$\omega_{1..n} = 0.0 - 1.0$	-1473.073	2	0.4055
$p = 0.0427$ $q = 0.0918$				
M8	$\omega_{1..n} = 0 - 39.31$ $\omega_s = 39.31$	-1472.1703		
	$p_0 = 0.9949$ $p = 0.1358$ $q = 0.3206$			
Ortholog cluster 7, n=3				
M7	$\omega_{1..n} = 0 - 0.66$	-2143.579	3	0.2107
	$p = 0.24$ $q = 1.41$			

M8	$\omega_{1..n} = 0 - 7.54$ $\omega_s = 7.54$	-2142.022		
	$p_0 = 0.9806$ $p = 0.2931$ $q = 2.38$			
Ortholog cluster 8, n=3				
M7	$\omega_{1..n} = 0.03 - 0.44$	-2272.0348	1	0.6358
	$p = 0.7184$ $q = 1.4288$			
M8	$\omega_{1..n} = 0.03 - 1.0$ $\omega_s = 1$	-2271.582		
	$p_0 = 0.9937$ $p = 0.87$ $q = 1.8324$			
Ortholog cluster 9, n=3				
M7	$\omega_{1..n} = 0.05 - 0.5$	-2700.1682	3	0.2593
	$p = 2.088$ $q = 6.4372$			
M8	$\omega_{1..n} = \omega_s = 1$	-2698.8186		
	$p_0 = 0.8478$ $p = 14.815$ $q = 99.0$			
clade A, n=9				
M7	$\omega_{1..n} = 0.0013 - 1.32$	-5762.6724	3	0.194
	$p = 0.45$ $q = 1.46$			
M8	$\omega_{1..n} = 0.0013 - 1.32$ $\omega_s = 1.32$	-5761.0327		
	$p_0 = 0.9399$ $p_1 = 0.06$ $p = 0.5297$ $q = 2.229$			
clade B, n=31				
M7	$\omega_{1..n} = 0.008 - 0.846$	-11483.901	12	0.0023 $**$
	$p = 0.7042$ $q = 0.2957$			
M8	$\omega_{1..n} = 0.0009 - 5.09$ $\omega_s = 5.092$	-11477.855		
	$p_0 = 0.9895$ $p_1 = 0.0105$ $p = 0.451$ $q = 1.247$			
clade C, n=39				
M7	$\omega_{1..n} = 0.0032 - 0.802$	-14234.6712	8	0.016 $**$
	$p = 0.566$ $q = 1.4528$			
M8	$\omega_{1..n} = 0.0044 - 1.29$ $\omega_s = 1.29$	-14230.5427		
	$p_0 = 0.9626$ $p_1 = 0.0374$ $p = 0.64$ $q = 1.95$			
Cleome, n=8				
M7	$\omega_{1..n} = 0.017 - 0.784$	-4806.20	13	0.0013 **
	$p = 0.8697$ $q = 1.8412$			
M8	$\omega_{1..n} = 0.041 - 1.46$ $\omega_s = 1.46$	-4799.57		
	$p_0 = 0.89$ $p_1 = 0.11$ $p = 1.711$ $q = 5.684$			

Positively selected sites: [1] 245 Q $p = 0.982$*

Table 1.7: dN/dS estimates and tests for selection performed with PAML using different site and branch-site models.

Model	$\omega_{1...n}, p_{\omega>1}$ †	ℓ	Test for selection	P
Short tree 3, n=45				
M8 ωfree	0.004-1.29, 0.0856	-20256.5962		
M8A $\omega = 1$	0.007-1, 0	-20259.3748	M8 vs. M8A	0.0621
M8B $\omega > 1$	0.005-1.4, 0.0827	-20256.5871	M8 vs. M8B	0.9909
M8C $\omega < 1$	0.0038-0.4, 0	-20266.3869	M8 vs. M8C	5.6$E-05$**
Model A	ω_0,$\omega_{2a} = 0.13$ ω_1,$\omega_{2b} = 1$	-20484.61		
	$\omega_0 = 0.13$ $\omega_1 = 0.55$ ω_{2a},$\omega_{2b} = 2.68$		MA vs MB	2.71$E-38$** [1]
$p_0 = 0.5929$ $p_1 = 0.3639$ $p_2a = 0.0267$ $p_2b = 0.0164$				
Model B	ω_0,$\omega_{2a} = 0.08$ ω_1,$\omega_{2b} = 0.55$	-20398.11		
	$\omega_0 = 0.08$ $\omega_1 = 0.55$ $\omega_{2a}, \omega_{2b} = 2.71$		M8 vs. MB	3.478E-62
$p_0 = 0.528$ $p_1 = 0.4407$ $p_2a = 0.0169$ $p_2b = 0.014$				
clade A, n=9				
M8 ωfree	0.0013-1.32, 0.06004	-5761.033		
M8A $\omega = 1$	0.0013-1	-5761.3202	M8 vs. M8A	0.75
M8B $\omega > 1$	0.002-1.4, 0.05727	-5761.046	M8 vs. M8B	0.98
M8C $\omega < 1$	0.001-0.4	-5761.3202	M8 vs. M8C	0.173
Model A	ω_0,$\omega_{2a} = 0.096$ ω_1,$\omega_{2b} = 1$	-5793.43		
	$\omega_0 = 0.096$ $\omega_1 = 1$ ω_{2a},$\omega_{2b} = 1.407$		MA vs MB	4.35$E-10**$
$p_0 = 0.713$ $p_1 = 0.2624$ $p_2a = 0.018$ $p_2b = 0.0066$				
Model B	ω_0,$\omega_{2a} = 0.0406$ ω_1,$\omega_{2b} = 0.42$	-5771.472		
	$\omega_0 = 0.0406$ $\omega_1 = 0.4198$ ω_{2a},$\omega_{2b} = 7.16$		M8 vs MB	2.92$E-5**$
$p_0 = 0.5334$ $p_1 = 0.4526$ $p_2a = 0.0076$ $p_2b = 0.0064$				
clade B, n=31				
M8 ωfree	0.0009-5.09, 0.01054	-11477.85		
M8A $\omega = 1$	0.003-1	-11479.34	M8 vs. M8A	0.226
M8B $\omega > 1$	0.002-1.4, 0.04821	-11479.68	M8 vs. M8B	0.1609
M8C $\omega < 1$	0.001-0.4	-11484.204	M8 vs. M8C	0.0017**
Model A	ω_0,$\omega_{2a} = 0.11$ ω_1,$\omega_{2b} = 1$	-11572.26		
	$\omega_0 = 0.11$ $\omega_1 = 1$ ω_{2a},$\omega_{2b} = 10.17$		MA vs MB	4.36$E-27**$ [2]
$p_0 = 0.6842$ $p_1 = 0.277$ $p_2a = 0.0276$ $p_2b = 0.0112$				
Model B	ω_0,$\omega_{2a} = 0.055$ ω_1,$\omega_{2b} = 0.46$	-11511.56		
	$\omega_0 = 0.055$ $\omega_1 = 0.46$ ω_{2a},$\omega_{2b} = 2.19$		**M8** vs MB	2.29$E-15**$
$p_0 = 0.5184$ $p_1 = 0.394$ $p_2a = 0.0498$ $p_2b = 0.0378$				
clade C, n=39				
M8 ωfree	0.004-1.29, 0.03743	-14230.54		
M8A $\omega = 1$	0.006-1	-14231.408	M8 vs. M8A	0.4208
M8B $\omega > 1$	0.002-1.4, 0.03384	-14230.609	M8 vs. M8B	0.1609
M8C $\omega < 1$	0.004-0.4	-14235.084	M8 vs. M8C	0.0107**
Model A	$\omega_0, \omega_{2a} = 0.13$ $\omega_1, \omega_{2b} = 1$	-14350.87		
	$\omega_0 = 0.13$ $\omega_1 = 1$ $\omega_{2a}, \omega_{2b} = 1.63$		MA vs MB	2.12$E-31$** [3]
$p_0 = 0.6692$ $p_1 = 0.2639$ $p_2a = 0.048$ $p_2b = 0.019$				
Model B	$\omega_0, \omega_{2a} = 0.07$ $\omega_1, \omega_{2b} = 0.47$	-14280.239		
	$\omega_0 = 0.07$ $\omega_1 = 0.47$ $\omega_{2a}, \omega_{2b} = 2.44$		**M8** vs MB	2.61$E-22$
$p_0 = 0.5314$ $p_1 = 0.4199$ $p_2a = 0.0272$ $p_2b = 0.0215$				

Positively selected sites identified with BEB: [1] 80 S $p = 0.963$*, 133 S $p = 0.999$**, 164 T $p = 0.953$*; [2] 139 Q $p = 0.965$*, 253 T $p = 0.974$**; [3] 151 G $p = 0.983$*, 247 S $p = 0.971$*

†The values from $\omega_{1...n}$ indicate the values of ω can take in the β distribution.

1.4.5 Mapping selected sites onto protein structure and inference on function

We mapped selected sites predicted by the different PAML tests onto the three dimensional structure of the protein encoded by At1g07440 (PDB code Q9ASC2, Figure 1.9). As expected, two of the selected sites, site 133 (serine) detected in the analysis for the short tree and site 151 (glycine) detected in the analysis of clade C, are both close to the substrate binding site of the protein and could therefore be responsible for functional modifications. In the MCMC analysis of posterior probabilities and functional importance of changes at each amino acid site, we found six sites that could be experiencing evolutionary rate shifts after duplications ($P(S_1) > 0.5$, black rectangles on Figure 1.10). Two of these sites (first two positions marked in the alignment in Figure 1.10) are in the region corresponding to the cofactor (NADP) binding site. The first site (glycine 18) and the two following sites (serine 78 and methionine 86 respectively) are conserved in the paralogs of clade B, but very diverse in the rest of the paralogs (Figure 1.10). This follows the pattern from the Bayesian tree, where clade C contains more recent duplications and appears more variable (Figure 1.3). The serine (78) was also detected as positively selected in the branch-site PAML analysis (serine 133, its position varies because of the different alignment gaps, Table. 1.7). The next variable site is particularly interesting, site 149, a phenylalanine conserved in clade C, but is variable in all other paralogs. This site maps 2 residues upstream from the substrate binding site and is very probably responsible for functional changes. The next site detected by this analysis is methionine 172 in clade B and a lysine in clade C. This site maps 6 residues upstream from the proton donor site and might also confer functional variability, as seen later in the 3D analysis. The last site detected as a Type I change is a valine 259, which was conserved in clade C but variant in the rest of the clades. This site was not detected in any of the PAML analyses performed.

The Type II divergence statistic finds variant sites possibly affecting function more frequently than the Type I divergence statistics and the PAML analysis. In the amino acid sequence four sites are found in the NADP binding region (amino acids 27-28 and 40-41, posterior probability is $P(S_1 > 0.5$, Figure 1.10) and another variant site seven residues upstream from the binding site (site 158 in Figure 1.10). The other variable sites, some of them in α-helix regions, could be false positives, except for those found also in the PAML analysis, which are physically close to the active site in the protein structure. These last results are analyzed in more detail in the discussion.

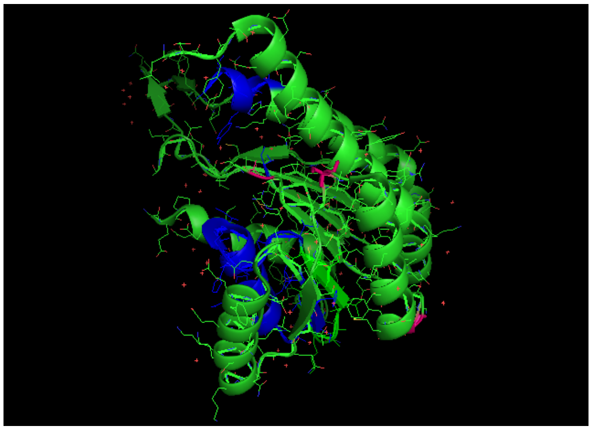


Figure 1.9: Three-dimensional structure of the tropinone-reductase-like protein from *Arabidopsis thaliana* encoded by At1g07440. The plot was generated with MacPYMOL. In blue are the NADP binding site (bottom left), the substrate binding and proton acceptor (upper). In pink are the selected sites detected with PAML, which also coincide with those detected by DIVERGE.

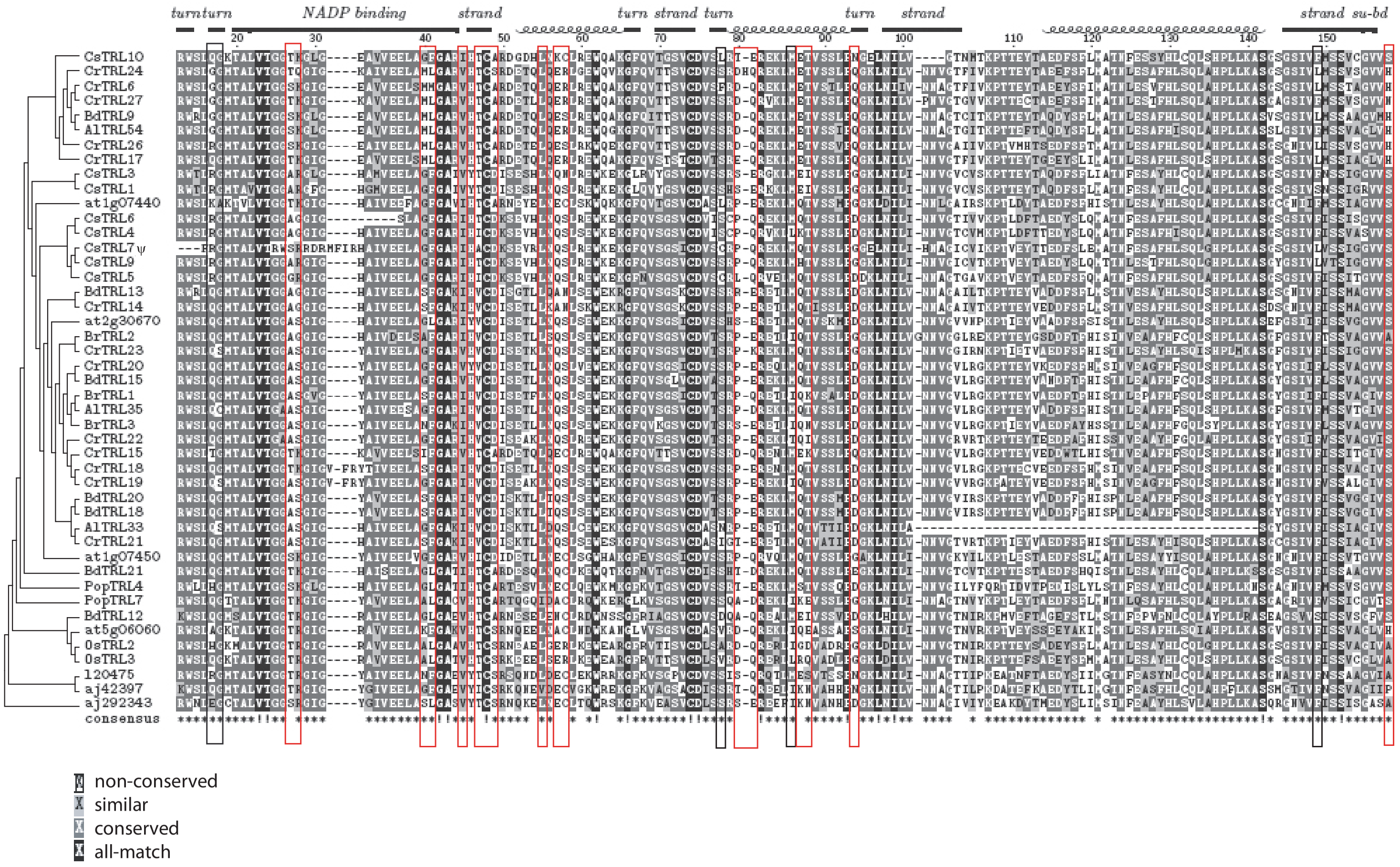


Figure 1.10: Alignment of conserved sites among paralogs. Black rectangles denote sites with Type I divergence ($p > 0.5$), and red rectangles denote sites with Type II divergence ($p > 0.5$). Shown are 5' residues 1-158 (159- 265 are shown in the next page). On top of the alignment the structural features of the protein can be seen. Abbreviations: su-bd= substrate binding; prot accept= proton acceptor.

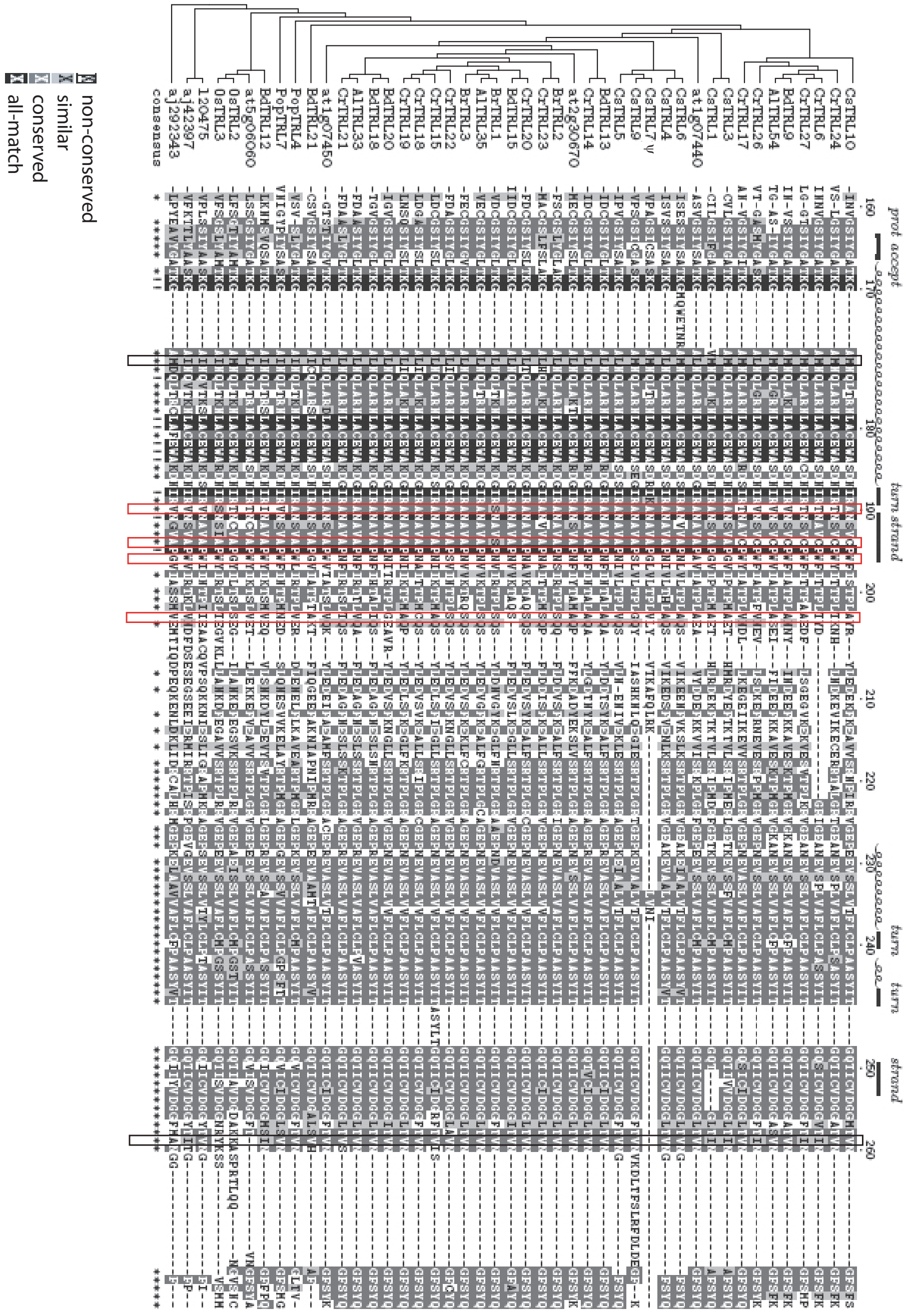


Figure 1.11: Alignment of conserved sites among paralogs. Black rectangles denote sites with Type I divergence ($p > 0.5$), and red rectangles denote sites with Type II divergence ($p > 0.5$). Shown are residues 159- 265, 3'. On top of the alignment the structural features of the protein can be seen. Abbreviations: su-bd= substrate binding; prot accept= proton acceptor.

1.5 Discussion Chapter 1

Structure of the TRL gene cluster in six Brassicaceae and *Cleome spinosa*

BAC sequences allowed us to comfirm that the region containing the TRLs is syntenic in three of the five species where the cluster was isolated. The TRL cluster flanking genes, coding for ion channels, pumilio Mpt5 proteins and glutathione S-transferases were found together with the complete number of TRLs in all of the species studied, with the exception of *Boechera divaricarpa* and *Arabidopsis cebennensis* (Figure 1.2). *Boechera divaricarpa* was identified as a hybrid and possibly tetraploid (50). The BAC might have suffered an expansion of the sequenced region, and therefore no BAC had both flanking sequences. Another possibility is that the position of the flanking genes has changed due to recombination. We managed to reconstruct the syntenic order of the TRLs and its flanking genes after sequencing three BACs. No BAC contained the whole region, since the BAC library of this species had inserts that were on average 64 Kb, and the region containing the TRL genes appears to span more than 80 Kb (Figure 1.2). In the case of *A. cebennensis*, the flanking and/or central regions where present in the hybridizations to the BACs, but after one BAC was sequenced, only a fraction of orthologous TRL genes was found in a sequence of 134.57 Kb of length, the second largest sequenced BAC. Either genome expansion or unequal recombination could have lead to the loss of all the posterior TRL genes. We find this last hypothesis to be less probable, as it would imply the loss of a large region. The large number of transposons (16.8%, Table 1.3) and pseudogenes in this BAC would support the first hypothesis. The upstream genes of *A. cebennensis* could have become pseudogenes, since the total number of genes and pseudogenes adds up to twelve functional TRLs (Figure 1.2), close to the number of TRLs contained in the other *Arabidopsis* species. The retainment of more pseudogenes in the genus could be a result of small population sizes (84). Inbreeding and small populations are supposed to aid in eliminating non-essential genes from genomes (8), and therefore it would be useful to obtain population genetic data for the species.

1.5.1 Estimation of the tempo and mode of duplication from phylogeny

By analyzing TRL genes in multiple species, we can elucidate mechanisms and selective events that would not have been detected by looking at orthologs of one or two species only. From the phylogeny constructed with all TRL genes (Figure 1.3), it is clear that these genes are undergoing gene birth and death. This dynamic is characterized by gains, issued from gene duplications, and losses, evidenced by missing loci and pseudogenes. Since we know approximate split times for the species, shown in Figure 1.1, we can estimate rates of TRL gene gain and loss. Members of the TRL gene family can be seen as duplicating quickly at 0.65 per M.Y, these duplication rates are higher than those in the literature estimated for model organisms (0.001-0.03 per MY (10)) and for duplicated genes in wheat (0.0029/locus per MY (9)). These duplication rates are not constant, and were probably higher at some point in the phylogenetic history. Most probably the ancestor of the Cleomaceae-Brassicaceae already contained at least two TRL copies. After initial duplication of TRLs either selective advantage was high for the gene duplicates, or relaxed selective constraints allowed for TRL duplications. The PAML results support the first hypothesis, since TRL genes appear to be preserved rather than lost, and most are subject to negative selection possibly due to functional importance. Relaxed selection is less probable, since maintaining duplicate gene loci might be disadvantageous, locally through non-equal recombination and to the organism through genome size expansion, as we discuss later.

Gene loss rates are approximately three times lower than the gene gain rate (0.25 per MY), something also observed in grasses (0.01/locus x MY in (9)). This supports the functional importance of TRLs. Partial losses or pseudogenization have been more than twice as frequent complete gene losses, seven in total (Table 1.8). Selection might not be very efficient in eliminating genes with no utility due to demography (8). Another possibility is that the partial transcripts we identify as pseudogenes are kept for complementary functions, but we did not test this.

TRLs have been dynamic among chromosomes, as can be seen from *Arabidopsis thaliana*. Apparently, At2g29260, located on chromosome 2, together with At5g06060, the one on chromosome 5, are ancestral, as both can be localized in clade A together with TRLs from rice and other species (Figure 1.3). Interestingly, TRLs on chromosome 1 appear to be more closely related to clade C in the phylogeny (Figure 1.3), and might have been a translocation to this chromosome from the ancestor of clade C genes that duplicated posteriorly. Homologs to At1g07440 and At1g07450 on different chromosomes might be present in all the other Brassicaceae, which we cannot know, since we don't have the complete genome sequences of these other species.

1.5.2 Duplication dynamics from genomic comparison

The most likely way to obtain two large clades of duplicated genes (clades B and C, Figure 1.3) is through non-homologous recombination of adjacent loci, see (Figure 1.5). Since all genes have introns, we can rule out retrotransposons as the mediators of transposition, as this process undergoes a reverse-transcription. Non-homologous recombination could produce clades of recently duplicated paralogous genes that are physically close, as observed specially in clade C (Figure 1.3 and Figure 1.5). It appears that the ancestral genes forming each of the two large clades already duplicated in the ancestor of *Boechera spp.-Capsella spp.* and of *Arabidopsis spp.*, since all Brassicaceae studied have orthologs in the major clades. This observation supports the hypothesis that most of the genes in clade B and C have existed at least in the last 10 ±5 MY. Unequal crossing-over as an ongoing mechanism among gene duplicates is supported by the fact that pseudogenes have more often been generated by loss of exons, rather than point mutations (frame shifts). Nevertheless, we cannot confirm this with the available data and more BACs need to be screened and/or sequenced.

An unequal sister chromatid recombination in the ancestor of the Brassicaceae-Cleomaceae, might have caused a duplication, which generated the ancestors of

Locus	Eq. in *A. thaliana*	Duplication	Gain	Loss	Pseudo-	Time
1	At2g29150	0	1	0	0	Brassicaceae ancestor (?)
2	At2g29170	1	0	0	0	Boechera-Capsella
3	At2g29260	0	0	0	0	ancestral
4	none	1	1	0	1	gain, Capsella, after loss of function
5	At2g29280	0	1	0	1	gain, Brassica, loss of function in *A. thaliana*
6	At2g29290	0	1	0	0	after Brassica
7	At2g29300	1	0	0	0	in *A. thaliana*
8	At2g29310	0	0	1	0	lost in *Capsella rubella*
9	At2g29320	0	1	0	0	in *A. thaliana*
10	At2g29320	0	1	0	0	in *A. thaliana*
11	At2g29330	1	0	0	1	dupl in *A. lyrata*, loss of function in Boechera
12	none	2	0	1	1	lost in Arabidopsis ancestor, duplicated in *Capsella rubella*
13	At2g29340	2	0	0	0	dupl. in *A. lyrata*, emerged after Brassica
14	none	0	1	1	0	lost in *A. thaliana*, *Boechera* (?)
15	At2g29350	0	0	1	0	lost in *Boechera*
16	none	0	0	1	1	lost in *A. lyrata* and *A. thaliana*
17	At2g29360	0	1	0	0	Brassicaceae ancestor
18	At2g29370	0	1	0	1	Brassicaceae ancestor, loss of function in Brassica?
Total		7	9	5	6	

Table 1.8: Duplications, gains, losses and pseudogene counts in TRL loci. With *A. thaliana* as the reference genome.

clade B and clade C genes (Figure 1.3). *Cleome spinosa* genes are found in both clades forming paralogous groups. Duplication of the TRLs in the *Cleome spinosa* lineage appears to have occurred independently from the duplications in Brassicaceae, after an initial duplication in the common ancestor. We tested for gene conversion using GENECONV (results not shown), since this process would falsely indicate genes to be recently duplicated within a species (14), but recombination among the genes did not appear to be significant. It is noteworthy that no *Cleome spinosa* gene is present in the ancestral clade A. Since this species is polyploid, as shown by (50), the ancestral gene might be located in another homeologous segment. Another possibility is that this gene was lost, but we cannot observe this with the available data.

A further non-homologous recombination of gene duplicates within clade B separated genes physically (up- and downstream genes, both in clade B (red lines in Figure 1.5), but this must have occurred in the common ancestor of all *Arabidopsis spp.* and *Boechera spp.-Capsella spp.*, since all Brassicaceae contain the upstream TRLs in anti-sense, except *Brassica rapa* and *A. cebennensis* (see Figure 1.2). *B. rapa* also contains one inverted TRL upstream of the other TRLs, but this TRL is more similar to the genes in cluster C. We might not have found all of the paralog TRL genes for *A. cebennensis*, since the BAC library might not have contained these genes, as discussed above.

With the exception of *Capsella rubella*, transposon insertions are numerous for four of the BACs sequenced in *A. lyrata*, *A. cebennensis* and *Boechera divaricarpa*, as discussed above. Transposable elements (TEs) might have caused some of the duplications. TEs appear to be active, as they even interrupt ORFs; note the case of one TRL from *A. thaliana*, At2g29170, where one LTR-retrotransposon is inserted into intron 2 (Figure 1.2). This gene might have become a pseudogene after this insertion, which is supported by the fact that it has no unique MPSS signature, and microarray data in Genevestigator show slight expression throughout plant developmental stages (not shown). We need experimental confirmation by qRT-PCR to prove this last observation.

A possibility is that selection tolerates transposon insertions into ORFs if these are redundant. Particularly in selfing species, we would expect efficient removal of transposon insertions, as discussed in (85). However, *A. thaliana* and *A. cebennensis*, both selfers, do contain more transposon insertions than do other species in the genomic region sequenced. In contrast, *Capsella rubella*, a highly selfing species, contains only one repetitive element in the same sequenced region. Therefore, retention of transposable elements in the TRL region might not necessarily be explained by differences in recombination, but instead be relict of local genome dynamics.

1.5.3 Is selection playing a role in retaining multiple TRL copies?

The prevailing explanation for the preservation of duplicate genes in plants is functional diversification of genes and proteins (86; 87). TRLs are more probably involved in plant secondary metabolism (Table 1.1). Gene duplication has been shown to be a mechanism leading to the evolution of new functions in genes from plant secondary metabolism. Some examples include chalcone synthases (88) and MAM genes (19). Differential selection of gene duplicates might occur at different stages, but it has been proposed to occur more often in the early stage of duplication (89). When we looked for differences in selection among orthologous groups, we found that two of the orthologous groups evolve at rates different from the mean (Table 1.5). But the fact that $\omega < 1$ indicates that they nevertheless evolve under negative selection. Furthermore, when varying ω and the number of selected sites (M7 and M8, Table 1.6), we obtained tests of selection that were significant in half of the cases. In most cases, they were significant for $\omega < 1$. We identified only one case of positive selection within orthologous groups (orthologous group 1 in Table 1.6), a group that contains the orthologous genes At2g29150, AlTRL54, BdTRL9 and CrTRL5. This group might have diversified recently and it might be interesting to test experimentally if the function has differentiated among orthologs.

We detected positive selection (Table 1.7), with $\omega > 1$ at the branches leading to both clades B and C (bottom trees in Figure 1.8) after the separation from the 'ancestral' clade (clade A) and the other TRLs at the base of the tree (see Figure 1.3). Therefore, diversification of function in tandemly duplicated TRLs appears to have occurred shortly after the split from the ancestral Brassicaceae TRL sequences. We expected this, since these genes have been preserved in multiple copies over multiple species. Such preservation is more probably because genes have different functions, as we can observe variable expression of TRLs in *A. thaliana* in the available databases((MPSS, 36) and microarray data (37)). After this initial duplication, the orthologs evolved under negative selection, something that appears to be common to duplicate genes, according to (89). The best fit of the branch-site analysis was for MB (Table 1.6), which was always significantly different to MA. We defined two categories of rates of evolution in MB: branches before the duplication (background) evolve neutrally and some under negative selection, and in branches after the duplication (foreground) positive selection occurs.

1.5.4 Possible differential functions of the TRL copies

The PAML analysis appears to support functional divergence after the initial duplication. MPSS and microarray chip databases available online provide a tool to test indirectly for neo-functionalization. Since orthologous groups were subject to negative selection after the initial duplication, we may use *A. thaliana* microarray data to infer the function of the orthologs (Figure 3 in the Introduction). It is necessary to look at more than one method for comparison of expression patterns, especially since MPSS is a more sensitive method than microarrays, which could be subject to cross-hybridization (36). When looking at MPSS and microarray data, we find that the expression patterns of the TRLs vary in different plant tissues throughout development (Figure 3) indicating neo-functionalization at some point in their history. It appears that genes that are found in the 'ancient' clade, At5g06060 and At2g29260 (clade A, Figure 1.10), are constitutively expressed in different developmental stages of the plant, although at different magnitudes in different tissues. Specialization of function would thus occur after the first duplication, which is confirmed by the differential expression observed in the other TRLs, including those on chromosome 1.

Exon-intron counts support differential evolution in coding and non-coding sequences after duplication (Figure 1.6 and Table 1.4). Furthermore, alternative splice variants have been found for two of these genes: At2g29340 and At2g29350 (SAG13) (36) and they do not show similar or correlated expression throughout the plants' development (Figure 3). This shows that function diversification and alternative splice variants can evolve in parallel lineages, since both genes have evolved independently (see Figure 1.3), despite their physical proximity. Further functional experiments, for instance, mutant screening, are needed to confirm the role of these enzymes in the multiple Brassicaceae and the closely related Cleomaceae.

We would expect the elimination of non-functional genes in organisms with a relatively small genome size to be taking place by natural selection, such as *Arabidopsis spp.* and *Capsella rubella* (described in (83)). This form of selection would be evident as a bias for retaining functional genes, something that has been suggested to affect, for instance, sequence evolution of reverse transcriptases from retrotransposons by (90). Since all of the species contain more or less the same number of TRL pseudogenes, except for *A. cebennensis*, a bias for retaining functional genes depending on genome size does not appear present. A small test of correlation between the available genome sizes and pseudogene number shows that this correlation does not appear to be present ($Pearson r = -0.5$).

1.5.5 Inferences of possible TRL functions from combined data

TRLs appear to respond to environmental stimuli and may be involved in signalling pathways (Table 1.1). It is not clear if TRLs can be redundant in function, as some have been identified as responding to the same factors (Table 1.1). Although tropinone reductases were initially described in plants that produce alkaloids as enzymes leading to the synthesis of these compounds, a recent phylogenetic study of the presence of alkaloids in the Brassicaceae by (91) shows that *A. thaliana* and other Brassicaceae have either lost or never gained the ability to produce tropane alkaloids. Unfortunately, no tests for tropane alkaloids were performed on any of the other species studied and therefore we do not know if the presence of the TRLs in these plants is related to tropane alkaloid production. *A. thaliana* might have lost its capacity to produce alkaloids after losing one of the upstream enzymes necessary for conveting of ornithine to putrescine, ornithine-decarbolxylase (ODC), as shown in (30), which might limit the production of putrescine, a precursor of tropine. This is consistent with evidence from chemical studies, which show that synthesis of tropane alkaloids is less efficient from arginine than from ornithine (92).

If the TRL enzymes in these plants were never used for alkaloid synthesis, another possibility is that they all play intermediary roles in similar pathways. The presence of TRL enzymes, might have conferred a flexibility used to respond to stress. Plants with small genomes such as *Arabidopsis spp.*, have lost large fractions of the genome. So enzymes of similar function might have been replaced by the TRLs.

Duplicate genes involved in secondary metabolism and plant defense often show differential constraints; these are viewed as positive selection in (19) or as differential selection in (93). We did not expect to find gene redundancy, as TRLs are very variable. Redundancy has been observed primarily in developmental genes and genes of primary metabolism (22; 23), since these genes are essential in an organism's life.

Our results show that positive selection precedes duplications of secondary metabolism genes. Once the duplications happen, negative selection is responsible for retaining these genes in large numbers. This would be the case of the TRLs in the Brassicaceae and, probably also, in the Cleomaceae, represented in our study by the species *Cleome spinosa*. We think that TRLs have also acquired different functional importance after the initial duplication in this last species and have therefore been retained in multiple copies. It is clear from the phylogeny that these genes duplicated and evolved independently of Brassicaceae TRLs. Parallel duplication together with function diversification has previously been shown to occur in yeast gene families by (94) and other plant genes of developmental importance, such as the MADS-Box (95). This study shows that this phenomenon occurs in secondary metabolism genes within a plant family, as had been shown in MAM genes (19).

Most probably, diversification of function happened after the initial duplication, since we found positive selection occurring at this point in the phylogenetic analyses. Also, a brief inspection of MPSS and microarray data in databases, discussed in the previous section, support this hypothesis. A further dataset complementing our interpretation of the microarray data was published recently, the ATTED-II database from Riken (96). Most of the cellular targets of TRLs, were found to have different gene targets (see Table. 1.9).

Table 1.9: Cellular targets and possible functions of TRL.

Locus	Target *	Function †
At1g07440	O,N	putative TR, TDH
At1g07450	O,N	putative TR, TDH
At2g29150	O,C	putative TR, TDH
At2g29170	O,C	SDR family protein, putative TR
At2g29260	C,C	putative TR, TDH
At2g29290	O,C	putative TR, TDH
At2g29300	S,C	putative TR, TDH
At2g29310	S,C	putative TR, TDH
At2g29320	M,P	putative TR, TDH
At2g29330	O,C	putative TR, TDH
At2g29340	O,E	SDR family protein, putative TR
At2g29350	O,C	putative TR, TDH (SAG13)
At2g29360	O,Y	putative TR, TDH
At2g29370	O,C	putative TR, TDH
At2g30670	O,N	putative TR, TDH
At5g06060	O,Y	putative TR, TDH

Adapted from the ATTED-II project; targets are from TargetP and WolfPSORT (refs. in (96) * C=chloroplast ; E= endoplasmic reticulum; N=nuclear; O=others (unknown); P=peroxisome; S= secretory; Y=cytoplasm.
†TR=tropinone reductase; TDH=tropinone dehydrogenase; SDR=short-chain dehydrogenase/reductase.

1.5.6 Conclusions

Our results support the previous hypothesis of (27), namely that small changes in the tropinone reductase proteins might cause large changes in substrate-binding ability. According to the analysis with DIVERGE (Figure 1.10) six type I changes underlie the separation of the gene duplicates in clades B and C. An additional mechanism for the TRL gene family to diversify in expression and cellular gene targets, might be variation in cis-regulation (95). Regulatory variation, together with the changes close to the active sites, would provide new functions, and selection might be willing to retain the duplicate gene copies in the genome.

The region containing the in tandem TRL gene family, which is homologous to chromosome 2 of *A. thaliana*, is on average 102 Kb long (equivalent to 0.4 cM). We would expect the genes to be linked in *A. thaliana*. This region might nevertheless be affected by recombination, especially if the richness of transposable elements and duplicated genes promotes unequal (meiotic) recombination, as has been shown for plant genes tandemly duplicated (12). Increased recombination might be important for plants that are predominantly selfing, such as *A. thaliana*, *A. cebennensis* and *Capsella rubella*.

The role of evolutionary divergence in protein family evolution might not readily be detected by DNA and protein sequence comparison, specially if these genes come from only one species. In this study we show that duplication dynamics over multiple species can explain retainment of multiple copies of a gene. Last but not least, a comparative study of closely related species can distinguish the past and current dynamics of gene families and differential evolution vs. parallel evolution, both evolutionarily distinct processes.

Acknowledgments We thank Mariana Mondragón-Palomino and John Parsch for helpful comments on the manuscript. Henriette Ringys-Beckstein, Domenica Schnabelrauch, Claudia Lange and Claudia Wenderoth for assistance in the lab; and Thomas Mitchell-Olds for providing technical facilities. This work was supported by the Max Planck Society, a grant of the Human Frontiers in Science to KJS and a grant from the German Academic Exchange (DAAD) to ANQ.

Chapter 2

Evolution of non-coding sequences and gene expression after gene duplication: an insight from the tandemly duplicated tropinone-reductase-like gene family in *Brassicaceae*

T. Eichner*, A. Navarro-Quezada*, K.J. Schmid

**These authors contributed equally to this work.*

17th October 2007

Abstract Chapter 2

Background To learn about the consequences of gene duplication on non-coding sequences, we looked for correlated evolution of protein and promoter/intron sequences of the genes from the multi-copy tropinone reductase-like (TRL) gene family. We did phylogenetic reconstructions to compare rates and patterns of *cis*-regulatory and non-coding sequence evolution with changes in TRL coding sequences from six Brassicaceae (*Arabidopsis lyrata*, *Arabidopsis cebennensis*, *Boechera divaricarpa*, *Capsella rubella*, *Brassica rapa*) and an outgroup *Cleome spinosa* using phylogenetic methods. We obtained expression data from *Arabidopsis thaliana* microarray databases to detect co-expression of TRLs, and performed qRT-PCR experiments to detect if expression is correlated or diverges among orthologous and paralogous tropinone reductase-like genes recently separated by speciation (from *Arabidopsis thaliana* and *Arabidopsis lyrata*).

Results We found *cis*-promoter and protein genetic distances to be significantly correlated according to Mantel tests and to share clusters 58% of the time. The clusters in the promoter tree frequently contained recently duplicated orthologous genes. When detecting regulatory motifs using FootPrinter there was less conservation of gene clustering compared to the coding sequences (only 32% of the paired comparisons share motifs). From publicly available microarray data, we learned that expression varies significantly among *A. thaliana* TRL genes subject to 166 experimental conditions, despite correlation of coding and non-coding sequences of TRLs. We found significant differences in the expression of TRLs within *A. thaliana* and *A. lyrata* and among these species when plants were treated with salicylic acid, *Pseudomonas syringae* and cold.

Conclusions From phylogenetic analyses of promoters of the tropinone-reductase-like gene family, we learned that the promoters have not evolved at the same rates as the protein coding regions they regulate, except for genes in species of recent origin ($<$ 10 MY). Expression profiles display different patterns in genes with closely related coding sequences and regulatory motifs. We propose that variability in *cis*-regulatory regions and interaction of transcripts, together with few changes in protein sequences, are needed to provide the variation that allows multiple conserved copies of a gene to be preserved in the genome for more than 20 MY.

2.1 Introduction Chapter 2

The relationship between regulatory and structural evolution has not yet been elucidated (97). Gene families provide interesting material with which to study this question, as gene replicates act as non-allelic variants, on which selection can act independently. Non-allelic variation might be important in organisms in which recombination does not occur frequently, as for instance, in selfing plant species. In this study we analyzed the relationship of proteins with non-coding sequences and expression data for a gene family in seven closely related species of Brassicaceae and a closely related outgroup *Cleome spinosa*. We studied the tropinone-reductase-like (TRL) gene family. A previous study had isolated all the TRL genes in an homeologous region of six species: *Arabidopsis lyrata*, *Arabidopsis cebennensis*, *Boechera divaricarpa*, *Capsella rubella*, *Brassica rapa* and *Cleome spinosa*. TRL enzymes, like alcohol-dehydrogenases (ADH), are short-chain dehydrogenases. TRL enzymes are involved in alkaloid biosynthesis in Solanaceae. In *Arabidopsis thaliana* and other Brassicaceae species, these enzymes display a high degree of conservation ($>$ 64% similarity) but the function is not yet known, since tropane alkaloids have not been detected either in *Arabidopsis thaliana*, or in any of the other studied species (28). TRLs from the studied Brassicaceae have previously been identified to be subject to a birth and death dynamic and positive selection, which predicts the evolution of new functions (neofunctionalization) (14).

Evidence for new functions can be found in divergence of regulatory regions and expression patterns. We looked for a correlation of promoter regions with protein coding regions of TRLs. A correlation would suggest parallel evolution of coding sequences and promoters and/or for partial subfunctions. The absence of correlation would suggest new functions. Diversification at the promoter level can be further explored by identifying the regulatory motifs in upstream regions and displaying them on a phylogeny. As for promoters, intron sequences will be constrained if they provide some important structural and/or biological function. Sub- or neofunctionalization might be evident as a parallel evolution of promoter and proteins, but diversification of regulatory motifs. Another possibility is that genes share regulatory motifs but have diverged after duplication.

The availability of microarray and EST data complements the search for evolutionary important regions, but the relationship of expression with protein evolution is not yet clear. On one hand, experimental efforts have shown that functional evolution of orthologous genes might not be directly related to protein sequence evolution (98). On the other hand, analyses of sequence evolution in fruit-flies and other animal species suggest a correlation of protein and expression (99; 100), and DNA sequence and expression (101). A recently published study comparing structural with regulatory and functional evolution of two species of the nematode *Caenorhabditis* found that protein and *cis*-regulatory evolution in paralogous genes is coupled (102). In another study on plants (103), gene clustering has been found to affect the evolution of promoter sequences and expression.

Diversity in expression contributing to the preservation of gene duplicates has been shown for other plant gene families by (104). This study also shows expression patterns to be helpful in detecting neo- and sub-functionalization events. Taking advantage of well maintained databases of microarray data for *Arabidopsis thaliana*, e.g. (37; 105), we tested the relationship of gene and promoter sequence evolution with expression patterns for this species. We complemented this information with expression data obtained by quantitative RT-PCR of three TRL genes from *A. thaliana* Col-0 and *A. lyrata spp. lyrata* to investigate whether orthologous genes conserve expression among species. Correlating the previous results of coding- and non-coding divergence with diversity in expression levels, we looked for evidence of the evolution of partial subfunctions or completely new functions of the tropinone-reductase-like genes.

2.2 Materials and Methods Chapter 2

2.2.1 Phylogenetic analysis of upstream regulatory regions

Regulatory regions from TRL genes were isolated from previously sequenced and annotated TRLs from six BAC sequences (106) belonging to *Arabidopsis lyrata*, *Arabidopsis cebennensis*, *Boechera divaricarpa*, *Capsella rubella*, *Brassica rapa* and *Cleome spinosa* (Cleomaceae). The annotated sequences were in Genbank format and most of the programs used for analysis are available online. We isolated 72 upstream regions (5') of functional TRL genes (i.e. with complete open reading frames) from the annotated BACs. The maximum length of an upstream region was set to 2 kb in the first analysis. After observing the results from the analysis with FootPrinter (107) (described in the next section), we reduced the upstream region to 1 Kb increasing the number of detected motives per analyzed length. If a gene was found earlier than 1-2 Kb in the 5' region, the region analyzed as a promoter was shorter. Since genes are found in both coding strands, the promoters of genes in the non-sense strand were reverse-complemented. As an outgroup for the phylogenetic reconstruction we used a tropinone- reductase-like gene from rice (Genbank Accession: OJA1364E02.17).

Pairwise alignments of promoters from tandem and segmental duplications were computed using DIALIGN (108; 109). Pairwise alignments were performed to avoid length variation of the alignment and changes in the distance matrices due to insertion of numerous gaps, which occur frequently when aligning sequences with low conservation, such as upstream regions. Two other different alignment methods (T-Coffee (110) and ClustalW (111)) were compared to DIALIGN, but DIALIGN performed better, showing 30% of the time higher bootstrap values (all above the 50 % threshold).

To calculate distances among promoter sequences, we used different approaches: 1) the number of similarities in relation to the alignment length, 2) the number of similarities in relation to the total length of both aligned sequences, and 3) the number of similarities divided by the length of the shorter sequence. The first approach is sensitive to the insertion of gaps, the second will give values close to zero if one of the compared sequences is very short, and the third approach avoids these errors, but will bias the comparison for high similarity, since we observed that upstream regions are more similar the closer they are to the transcription start. Using DIALIGN it was possible to include marked motifs to calculate the distance. The obtained similarities were converted to distances and put into $N \times N$ matrices. To evaluate the distance matrices the program NEIGHBOR (73) from the PHYLIP package was used, which implements the neighbor-joining method (112). We defined the rice sequence as the outgroup. The UPGMA method provided a similar result. To test for the stability of the trees, a bootstrapping was performed by SEQBOOT (73), which produces multiple datasets. After the generation of 1000 bootstrapped matrices using NEIGHBOR, we created a consensus tree using the CONSENSE (73) program.

2.2.2 Identification of regulatory motifs using phylogenetic footprinting

Phylogenetic footprinting is a method used to detect regulatory elements among different species. The method, implemented in the program FootPrinter, is supposed to identify motifs in suspected regulatory regions. FootPrinter (107) does not rely on the information of a multiple alignment, but searches regulatory motifs directly on the sequences, taking into account their phylogenetic relationship. The motif length can be set, as well as the minimal number of mutations allowed for a portion of the tree. For our analysis, we did not allow any mutations to occur, and motif length was set to 12. These parameters detected an amount of motifs that could easily be scored. We checked if regulatory regions were inverted, using dot-plots to compare promoter regions of paralogs (not shown).

2.2.3 Analysis of intron regions

To investigate the relationship of intron and protein sequence evolution, we isolated all the introns from the annotated TRL genes and performed pairwise alignments in the same way as for the upstream regions. We performed an AVID alignment-based VISTA analysis (113) and phylogenetic analysis in the same way as for the promoter sequences, to test if intronic sequences are conserved among species. VISTA is a program for visualizing several DNA sequence alignments of arbitrary length on the same scale. Conserved coding (exon) and non-coding (CNS) as well as untranslated (UTR) sequences are marked with different colors and percent identity is displayed on the y-axis (see Figure 2.4). We used GenomeVista with the *Arabidopsis thaliana* genome as a reference. Comparative sequence analysis has enabled regulatory non-coding regions and location of coding exons to be identified and located. Distances of intronic sequences were calculated and phylogenetic reconstructions were performed in the same way as for the promoters. We compared branch lengths of the phylogenetic reconstructions within orthologous groups.

2.2.4 Relationship of protein and non-coding sequences

To test for correlated evolution of protein amino acid sequences with promoter sequences we counted the number of groups that contained the orthologous genes. We refer to these as conserved clusters of orthologs, **CCOs**. To detect correlation between coding and non-coding sequences, we performed a Mantel test using the program zt (114) of the distance matrices obtained for promoters and introns. The following options were used: -s for simple Mantel test, and 1000 random permutations.

2.2.5 Expression analysis of the TRL genes using AtGenExpress Data

For the expression analyses we obtained microarray data from the AtGenExpress database at http://www.weigelworld.org for the ATH1 Chip (22,810 genes). We only used data from experiments on wild types of *A. thaliana* var. Col-0. Around 300 experiments could be found in this database that proved useful for testing differences in expression due to abiotic and biotic stress. From the same site, we obtained descriptions of the experiments, that are needed for the standardization of the data as described in (115). Downloaded data were already *gcRMA* normalized (116; 117). The gcRMA function normalizes the data using the Robust Multi-Array Average (RMA) expression measure taking into account the gene chip variation in probe hybridization (117). We used directly the data for further standarizing relative expression, as gcRMA normalized data are very robust. It is necessary to 1) calculate of the geometrical mean for the three replicates of each hybridization experiment and save the means into a MySQL database, and 2) obtain the ratio $T_i = R_i/G_i$ (118), where R_i is the value of the probe and G_i is the reference or control experiment (e.g. darkness for light induction). To find genes that were significantly differently expressed, the arithmetic mean μ and standard deviation σ were calculated from the $log_2(T_i)$ distribution, where T_i is the relative expression. Values lying above the $1,96 \cdot \sigma$ (95 % confidential) from the mean are considered as differentially expressed and were visualized in an R-I (ratio vs. intensity) plot (shown in Figure 2.10). Furthermore, Z-values were estimated using a sliding window approach, with a window size of 500 genes along the x-axis. For all N_{local} datapoints within the window the mean and standard deviation were calculated using equation 2.1(modified from (115)).

$$\sigma^{local}_{log_2(T_i)} = \sqrt{\frac{1}{N_{local}-1}\sum_{j=1}^{N_{local}}(log_2(T_i)_j - \overline{log_2(T_i)})^2} \tag{2.1}$$

Z-values for standard deviations of the data points from the mean were estimated as described in (118). To estimate Z-values for 22,810 genes for each of the experiments, we used the previously generated MySQL database. The significantly differently expressed genes were those with $\left|Z_i^{local}\right| > 1.96$.

2.2.6 Co-expression analysis of the TRL genes

We calculated the Euclidean distance $D(A, B)$ between all TRL genes on the ATH1 chip. The resulting distance matrices were normalized using all the TRL genes. The transformed distance matrix was visualized as a distance tree using SEQBOOT, NEIGHBOR and CONSENSE (73), as described previously for the promoter trees. Clades with bootstrap values over 70 % were further analyzed as candidates for genes that are co-regulated by Pearson correlation coefficients (PCC).

PCCs were estimated as described in (115) to analyze the relationship between elements in the matrices. To compare PCCs and Euclidean distance reconstructions, we used the program Treejuxtaposer (119) which displays differences in a graphic way (using default parameters). This software identifies branches where the trees differ and marks them in red.

Another way of visualizing correlated expression is using Z-value plots, which we did for all possible co-regulated TRLs, as shown in (115). To further detect genes co-regulated with TRL genes, we calculated the Euclidean distance $D(A, B)$ between all genes on the ATH1 chip.

2.2.7 Expression analysis of three TRL genes with quantitative real-time PCR

We analyzed expression of three TRL genes from *A. thaliana* and *A. lyrata* to gain insight into the correlation of differences in gene- and promoter regions with differences in gene expression. We chose genes that had been reported in the *A. thaliana* literature to be associated with senescence (At2g29350 or SAG13) or responsive to salicylic acid (33; 32), pathogens (At2g29350, upregulated and At2g29290, downregulated) (35), or cold treatment (At2g29340) (61). These genes were also chosen because their microarray expression profiles displayed either a negative (At2g29350 and At2g29290) or positive (At2g29350 and At2g29340) PCC.

Primers for the genes At2g29290 and At2g29350 from *A. thaliana* were designed using the software Beacon Designer (Palo Alto, CA, USA). The primer sequence for At2g29340 (CATMA2a27740) was obtained from the CATMA project (120). Primers were designed for all *A. thaliana* TRL orthologs in *A. lyrata spp. lyrata*, AlTRL29 and AlTRL37, respectively. PCR products were cloned and sequenced to confirm the gene specificity of each primer, and only gene-specific primers were used. We failed to find specific primers for AlTRL35, the ortholog of At2g29340. Instead, data were obtained for another *A. lyrata spp. lyrata* TRL gene (AlTRL54, orthologous to At2g29150) to have another independent sample.

A. lyrata spp. lyrata seeds were obtained from Dr. Maria Clauss (Max Planck Institute for Chemical Ecology). Seeds for this species were kept on wet filters in the fridge for three weeks prior to germination. Both *A. lyrata spp. lyrata* and *A. thaliana* Col-0 seeds were planted and kept at 4°C for three days to aid germination; after that the pots were transferred to short-day growth chambers at 21°C. Five-week-old plants were cold-treated at 4°C and samples were taken at time intervals according to those described in AtGenExpress: 1, 6, 12 and 24 h (105). AtGenExpress time intervals were also used for salicylic acid (SA) treatment (4, 28 and 52h). The treatment involved spraying salicylic acid at 0.3 mM dissolved in 0.02% silwet (pH 7), controls were sprayed with 0.02% silwet only. *Pseudomonas syringae* DC3000 was kindly provided by Dr. Beate Völksch (Friedrich-Schiller-University, Jena, Germany). We injected the overnight stationary culture into 5-week-old-plant leaves at a concentration of $1 \times 10^{-}6$ bacteria/ml using a syringe without a needle, according to Dong et al. (121). Time collection points for pathogen treated leaves were 6, 12 and 24 h.

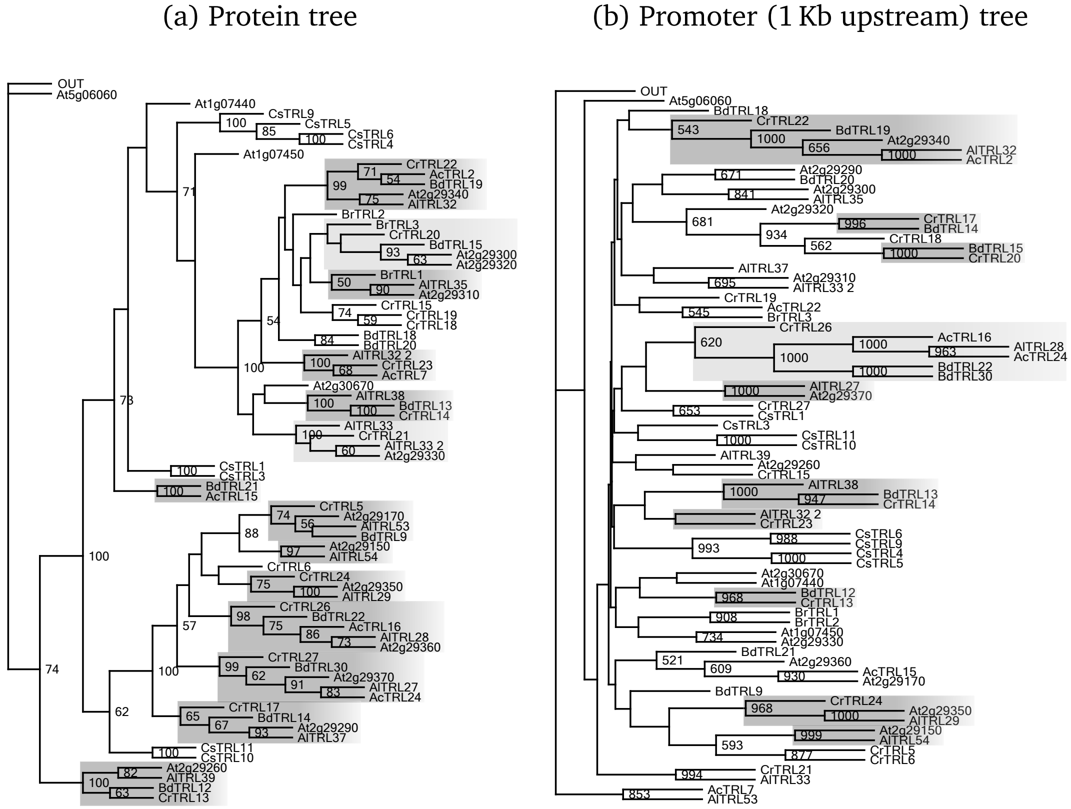
Total RNA was extracted from leaves using the RNA Plant Mini Kit from QIAGEN including DNAse for on-column digestion (QIAGEN). The OmniscriptRT Kit from QIAGEN was used for first strand cDNA synthesis. RNAseOUT (Invitrogen Karlsruhe, Germany) was added to the reaction. We performed quantitative (qPCR) using the Brilliant SYBR $Green\overline{TM}$ (Molecular Probes, Eugene, Oregon, USA) and followed the protocol described in the manual of the kit (Brilliant SYBR Green Stratagene, La Jolla, CA, USA). Primers used for qPCR were HPLC purified, and the final concentration in the reaction was 100nM. We measured efficiency for each primer and determined relative change using the efficiency corrected model (122). Reference genes used were adenine phosphoribosyltransferase (APT1) or RNA polymerase 2 large subunit (RP2ls). Both demonstrated stable expression between control and treatment conditions to within $\pm 2C_{ts}$ (122).

2.3 Results Chapter 2

2.3.1 Comparative genomics of TRL promoters

Promoter sequences are assumed to evolve faster than protein coding sequences (123; 124) and sometimes independently from each other- with the exception of transcription factor binding sites (124; 125). Furthermore, upstream regions are not necessarily conserved, and therefore inappropriate for global alignments. Therefore, obtaining a correct alignment was necessary to build the phylogenetic tree. By evaluating different alignment methods, we found the highest number of clades sharing orthologs using the DIALIGN-algorithm. This was the case for both protein- and promotor- based trees (126). The phylogenetic relationships of the TRL protein sequences (Figure 2.1) allowed the comparisons between coding and non-coding sequences to be performed. We found that although 10 conserved-clusters of orthologs (**CCO**'s) out of 14 are delineated, these include fewer genes than the equivalent clusters in the protein tree. Only 58% of the orthologous genes are also orthologous at the promoters. This is in agreement with faster evolutionary rates expected for promoter sequences. Also, as in the protein tree, the *Cleome spinosa* promoters all cluster apart from the *Brassicaceae* genes, confirming independent duplication and evolution of the TRLs after the plant families split, as observed in a previous study (106).

In the analysis using the Mantel test we observed that all matrices showed significant correlations (Table 2.3). The best correlation was between the protein and intron 4 distance matrices ($r = 0.5809$). The intron 4 tree had the second largest number of **CCO**'s containing 60% of the genes (9 out of 14, Figure 2.2). The second best correlation is with the promoter ($r = 0.4216$), which is confirmed by the presence of 10 **CCO**s in the promoter tree (out of 14, Figure 2.1). All positive correlation values are significant according to a one-tailed *Z-distribution* (number of permutations=1000) test, we expect the best correlation to be close to one. Therefore, non-coding sequences appear to be more similar to their protein sequences than expected by chance.

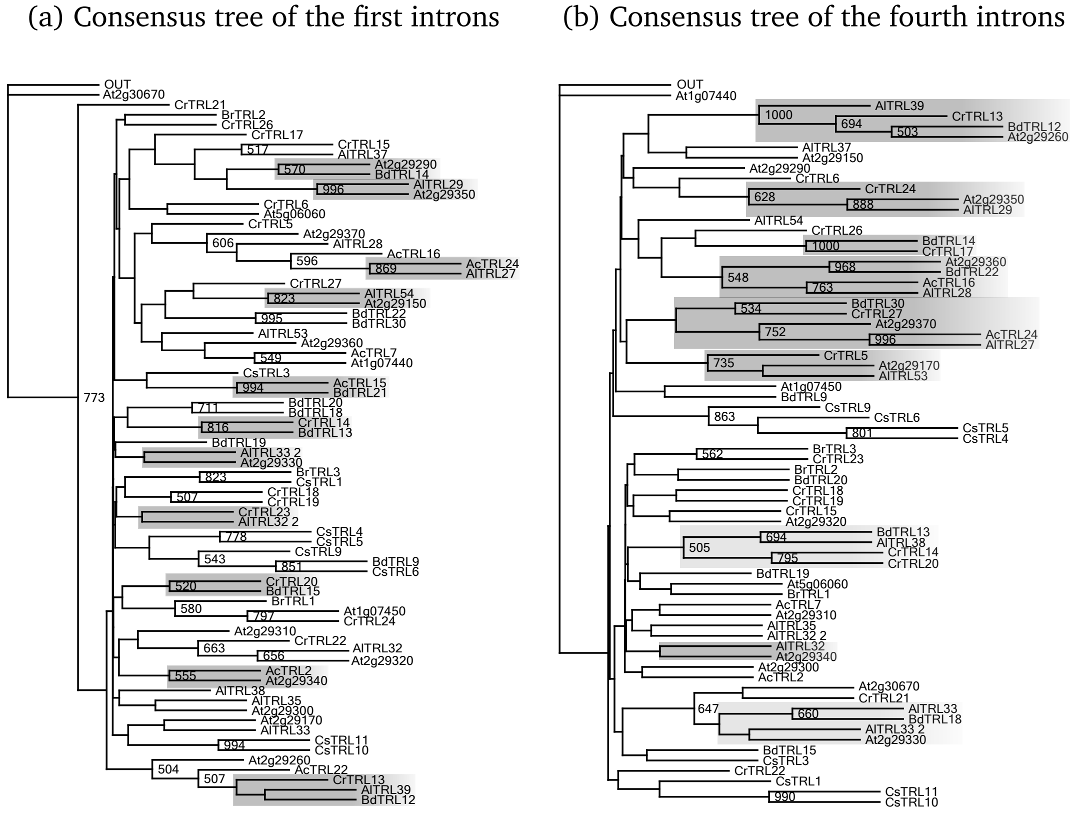

(a) Protein tree

(b) Promoter (1 Kb upstream) tree

Figure 2.1: Phylogenetic relationship of TRL genes and promoters. Orthologous groups, defined as phylogenetic reconstruction of TRL genes, are shaded gray. Dark gray shading demarcates panorthologs, light-gray shading imparalogs.

Table 2.1: Intronic lengths and branch length comparison method within orthologous groups. Longer branches and more variability in length are equivalent to fast evolution, because of indels.

	Intron 1	**Intron 2**	**Intron 3**	**Intron 4**
Average intronic length (bp)	128.34	137.61	147.03	99.43
STD Deviation	71.46	48.84	227.61	19.07
Average branch length	0.207	0.193	0.204	0.188


(a) Consensus tree of the first introns

(b) Consensus tree of the fourth introns

Figure 2.2: Phylogenetic relationship of TRL gene introns 1 and 4. Orthologous groups, defined on phylogenetic reconstruction of TRL genes, are shaded gray. Dark gray shading demarcates panorthologs, light-gray shading imparalogs.

An interesting group, since it appeared to conserve its clustering in proteins, intron 4 and promoter tree, is the one including the At2g29340 paralogs (upper clade in Figure 2.1). In the FootPrinter analysis the five genes forming this clade have at least two shared motifs. One of the identified motifs is the POLLEN1LELAT52, which is expressed in pollen development, indicating TRLs play a role in development. This is confirmed by expression data of At2g29340 from the AtGenExpress database (105), where this gene displays higher than the mean normalized expression in the flower organs (Figure 2.3).

2.3.2 Comparative genomics of TRL intronic regions

Intron sequences have been shown to contain regulatory sequences (127; 128), so we tested for correlated evolution of coding sequences and introns by using phylogenetic methods. TRL genes depict various numbers and lengths of the gene components (Table 2.1); most TRL genes display the pattern of 5 exons and 4 introns, but 12 % of the genes analyzed have a different exon- intron count (Navarro-Quezada, unpublished). Two TRLs of *A. thaliana* have been shown to have alternative splicing of the first intron retention type, and this might restrict the evolution of intron 1 (129). The Mantel test of correlation (130) of the distance matrices, shows the best correlation to be between the coding sequence and intron 4 (r=0.581), but intron 1 also showed a relatively high correlation (r=0.402). Conservation of intron 4 is also supported in the VISTA analysis. A VISTA alignment of *A. thaliana* and *Capsella rubella* is depicted in Figure 2.4. Intron 4 has the lowest variation in length but also appears to evolve slowly (Table. 2.1); therefore indels might be restricted for this intron, indicating structural importance.

GenomeVista made possible a visual and qualitative comparison from which we observed that conserved non-coding sequences (CNS) are shared for both UTRs and introns between most of the Brassicaceae and *A. thaliana*. The number of CNS decreased with phylogenetic distance to the reference species (see Brassicaceae phylogeny in Figure 1.1, Chapter 1), as we had expected. Intron 4 was often conserved among the *A. thaliana* genome and the analyzed Brassicaceae. It was

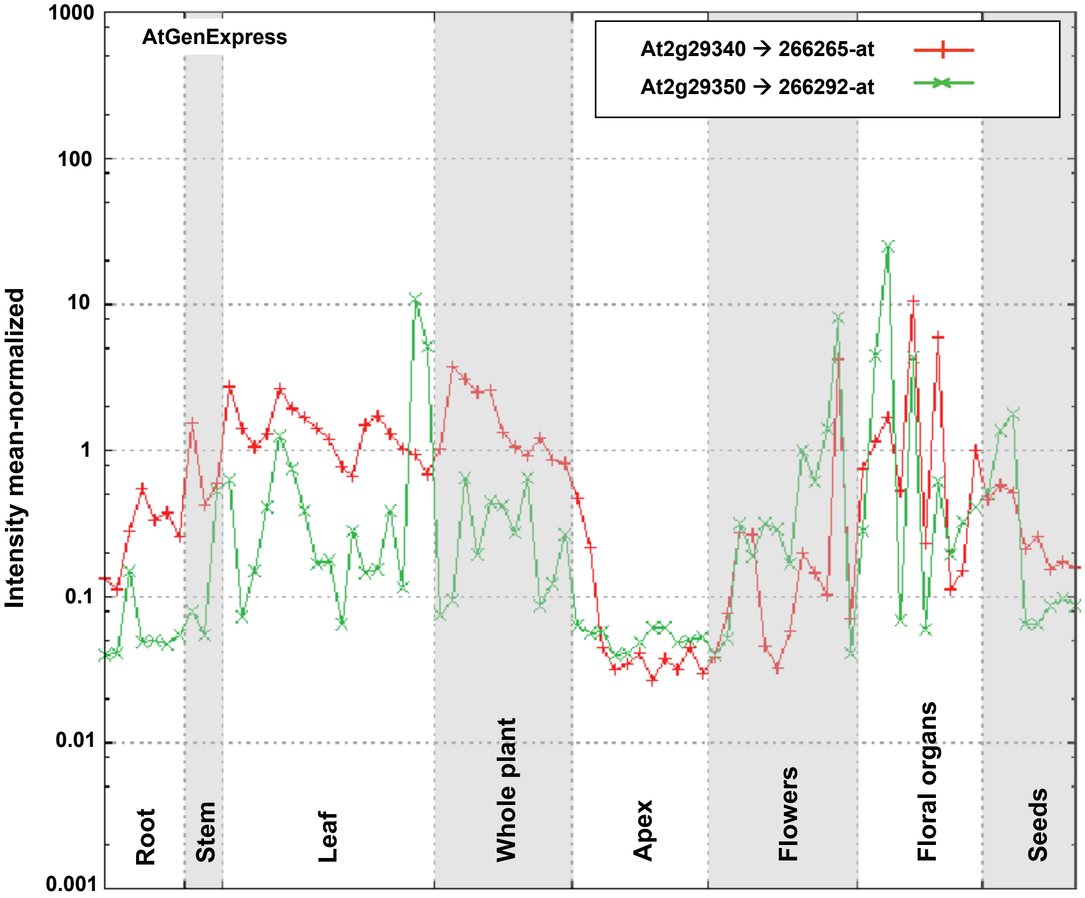


Figure 2.3: Mean normalized values of two co-regulated genes: At2g29340 and At2g29350 throughout the plants development. Data and graph were obtained from the AtGenExpress Visualization Tool (AVT) at (105).

Conserved intron count						
Species	intron 1	intron 2	intron 3	intron 4	intron 5	
A. lyrata	5	5	3	6	0	
A. cebennensis	1	-	1	4	0	
Boechera divaricarpa	3	2	3	3	1	
Capsella rubella	5	3	4	6	0	
Brassica rapa	-	-	-	1	0	
total	14	10	11	20	1	56

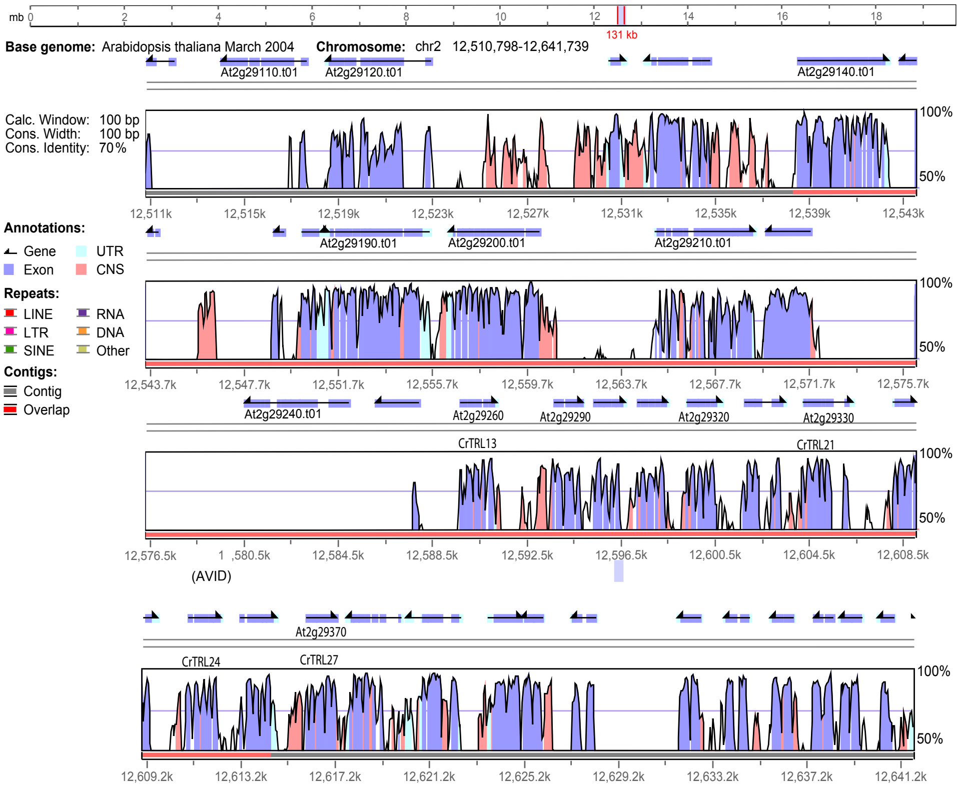


Figure 2.4: Identity graphs from the comparison of the sequenced BAC from *Capsella rubella* and *A. thaliana* using GenomeVista. Genes from *A. thaliana* are blue bars above the identity graphs. Percent indentity is plotted on the y-axis. The window-size is 100 bp. TRL gene names were added by the authors.

roughly twice as conserved than introns 2 and 3 (20 vs. 10 and 11 out of 56 times). 25 % of the times intron 1 was conserved, and only once did we find a conserved fifth intron. The species *Cleome spinosa* shared the fewest CNS with *A. thaliana* as expected.

2.3.3 Identification of regulatory elements by phylogenetic shadowing

In the FootPrinter and PLACE analyses we expected to identify shared regulatory elements among recently duplicated genes and among orthologs. From the tree and figure obtained with FootPrinter (Figure 2.5), we see that motifs identified are shared either within orthologous groups, and more often among closely related gene duplicates (17 out of the 51 shared motif stretches, 32%). The genes belonging to *A. thaliana* and *A. lyrata* shared regulatory motifs more often than the other species (7 out of 17 shared motifs). Unrelated genes did not share motifs among them.

2.3.4 Analysis of co-expression of TRLs

The consensus tree obtained for the expression of TRL genes in *Arabidopsis thaliana* (Figure 2.6) depicts three terminal clades with bootstrap values over 70 %. These are the clades which were further analyzed as candidates for genes that are co-regulated using Z-value plots over experiments. We found genes displaying PCC correlation values close to $+1$, whenever both genes had almost null expression 2.2. This effect was lower though, when more experiments were added to the analysis. We confirmed that the same gene pairs cluster together using PCC to Euclidean distance trees among expression values (Figure 2.7). Both genes At2g29290 and At2g29310 show the best clustering value, with a bootstrap value of 98,7 % (Figure 2.6), and the highest PCC value (Table 2.2, p=0.0011). Furthermore, when plotting *Z*-values over all experiments for both TRL genes (Figure 2.8), we can see a correlation in the expression pattern of these two genes for all the ana-

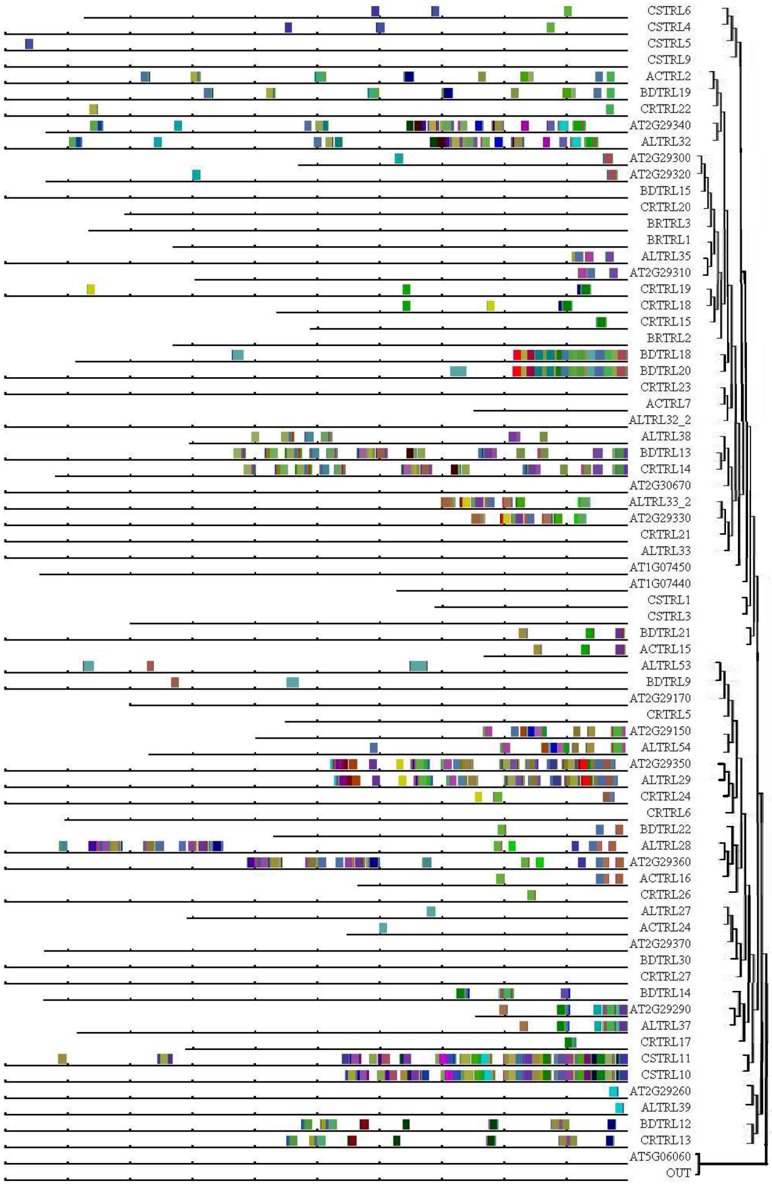


Figure 2.5: Motifs identified with FootPrinter. We identified 362 motives, using a minimal length of 12 nucleotides and no mutation. On the right, the promotor tree is displayed.

lyzed experiments. A good PCC value is also found for the cluster of At2g29340 and At2g29350 (Table 2.2, p=0.0059), but correlated expression in the Z-value plot is less evident (Figure 2.8).

	At5g06060	At1g07450	At1g07440	At2g29340	At2g29370	At2g29330	At2g29310	At2g29300	At2g29290	At2g29260	At2g29170	At2g29150	At2g29320	At2g29350	At2g29360	At2g30670
At5g06060	1	0.37495	0.26612	0.27148	0.25096	0.25521	0.38788	0.00106	0.39567	0.16013	0.14942	-0.01653	0.34157	-0.17633	0.32958	0.20174
At1g07450	0.37495	1	0.18100	0.04504	0.13786	0.08348	0.10252	0.09650	0.25471	0.02217	0.11800	0.12045	0.27424	-0.17579	0.27679	0.20678
At1g07440	0.26612	0.18100	1	0.17294	0.29467	0.06907	0.28467	0.31078	0.40585	0.14567	0.18899	0.29198	0.18265	-0.18404	0.17006	-0.01244
At2g29340	0.27148	0.04504	0.17294	1	0.17780	0.03410	0.21145	0.19784	0.17313	0.06547	0.00028	0.04445	0.19230	0.31996	0.07461	0.19394
At2g29370	0.25096	0.13786	0.29467	0.17780	1	0.02360	0.24817	0.33977	0.16980	0.14774	0.23631	0.23875	0.04871	-0.00480	0.14375	0.06906
At2g29330	0.25521	0.08348	0.06907	0.03410	0.02360	1	-0.15383	0.03110	-0.11198	0.15016	0.14302	-0.18999	0.09650	-0.30925	0.01511	0.09738
At2g29310	0.38788	0.10252	0.28467	0.21145	0.24817	-0.15383	1	0.09277	0.66399	0.01679	0.05761	0.10245	0.22252	-0.12834	0.26405	0.04432
At2g29300	0.00106	0.09650	0.31078	0.19784	0.33977	0.03110	0.09277	1	0.32301	-0.02013	0.26804	0.17124	0.02030	-0.13141	0.10940	-0.05843
At2g29290	0.39567	0.25471	0.40585	0.17313	0.16980	-0.11198	0.66399	0.32301	1	0.05541	0.25192	0.14706	0.30433	-0.23073	0.39760	0.02471
At2g29260	0.16013	0.02217	0.14567	0.06547	0.14774	0.15016	0.01679	-0.02013	0.05541	1	-0.05332	-0.03558	0.06936	-0.09999	0.11907	0.00489
At2g29170	0.14942	0.11800	0.18899	0.00028	0.23631	0.14302	0.05761	0.26804	0.25192	-0.05332	1	0.34945	-0.04606	-0.10442	0.19547	-0.06478
At2g29150	-0.01653	0.12045	0.29198	0.04445	0.23875	-0.18999	0.10245	0.17124	0.14706	-0.03558	0.34945	1	-0.05049	0.17722	0.10027	0.09923
At2g29320	0.34157	0.27424	0.18265	0.19230	0.04871	0.09650	0.22252	0.02030	0.30433	0.06936	-0.04606	-0.05049	1	-0.11922	0.22180	0.01922
At2g29350	-0.17633	-0.17579	-0.18404	0.31996	-0.00480	-0.30925	-0.12834	-0.13141	-0.23073	-0.09999	-0.10442	0.17722	-0.11922	1	-0.16567	0.26489
At2g29360	0.32958	0.27679	0.17006	0.07461	0.14375	0.01511	0.26405	0.10940	0.39760	0.11907	0.19547	0.10027	0.22180	-0.16567	1	0.00701
At2g30670	0.20174	0.20678	-0.01244	0.19394	0.06906	0.09738	0.04432	-0.05843	0.02471	0.00489	-0.06478	0.09923	0.01922	0.26489	0.00701	1

Table 2.2: Comparison of normalized PCC values among genes. Different degrees of red indicate co-regulation, different degrees of blue mean antagonistic expression.

In the analysis of PCCs including all the *A. thaliana* genes available on the ATH1 chip, we identified 40 genes in the genome that have a stronger PCC correlation with a TRL gene than any of the correlated pairs of TRLs (not shown). Our results are confirmed by online databases (see discussion).

Gene expression ratios for TRLs in co-regulated clusters helped us identify experimental factors that induce TRL genes significantly. The expression profile (*Z*-values) of At2g29340 and At2g29350 shows that this pair of genes with a significant PCC value responds similarly to pathogen induction, wounding and heat stress, but it appears to be expressed antagonistically in the rest of the experiments (Figure 2.8). We chose this pair of genes for further study in qPCR experiments to establish the relationship of these genes under three experimental conditions (see next paragraph). We identified significant induction of At2g29290 and At2g29310 by light of different wavelengths (Figure 2.8). Individually, At2g29290 appears to respond significantly to pathogen induction and cold, and to be downregulated

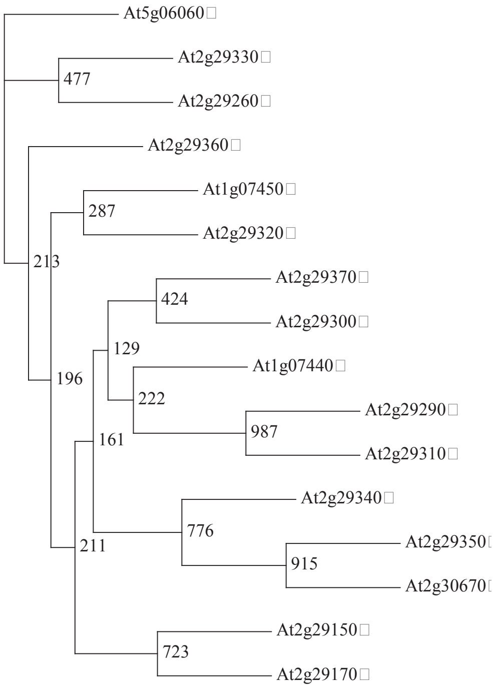
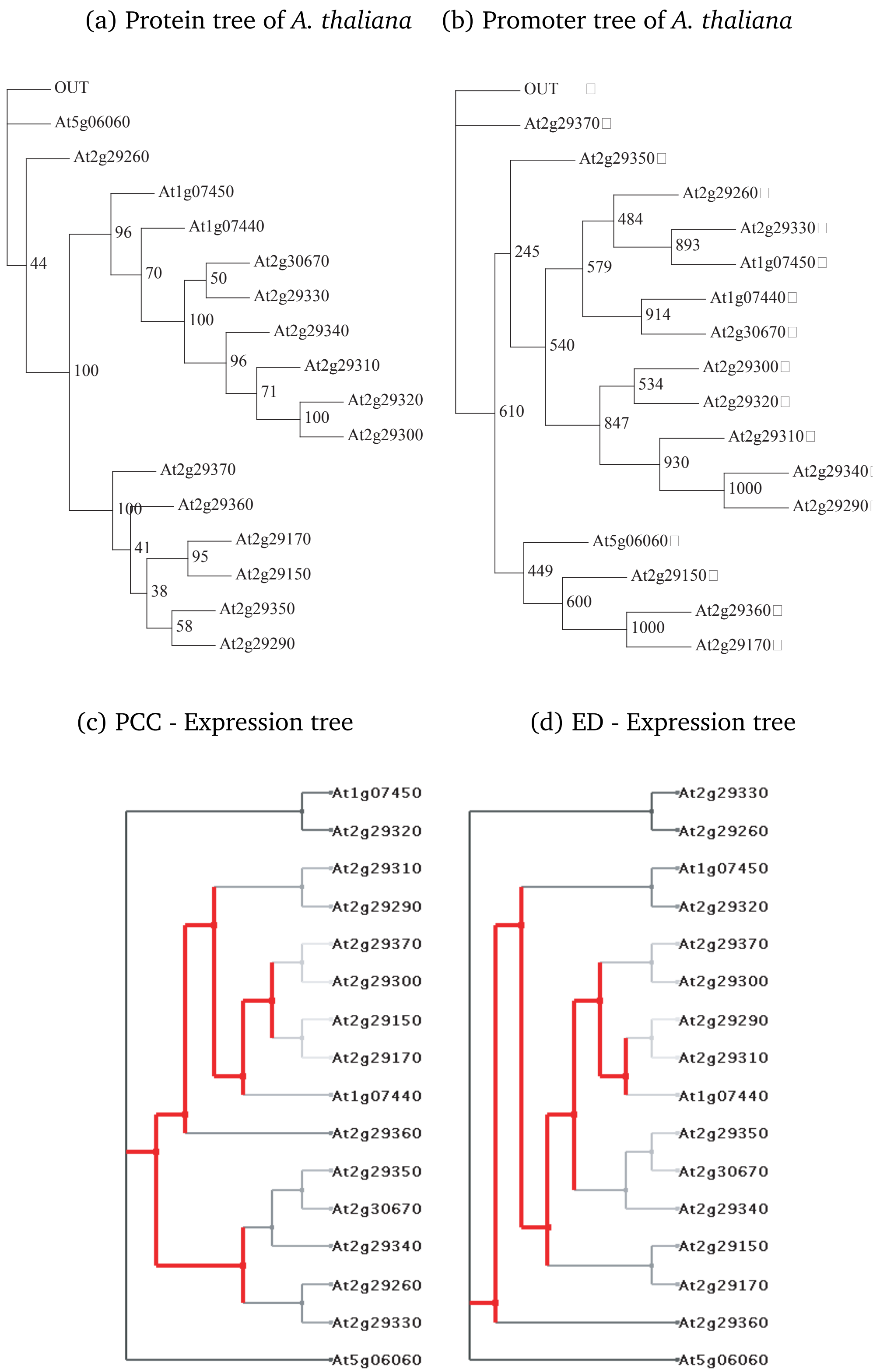


Figure 2.6: Neighbor joining consensus tree of the TRL expression profiles. Consensus tree based on Euclidean distances of *Z*-values, bootstrap is n=10000. Only four nodes display values >500. These are co-regulated genes that also had high PCC values and keep co-expression over multiple experiments, as shown later in paired expression profiles (Figure 2.8).


(a) Protein tree of *A. thaliana*

(b) Promoter tree of *A. thaliana*

(c) PCC - Expression tree

(d) ED - Expression tree

Figure 2.7: Comparison of expression trees based on PCCs and Euclidean distance of *Z*-values. The program Treejuxtaposer displays differences in red. Independent of the method, the terminal nodes of both trees are the same.

by osmotic stress, whereas At2g29310 is upregulated significantly by genotoxic, oxidative stress and UV-B.

We found that different types of stress changed the expression of TRLs (light induction, hormone and pathogen treatments). The effects of different treatments on pairwise recently duplicated genes of *A. thaliana* are summarized in Figure 2.11.

2.3.5 Transcript abundance of TRLs following pathogen stress, cold stress or salicylic acid (SA) treatment

A. lyrata is an outcrosser and therefore displays a higher number of heterozygous loci. It proved difficult to find ideal primers for amplifying single primer cDNAs, which in turn made it difficult to evaluate expression. Nevertheless, we found two primers for orthologous genes in *A. thaliana* that worked well in this species. An extra primer pair was tested for which we did not have the orthologous gene in *A. thaliana*. There is evidence that at least one of the TRLs, At2g29350 associated with cell death (*SAG13*) which is one of the best-known markers for senescence (33) is regulated by SA (131). Furthermore, induction by cold and heat in *Arabidopsis thaliana* has previously been reported (35; 61), consistent with the microarray data (Figure 2.8, and Figure 2.12). Given that *Pseudomonas syringae* infection induces senescence and SA response (121), it is not surprising that both treatments elicited expression of TRLs. After one to three days of being infected with *Pseudomonas syringae*, plants of both species showed yellow and brownish spots, signs of senescence, on the inoculated leafs. *A. lyrata* TRLs do not appear to respond in the same way as *A. thaliana* when induced by the pathogen, except for AlTRL54 after 24h of elicitation (Figure 2.9). All *A. lyrata* TRL genes tested responded to induction by SA after 28 and 52 hours respectively. AlTRl29 did not show any response to *Pseudomonas syringae* but was upregulated by SA in the qPCR experiments, whereas its ortholog, At2g29350, responded to SA and *Pseudomonas syringae* after 4h and 6h, respectively, of exposure.

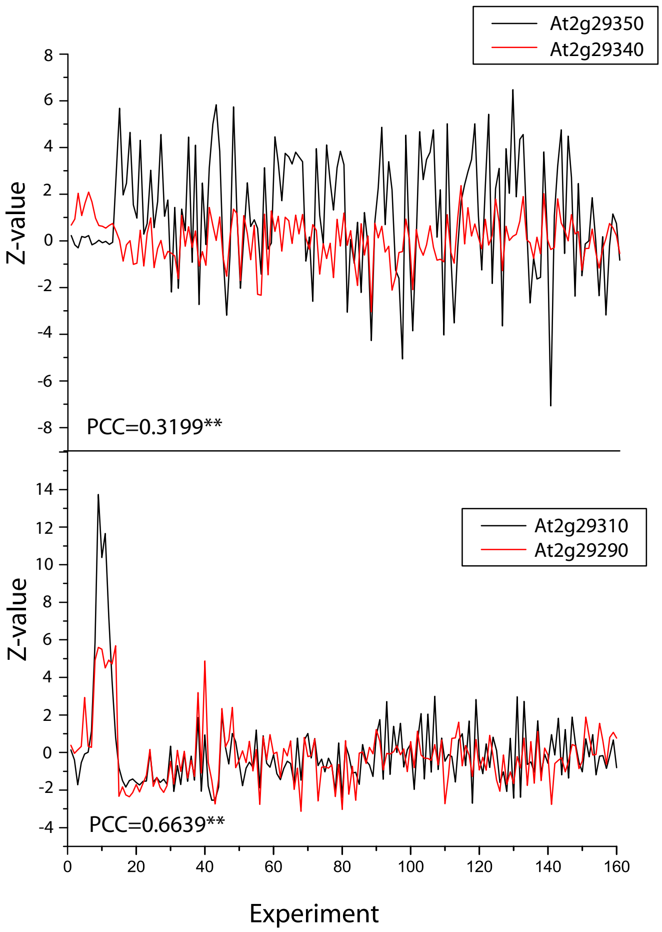


Figure 2.8: Expression profiles of corrected Z-values for microarray data from databases. Expression profiles shown are for A) co-regulated genes At2g29340 and At2g29350 and B) co-regulated genes At2g29290 and At2g29310. For A) there appears to be no significant correlation between these two genes, which appear rather antagonistic. In B) both genes show the best PCC correlation and highly similar expression profiles. They are significantly upregulated as a response to light (experiments 1-14). Other experiments are: 15- 44: pathogen induction; 45-56: cold; 57-68: osmotic; 69-80: salt; 81-94: drought; 95-106: genotoxic; 107-118: oxidative; 119-132: UV-B; 133-146: wounding; and 146-162: heat.

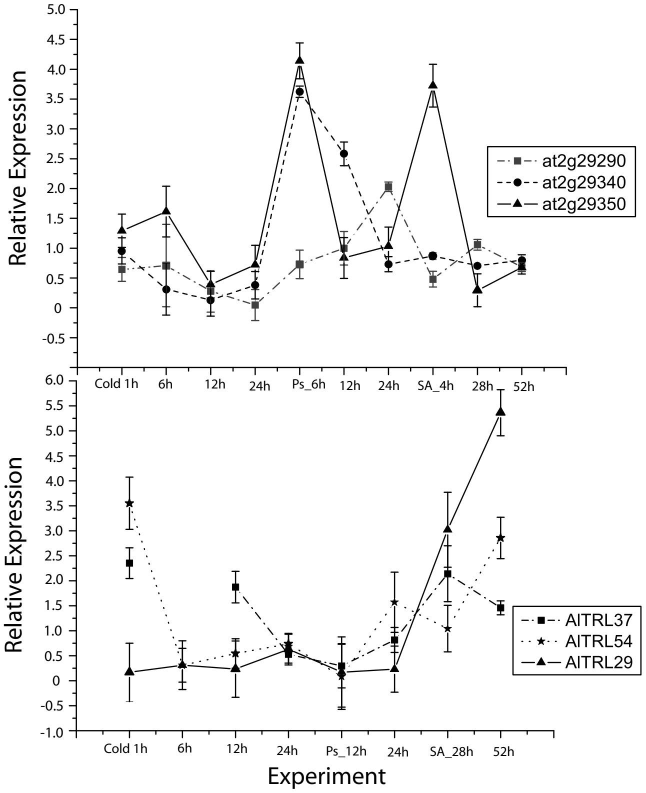


Figure 2.9: Relative expression of At2g29340, At2g29350 and At2g29290 (Upper panel) from *A. thaliana* and AlTRL29, AlTRL37 and AlTRL54 (bottom) from *A. lyrata* obtained by qPCR experiments. Three biological replicates and three technical replicates are averaged for each time-point; error bars display standard erros. Orthologs have the same symbols (At2g29350 and AlTRL29; At2g29290 and AlTRL37). Relative fold change to reference genes *ATP1* or *RP2ls* was estimated with the efficiency corrected model (122). At2g29350 responded to salicylic acid (SA) and *Pseudomonas syringae* DC3000, confirming microarray results. Its ortholog in *A. lyrata* responded only to SA. At2g29340 and At2g29290 also responded to *Pseudomonas* but at different times; the latter showing an expression profile opposite to At2g29350. *A. lyrata* TRLs showed responses to cold stress and to SA but *Pseudomonas* failed to significantly induce TRLs in this species.

The cold response of TRLs from *A. thaliana* correlated well with the microarray results. Most of the genes were either not differentially expressed or were downregulated. We did not find, as in the microarrays, that At2g29290 was upregulated after 1h of cold treatment. Its ortholog in *A. lyrata*, AlTRL37, was upregulated after 1h of cold, as was AlTRL54, but neither gene responded to longer exposures to cold.

2.4 Discussion Chapter 2

2.4.1 Identifying correlated evolution of proteins, *cis*-regulatory regions and other non-coding sequences

Protein sequence homology alone is insufficient for predicting of specific enzyme function. Therefore, more information about the genes and their evolution was necessary to provide insight into TRL expression and their associations. To investigate if coding and non-coding evolution are related in the TRL gene family, we compared phylogenetic relationships of protein, *cis*-regulatory regions (adjacent promoter) and intronic sequences. Phylogenetic analysis of promoters is still one of the methods of choice in bioinformatics, and phylogenetic trees are also used for studying co-evolution. Our hypothesis was that promoters and proteins would be evolving similarly if genes are subject to negative selection. Thus, if promoters are selected to retain their function we would expect them to parallel changes in their phylogeny. If promoters evolve differently, this would indicate the occurrence of new functions (through either neo- and/or subfunctionalization).

We found CCOs in the phylogenetic analysis, but 60% of these groupings include only paired orthologous groups (Figure 2.1). This shows that promoter sequences and motifs might be conserved in sequences that are more closely related, i.e. that have separated recently ($<<$ 10 MY). Therefore, pairwise alignments of promoter comparisons will only be effective within short evolutionary distances, something that is confirmed in our study by the motif analysis (Figure 2.5). From the FootPrinter analysis we learn that conserved promoter regions have been preserved for $\approx 5-10$ MY, since we find most of the conserved motifs between *A. lyrata* and *A. thaliana*, but also between *Capsella rubella* and *Boechera divaricarpa* (Figure 2.5). Although the search might already be constrained by the input tree used in FootPrinter, in this case the promoter tree, it is expected that shared regulatory elements are more often found among recently duplicated genes.

Regulatory regions can be found in first introns (127; 128). We found shorter branch lengths among orthologs in introns 4. From the VISTA analysis we saw that both introns 1 and 4 are conserved more often (Figure 2.4). Preservation of intron 1 can be explained by potential function, as they appear to be involved in alternative splicing, which has been shown to occur in two of the TRL genes, At2g29340 and At2g29350 (36). Intron 4 might have some structural importance, as they appear to be more conserved, but this needs to be tested experimentally.

2.4.2 Relationship of non-coding variation and expression of TRLs in Brassicaceae

Divergence in function might not only correlate with divergence in regulatory regions and motifs, but might be reflected in differences in expression of the TRL gene duplicates (95). From the databases we obtained data from microarrays and MPSS for *A. thaliana*. We also obtained qPCR data for this species and *A. lyrata*. We found expression profiles of TRLs in *A. lyrata* to be different from its orthologs in *A. thaliana*. Since *A. lyrata* is an outcrossing relative, different alleles of TRLs add another level of variation, which could increase the number of functions.

In general, we observed in the qPCR and microarray analysis that expression patterns of TRLs are divergent and do not necessarily reflect duplication history. Our results agree with the study done by Casneuf et al. (54). They found that genes that have emerged from local duplications have diverged more in expression than have gene products of large scale tandem duplications. This contradicts the pattern observed in other organisms, for instance *Drosophila*. Nevertheless, when we looked at *A. thaliana* microarray data only, a positive correlation exists between the distances of coding sequences and expression (Table 2.3), which might be partly explained by effects of the position in the genome.

Table 2.3: Mantel test of correlation between coding and noncoding regions of TRL genes.

Comparison	Correlation r	p-value
Coding vs. Promoter	0.422	0.0001
Coding vs. Intron1	0.402	0.0001
Coding vs. Intron2	0.288	0.0001
Coding vs. Intron3	0.080	0.0003
Coding vs. Intron4	0.581	0.0001
Coding (A.thaliana) vs. Expression (PCC)	0.668	0.0001
Coding (A.thaliana) vs. Expression (ED)	0.414	0.0001

2.4.3 Identification of putative functions of TRL genes through *in silico* analysis of co-regulation and by qPCR

As the microarray and qPCR analyses show, TRL have diversified their expression (Figs. 2.8 and 2.9). As we expect clustered genes to be co-regulated, i.e. those physically closer to share regulatory regions, we tested for co-regulation in the TRL gene cluster. Studies of gene order in different organisms have proved that physically linked genes might be in part co-regulated and that selection has played a role in conserving gene order (132). In *A. thaliana* a tendency for clustered genes to be co-regulated was described previously (133). Also, it implies that sequences of the regulatory regions might be similar, if genes have recently duplicated, which was shown in the FootPrinter analysis for some of the TRL genes. Numerous duplication events of TRL genes have occurred in the Brassicaceae as a consequence of the tandem duplication in the Brassicaceae-Cleomaceae ancestor, which occurred around 20 MY ago (106). Furthermore, we expected to detect epistatic interactions, such as complementary expression, if genes show the same expression profile over all experiments. If genes are expressed similarly but with opposite magnitudes, then one gene might be repressing the other.

With the PCC analysis we identified a few genes that are significantly corregulated, for instance At2g29290 and At2g29310 (Figure 2.2). As shown in Figure 2.8, both genes display a highly similar expression profile. These two TRL genes display the largest response in the experiments of varying light. The correlation of this gene pair is confirmed when browsing for the function of these genes in the ATTED-II database. There we found both genes to have targets in the chloroplast and to be related to thylakoid lumen proteins. At2g29290 is also co-expressed with phototrophic responsive enzymes (96).

A pair of antagonistic genes identified according to the PCC analysis and expression-profiles would be At2g29350 and At2g29290 (PCC=-0.23073, Figure 2.2). Both genes are very similar in their amino acid sequence and therefore share a more recent common ancestor with each other than with other TRL paralogs. Divergence in expression of this pair of genes is expected since they are not close in the promoter phylogeny and the FootPrinter (Figure 2.1). Therefore, we decided to perform quantitative real-time PCR (qRT-PCR) for both of the genes in the plants we treated with cold, salicylic acid and *Pseudomonas syringae*. These experiments confirmed what we had observed in expression profiles. Once one of the genes is upregulated, the other appears to respond in the opposite way (Figure2.9).

At2g29330 and At2g30670 have been reported to be responding to brassinosteroids (62), and they appear close in the phylogenetic analysis of both protein and promoter. Nevertheless, they do not display a significant PCC value and appear correlated only in reponse to pathogen (Figure 2.10 experiments 45-56).

2.4.4 Inferences about evolutionary processes by examining the link between protein, non-coding sequences and expression

We found agreement between the qPCR experiments with microarrays in most of the cases. Frequently, the induction happened at a later/earlier point in time. *Arabidopsis lyrata* TRL orthologs do not replicate the pattern of their *A. thaliana*

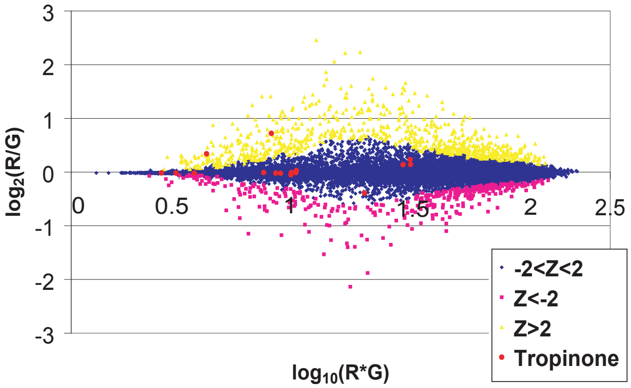


Figure 2.10: Graphic representation of the identification of differential gene expression. All significant Z-values for the experiment 'blue light' after 4h of exposure are shown as yellow dots. 'Darkness' is used as a reference, shown as pink dots. Non- significant values are blue dots, and red dots correspond to tropinone-reductase-like (TRL) genes. Two of them are significantly overexpressed, at2g29290 and at2g29310 $(Z - value) > 1 \cdot \sigma)$.

partners, and appear to have different expression patterns, as a response to cold stress and salicylic acid after long times of exposure (Figure 2.9, Fig 2.12). Since we have fewer data points for *Arabidopsis lyrata* in the SA and *Pseudomonas syringae* experiment, we might have missed the early responses that their orthologs display. Another possibility is that selfing lineages, like *A. thaliana* contain variable functional gene duplicates, which in the outcrosser *A. lyrata* is compensated for by allelic variants; additionally, some TRLs might not be functional. We cannot ascertain this by qRT-PCR.

Senescence can be induced by the hypersensitive response to pathogen *Pseudomonas syringae* (121); At2g29350 (SAG13) responded to this elicitor, as we observed in our qRT-PCR experiments. SAG13 has been reported in the literature to induce of senescence (33; 134). One *A. lyrata* TRL, AlTRL54, responded to *P. syringae* after 24h of exposure, although it was not the ortholog of At2g29350. This might be due to the missing data for the early responses for this species. At2g29350 responded to salycilic acid after 4 h of exposure. Regulation by SA was expected for this gene, as SA induces of senescence. The ortholog of this gene in *Arabidopsis lyrata* responded to SA after 28 and 52h of exposure. SA induction is consistent with the discovery of higher than random average frequency of WRKY promoter elements (yellow-green squares in Figure 2.5) (135). WRKY promoters have been associated with induced expression of SAG13 (136).

The FootPrinter analysis might help identify neo- and/or sub-functionalization events. In particular, when analyzing orthologous groups, different motifs can argue for divergence, if protein sequences are very conserved. One case of neo-functionalization among paralogs in the same species could be the case of AlTRL29 and AlTRL54. Both paralogs appear in neighboring groups both in the protein and in the promoter trees (see Figure 2.1). Nevertheless, their regulatory motifs differ and their expression has differentiated. We observe the same for its orthologs, At2g29350 and At2g29150 respectively (Figure 2.12), *in silico* and in their qRT-PCR expression profiles (see Figure 2.9).

In recent studies, sub-functionalization has been detected as a mutation in regulatory sequences of recently duplicated genes that have acquired transcriptional specialization, for instance, in different tissues (104). Often, comparisons are made over large evolutionary distances, for instance among fish and human. Our study spans shorter evolutionary distances which allows more confident comparisons to be made between coding and promoter evolution. Sub-functionalization might be happening in the case where expression is significantly correlated among genes, but one gene does not show expression that is significantly different from the mean for any of the experiments analyzed. As can be seen in Figure 2.8, expression can be correlated in two genes over multiple experiments, but they can differ in their magnitude of expression. One example is the At2g29340-At2g29350 gene pair, that appear co-regulated in qRT-PCR and microarray experiments. Their transcripts may complement each other. Another possible sub-functionalization in recent gene duplicates would be the case of At2g29150 and At2g29170 (Figure 2.11). The latter gene has significant expression only in a few experiments, but appears to show similar pattern of expression of its co-expressed gene. These are hypotheses that need experimental confirmation.

We found promoter, motifs and expression analysis to be useful in the study of gene families, as the combined analysis allowed us to set up hypotheses about the evolutionary dynamics and of possible interactions among the members of the gene family, that might otherwise not be evident from molecular evolutionary studies alone.

Acknowledgments

We thank John D'Auria, Jana Pastuschek and Michael Philips for assistance in the Lab; Mike Phillips also made useful comments for the qRT-PCR section. Thomas Mitchell-Olds and Jonathan Gershenzon provided technical facilities. This work was supported by the Max Planck Society, a grant of the Human Frontiers in Science to KJS, a grant from the German Academic Exchange (DAAD) to ANQ, and a grant from the Max Planck Society to TE.

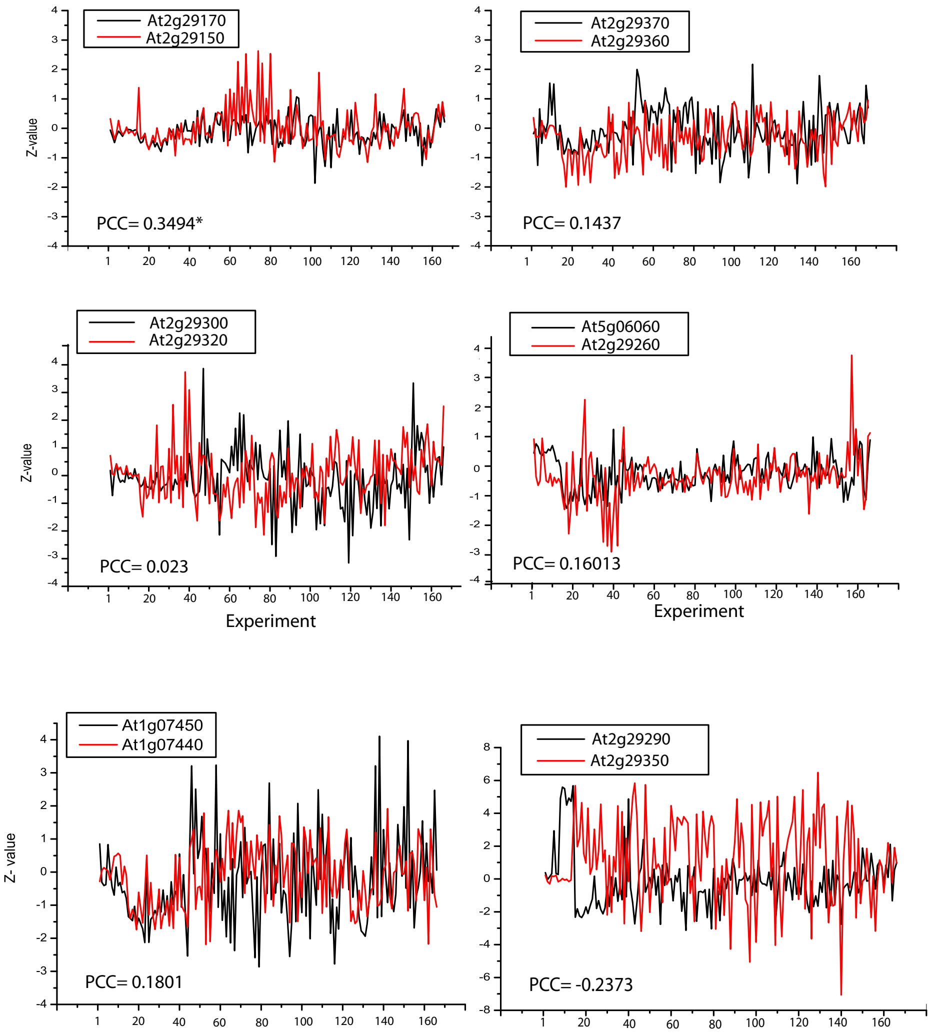


Figure 2.11: Comparison of expression profiles of paired genes from *A. thaliana*. All data were obtained from AtGenExpress. All axes are the same (-4;4), except for the comparison At2g29350 vs. At2g29290, where the y- axis goes from (-8;8) as both gene display higher differential expression than the other TRLs.

Cold stress

		1h	4h	12h	24h
at2g29290	A	n.a.	0.47	0.12	0.24
	B	2.40	-0.54	-0.98	-2.76
	C	-2.00	n.a.	n.a.	n.a.
	D	0.64	0.71	0.28	0.05
AlTRL37 (Ortholog)	D	2.35	n.a.	1.87	0.53
at2g29340	A	n.a.	0.64	1.20	0.43
	B	1.17	1.08	-0.80	-2.30
	C	n.s.	n.a.	n.s.	n.s.
	D	0.95	0.31	0.13	0.38
at2g29350	A	n.a.	1.00	2.42	0.63
	B	-0.63	0.15	0.61	0.53
	C	2.20	n.a.	n.a.	n.a.
	D	1.29	1.61	0.39	0.72
AlTRL29 (Ortholog)	D	0.17	0.31	0.23	0.63
at2g29150	A	n.a.	0.87	1.37	1.56
	B	0.69	-0.31	0.30	0.50
AlTRL54 (Ortholog)	D	3.55	0.31	0.55	0.74

SA

		4h	28h	52h
at2g29290	A	0.46	n.a.	n.a.
	B	-2.24	0.88	1.71
	D	0.48	1.06	0.68
AlTRL37 (Ortholog)	D	n.a.	2.14	1.46
at2g29340	A	1.29	n.a.	n.a.
	B	-0.97	-1.01	0.13
	D	0.87	0.71	0.80
at2g29350	A	1.57	n.a.	n.a.
	B	1.04	2.12	1.02
	D	3.72	0.29	0.68
AlTRL29 (Ortholog)	D	n.a.	3.02	5.36
at2g29150	A	0.43	n.a.	n.a.
	B	-1.96	-1.82	0.51
AlTRL54	D	n.a.	1.04	2.86

P. syringae

		2h	6h	12h	24h
at2g29290	A	n.a.	0.36	n.a.	n.a.
	B	-2.14	-1.71	n.a.	-2.35
	D	n.a.	0.73	1.00	2.03
AlTRL37 (Ortholog)	D	n.a.	n.a.	0.29	0.82
at2g29340	A		0.66		
	B	-1.00	-0.95	n.a.	0.00
	C	n.a.	n.a.	n.a.	n.a.
	D	n.a.	3.62	2.58	0.73
at2g29350	A	n.a.	11.96	n.a.	n.a.
	B	1.54	0.98	n.a.	4.65
	D	n.a.	4.14	0.84	1.03
AlTRL29 (Ortholog)	D	n.a.	n.a.	0.17	0.23
at2g29150	A	n.a.	2.17	n.a.	n.a.
	B	-0.54	-0.72	n.a.	-0.18
AlTRL54	D	n.a.	n.a.	0.08	1.57

Figure 2.12: **Relative expression values obtained with various methods compared to those obtained by qRT-PCR for co-regulated genes (from PCCs) At2g29340, At2g29350 and At2g29290. The colors follow the rules of expression heat-maps, red for induction and green for repression.** Z-value corrected expression is displayed for A) Genevestigator (37), B) microarrays (105), C) bbc Botany Array Resource (bbc, (137)) and data from (35) D) qRT-PCR; n.a.= not available; n.s.= not significant. In all databases cold induces the downregulation of At2g29340 and At2g29290; SA and *Pseudomonas* induce At2g29350. In the response to the plant pathogen, AlTRL54 and AlTRL37 follow a different expression pattern than their orthologs.

Chapter 3

Positive selection in tau GSTs duplicated *in tandem* in Brassicaceae

A. Navarro-Quezada and K.J. Schmid

17th October 2007

3.1 Abstract Chapter 3

We studied genes belonging to the tau glutathione-S-tranferases, a gene family special to plants, in six Brassicaceae species. The members of the tau GST gene family duplicated *in tandem* in all the studied species, ranging from four to nine copies. Tau GST genes radiated in Brassicaceae independently from other plants. These genes have been subject to positive selection, which supports that they have diversified in function. GSTs are important plant defense enzymes, and we had expected to see them evolving according to positive selection, although in the present they appear to be evolving slower due to co-regulation of closely related tau GST enzymes.

3.2 Introduction Chapter 3

In theory, gene families tend to cluster according to function (138). In the model plant *Arabidopsis thaliana* 10% of its genes are within large, co-expressed chromosomal regions. 40% of these co-expressed genes have also high sequence similarity (139). Functional similarity is true in the case of the sequenced BACs from Chapter 1. We found a gene family upstream from the tropinone reductase like (TRL) genes, the genes coding for glutathione-S-transferases (GST), which are enzymes typically associated with stress response (38).

Although less numerous as the TRL genes in the same genomic region, GST genes are numerous in plant genomes. The GST gene family is very large in all organisms where it has been analyzed. In *A. thaliana* 48 GST and GST-like genes are found. The functions of these enzymes are also largely different (140), they range from catalyzing detoxification reactions to binding flavonoid natural products prior to excretion. GSTs are involved in detoxifying and are often used as markers of stress. In insects, GSTs have been shown to play an important role in the detoxification of many substances including plant allelochemicals (141). GSTs are ubiquitous enzymes that bind glutathione to reactive oxygen molecules, which are toxic to cells, so that these molecules can be excreted. Two plant-specific GSTs have been described in the literature: phi- and tau GSTs (140). Only the first type of GSTs have been crystallized, and the structure appears conserved among phi GSTs and the other known GSTs (140). Enzymatic tests on both plant-specific GSTs have shown that the enzyme types respond differently to one of the major plant defense signals, salicylic acid. The response is also different among in tandem duplicated genes. Recent reviews indicate that a large number of GSTs might be due to the substrate specificity of each enzyme needed to detoxify diverse toxic compounds (142).

Tau class GSTs were first identified as being induced by auxins, and have recently been shown to be involved in the response to a variety of endogenous and exogenous stresses including pathogen attack, wounding, heavy metal toxicity, oxidative and temperature stress (143). Their genes contain a single intron at a conserved position. Using phylogenetic methods, we analyzed if tau GSTs evolve in a similar way as its neighboring genes, the TRLs, which underwent positive selection at the origin of the duplications.

3.3 Materials and Methods Chapter 3

3.3.1 Sequencing the tau GST genes on Chromosome 2

Tau GST genes were obtained in our effort to isolate the TRL genes from Brassicaceae, as they are localized 3' upstream from TRL genes in the species sequenced: *Arabidopsis lyrata, Arabidopsis cebennensis, Boechera divaricarpa, Capsella rubella*, and a more distantly related species, *Cleome spinosa* (Capparales). For the origin of the BACs, the sequencing and annotation procedure, we refer to the Materials and Methods section in Chapter 1. We learned from the literature that *Brassica rapa* contains at least one GST encoding gene upstream from the TRLs, although we did not have the sequence, thus we do not analyze GST genes for this species (64). Repetitive elements were detected with REPEATMASKER (`www.repeatmasker.org`).

3.3.2 Phylogenetic analysis

As for TRLs, (Chapter 1) alignment of tau GST genes was done initially with CLUSTALX and checked for misalignments by eye. Phylogenetic trees were constructed from both coding DNA and amino acid sequences using SEQBOOT to produce 100 bootstrapped sets and DNAML and PROML programs from the PHYLIP package, using default options for all 100 trees (73). We used the program NEIGHBOR to construct a neighbor-joining (NJ) tree of both DNA and amino acid sequences. Among all genes, we defined groups of orthologous genes as those that grouped together in a clade whose supporting branches have a bootstrap value higher than 50%. Neighbor-joining trees proved to be robust, and were used in further analyses. The outgroups used in the phylogenetic reconstruction were obtained by performing a search with the blastx program that compares six-frame conceptual translation products of a nucleotide query sequence (both strands) to a database of proteins. We identified those GSTs that were more closely related to the *A. thaliana* tau GSTs (68), which all came from the species *Glycine max* from the Leguminosae plant family at the time the sequences were obtained (May 2006) (144).

3.3.3 Tests for positive selection

Tests for positive selection were performed using PAML, as described in Chapter 1. We performed M7 and M8 for the complete GST tree and for the three major clades (labeled 1, 2 and 3 in Figure 3.2). For the branch-site analysis we labeled those branches at the base of clades 1, 2 and 3, that separate these clades from the other GST paralogs. Labeled branches are shown in Figure 3.3.

3.4 Results Chapter 3

3.4.1 The tau GSTs on Chromosome 2 in *A. thaliana* and their equivalents in other Brassicaceae

Among the six sequenced BACs, we identified four (in *B. divaricarpa*) to nine tau GSTs (in *C. rubella*) that are located upstream of the TRLs (Figure 3.1). One quarter (7 out of 28) of the *A. thaliana* tau GSTs are located in this genomic region(1), which is equivalent to 25% of the total number of tau GSTs (7 out of 28 (145)). We were able to identify all *A. thaliana* tau GSTs orthologs in *Arabidopsis lyrata*, *Arabidopsis cebennensis* and *Capsella rubella*. The last species contains GST genes that duplicated recently and independently from the *Arabidopsis spp.* (see Figure 3.2). We were not able to identify all orthologs for *B. divaricarpa*, as the BAC clones of this species have smaller inserts (discussed in Chapter 1).

Amino acid sequence similarity ranges from 34 to 99.7% among orthologs, from 37 to 90% among paralogs within a species and from 26 to 85% among paralogs from different species ('outparalogs' (146)).

Many GSTs are flanked by retrotransposons, a list of the transposons associated the gene family is shown in Table 3.1. These mobile elements might have contributed to their duplication. Nevertheless, as in the case of TRLs, *Capsella rubella* is the only species that does not contain any transposons in this genomic region, therefore this is not a possible duplication mechanism for this species. Furthermore, all of the tau GST genes have one intron, as is expected for these type of genes, and therefore duplication mediated by reverse transcription can be ruled out. As in the case of TRL genes, we see that transposons are active in this genomic region. They interrupt ORFs, as in the case of one TRL from *A. lyrata*, AlGSTU1, where one LTR-retrotransposon is inserted into intron 2 (Figure 1.2).

Genes closer in the phylogeny are also closer physically, which confirms subsequent duplication (Figure 3.1). The patterns of divergence in the phylogeny are consistent with single gene duplications that occurred at various times, probably by non-equal recombination.

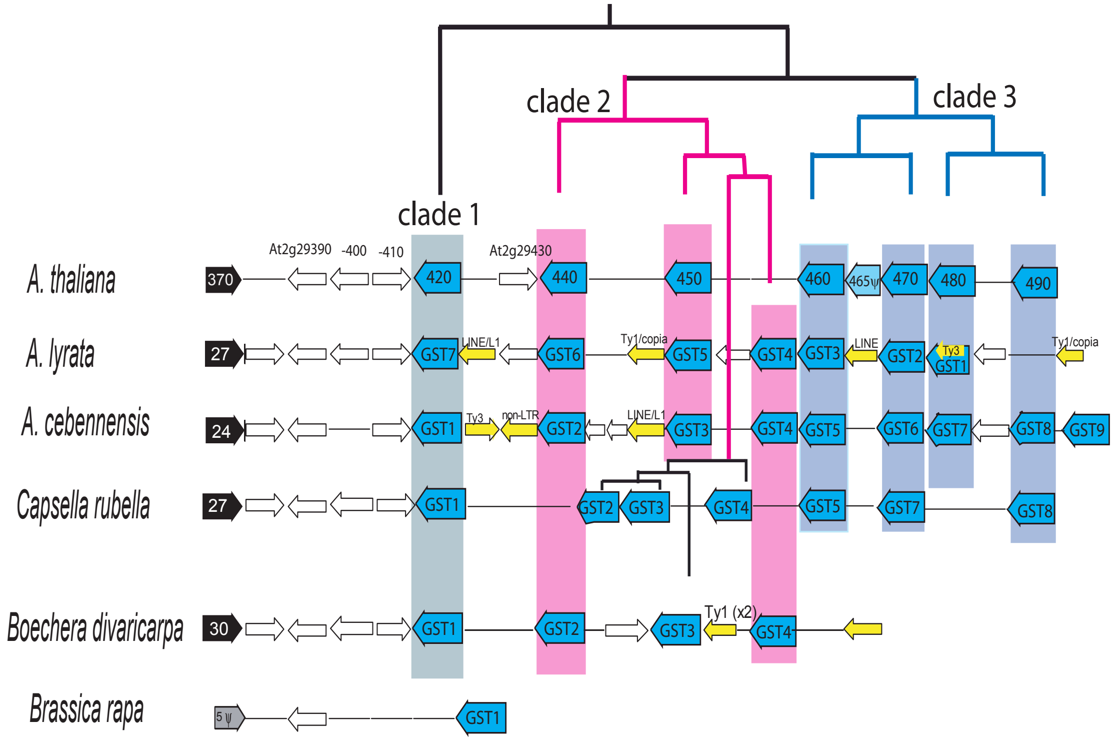


Figure 3.1: The tau glutathione-S-reductase-like genes *in tandem* from studied Brassicaceae, aligned according to orthology. The *A. thaliana* genes are used as a reference and shadowing indicates groups of orthologs. The phylogeny reflects the succession of the genes on the chromosome, and the order probably reflects the duplication history. No inversions or translocations have occurred, as in the TRLs. Black lines lead to the clade containing the genes in cluster 1. Pink lines (middle) indicate clade 2 and blue (left) lines clade 3. Tau GST genes are depicted by blue arrows. The ORF of the first tau GST from the left from *A. lyrata* is interrupted by a transposon, indicating AlTRL1 might be a pseudogene. Fine lines connect paralogs from *Capsella rubella*.

Table 3.1: Transposable elements identified within the tau GST gene cluster

Species	Ty1-copia	Ty3-Gypsy	LINE	Other	number
A. lyrata	2	1	2	n.a.	5
A. cebennensis	0	1	1	non-LTR	3
A. thaliana	0	0	0	n.a.	0
C. rubella	0	0	0	n.a.	0
B. divaricarpa	5	0	0	n.a.	5
B. rapa	0	0	0	n.a.	0
C. spinosa	0	0	0	non-LTR	1

3.4.2 Phylogenetic analysis of GSTs

The neighbor-joining tree of GSTs constructed from amino acid sequences had high values of bootstrap supporting the branches, and therefore we worked with this tree for further analysis. As observed in the TRLs, the tau GSTs from Brassicaceae have independent duplication histories from their homeologous GSTs in the outgroup species *Cleome spinosa* and *Glycine max* (Figure 3.2). Tau GSTs have differentiated in three major clades (numbers 1 to 3 in Figure 3.2). Most of the subclades we find within these major clades contain one representative of each of the Brassicaceae species studied, except major clade 3 that does not contain any *Boechera divaricarpa* sequence. Since we have smaller sequences inserted into the *B. divaricarpa* BACs, and these GSTs are at the end of the BAC we think we do not have all tau GSTs for this species.

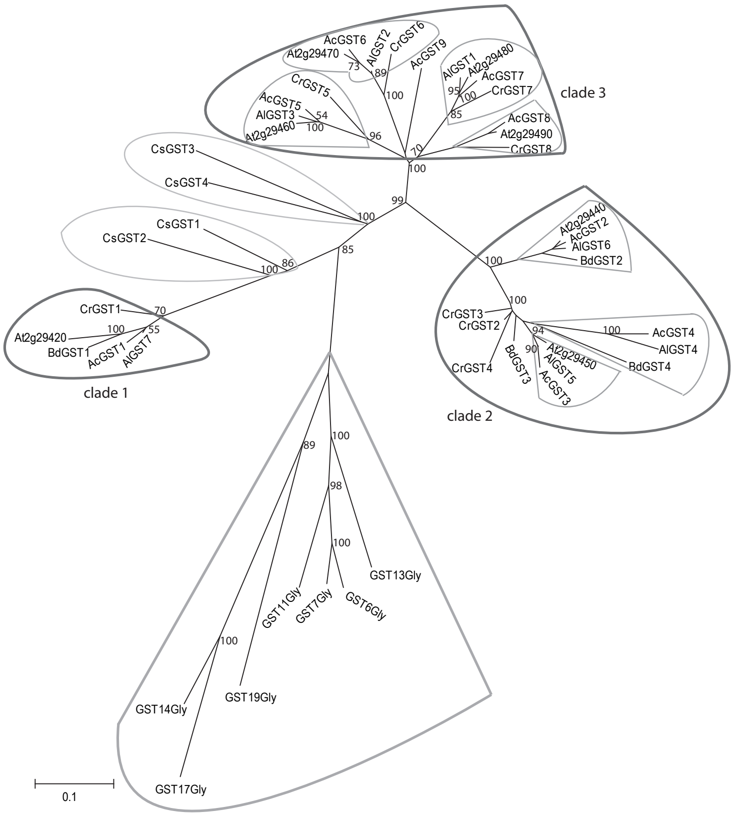
The phylogeny accords with gene birth and death and divergent evolution. Some species appear to have had recent duplications supporting gene birth, as is the case for *Capsella rubella*. Although there are few pseudogenes as evidence of gene death, one pseudogene is present in *A. thaliana* (light blue arrow in Figure 3.1) and one in *A. lyrata*, whose ORF is interrupted by a retrotransposon (Figure 3.1). Also Tau GST do not display large fluctuations in copy numbers among species as do TRLs. Nevertheless, tau GST orthologs appear to be lost in some of the subclades (grey lines in Figure 3.2), particularly in major clade 2. This can also be seen more in the depiction of gene relationships and orthology of genes in clade 2 (purple lines Figure 3.1), where *A. thaliana* appears to have lost one of the paralogs as well as *Capsella rubella*. Some paralogs of this last species appear to evolve independently from the paralogs in clade 2 of the other tau GSTs from the sampled Brassicaceae, although the phylogeny is not well resolved at the branches leading to these genes, as show the low bootstrap support (CrGST2, CrGST3 and CrGST4 in the tree in Figure 3.2).

3.4.3 Detecting positive and negative selection in the different parts of the tree

When testing for selection in each separate clade in the phylogenetic tree (clades 1-3 in Figure 3.2), we found that these appear evolving in a conserved fashion with some sites evolving close to neutrality. M7 and M8 are equivalent, with ω values between 0- 0.7 (Table 3.2). When we analyze the complete tree using tests M7 and M8, we found M8 has a better probability than M7, where M8 estimates one extra site with $\omega = 1.04$ suggesting neutral evolution or slightly positive selection (Table 3.3).

From the previous results we observe that on the one hand it appears that positive selection is not playing an important role among all the studied tau GSTs. On the other hand, when we test the branch-site models for positive selection in foreground branches, these tests were significant and we detected four selected sites in two clades (Table 3.3). Only one of the positively selected sites remains the same when allowing positive selection after the split of two basal branches in clade 2 (branches 3 and 4 in Figure 3.3).

Positive selection appears to have occurred at the basal branch that splits part of clade 2 from the rest of the sequences in the tree, as MB has the highest probability when the sequences following the split of BdGST2 and its orthologs from the rest of the sequences (after square 3 in Figure 3.3). MA and MB have high likelihoods for all of the trees tested with different foreground branches, although M8 performed always better. Tau GSTs probably diversified in the past and in the present within clade 2, in the other clades tau GSTs are evolving according to negative selection.



5

Figure 3.2: **Neighbor-joining tree of tau GST amino acid sequences, obtained from BACs of *A. thaliana*, *A. lyrata*, *A. cebennensis*, *Capsella rubella*, *Boechera divaricarpa* and *Cleome spinosa*.** We indicate nodes with bootstrap values > 50 %. Light gray lines enclose orthologs, in the case of *Cleome spinosa* (CsGST), paralogs. GSTGly are GSTs from *Glycine max* obtained from GenBank.

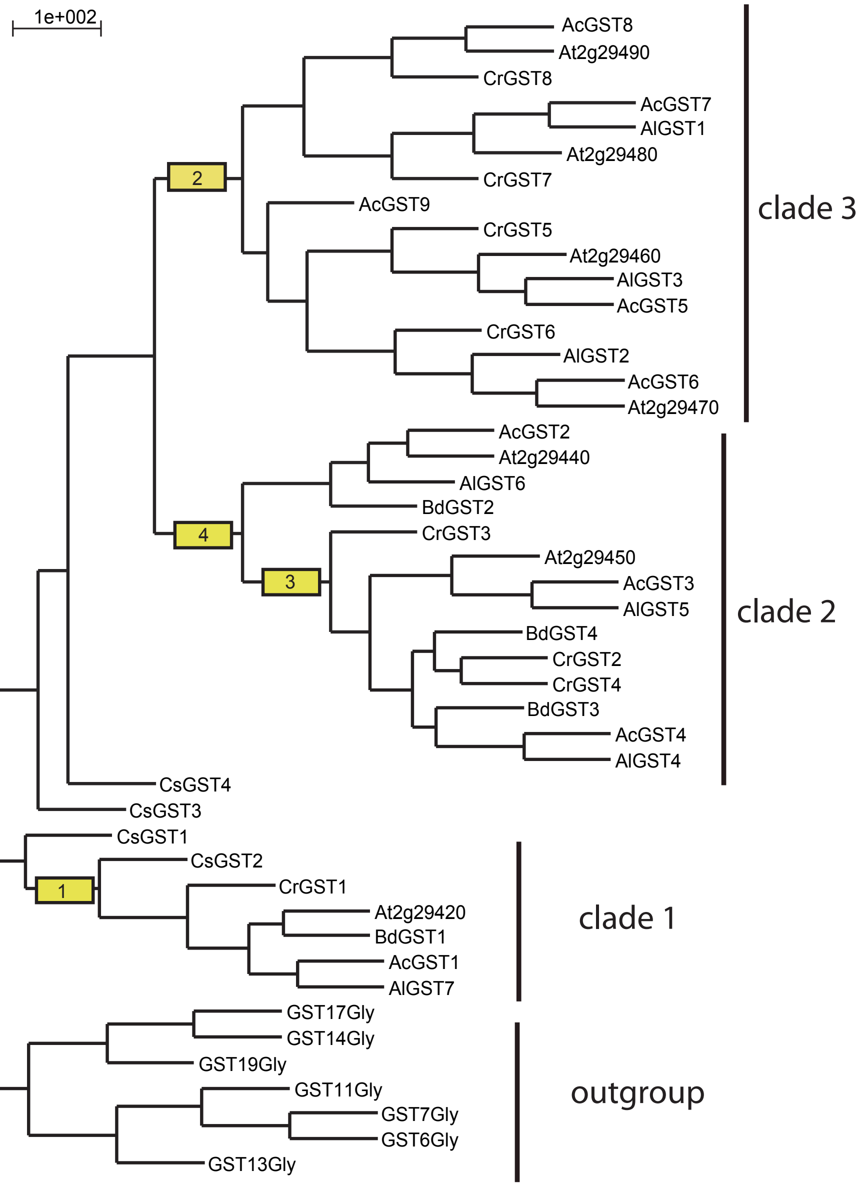


Figure 3.3: Tree indicating the branches after which positive selection was allowed using the PAML branch-site analysis MA and MB. Branches indicated by yellow squares, are foreground branches, as well as branches after this branch (going to the tips). The number inside the squares indicate the number given to the tree, in which this branches were labeled as foreground branches. All other branches except those following the yellow squares are background. The branches that were labeled were tested, as they separate clades from the rest of the sequences, for instance, clade 1 vs. rest of the tree.

Table 3.2: Parameter estimates and tests of selection using branch-site models for clades of closely related tau GSTs paralogs. Test for selection is M7 vs. M8.

Model	Parameter estimates	ℓ	$2\Delta l$	P
Clade 1, n=6				
M7	$\omega_{1..n} = 0 - 0.49$	-2.40418	5.32	0.0699
	$p = 0.6754$ $q = 3.525$			
M8	$\omega_{1..n} = 0.004 - 4.136$ $\omega_s = 4.136$	-2.4015		
	$p_0 = 0.9887$ $p = 0.786$ $q = 4.714$			
Clade 2, n=15				
M7	$\omega_{1..n} = 0.0112 - 0.791$	-3.7861	1.613	0.4464
	$p = 0.7628$ $q = 1.702$			
M8	$\omega_{1..n} = 0.0115 - 145.2$ $\omega_s = 145.2$	-3.7854		
	$p_0 = 0.996$ $p = 0.775$ $q = 1.7545$			
Clade 3, n=17				
M7	$\omega_{1..n} = 0.01 - 0.69$	-4.5896	4.802	0.0964
	$p = 0.8954$ $q = 2.4086$			
M8	$\omega_{1..n} = 0.02 - 1.59$ $\omega_s = 1.59$	-4.5872		
	$p_0 = 0.9715$ $p = 1.1121$ $q = 3.448$			

3.5 Discussion Chapter 3

Most detoxification enzymes, as are GSTs, have been studied in the context of their biochemistry. Recently, with the advent of genome data, comparative studies of gene families have become numerous, and this has included studies on stress responsive enzymes, as are CYPA and cytochrome P450 (CYP450) (147; 148). Some of these studies have identified positive selection as one of the mechanisms driving the evolution of these clustered genes, which have originated by tandem duplications (148). One recent study in a numerous gene family with functions ranging from signaling in development to detoxification, the CYP450 family, identified that those enzymes involved with pathogen and toxin recognition are unstable, evolving by positive selection rather than gene conversion (16).

model	κ	ω	tree length	no of pos. sel sites	ℓ	Test	2ΔL	p
M0	1.74	$\omega = 0.2429$	17.931		-13252.176			
M7	1.74	$\omega = 0.2786$	19.342		-13033.889	M7 vs. M8	11.42	0.0033**
M8	1.753	$\omega = 0.2836$ $\omega_s = 1.04$	19.309		-13028.178			
Labelled tree 1								
MA	1.84	$\omega_0 = 0.1969$ $\omega_1, \omega_{2a,2b} = 1.0$	18.9135		-13110.364	MA vs. MB	79.781	0***
MB	1.75	$\omega_0 = 0.1347$ $\omega_1 = 0.4392$ $\omega_{2a,2b} = 4.97$	18.944		-13070.473	MB vs. M8	84.591	0 ***
Labelled tree 2								
MA	1.84	$\omega_0 = 0.194$ $\omega_1 = 1$ $\omega_{2a,2b} = 1$	18.95		-13109.354	MA vs. MB	79.76	0**
MB	1.7092	$\omega_0 = 0.133$ $\omega_0 = 0.5245$ $\omega_{2a,2b} = 1.789$	21.619		-13069.474	MB vs. M8	82.59	0 **
Labelled tree 3								
MA	1.84	$\omega_0 = 0.189$ $\omega_1 = 1$ $\omega_{2a,2b} = 3.272$	19.984	4^1	-13091.218	MA vs. MB	81.77	0.00*
MB	1.764	$\omega_0 = 0.1273$ $\omega_1 = 0.5146$ $\omega_{2a,2b} = 2.886$	19.102		-1305.0336	MB vs. M8	44.3	0 **
Labelled tree 4								
MA	1.86	$\omega_0 = 0.1914$ $\omega_1 = 1$ $\omega_{2a,2b} = 5.4999$	19.083	4^2	-13094.7570	MA vs. MB	89.42	0.00**
MB	1.77	$\omega_0 = 0.1264$ $\omega_0 = 0.5136$ $\omega_{2a,2b} = 4.8754$	19.12		-13050.0448	MB vs. M8	43.7	0 **

Table 3.3: Parameter estimates and tests of selection using branch-site models for single ortholog clusters and clades.

Positively selected sites: [1] 51 D 0.994**, 156 V 0.995**, 186 I 0.977*, 192 I 0.977* ; [2] 14 D 0.996**, 34 V 0.999**, 35 E 0.989*, 51 D 0.970*

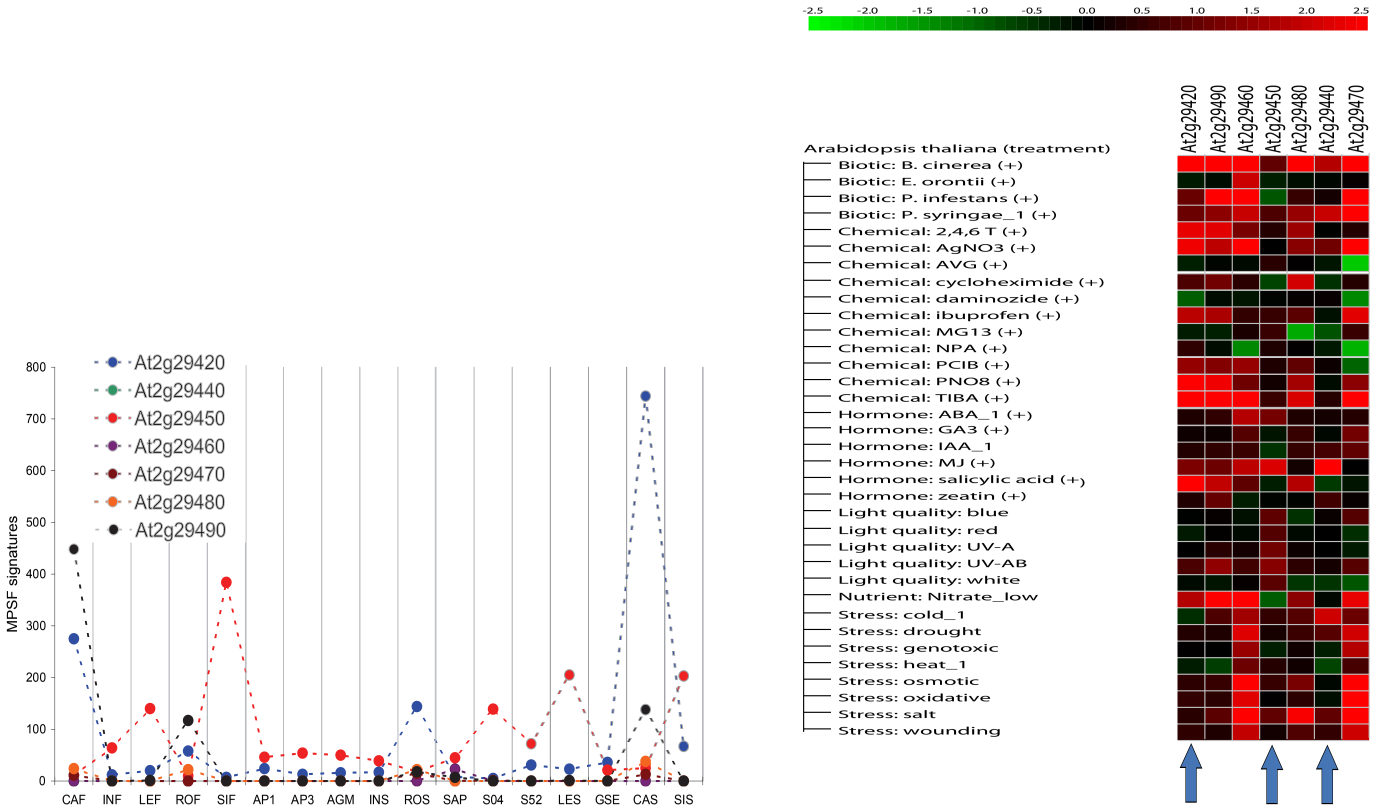
We do not know the functions of the studied tau GSTs, although we learned from microarray data available online that they all respond to infection by the fungus *Botritys cinerea*, to nitrate deprivation and salt stress (Figure 3.4, (37)). Many of these clustered tau GSTs respond to the infection by plant pathogens, as for instance *Pseudomonas infestans*, *P. syringae*. Furthermore, some of the tau GSTs respond to salicylic acid (At2g29420, At2g29480 and At2g29490), and others respond to methyljasmonate (At2g29440, At2g29450 and At2g29460). As these plant hormones have antagonistic functions in plant defense signaling (149), this might indicate the occurrence of subfunctionalization of the tau GSTs. Two closely related tau GSTs, At2g29460 and At2g29470, respond to osmotic, oxidative, salt and wounding stress more strongly than others, as can be seen in Figure 3.4, which shows a graphical depiction of a microarray heatmap and MPSS expression values.

We obtained a list of function and gene targets of the tau GSTs from *A. thaliana* from the ATTED-II database (96). We can observe that most tau GSTs are in cytosol, and only two in the chloroplast (see Table. 3.4). These last two enzymes, At2g29460 and At2g29470, appear to be co-expressed. A further group of co-expressed tau GSTs are those mentioned previously, which respond to salicylic acid (At2g29420, At2g29480 and At2g29490). Contrary to TRLs where co-regulation was only the case for a few genes, the upstream tau GSTs appear to be significantly co-regulated.

Table 3.4: Cellular targets and possible functions of the tau GSTs in chromosome 2 from *A. thaliana*.

Locus	Target *	Function (synonym)
At2g29440	O,Y	glutathione S-transferase, putative ((ATGSTU6))
At2g29450	O,Y	glutathione S-transferase (103-1A) ((AT103-1A, ATGSTU5))
At2g29460	O,C	glutathione S-transferase, putative ((ATGSTU4))
At2g29470	O,C	glutathione S-transferase, putative ((ATGSTU3))
At2g29480	O,Y	glutathione S-transferase, putative ((ATGSTU2))
At2g29490	O,Y	glutathione S-transferase, putative ((ATGSTU1))

Adapted from the ATTED-II project; targets are from TargetP and WolfPSORT (refs. in (96)) * C=chloroplast ; Y=cytoplasm.


(a) MPSS

(b) AtGenExpress microarray heat map

Figure 3.4: A)Expression of GST genes along different plant development stages according to MPSS. MPSS signatures are single hit 17bp signatures, if more than one probe was used the frequencies were pooled. Specific signatures are found for all GSTs. SIS= silique 24h post fertilization; CAS= callus actively growing; GSE= germinating seedlings; LES= Leaves (21 days); S52, S04= leaves after 52 and 4 hours salicylic acid treatment respectively; SAP= infloresence; ROS= root; INS= inflorescence; AGM= agamous inflorescence; AP1, AP3= inflorescences different stages; SIF, ROF, LEF, INF, CAF= siliques, roots, leaves, inflorescence and callus respectively, measured with classic MPSS. B) Heat map of AtGenExpress array data obtained from Genevestigator (37). Arrows point to genes that are highly expressed in different developmental stages, but are less induced by stress.

Our results from the analysis for positive selection support what has previously been reported for large gene families duplicated in tandem. We found that genes coding for enzymes that are clustered underwent positive selection after they had duplicated, as we could see over the whole tree and particularly for those genes in clade 2. Despite conservation of their overall exon- intron number and amino acid sequence similarities that can go up to 85% and 95% among outparalogs and paralogs respectively within a species, these enzymes have diverged in function, and this might be the reason they are preserved in the genome. The same mode of evolution was shown for a gene family that is upstream from the tau GSTs in the genome, the tropinone-reductase-like (TRL) gene family (106).

Although the mode of evolution of tau GST genes accords rather to birth and death than divergent evolution, it appears that they have not lost their function as often, as they might perform essential functions for the plants. Also, slower evolution of tau GSTs within clades, as compared to TRL genes, might be due to function complementation of these enzymes, as we learned from the co-expression database from the ATTED-II project (96) and as it appears to be the case in transcripts from the microarray databases (37) (Figure 3.4).

Conclusions

Chapter 1

We identified a large gene family containing from 4 to 15 members duplicated in tandem that had been preserved in a syntenic region in species separated by more than 20 MY. This gene family encoded short-chain dehydrogenases, highly similar to tropinone reductases from Solanaceae. The function of these tropinone-reductase-like (further: TRL) enzymes in Brassicaceae is puzzling, given that in tests for production of tropane alkaloids from their substrates in *Arabidopsis thaliana* and other Brassicaceae, no tropane alkaloids are formed (28). In this thesis we identified the presence of multiple copies of TRLs not only in two *Arabidopsis spp.*, *A. lyrata* and *A. cebennensis*, but also in three further relatives, *Boechera divaricarpa*, *Capsella rubella* and *Brassica rapa*, as well as in another species from the Cleomaceae family, *Cleome spinosa*. This hints at the preservation of these genes in the plants' genome due to functional importance. In Chapter 1, the phylogenetic reconstruction led to the conclusion that these genes are evolving according to gene birth-and-death, as we find most of them in groups of orthologs. Futhermore, most of the orthologs appear subjected to negative selection, indicating that they are preserved due to functional importance. We identified positive selection using phylogenetic analysis of maximum likelihood (PAML). The M8 test (PAML) identified $\approx$ 8.5% of sites evolving under positive selection. When we contrasted rates of evolution in the tree, we found ancestral TRL genes to be evolving under evolutionary constraints. Positive selection was detected in branches separating

’ancient’ TRLs from the ’modern’ clades, containing the genes generated from local duplications. This observations led us to propose that TRLs have been preserved in the Brassicaceae genome, as they differentiated early in their duplication history, which could have led to different functions. One possible secondary function is the reduction of tropine to produce tropane alkaloids, which was shown in some Brassicaceae and is the normal function in Solanaceae.

Chapter 2

We learned from the phylogenetic analyses of promoters of the tropinone-reductase-like gene family that they have not evolved at the same rates as the protein coding regions they regulate, except for genes in species of recent origin ($<$ 10 MY). Introns 1 and 4 appear to be constrained in their evolution, their phylogeny being significantly correlated with the gene phylogeny. Patterns from expression profiles are different in genes with closely related coding sequences and, in some cases, with conserved regulatory motifs. We propose that variability in cis-regulatory regions and interaction of transcripts, are needed to provide the variation that allows the preservation of multiple conserved copies of TRLs in the genome for more than 20 MY.

Chapter 3

The GSTs are important plant defense enzymes. As other plant defense enzymes (for instance, LRRs), we expected to see the tau GST gene family upstream from TRLs evolving according to positive selection in the present. The genes coding for tau GST enzymes radiated in Brassicaceae independently from tau GSTs from other plants. These genes follow a very similar evolution to the TRLs. They appear to be undergoing gene birth and death, although they are more conserved in gene order than TRL genes. Some tau GST genes have been subjected to positive selection, which supports that they have diversified in function. But most of the tau GSTs are evolving according to negative selection, which indicates the importance of conservation of function of these enzymes.

General conclusions

We found that TRLs and GSTs, both enzymes participating in stress response, are clustered in Brassicaceae genomes. These clusters in the genome are not necessarily due to regulation linkage, since regulatory sequences differ among neighboring and closely related genes, as is the case in TRLs. Thus, TRL (and possibly GST) genes are not clustered due to unity in function, but might reflect their origin through non-equal recombination. A study of the population genetics could elucidate if these clusters are dynamic in populations, and functional studies will be needed to identify the divergence in enzyme function suspected from protein sequence evolution and differential expression.

An open question left from this thesis is whether one of the functions that characterized tropinone-reductase-like enzymes, which is the reduction of tropinone to form tropane alkaloids, is a secondary acquired function. As some Brassicaceae are capable of producing tropane alkaloids (shown in Figure 2), tropinone-reductase function might have been selected in each species depending on the individual requirements of the plant. Therefore, it would be interesting to test if the other studied Brassicaceae are capable of producing tropane alkaloids from crude plant extracts.

Our study proved to be a good first step in the identification of selection and evolutionary dynamics of a region containing genes duplicated in tandem, which did not appear to have an essential function at a first glance, such as TRLs. Furthermore, it proved the importance of comparing gene families among multiple species, as this allowed us to reconstruct the duplication history and to gain insight into the duplication dynamics.

Appendix 1

Tests for recombination: GENECONV and Splitstree

GENECONV by (77) is a command line software that accepts aligned sequences of DNA or protein and carries out a test with 10,000 permutations to detect apparent gene conversion events. Permutations consists in shuffling polymorphic sites and assigning scores to sites that are more frequent than expected by chance. The test is based on imbalances in the distribution of agreement of paired sequences, and it automatically controls for variable mutation rates in the genome. The method is sensitive even when monophyletic samples are chosen. The logic behind the program is that if there has been no gene conversion since the most recent common ancestor of the sequences, the distribution of bases at silent polymorphic sites have been determined by independent neutral mutation at all sites within the same pedigree. It estimates the proportion of times a sequence fragment is repeated.

This program was used with aligned coding sequences to detect possible gene conversion among *A. thaliana*, *Capsella rubella* and *Cleome spinosa* genes. We run GENECONV on the command line with the following parameters: -Skip_indels -Dumpall /sp /lp (both produce lists of significant fragments involved in recombination). Seqtype was set to SILENT (silent polymorphisms of coding sequences are taken into account). The out put calculates shared fragments for single sites and runs of sites. The longest fragment shared after permutation in *Cleome spinosa* TRLs, where paralogs group always together, is between CsTRL9 and pseudogene CsTRL8, which might be a relic of non-equal recombination. With two exceptions, TRLs of this species do not appear to have been subject to significant gene conversion. Gene conversion was detected in *A. thaliana*, but the fragments are not long, and this might be due to 49 shared polymorphic sites, as in the case of At2g29150 and At2g29370. Other cases are At2g29320 and At2g29310; these genes share 39 polymorphic sites and are neighboring genes, but also closely related phylogenetically.

Splits Tree (150) is a program that uses phylogenetic networks. The method is based in on the mathematical method of split decomposition. For ideal data, this method gives rise to a tree, whereas less ideal data are represented by a tree-like network that may indicate evidence for different and conflicting phylogenies. The tree-like network or reticulate tree is used to display events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss, which can affect normal tree reconstructions. It estimates recombination from the network constructed. We used the program Splitstree4 available online (www.splitstree.org) on MacOSX. We constructed network trees for clades B and C of Brassicaceae TRLs. Using Splitstree we performed a phi-test for recombination. It did not find significant recombination for clade B (p= 0.666465) or clade C (p=0.9984). As an example of a reticulate network, the reconstruction with clade C is shown in Fig. 4.

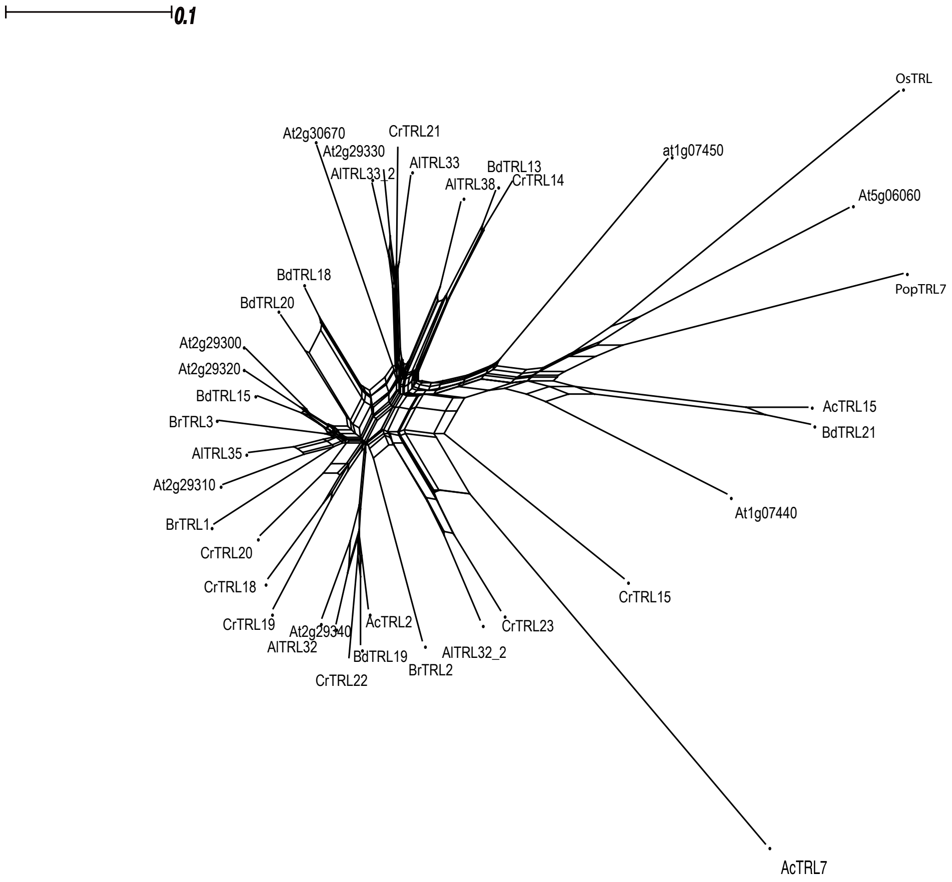


Figure 4: Network phylogenetic reconstruction of genes on clade C built with the program Splitstree4. Terminal branches of the tree are well separated, which supports the absence of recombination.

Appendix 2

Accession numbers of sequenced BACs

Table 4: Accession numbers of sequenced BACs in the Genbank database

Species	Accesion number
Arabidopsis lyrata	EU162608
Arabidopsis cebennensis	EU162612
Boechera divaricarpa 1	EU180847
Boechera divaricarpa 2	EU162610
Boechera divaricarpa 3	EU180848
Capsella rubella	EU162611
Cleome spinosa	EU162609

Bibliography

[1] The Arabidopsis Genome Initiative, “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana,” *Nature*, vol. 408, pp. 796–815, 2000.

[2] A. F. Anne E. Hall and D. T. Preuss, “Beyond the Arabidopsis genome: Opportunities for comparative genomics,” *Plant Physiol*, vol. 129, no. 4, pp. 1439–1447, 2002.

[3] K. Boivin, A. Acarkan, R.-S. Mbulu, O. Clarenz, and R. Schmidt, “The Arabidopsis Genome Sequence as a Tool for Genome Analysis in Brassicaceae. A Comparison of the Arabidopsis and Capsella rubella Genomes,” *Plant Physiol*, vol. 135, no. 2, pp. 735–744, 2004.

[4] L. Lukens, F. Zou, D. Lydiate, I. Parkin, and T. Osborn, “Comparison of a Brassica oleracea Genetic Map With the Genome of Arabidopsis thaliana,” *Genetics*, vol. 164, no. 1, pp. 359–372, 2003.

[5] S. I. Wright, C. B. Y. Kenneth, M. Looseley, and B. C. Meyers, “Effects of Gene Expression on Molecular Evolution in Arabidopsis thaliana and Arabidopsis lyrata,” *Mol Biol Evol*, vol. 21, no. 9, pp. 1719–1726, 2004.

[6] M. E. Schranz, B.-H. Song, A. J. Windsor, and T. Mitchell-Olds, “Comparative genomics in the Brassicaceae: a family-wide perspective,” *Curr Opin Plant Biol*, vol. 10, no. 2, pp. 168–175, 2007.

[7] B. R. M. S. I. and Wright, “Selective Constraints on Codon Usage of Nuclear

Genes from Arabidopsis thaliana," *Mol Biol Evol*, vol. 24, no. 1, pp. 122–129, 2007.

[8] M. Lynch, M. O'Hely, B. Walsh, and A. Force, "The Probability of Preservation of a Newly Arisen Gene Duplicate," *Genetics*, vol. 159, no. 4, pp. 1789–1804, 2001.

[9] J. Dvorak and E. D.Akhunov, "Tempos of Gene Locus Deletions and Duplications and Their Relationship to Recombination Rate During Diploid and Polyploid Evolution in the Aegilops-Triticum Alliance," *Genetics*, vol. 171, no. 1, pp. 323–332, 2005.

[10] M. Lynch and J. S. Conery, "The Origins of Genome Complexity," *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003.

[11] M. Hurles, "Gene Duplication: The Genomic Trade in Spare Parts," *PLos Biology,e206*, vol. 2, no. 7, p. e206, 2004.

[12] D. Leister, "Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes," *Trends Genet*, vol. 20, no. 3, pp. 116–122, 2004.

[13] P. W. Hedrick, "Evolutionary genetics of the major histocompatibility complex," *Am Nat*, vol. 143, no. 6, pp. 945–964, 1994.

[14] M. Nei and A. P. Rooney, "Concerted and birth-and-death evolution of multigene families," *Annu Rev Genet*, vol. 39, no. 1, pp. 121–152, 2005.

[15] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y.-l. Yan, and J. Postlethwait, "Preservation of Duplicate Genes by Complementary, Degenerative Mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[16] J. H.Thomas, "Rapid birth and death evolution specific to xenobiotic cytochrome p450 genes in vertebrates," *PLoS Genetics*, vol. 3, no. 5, 2007.

[17] B. Negre, S. Casillas, M. Suzanne, E. Sanchez-Herrero, M. Akam, M. Nefedov, A. Barbadilla, P. de Jong, and A. Ruiz, "Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex," *Genome Res*, vol. 15, no. 5, pp. 692–700, 2005.

[18] S. Aubourg, A. Lecharny, and J. Bohlmann, "Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana," *Molecular Genetics and Genomics*, vol. 267, no. 6, pp. 730–745, 2002.

[19] M. Benderoth, S. Textor, A. J.Windsor, T. Mitchell-Olds, J. Gershenzon, and J. Kroymann, "Positive selection driving diversification in plant secondary metabolism," *PNAS*, vol. 103, no. 24, pp. 9118–9123, 2006.

[20] J. Yang, H. Gu, and Z. Yang, "Likelihood analysis of the chalcone synthase genes suggests the role of positive selection in morning glories (ipomoea)," *Journal of Molecular Evolution*, vol. 58, no. 1, pp. 54–63, 2004.

[21] P. M. Schenk, K. Kazan, I. Wilson, J. P.Anderson, T. Richmond, S. C. Somerville, and J. M.Manners, "Coordinated plant defense responses in Arabidopsis revealed by microarray analysis," *PNAS*, vol. 97, no. 21, pp. 11655–11660, 2000.

[22] J. A. Cotton and R. D. M. Page, "Multiple Mechanisms Promote the Retained Expression of Gene Duplicates in the Tetraploid Frog Xenopus laevis," *PLos Genetics*, vol. 2, no. 4, p. e56, 2006.

[23] N. Maltsev, E. M. Glass, G. Ovchinnikova, and Z. Gu, "Molecular Mechanisms Involved in Robustness of Yeast Central Metabolism against Null Mutations," *J Biochem (Tokyo)*, vol. 137, no. 2, pp. 177–187, 2005.

[24] B. E. Shakhnovich and E. V. Koonin, "Origins and impact of constraints in evolution of gene families," *Genome Res*, vol. 16, no. 12, pp. 1529–1536, 2006.

[25] R. Keiner, H. Kaiser, K. Nakajima, T. Hashimoto, and B. Dräger, "Molecular cloning, expression and characterization of tropinone reductase II, an enzyme

of the SDR family in Solanum tuberosum (L.)," *Plant Mol Biol*, vol. 48, no. 3, pp. 299–308, 2002.

[26] K. Nakajima, T. Hashimoto, and Y. Yamada, "Two tropinone reductases with different stereospecificities are short-chain dehydrogenases evolved from a common ancestor," *Biochemistry*, vol. 90, pp. 9591–9595, 1993.

[27] K. Nakajima, A. Yamashita, H. Akama, T. Nakatsu, H. Kato, T. Hashimoto, J. Oda, and Y. Yamada, "Crystal structures of two tropinone reductases: Different reaction stereospecificities in the same protein fold," *PNAS*, vol. 95, no. 9, pp. 4876–4881, 1998.

[28] A. Brock, T. Herzfeld, R. Paschke, M. Koch, and D. Birgit, "Brassicaceae contain nortropane alkaloids," *Phytochemistry*, vol. 67, no. 18, pp. 2050–2057, 2006.

[29] R. Keiner, *Calystegine in Solanum tuberosum L.-Klonierung, Expression und Charakterisierung der Tropinonreduktasen I und II, putativer Enzyme des Tropanalkaloidstoffwechsels*. PhD thesis, Martin-Luther Universität Halle-Wittenberg, 2001.

[30] K. D. Allen, "Assaying gene content in arabidopsis," *PNAS*, vol. 99, no. 14, pp. 9568–9572, 2002.

[31] P. J. Facchini, D. A. Bird, and B. St-Pierre, "Can Arabidopsis make complex alkaloids?," *Trends in Plant Sci*, vol. 9, no. 3, 2004.

[32] K. N. Lohman, S. Gan, M. John, and R. Amasino, "Molecular Analysis of natural leaf senescence in Arabidopsis thaliana," *Physiol Plantarum*, vol. 92, pp. 322–328, 1994.

[33] J. D. Miller, R. N. Arteca, and E. J. Pell, "Senescence-Associated Gene Expression during Ozone-Induced Leaf Senescence in Arabidopsis," *Plant Physiol*, vol. 120, pp. 1015–1024, 1999.

[34] T. Meyer, M. Burow, M. Bauer, and J. Papenbrock, "Arabidopsis sulfurtransferases: investigation of their function during senescence and in cyanide," *Planta*, vol. 217, no. 1, pp. 1–10, 2003.

[35] H. Kim, E. C. Snesrud, B. Haas, F. Cheung, C. D. Town, and J. Quackenbush, "Gene Expression Analyses of Arabidopsis Chromosome 2 Using a Genomic DNA Amplicon Microarray," *Genome Res*, vol. 13, pp. 327–340, 2003.

[36] B. C. Meyers, S. S. Tej, T. H. Vu, C. D. Haudenschild, V. Agrawal, S. B. Edberg, H. Ghazal, and S. Decola, "The Use of MPSS for Whole-Genome Transcriptional Analysis in Arabidopsis," *Genome Res*, vol. 14, no. 8, pp. 1641–1653, 2004.

[37] L. H. P. Zimmermann, M. Hirsch-Hoffmann and W. Gruissem, "GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox," *Plant Physiol*, vol. 136, pp. 2621–2632, 2004.

[38] K. A. Marrs, "The functions and regulation of glutathione s-transferases in plants," *Ann Rev Plant Phys*, vol. 47, no. 1, pp. 127–158, 1996.

[39] C. Frova, "The plant glutathione transferase gene family: genomic structure, functions, expression and evolution," *Physiol Plantarum*, vol. 119, no. 4, pp. 469–479, 2003.

[40] E. Nutricati, A. Miceli, F. Blando, and L. D. Bellis, "Characterization of two arabidopsis thaliana glutathione s-transferases," *Plant Cell Rep*, vol. Volume 25, pp. 997–1005, 2006.

[41] W. J. Swanson, R. Nielsen, and Q. Yang, "Pervasive Adaptive Evolution in Mammalian Fertilization Proteins," *Mol Biol Evol*, vol. 20, no. 1, pp. 18–20, 2003.

[42] Z. Yang and R. Nielsen, "Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages," *Mol Biol Evol*, vol. 19, no. 6, pp. 908–917, 2002.

[43] J. P. Bielawski and Z. Yang, “Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution,” *Journal of Molecular Evolution*, vol. 59, no. 1, pp. 131–132, 2004.

[44] Z. Yang and R. Nielsen, “Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models,” *Mol Biol Evol*, vol. 17, no. 1, pp. 32–43, 2000.

[45] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, “Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites,” *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[46] W.-H. Li, *Molecular Evolution*. Sinauer, Sunderland MA, 1997.

[47] B. D. Romain Koszul, Sandrine Caburet and G. Fischer, “Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments,” *EMBO Journal*, vol. 23, no. 4, pp. 234–243, 2004.

[48] J. Wendel, “Genome evolution in polyploids,” *Plant Mol Biol*, vol. 42, pp. 225–249, 2000.

[49] G. Blanc, K. Hokamp, and K. H. Wolfe, “A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome,” *Genome Res*, vol. 13, no. 2, pp. 137–144, 2003.

[50] M. E. Schranz and T. Mitchell-Olds, “Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae,” *Plant Cell*, vol. 18, no. 5, pp. 1152–1165, 2006.

[51] S. Henikoff, “Rapid Changes in Plant Genomes,” *Plant Cell*, vol. 17, no. 11, pp. 2852–2855, 2005.

[52] G. Blanc and K. H. Wolfe, “Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution.,” *Plant Cell*, vol. 16, no. 7, pp. 1679–1691, 2004.

[53] F. Yu, P. C. Sabeti, P. Hardenbol, Q. Fu, B. Fry, X. Lu, S. Ghose, R. Vega, A. Perez, S. Pasternak, S. M. Leal, T. D. Willis, D. L. Nelson, J. Belmont, and R. A. Gibbs, "Positive selection of a pre-expansion CAG repeat of the human SCA2 gene.," *PLoS Genet*, vol. 1, no. 3, p. e41, 2005.

[54] T. Casneuf, S. D. Bodt, J. Raes, S. Maere, and Y. V. de Peer, "Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant arabidopsis thaliana," *Genome Biology*, vol. 7, no. 2, p. R13, 2006.

[55] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.-L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dujardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.-C. Lepla, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouz?, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.-J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. V. de Peer, and D. Rokhsar, "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).," *Science*, vol. 313, no. 5793, pp. 1596–1604, 2006.

[56] A. Ratzka, H. Vogel, D. Kliebenstein, T. Mitchell-Olds, and J. Kroymann, "Disarming the mustard oil bomb.," *PNAS*, vol. 99, no. 17, pp. 11223–8, 2002.

[57] A. Fiebig, R. Kimport, and D. Preuss, “Comparisons of pollen coat genes across Brassicaceae species reveal rapid evolution by repeat expansion and diversification.,” *PNAS*, vol. 101, no. 9, pp. 3286–3291, 2004.

[58] M. Schein, Z. Yang, T. Mitchell-Olds, and K. J. Schmid, “Rapid evolution of a pollen-specific oleosin-like gene family from Arabidopsis thaliana and closely related species.,” *Mol Biol Evol*, vol. 21, no. 4, pp. 659–669, 2004.

[59] T. Newman, F. J. de Bruijn, P. Green, K. Keegstra, H. Kende, L. McIntosh, J. Ohlrogge, N. Raikhel, S. Somerville, M. Thomashow, E. Retzel, and C. Somerville, “Genes Galore: A Summary of Methods for Accessing Results from Large-Scale Partial Sequencing of Anonymous Arabidopsis cDNA Clones,” *Plant Physiol*, vol. 106, no. 4, pp. 1241–1255, 1994.

[60] M. T. Nishimura, M. Stein, B.-H. Hou, J. P. Vogel, H. Edwards, and S. C. Somerville, “Loss of a Callose Synthase Results in Salicylic Acid-Dependent Disease Resistance,” *Science*, vol. 301, no. 5635, pp. 969–972, 2003.

[61] S. Fowler and M. F. Thomashow, “Arabidopsis Transcriptome Profiling Indicates That Multiple Regulatory Pathways Are Activated during Cold Acclimation in Addition to the CBF Cold Response Pathway,” *Plant Cell*, vol. 14, no. 8, pp. 1675–1690, 2002.

[62] H. Goda, S. Sawa, T. Asami, S. Fujioka, Y. Shimada, and S. Yoshida, “Comprehensive Comparison of Auxin-Regulated and Brassinosteroid-Regulated Genes in Arabidopsis,” *Plant Phys*, vol. 134, pp. 1555–1573, 2004.

[63] S. Rozen and H. Skaletsky, “Primer3 on the www for general users and for biologist programmers,” in *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (S. Krawetz and S. Misener, eds.), pp. 365–386, Humana Press, Totowa, NJ, 2000.

[64] K. Murase, H. Shiba, M. Iwano, F.-S. Che, M. Watanabe, A. Isogai, and S. Takayama, “A Membrane-Anchored Protein Kinase Involved in Brassica Self-Incompatibility Signaling,” *Science*, vol. 303, no. 5663, pp. 1516–1519, 2004.

[65] B. Ewing and P. Green, “Base-calling of automated sequencer traces using phred. ii. error probabilities,” *Genome Res*, vol. 8, pp. 186–194, 1998.

[66] D. Gordon, C. Abajian, and P. Green, “Consed: A Graphical Tool for Sequence Finishing,” *Genome Res*, vol. 8, no. 3, pp. 195–202, 1998.

[67] D. Gordon, C. Desmarais, and P. Green, “Automated Finishing with Autofinish,” *Genome Res*, vol. 11, no. 4, pp. 614–625, 2001.

[68] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman, “Basic Local Alignment Search Tool,” *J Mol Biol*, vol. 215, no. 3, pp. 403–410, 1990.

[69] E. Birney, M. Clamp, and R. Durbin, “GeneWise and Genomewise,” *Genome Res*, vol. 14, no. 5, pp. 988–995, 2004.

[70] A. Lukashin and M. Borodovsky, “GeneMark.hmm: new solutions for gene finding,” *Nucleic Acids Res*, vol. 26, no. 4, pp. 1107–1115, 1998.

[71] C. Burge and S. Karlin, “Prediction of complete gene structures in human genomic dna,” *J Mol Biol*, vol. 268, pp. 78–94, 1997.

[72] S. Lewis, S. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C. Wiel, L. Bayraktaroglu, E. Birney, M. Crosby, J. Kaminker, B. Matthews, S. Prochnik, C. Smith, J. Tupy, G. Rubin, S. Misra, C. Mungall, and M. Clamp, “Apollo: a sequence annotation editor,” *Genome Biology*, vol. 3, no. 12, pp. research0082.1–0082.14, 2002. This article is part of a series of refereed research articles from Berkeley Drosophila Genome Project, FlyBase and colleagues, describing Release 3 of the Drosophila genome, which are freely available at http://genomebiology.com/drosophila/.

[73] J. Felsenstein, “PHYLIP Inference Package version 3.64,” 2005.

[74] J. Huelsenbeck, F. Ronquist, R. Nielsen, and J. Bollback, “Bayesian inference of phylogeny and its impact on evolutionary biology,” *Science*, vol. 294, pp. 2310–2314, 2001.

[75] Z. Yang, “PAML: a program package for phylogenetic analysis by maximum likelihood,” *Comput. Appl. Biosci.*, vol. 13, no. 5, pp. 555–556, 1997.

[76] Z. Yang, “On the varied pattern of evolution of 2 fungal genomes: A critique of hughes and friedman,” *Mol Biol Evol*, vol. 23, pp. 2279–2282(4), 15 December 2006.

[77] S. Sawyer, “Statistical tests for detecting gene conversion,” *Mol Biol Evol*, vol. 6, no. 5, pp. 526–538, 1989.

[78] C. for eukaryotic structural genomics (CESG), “X-ray structure of putative tropinone reductase from Arabidopsis thaliana At1g07440. X-RAY CRYSTALLOGRAPHY (2.1 ANGSTROMS),” 2005.

[79] W. L. DeLano, “The PyMOL Molecular Graphics System ,” 2005.

[80] X. Gu and K. V. der Velden, “DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family ,” *Bioinformatics*, vol. 18, no. 3, pp. 500–501, 2002.

[81] X. Gu, “Maximum-Likelihood Approach for Gene Family Evolution Under Functional Divergence,” *Mol Biol Evol*, vol. 18, no. 4, pp. 453–464, 2001.

[82] S. Rama Devi, X. Chen, D. J. Oliver, and C. Xiang, “A novel high-throughput genetic screen for stress-responsive mutants of arabidopsis thaliana reveals new loci involving stress responses,” *Plant J*, vol. 47, no. 4, pp. 652–663, 2006.

[83] J. S. Johnston, A. E. Pepper, A. E. Hall, Z. J. Chen, G. Hodnett, J. Drabek, R. Lopez, and H. J. Price, “Evolution of Genome Size in Brassicaceae,” *Ann Bot*, vol. 95, no. 1, pp. 229–235, 2005.

[84] I. A. Al-Shehbaz, “Taxonomy and Phylogeny of Arabidopsis (Brassicaceae),” in *The Arabidopsis Book* (C. Somerville and E. Meyerowitz, eds.), American Society of Plant Biologists, 2002.

[85] S. Wright and D. Finnegan, “Genome evolution: Sex and the transposable element,” *Curr Biol*, vol. 11, no. 8, pp. R296–R299, 2001.

[86] T. Ohta, “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana,” *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[87] X. Yang, G. A. Tuskan, and M. Z.-M. Cheng, “Divergence of the Dof Gene Families in Poplar, Arabidopsis and Rice Suggests Multiple Modes of Gene Evolution after Duplication,” *Plant Physiol*, p. pp.106.083642, 2006.

[88] M. L. Durbin, B. McCaig, and M. T. Clegg, “Molecular evolution of the chalcone synthase multigene family in the morning glory genome,” *Plant Mol Biol*, vol. 42, no. 1, pp. 79–92, 2000.

[89] I. K. Jordan, Y. Wolf, and E. Koonin, “Duplicated genes evolve slower than singletons despite the initial rate increase,” *BMC Evolutionary Biology*, vol. 4, no. 1, p. 22, 2004.

[90] A. Navarro-Quezada and D. J. Schoen, “Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes,” *PNAS*, vol. 99, no. 1, pp. 268–273, 2002.

[91] D. Birgit, “Tropinone reductases, enzymes at the branch point of tropane alkaloid metabolism,” *Phytochemistry*, vol. 67, no. 4, pp. 327–337, 2006.

[92] S. Nyman, “Incorporation of arginine, ornithine and phenylalanine into tropane alkaloids in suspension-cultured cells and aseptic roots of intact plants oi Atropa belladonna,” *J Exp Bot*, vol. 45, no. 7, pp. 979–986, 1994.

[93] M. Mondragon-Palomino, B. C. Meyers, R. W. Michelmore, and B. S. Gaut, “Patterns of Positive Selection in the Complete NBS-LRR Gene Family of Arabidopsis thaliana,” *Genome Res*, vol. 12, no. 9, pp. 1305–1315, 2002.

[94] A. L. Hughes and R. Friedman, “Parallel Evolution by Gene Duplication in the Genomes of Two Unicellular Fungi,” *Genome Res*, vol. 13, no. 5, pp. 794–799, 2003.

[95] S. De Bodt, G. Theissen, and Y. Van de Peer, "Promoter Analysis of MADS-Box Genes in Eudicots Through Phylogenetic Footprinting," *Mol Biol Evol*, vol. 23, no. 6, pp. 1293–1303, 2006.

[96] T. Obayashi, K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito, and H. Ohta, "ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis," *Nucleic Acids Res*, vol. 35, no. suppl.1, pp. D863–869, 2007.

[97] E. P. C. Rocha, "The quest for the universals of protein evolution," *Trends Genet*, vol. 22, no. 8, pp. 412–416, 2006.

[98] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. Bergman, J. Nathan, and M. Kreitman, "Functional Evolution of a cis-Regulatory Module," *PLoS Biology*, vol. 3, no. 4, pp. e93, 588–598, 2005.

[99] B. Lemos, B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl, "Evolution of Proteins and Gene Expression Levels are Coupled in Drosophila and are Independently Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions," *Mol Biol Evol,*, vol. 22, no. 5, pp. 1345–1354, 2005.

[100] B. Lemos, C. D. Meiklejohn, M. Cáceres, and D. L. Hartl, "Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories," *Evolution*, vol. 59, no. 1, pp. 126–137, 2005.

[101] K. D. Makova and W.-H. Li, "Divergence in the Spatial Pattern of Gene Expression Between Human Duplicate Genes," *Genome Res*, vol. 13, no. 7, pp. 1638–1645, 2003.

[102] C. I. Castillo-Davis, D. L. Hartl, and G. Achaz, "cis-Regulatory and Protein Evolution in Orthologous and Duplicate Genes," *Genome Res.*, vol. 14, no. 8, pp. 1530–1536, 2004.

[103] X.-Y. Ren, M. W. Fiers, W. Stiekema, and J.-P. Nap, "Local Coexpression Domains of Two to Four Genes in the Genome of Arabidopsis," *Plant Physiol*, vol. 138, no. 2, pp. 923–934, 2005.

[104] J. M. Duarte, L. Cui, P. K. Wall, Q. Zhang, X. Zhang, J. Leebens-Mack, H. Ma, N. Altman, and C. W. de Pamphilis, "Expression Pattern Shifts Following Duplication Indicative of Subfunctionalization and Neofunctionalization in Regulatory Genes of Arabidopsis," *Mol Biol Evol*, vol. 23, no. 2, pp. 469–478, 2005.

[105] D. Weigel, K. Lab, P. Lab, H. Lab, and N. K.-D. Lab, "Atgenexpress microarray data." Website contains Abiotic, Pathogen, Ecotype, Light treatment Data.

[106] A. Navarro-Quezada, S. Gebauer-Jung, and K. Schmid, "Adaptive radiation of a tandemly repeated short chain dehydrogenase encoding gene family in Brassicales," 2007.

[107] M. Blanchette and M. Tompa, "Footprinter: a program designed for phylogenetic footprinting," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3840–3842, 2003.

[108] B. Morgenstern, K. Frech, A. Dress, and T. Werner, "DIALIGN: Finding local similarities by multiple sequence alignment," *Bioinformatics*, vol. 14, no. 3, pp. 290–294, 1998.

[109] G. Haberer, T. Hindemitt, B. C. Meyers, and K. F. X. Mayer, "Transcriptional Similarities, Dissimilarities, and Conservation of cis-Elements in Duplicated Genes of Arabidopsis," *Plant Physiol*, vol. 136, pp. 3009–3022, 2004.

[110] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment," *Mol Biol+*, vol. 302, pp. 205–217, 2000.

[111] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[112] N. Saitou and M. Nei, “The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees,” *Mol Biol Evol*, vol. 4, no. 4, pp. 406–425, 1987.

[113] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak, “VISTA: visualizing global DNA sequence alignments of arbitrary length,” *BIOINFORMATICS*, vol. 16, no. 11, pp. 1046–1047, 2000.

[114] E. Bonnet and Y. V. de Peer, “zt: a software tool for simple and partial Mantel tests,” *Journal of Statistical software*, vol. 7, no. 10, pp. 1–12, 2002.

[115] H. C. Causton, J. Quackenbush, and A. Brazma, *Microarray, Gene Expression Data Analysis*. Blackwell Publishing, 2003.

[116] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, and J. U. Lohmann, “A gene expression map of Arabidopsis thaliana development,” *Nat Genet*, 2005.

[117] Z. Wu and R. A. Irizarry, “Preprocessing of oligonucleotide array data,” *Nature Biotechnol*, vol. 22, pp. 656 – 658, 2004.

[118] J. Quackenbush, “Microarray data normalization and transformation,” *Nature Genet suppl*, vol. 32, pp. 496–501, 2002.

[119] T. Munzner, F. Guimbretiere, S. Tasiran, L. Zhang, and Y. Zhou, “TreeJuxtaposer: Scalable Tree Comparison using Focus + Context with Guaranteed Visibility,” *Acm Transactions on Graphics*, pp. 453–462, 2003.

[120] P. Hilson, J. Allemeersch, T. Altmann, S. Aubourg, A. Avon, J. Beynon, R. P. Bhalerao, F. Bitton, M. Caboche, B. Cannoot, V. Chardakov, C. Cognet-Holliger, V. Colot, M. Crowe, C. Darimont, S. Durinck, H. Eickhoff, A. F. de Longevialle, E. E. Farmer, M. Grant, M. T. Kuiper, H. Lehrach, C. Leon, A. Leyva, J. Lundeberg, C. Lurin, Y. Moreau, W. Nietfeld, J. Paz-Ares, P. Reymond, P. Rouze, G. Sandberg, M. D. Segura, C. Serizet, A. Tabrett, L. Taconnat, V. Thareau,

P. Van Hummelen, S. Vercruysse, M. Vuylsteke, M. Weingartner, P. J. Weisbeek, V. Wirta, F. R. Wittink, M. Zabeau, and I. Small, "Versatile Gene-Specific Sequence Tags for Arabidopsis Functional Genomics: Transcript Profiling and Reverse Genetics Applications," *Genome Res.*, vol. 14, no. 10b, pp. 2176–2189, 2004.

[121] X. Dong, M. Mindrinos, K. R. Davis, and F. M. Ausubel, "Induction of Arabidopsis Defense Genes by Virulent and Avirulent Pseudomonas syringae Strains and by a Cloned Avirulence Gene," *Plant Cell*, vol. 3, no. 1, pp. 61–72, 1991.

[122] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time RT-PCR," *Nucleic Acids Res*, vol. 29, no. 9, pp. e45–, 2001.

[123] M. S. Taylor, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and C. A. M. Semple, "Heterotachy in Mammalian Promoter Evolution," *PLoS Genetics*, vol. 2, no. 4, 2006.

[124] M. Z. Ludwig, "Funtional Evolution of Non-coding DNA," *Curr Opin Genet Dev*, vol. 12, no. 6, pp. 634–639, 2002.

[125] M. A. Koch, B. Weisshaar, J. Kroymann, B. Haubold, and T. Mitchell-Olds, "Comparative Genomics and Regulatory Evolution: Conservation and Function of the Chs and Apetala3 Promoters," *Mol Biol Evol*, vol. 18, no. 10, pp. 1882–1891, 2001.

[126] T. Eichner, "Bioinformatische Analyse von Promotorregionen und Genexpressionsprofilen der Tropinone-reductase like Genfamilie von Arabidopsis thaliana," Master's thesis, Friedrich Schiller Universitaet Jena, 2006.

[127] C. Chang and T. Sun, "Characterization of cis-regulatory regions responsible for developmental regulation of the gibberellin biosynthetic gene GA1 in Arabidopsis thaliana," *Plant Mol Biol*, vol. 49, no. 6, pp. 579–589, 2002.

[128] G. Marais, P. Nouvellet, P. D. Keightley, and B. Charlesworth, “Intron size and exon evolution in drosophila,” *Genetics*, vol. 170, pp. 481–485, 2005.

[129] K. Iida, M. Seki, T. Sakurai, M. Satou, K. Akiyama, T. Toyoda, A. Konagaya, and K. Shinozaki, “Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences,” *Nucleic Acids Research*, vol. 32, no. 17, pp. 5096–5103, 2004.

[130] B. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*. London: Chapman and Hall, 1991.

[131] K. Morris, S. A. H.-Mackerness, T. Page, F. C. John, A. M. Murphy, J. P. Carr, and V. Buchanan-Wollaston, “Salicylic acid has a role in regulating gene expression during leaf senescence,” *Plant J*, vol. 23, no. 5, pp. 677–685, 2000.

[132] T. J. Vision, “Gene order in plants: a slow but sure shuffle,” *New Phytologist*, vol. 168, no. 1, pp. 51–60, 2005.

[133] E. J. B. Williams and D. J. Bowles, “Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana,” *Genome Res*, vol. 14, pp. 1060–1067, 2004.

[134] L. M. Weaver, S. Gan, B. Quirino, and R. M. Amasino, “A comparison of the expression patterns of several senescence-associated genes in response to stress and hormone treatment,” *Plant Mol Biol*, vol. 37, no. 3, pp. 455–469, 1998.

[135] T. Eulgem, P. J. Rushton, S. Robatzek, and I. E. Somssich, “The WRKY superfamily of plant transcription factors,” *Trends in Plant Sci*, vol. 5, no. 5, pp. 199–206, 2000.

[136] Y. Noutoshi, T. Ito, M. Seki, H. Nakashita, S. Yoshida, Y. Marco, K. Shirasu, and K. Shinozaki, “A single amino acid insertion in the wrky domain of the arabidopsis tir-nbs-lrr-wrky-type disease resistance protein slh1 (sensitive to

low humidity 1) causes activation of defense responses and hypersensitive cell death," *Plant J*, vol. 43, no. 6, pp. 873–888, 2005.

[137] K. Toufighi, S. M. Brady, R. Austin, E. Ly, and N. J. Provart, "The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses," *Plant J*, vol. 43, pp. 153–163, 2005.

[138] L. D. Hurst, C. Pal, and M. J. Lercher, "The evolutionary dynamics of eukariotic gene order," *Nature Rev Genet*, vol. 5, no. 4, pp. 299–310, 2004.

[139] S. Zhan, J. Horrocks, and L. N. Lukens, "Islands of co-expressed neighbouring genes in arabidopsis thaliana suggest higher-order chromosome domains," *Plant J*, vol. 45, no. 3, pp. 347–357, 2006.

[140] D. Dixon, A. Lapthorn, and R. Edwards, "Plant glutathione transferases," *Genome Biology*, vol. 3, no. 3, pp. reviews3004.1–reviews3004.10, 2002.

[141] F. Francis, N. Vanhaelen, and E. Haubruge, "Glutathione s-transferases in the adaptation to plant secondary metabolites in the *Myzus persicae aphid*," *Arch Insect Biochem*, vol. 58, no. 3, pp. 166–174, 2005.

[142] R. Edwards, D. P. Dixon, and V. Walbot, "Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health," *Trends in Plant Sci*, vol. 5, no. 5, pp. 193–198, 2000.

[143] U. Wagner, R. Edwards, D. P. Dixon, and F. Mauch, "Probing the diversity of the arabidopsis glutathione s-transferase gene family," *Plant Mol Biol*, vol. 49, no. 5, pp. 515–532, 2002.

[144] B. McGonigle, S. J. Keeler, S.-M. C. Lau, M. K. Koeppe, and D. P. O'Keefe, "A Genomics Approach to the Comprehensive Analysis of the Glutathione S-Transferase Gene Family in Soybean and Maize," *Plant Physiol*, vol. 124, no. 3, pp. 1105–1120, 2000.

[145] A. Moons, "Osgstu3 and osgtu4, encoding tau class glutathione s-transferases, are heavy metal- and hypoxic stress-induced and differentially salt stress-responsive in rice roots," *FEBS Lett*, vol. 553, no. 3, 2003.

[146] E. L. L. Sonnhammer and E. V. Koonin, "Orthology, paralogy and proposed classification for paralog subtypes," *Trends Genet*, vol. 18, no. 12, pp. 619–620, 2002.

[147] H. Doddapaneni, R. Chakraborty, and J. S. Yadav, "Genome-wide structural and evolutionary analysis of the p450 monooxygenase genes (p450ome) in the white rot fungus phanerochaete chrysosporium : Evidence for gene duplications and extensive gene clustering," *BMC Genomics*, vol. 6, 2005.

[148] H. M. H. Goldstone and J. J. Stegeman, "A revised evolutionary history of the cyp1a subfamily: Gene duplication, gene conversion, and positive selection," *J Mol Evol*, vol. 62, no. 6, pp. 706–718, 2006.

[149] B. N. Kunkel and D. M. Brooks, "Cross talk between signaling pathways in pathogen defense," *Curr Opin Plant Biol*, vol. 5, no. 4, pp. 325–331, 2002.

[150] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, 1998.

Aura R. Navarro-Quezada

Curriculum Vitae

Contact

anavarro@ice.mpg.de

http://www.ice.mpg.de/dbs-staff/hopa/auna2601/web/main_en.htm

General Research Interests

Population Genetics, Genomics, Development, Host-parasite interactions.

Education

Ph.D., Biology, in progress (expected Summer 2007)
Ludwig Maximilians-Universität München
Munich, Germany and
'International Research School on Exploration of Ecological Interactions with Chemical and Molecular Techniques' at the Max Planck Institute of Chemical Ecology, Jena, Germany

Thesis: *Molecular Evolution of tropinone-reductase-like and tau GST genes duplicated in tandem in Brassicaceae*
Adviser: Dr. Karl Schmid- Prof. John Parsch
Committee: Prof. Wolfgang Stephan, Prof. John Parsch, Prof. Susanne Renner, Prof. Dario Leister

M.S. Biology, October 2001
McGill University
Montreal, Canada

Thesis: *Sequence Evolution of copia-like Retrotransposons in diverse Plant Genomes*
Adviser: Prof. Daniel J. Schoen
Thesis Reviewed: Prof. Graham Bell, Prof. Damian Labuda

GPA:3.7

B.S., Biology, June 1999
Universidad Nacional Autónoma de México (UNAM)
Mexico City, Mexico
Honors Thesis: *Population Genetics of* Agave subsimplex*,* Agave desertii *and* Agave cerulata *using RAPD markers*
Adviser: Prof. Luis E. Eguiarte
Average grade: 96%

Deutsches Abitur, May 1994
Deutsche Schule Alexander von Humboldt, Mexico City, Mexico.
Emphasis on German and Mathematics
Average grade: 2.0

Stipends

Deutsches Akademisches Austauschdienst- DAAD
Awarded in 2004-present, 780-1140 Euro/month

Consejo Nacional de Ciencia y Tecnologia, CONACyT
Awarded in 1999-2001 and 2003-2004, 9000 USD/year

Chaire du Canada en Génomique Forestière
Awarded in 2002-2003, 12000 USD/Year

Fundacion UNAM Support for Undergraduate Studies
Awarded in 1996 to 1999

Teaching Experience

1/2000–5/2000 Teaching Assistant in Genetics
McGill University Montreal
Canada

9/2001–12/2001 Teaching Assistant in Practical Cellular Biology
Université de Montréal
Montreal, Canada

9/2001–12/2001 Teaching Assistant in Botany Laboratory
Université de Montréal
Montreal, Canada

Publications

Peer Reviewed Journal Papers

Khasa, D., Pollefeys P., **Navarro-Quezada A.**, Perinet P. and J. Bousquet. 2005. Species-specific microsatellite markers to monitor gene flow between exotic poplars and their natural relatives in eastern North America. Molecular Ecology Notes 5 (4): 920-923.

Navarro- Quezada, A., González- Chauvet R., Molina-Freaner F. and L.E. Eguiarte. 2003. Genetic differentiation of the Agave deserti (Agavaceae) complex of the Sonoran desert. Heredity 90: 220-227.

Navarro- Quezada, A. and Daniel J. Schoen. 2002. Sequence evolution and copy number of Ty1- copia retrotransposons in diverse plant genomes. Proceedings of the National Academy of the Sciences 99: 268-273.

Diaz-Barriga, F., **Navarro- Quezada, A.**, Grijalva, M. I., Grimaldo, M., Loyola-Rodriguez, J.P. and M.D. Ortiz. 1997. Endemic Fluorosis in Mexico. Fluoride 30: 233- 239.

Submitted Papers or in process

Navarro- Quezada, A., Gebauer-Jung, S. and K. J. Schmid. Of birth and death in the (gene) family: Evolution of TR-like genes in Brassicaceae. To be submitted to Genome Research.

Eichner, T., **Navarro- Quezada, A.** and K. J. Schmid. Evolution of promoter regions in the tropinone reductase gene family. To be submitted to Genome Biology.

Research Experience

Graduate Work

Presentations at Workhops and Conferences

- 10th Meeting of Evolutionary Biology in Marseille, France. Best Student Poster Award. Short Oral Presentation. September 2006.
- 'Otto Warburg International Summer School on Networks and Regulation' at the Max Planck Institut for Molecular Genetics, Berlin, Germany. 2005. Poster presentation.
- ESEB (European Society for Evolutionary Biology) in Krakow, Polen. 2005. Poster presentation.
- 'Summer School for Molecular Evolution and Diversity' in Nottingham University, U.K. Funded by the BBSRC. 2004. Oral Presentation.
- Oral Presentation at the Congress of the ACFAS (Francophone Association for Scientific Knowledge) on 'Development and optimisation of microsatellite markers in poplars (Populus spp.)'. May 2002.
- Congress of the American Society for Evolution. 2000. Poster presentation.
- 11th Latinamerican Congress of Botany. 1998. Poster presentation.

Undergraduate Work

- Honors Thesis Project funded by UNAM: Genetic Diversity in three *Agave spp.* from Baja California, Mexico. The project included one field collection and implementation of molecular markers (RAPDs).

- Participated in the CONACyT funded Summer of Science at the Faculty of Medicine in the Universidad de San Luis Potosí in 1996 and 1997. Performed research on Human Toxicology.

Miscellaneous

Citizenship: Mexican

Languages: Spanish, English, German and French.

17th October 2007

Acknowledgements

The thesis work was financed by the Deutsches Akademisches Austauschdienst (DAAD), the Max Planck Gesellschaft and the first months in Munich by the mexican government through the Consejo Nacional de Ciencia y Teconologia (CONACyT). For the lab work, funds came from the Max Planck Gesellschaft and a project to my supervisor from Human Frontiers in Science.

More acknowledgements

Thanks to Theresia Eichner who was a great colleague and huge help with the microarray analysis. It was fun to collaborate with her, but also with Steffi Gebauer-Jung, who introduced me a bit to python, was always ready to build new scripts and correct the old ones, and answered my unix questions patiently.
Further thanks to Tom Mitchell-Olds, even if we had few discussions, he was able to transmit his enthusiasm for plant population genetics, and for giving me the opportunity to join his group at the MPI in Jena. I also thank John Parsch, who always was ready to give me good and clear advise, personally or over email, who read over my DAAD proposals and supported my thesis at the LMU. Thanks to Karl Schmid, who had the idea of working with the tropinone-reductase-like gene family, and supported me for three months (from an Emmy-Noether Grant) before I obtained the DAAD financing. He was a critical adviser, and made helpful comments, specially for making better figures. Thanks to Anne Kupzoc, Brad Rauh,

Amy Lawton-Rauh, Maria Clauss, Sebastian Ramos-Onsins, Juliette DeMeaux, Eric Schranz, Aaron Windsor, Jürgen Kroyman, Heiko Vogel, Ryan Oyama, Chris Wheat and Alice Schumate for either scientific discussions and/or answering technical questions. Jette Ringy-Beckstein was a great lab assistance, and also the reliable Tina Riedel and Claudia Wenderoth. Bao-Hua Song for discussion and great company in the ESEB trip to Krakow.

Further thanks to my colleagues in Munich, for comments on my presentations, specially: Laura Rose and John Baines. For more answers on technical questions, I thank Pleuni Pennings and Ann Arunyawat, and the rest of the colleagues in Munich for 'hilfsbereitschaft'. Also to the International Max Planck Research School people in Jena, always helpful, specially Karin Groten. For critical advise, Professors Ian Baldwin and David Heckel. Dr. Heckels group (Alexie, etc.), for insightful discussions at the group seminar. Thanks to Mariana Mondragón-Palomino who made critical comments to Chapter 1 on this thesis and for the chats.

Thanks to the people in my committee: Wolfgang Stephan, Susanne Renner and Dario Leister.

Discussions with neurobiologist Amaya Miquelajauregui and microbiologist Laila Partida, made research (and Mittel-Deutschland) much more fun. Manu Gaquerel, who I met in the last year of my thesis, gave me support, lots of good suggestions and helped me improve the figures in this thesis. He helped with insight into plant physiology and french culture. Last but not least to Andrea Navarro-Quezada, my parents Hugo and Rocio who supported me (thanks to webcam).