Ghent University
Faculty of Sciences
University Department of Molecular Genetics
VIB Department of Plant Systems Biology
Research group of Bioinformatics and Evolutionary Genomics

Power in numbers: *In silico* analysis of multigene families in *Arabidopsis thaliana*.

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Sciences, Biotechnology

September 2003

Jeroen Raes

Promotor: Prof. Dr. Yves Van de Peer

Het heeft eventjes geduurd en het heeft heel wat doorzettingsvermogen gevraagd, maar hier zijn we dan: hoofdstukken geschreven, inhoudstabel zonder fouten (hoop ik) en klaar voor een hoop printproblemen. En als je dan terugkijkt op al die jaren, besef je pas hoeveel mensen een kleinere of grote bijdrage hebben geleverd om tot hier te geraken. Wat nu volgt is een schamele poging om al die mensen te bedanken. In chronologische volgorde leek mij het handigst.

Yves, toen ik een eeuwigheid geleden binnenstapte in dat kleine burootje op de UIA had ik geen flauw idee waar ik aan begon… maar na een half uurtje van jouw enthousiasme over de wondere wereld van evolutieonderzoek was ik verkocht. Bedankt voor die woelige eerste jaren van leugenpaleis-imitaties, koffies met cognac en bijzonder onwetenschappelijke postkaartjes (science? what science?).

This naive and happy period was unfortunately followed by a few setbacks (which I'll describe in full detail after a beer or 2), which forced me to look for a different lab to be able to stay in science… which brought me to the lab in Gent (a much more international place, hence the language change).

Merci Pierre, pour m'avoir donné une chance quand je n'en avais plus aucune. Une toute petite phrase, mais grâce a ce geste de ta part, je suis là où je suis maintenant, et pour cela je te remercie beaucoup. Ook veel dank aan Marc Zabeau, voor zijn geloof in bioinformatica en om de groep (en mij erbij) de kans te hebben gegeven uit te groeien wat we nu zijn.

*Merci au club français de cette ancienne période, pour m'avoir appris plein de choses sur la Cuisine la Plus Meilleure du Monde, mais surtout sur mon propre pays (mais non c'est pas dégueu les frites à la mayonnaise!): Magali (j'attends toujours mes anchois "faits maison"), Toto.pl …euh Patrice (le Dieu de perl – blip blip !), Cathérine (non, pas la mienne), Vincent (pour m'avoir appris le "vrai" français… s*le p*t*ss*!), Sylvie (pour des conneries générales mais très amusantes), Stéphanie (viva Ché!), Enric (d'accord: demi-français) et finalement un très grand merci à toi Seb, pour tout ton soutien et ton amitié (et conneries générales – café?).*

After two years of virtually working in France, my former advisor decided he couldn't live without me and became the new head of our team… which gives him exclusive rights to two thank yous…

Bedankt Yves, ik ken weinig zo'n gedreven en enthousiaste wetenschappers, die er tegelijkertijd in slagen om zo'n goeie en sympathieke baas te zijn. Bedankt voor al je steun en vertrouwen, voor alle weekends en avonden die je aan papers, projecten en e-mail discussies gespendeerd hebt en vooral voor alles wat je me geleerd hebt - meest van al nog die dingen die niets met bioinformatica te maken hebben.

And of course a big thanks to all the current members of the BioInformatics team, a great gang of enthusiastic and fun people (unlike our reputation ;-): Stefanie (for secret eclairekes), Steven I (for general entertainment and Ghent dialect lessons), Jan (for the espresso – eternal gratitude!), Tine (for unspoilt young enthusiasm), Kobe (gastronomic partner in crime), Francis (fancy webdesign and many future programming tips), Sven (for crashing the cluster once in a while), Yvan (badminton partner in crime), Guy (for future brilliant algorithms), Cedric (a close second to yours truly in Monty Python quotes), Logistics Gevers (and especially the great guy behind the multinational), Belettering Vandepoele (CorelDraw guru and Westvluts teacher), Casanova De Vos (the name says it all), Carine (wet-lab stuff emergency helpdesk), Eric (the non-cheese-eating French guy – can you imagine?), Stephane (for late night crashing, wild drives to the train station and many other things...thanks Stephane for always being ready to help), Lieven (for teaching me the alumeur trick – I'm almost there!), Steven II (travel tips and coffee champion), Steven III (with the unbreakable camera), Dirk II (now you see me, now you don't!), and the numerous new guys and girls who will probably have arrived by the time people are actually reading this. Also all the other people form the front building whom I've relied on for so many things: Luc, Roland, Hendrik, Philippe, Martine, Jacques, Diane, Christine, Hilde (the old and the new) and Ann.

Doing science is collaborating, and this lab gives you plenty of opportunity to do that... thanks to all of you wet guys and girls to initiate me in the wonderous world of lab magic: Antje, Jørgen, Wout, Lieven, Juan-Antonio, Peter, Roos and Gerda. Thanks also to Dirk Inzé, for his (succesful) efforts to position this lab at the world top of plant research - something we all benefit from. A further big thank you to all the great friends I made in the lab – you know who you are (think 4 a.m. at the VIB seminar, think Fortgeschrittenenkurs) and also those from the real outside world for continuous distraction from work.

Tenslotte wil ik nog het thuisfront bedanken: grootouders (de trouwste supportersploeg die er is), Jos, Ilse en Sophie (aangetrouwd maar minstens even enthousiast), Melissa, a.k.a. Zus (+aanhangsel) en vooral mijn ouders: bedankt om mij al die jaren te steunen en mijn ding te laten doen – zonder jullie was het nooit gelukt.

Het schoonste volk komt altijd laatst: bedankt Katrien, voor alles.

Table of contents

Summary

The completion of the genome sequence of the model plant *Arabidopsis thaliana* represented a milestone in plant molecular biology. For the first time, the complete blueprint of a plant was available. Thanks to the automated gene prediction procedure set up by the different sequencing consortia, the full complement of *Arabidopsis* genes was rapidly made available to the scientific community. The first analysis of these genes further emphasised the notion that genes should not be considered as individual entities, as strong evidence was found for the ubiquitous existence of families of structurally and functionally related genes, having evolved from a common family ancestor through gene duplication and divergence. Consequently, it was realised that the function of these genes should be studied in the context of the whole family, as multiple (partially) redundant genes might be involved in the same process. Unfortunately, early studies evaluating the used gene prediction methods quickly showed that the prediction quality of the first annotation of the *Arabidopsis* genome was far from perfect, thus compromising further functional studies of the genes and their respective families.

In this thesis, we have taken advantage of the sequence conservation between family members to improve the structural and functional annotation of genes. This approach was the foundation of the Génoplante GeneFarm *Arabidopsis* re-annotation project, which was coordinated from within the bioinformatics team. In the framework of this project, we have tried to conceive a rigorous methodology for family-wise manual annotation, and developed a semi-automated routine to speed up this process. This method was successfully applied to the MYB family of transcription factors, of which 137 genes encoding members of this family were found and annotated in the *Arabidopsis* genome (chapter 2).

On the other hand, gene families did not merely serve as a means to improve annotation. As described in this work, the annotation of gene families can also be a first step in the analysis of the function of the different members and of the evolution of the family as a whole. As expected, this manual annotation repeatedly confirmed the poor quality of the publicly available (automated) gene prediction data.
Three gene family annotation studies are described in this work. The first one, done in close collaboration with the tree biotechnology research group led by Prof. W. Boerjan, is a study aimed at the elucidation of the toolbox necessary for monolignol biosynthesis in plants (chapter 3). As a first step towards this goal, we have surveyed the *Arabidopsis* genome *in silico* and identified and annotated all the homologues of all the monolignol biosynthesis genes known to date. Subsequently, the expression of these 34 genes was analysed by different methods. First, by RT-PCR on a complete tissue panel, second, by a refined *in silico* analysis of all *Arabidopsis* EST libraries, and third, by assembling all existing expression data available for these genes that is scattered in the literature. Then, we have carried out thorough phylogenetic analyses on each of the gene families as well as *in silico* promoter analyses. By integrating the results with the extensive expression data, we have identified 10 genes that are most likely involved in developmental lignification in the vascular tissues. Furthermore, we have identified a possible link between the biosynthesis of G lignin and the presence of the AC promoter element.

The second study, in collaboration with Prof. G. Theissen of the University of Jena, focused on the annotation and evolutionary analysis of the type I MADS-box genes in *Arabidopsis* and rice, which were largely uncharacterised before then (as opposed to its well-known type II sister class involved in flower development). In this study, we annotated all 47 members of the type I MADS-box gene family in *Arabidopsis thaliana* and exerted a thorough analysis of the C-terminal regions of the translated proteins. On the basis of conserved motifs in the C-terminal region, we were able to classify the gene family into three main groups, two of which could be further subdivided. Additional phylogenetic analysis revealed a significantly different dynamic of evolution in plant type I genes in comparison to animal type I (SRF) and plant type II (MIKC-type) genes (chapters 4 and 5).
The third study, in close collaboration with the cell cycle group led by Prof. D. Inzé, aimed at the characterisation of all core cell cycle genes in the *Arabidopsis* genome (chapter 6). In total, 61 genes were identified belonging to seven families of cell cycle regulators, of which 30 were new or corrections of the existing annotation. Phylogenetic analysis of these families allowed the determination of several subclasses. In addition, a new class of putative cell cycle regulators was found which are probably competitors of E2F/DP transcription factors, mediating the G1-to-S progression.

In order to further investigate the evolutionary history of these families, we wanted to relate their expansion with the large-scale duplication or polyploidy events that were postulated to have shaped the *Arabidopsis* genome as it is today. To do this, a tool (ADHoRe) to detect homologous regions within or between genomes was developed within the bioinformatics team (see addendum I). ADHoRe was used to reanalyse previously described duplicated regions found in the *Arabidopsis* genome, pointing at several large-scale duplication events in the evolutionary history of this model plant. Furthermore, the date of divergence of these duplicated blocks was determined based on silent substitution estimations between the paralogous genes and, where possible, by phylogenetic reconstruction. Based on these analyses, it was shown that previously used methods based on averaging protein distances of heterogeneous classes of duplicated genes lead to unreliable conclusions and that a large fraction of blocks duplicated much more recently than assumed previously. We found clear evidence for one large-scale gene or even complete genome duplication event somewhere between 70 to 90 million years ago. Traces pointing to a much older (probably more than 200 million years) large-scale gene duplication event could be detected as well and were later confirmed by other studies in our group (chapter 7).
Although in theory these genome duplications are hypothesised to have an important impact on the evolution of the duplicated genes and the species as a whole, the correlation of these events with the evolution of the investigated families does not allow drawing general conclusions. The future analysis of duplicated genes at the regulatory level, combined with an in-depth analysis of subtle functional shifts at the protein level (chapter 8), will hopefully allow to further clarify the impact of gene duplication on the complex system of processes that define plants.

Samenvatting

De voltooiing van de genoomsequentie van de modelplant *Arabidopsis thaliana* vormde een mijlpaal in de moleculaire plantenbiologie. Voor het eerst was de volledige blauwdruk van een plant beschikbaar. Dankzij de geautomatiseerde genpredictie-procedure, op punt gesteld door de verschillende sequeneringsconsortia, werd het volledige gamma van *Arabidopsis* genen snel toegankelijk gemaakt voor de wetenschappelijke gemeenschap.
De eerste analyse van deze genen versterkte verder het idee dat genen meestal niet beschouwd moeten worden als individuele entiteiten, aangezien families van structureel en functioneel verwante genen, onstaan uit een gemeenschappelijke voorouder door genduplicatie en divergentie, alomtegenwoordig bleken te zijn. Bijgevolg groeide het besef dat de functie van deze genen in de context van de gehele familie moest bestudeerd worden, aangezien meerdere (partieel) redundante genen in eenzelfde proces betrokken zouden kunnen zijn. Helaas toonden vroege studies die de gebruikte genpredictiemethoden evalueerden reeds snel aan dat de kwaliteit van de eerste annotatie van het *Arabidopsis* genoom verre van uitmuntend was, wat bijgevolg de verdere functionele studie van genen en hun respectievelijke families compromitteerde.

In dit proefschrift hebben we gebruik gemaakt van het behoud van sequentiesimilariteit tussen leden van een familie om de structurele en functionele annotatie van genen te verbeteren. Deze aanpak was de grondslag van het Génoplante GeneFarm *Arabidopsis* herannotatieproject, dat gecoördineerd werd vanuit de onderzoeksgroep bio-informatica. In het kader van dit project is er gepoogd om een rigoureuze methodologie voor familiegewijze manuele annotatie op te stellen, en is bovendien een semi-automatische procedure ontwikkeld om dit proces te versnellen. Deze methode is met succes toegepast op de MYB familie van transcriptiefactoren, waarbij 137 genen die leden van deze genfamilie encoderen gedetecteerd en geannoteerd werden in het *Arabidopsis* genoom (zie hoofdstuk 2).

Genfamilies dienden echter niet louter als middel om annotatie te verbeteren. Dit proefschrift beschrijft ook de annotatie van genfamilies als een eerste stap in de functionele analyse van de verschillende leden en het onderzoek naar de evolutie van de familie in zijn geheel. Zoals verwacht, bevestigde deze manuele annotatie herhaaldelijk de bedenkelijke kwaliteit van de publiek beschikbare (geautomatiseerde) genpredictie data.
Drie annotatiestudies van genfamilies worden in dit proefschrift beschreven. De eerste, uitgevoerd in nauwe samenwerking met de onderzoeksgroep biotechnologie van bomen o.l.v. Prof. W. Boerjan, is een studie die tot doel had de genetische 'toolbox', nodig voor biosynthese van monolignolen in planten, op te helderen (hoofdstuk 3). In een eerste stap hebben we het *Arabidopsis* genoom *in silico* gescreend en alle homologen van alle tot nu toe gekende monolignol biosynthese genen geannoteerd. Vervolgens werd de expressie van deze 34 genen geanalyseerd op basis van verschillende methoden: Ten eerste, door middel van RT-PCR op een compleet weefselpaneel, ten tweede, aan de hand van een doorgedreven *in silico* analyse van alle *Arabidopsis* EST collecties en ten derde, door alle

expressiedata die verspreid in de literatuur beschikbaar is voor deze genen samen te voegen.
Voorts werd van elke genfamilie een grondige fylogenetische analyse uitgevoerd evenals een *in silico* promoteranalyse. Door deze resultaten te integreren met expressiedata werden 10 genen geïdentificeerd die zeer waarschijnlijk betrokken zijn in ontwikkelingsgebonden lignifiëring in vasculaire weefsels. Bovendien hebben we een mogelijk verband kunnen leggen tussen G lignine biosynthese en de aanwezigheid van een AC promoterelement.
De tweede studie, in samenwerking met Prof. G. Theissen van de Universiteit van Jena, focuste op de annotatie en evolutionaire analyse van type I MADS-box genen in *Arabidopsis* en rijst, die totnogtoe grotendeels ongekend waren (in tegenstelling tot de welbekende type II klasse die betrokken is in bloemontwikkeling). In deze studie werden alle 47 leden van de type I MADS-box genfamilie in *Arabidopsis thaliana* geannoteerd en werd een grondige analyse van de C-terminale regio's van de respectievelijke proteinesequenties uitgevoerd. Op basis van geconserveerde motieven in de C-terminale regio werd de genfamilie in drie groepen ingedeeld, waarvan twee verder konden worden opgedeeld. Aanvullende fylogenetische analyse heeft geleid tot de ontdekking van een aanzienlijk verschillende evolutiedynamiek in plant type I genen in vergelijking tot dierlijke type I (SRF) en plant type II (MIKC-type) genen (hoofdstukken 4 en 5).
De derde studie, in nauwe samenwerking met de celcyclus onderzoeksgroep o.l.v. Prof. D. Inzé, had tot doel om alle 'core' celcyclusgenen te karakteriseren in het *Arabidopsis* genoom (hoofdstuk 6). In totaal werden 61 genen geïdentificeerd, behorend tot zeven families van celcyclus regulerende genen, waarvan 30 nieuw of correcties van bestaande annotatie. Fylogenetische analyse van deze families liet toe verschillende subklassen te definiëren. Bovendien werd een nieuwe klasse van vermeende celcyclus regulerende genen ontdekt, die waarschijnlijk 'competitors' zijn van de E2F/DP transcriptiefactoren, die de G1-naar-S transitie reguleren.

Om de evolutionaire geschiedenis van deze families verder uit te diepen, hebben we gepoogd om hun expansie te correleren met grootschalige gen- of genoomduplicaties die verantwoordelijk geacht worden voor de huidige structuur van het *Arabidopsis* genoom. Hiertoe werd een tool (ADHore) ontwikkeld binnen de bioinformatica onderzoeksgroep (zie addendum I). ADHoRe werd gebruikt om reeds beschreven gedupliceerde regio's in het *Arabidopsis* genoom, die wezen op meerdere grootschalige genduplicaties in de evolutiegeschiedenis van deze modelplant, te heranalyseren. Bovendien werd de divergentiedatum van deze gedupliceerde regio's bepaald aan de hand van schattingen van het aantal synonieme substituties tussen de paraloge genen en, waar mogelijk, aan de hand van fylogenetische reconstructies. Op basis van deze analyses kon aangetoond worden dat eerder gebruikte methoden, gebaseerd op gemiddelde proteïne evolutie-afstanden binnen heterogene groepen van gedupliceerde genen, tot onbetrouwbare resultaten leiden, en dat een groot aantal van de regio's veel recenter gedupliceerd was dan oorspronkelijk werd aangenomen. Er werden duidelijke aanwijzingen gevonden voor een grootschalige of zelfs complete genoomduplicatie zo'n 70 tot 90 miljoen jaar geleden. Er werden verder nog aanwijzingen gevonden voor een veel oudere duplicatiegebeurtenis (waarschijnlijk meer dan 200 miljoen jaar geleden) en dit werd bevestigd door latere studies binnen onze onderzoeksgroep (hoofdstuk 7).

Alhoewel deze genoomduplicaties in theorie verondersteld worden een belangrijke invloed te hebben op de evolutie van de gedupliceerde genen en de soort op zich, laat de correlatie van deze gebeurtenissen met de evolutie van de onderzochte genfamilies niet toe om algemene conclusies te trekken. De toekomstige analyse van gedupliceerde genen op het regulatorische niveau, gecombineerd met een diepgaande analyse van subtiele functionele verschuivingen op het proteïneniveau (hoofdstuk 8), zal hopelijk toelaten om de impact van genduplicatie op het complexe systeem van processen waaruit een plant bestaat, te verduidelijken.

[Chapter 1]

Introduction

In the late sixties, the pioneering work of Margaret Dayhoff and colleagues showed that genes should not be considered as individual entities, but that they can be grouped in families, based on their common evolutionary origin and subsequent sequence or structural similarities. Since then, mostly due to the advent of high-throughput sequencing approaches, an enormous abundance of related genes was discovered. Functional analysis of the genes constituting these families showed that, besides their sequence similarity, a strong functional relatedness, or even redundancy, existed between family members. On the other hand, in many of the cases, a rich diversity was observed in the processes in which the different family members acted. It became clear that, when investigating the function of a gene, this should be done in he context of the whole family, as a complex balance between divergence and redundancy between family members contributes to the role of each individual gene.

The study of gene families has an important role in a wide range of research domains, as the comparison of members of a family is a rich source of information. First of all, it allows the transfer of functional information between members. As such, it is used to assign functions to newly detected members in one species through their orthology with functionally characterised members in another (Eisen, 1998). Furthermore, through within-family sequence comparison, functionally important regions and residues can be detected in proteins, as well as in up- and downstream regulatory regions. Secondly, degenerate as well as specific primers and probes for diverse functional studies (RT-PCR, Northern, micro-arrays, etc.) are designed by comparing the different members of a family. Thirdly, the prediction of protein structure is often based on comparison with family members for which the structure has been determined experimentally (Mount, 2001). Finally, comparing gene family members can provide other kinds of information: the analysis of families, for example, has proven its use in addressing diverse evolutionary problems such as adaptation to new nutritional niches (Zhang et al., 2002) or the role of polyploidy in vertebrate genome evolution (Gu et al., 2002; Friedman and Hughes, 2003).

In many applications, it is of great importance to have an exhaustive dataset to come to reliable conclusions. For example, for knockout studies, it is necessary to know the whole set of homologous proteins found within the organism to be able to reliably interpret the observed (absence of) phenotypes.

With the arrival of the complete genome sequence of *Arabidopsis thaliana* (AGI, 2000), this has become possible. The genome has provided us with a first glimpse at the diversity of plant gene families and has shown us the similarities and differences in gene content between plants and animals. Since then, many studies have explored these families in depth, investigating the evolution, expression and function of the different members. However, some of these studies also have taught us that gene families can be very different in different species, with cases of species-specific family expansion (e.g. the actin family in *Petunia* containing up to 200 members where *Arabidopsis* has only eight;

Baird and Meagher 1987; Meagher et al., 2000), or complete absence of genes in some species (e.g. soluble adenylyl cyclase, which is present in human, rodents, *Dictyostelium* and bacteria, but absent in *Drosophila* ,*Caenorhabditis*, *Arabidopsis* and *Saccharomyces*; Roelofs and Van Haastert, 2002). Therefore, despite general expectations, it seems that the transfer of function from *Arabidopsis* genes to homologs in commercially or scientifically important plant species will not always be that trivial, because of the uncertainty on whether one is investigating the "real" ortholog in these plants. Consequently, large-scale genome sequencing programs as conducted for species such as *Populus* (Wullschleger et al., 2002), *Brassica* and *Medicago*, remain of crucial importance to take the full diversity of the family into account when inferring gene function in these species.

Duplication, duplication: the origin of gene families

By definition, gene families arise through duplication and subsequent divergence of genes (Dayhoff, 1974; Zuckerkandl, 1975; Ohta, 1990). Duplication seems to occur very frequently in plants: the rate of origin of new duplications in *Arabidopsis* is estimated at 2.2 duplications per gene per billion years (Lynch and Conery, 2000; 2001), and polyploidy is a widespread phenomenon in the plant world (see further; Wendel, 2000). Therefore, one can hardly be surprised by the discovery of numerous and large gene families after the sequencing of the *Arabidopsis thaliana* and *Oryza sativa* genomes. About 65% and 77% of genes in these two respective genomes are believed to be part of a gene family (AGI, 2000; Goff et al., 2002). The precise timing and nature of the duplications responsible for the expansion of the majority of these families remains, however, unclear. The duplication of genes can occur through a number of different processes: local tandem duplications, (partial) chromosomal duplication, polyploidy and retrotransposition (Ohno, 1970; Fryxell, 1996; Hughes, 1999; Graur and Li, 2000).

Tandem duplication

Tandem duplication by unequal crossing-over, first described for the Bar locus in *Drosophila* (Sturtevant, 1925; Bridges, 1936), is a well-known process for the procreation of repetitive sequences and genes. One of the most spectacular examples of the tandemly duplicated genes are those encoding ribosomal RNAs, which occur in long tandem arrays of up to 700 copies in some eukaryotic genomes, depending of the type of rRNA molecule coded (Brown and Dawid, 1968; Cronn et al., 1996). But, now that an increasing number of families are characterised, it becomes clear that also for protein-coding genes, tandem duplication is an important process in the expansion of gene families (Long and Dawid, 1980). Some of the numerous examples in plants include the cell wall associated kinase-like (WAKL) family (Verica and Ye, 2002), NBS-LRR pathogen resistance genes (Meyers et al., 2003) and gluthatione transferases (Dixon et al., 2002).

Polyploidy

Also polyploidy is an important contributor to gene family expansion in plants, as it has been estimated that 95% of pteridophytes and up to 80% of angiosperms are polyploid (Masterson, 1994; Leitch and Bennet, 1997). In addition, comparative mapping studies and computational sequence analyses have shown that many plant species, such as *Arabidopsis* (Blanc et al., 2000; Simillion et al., 2002) and maize (Gaut and Doebley, 1997) are probably paleopolyploids, which returned to the diploid state in the course of time through gene silencing, mutation, rearrangements and loss (Wendel, 2000; Kellogg, 2003). Polyploidy can occur through different mechanisms. First of all, nondisjunction of chromosomes during the meiosis can result in diploid gametes, which produce polyploid zygotes when united. Secondly, nondisjunction of chromosomes during the first mitotic division of a diploid zygote can lead to a tetraploid organism (Snustad et al., 1997).
Likewise, a plant meristematic cell can form a tetraploid cell line, which, when detached from the plant and rooted, can give rise to a tetraploid plant. In contrast to the previous process of *auto*polyploidy, *allo*polyploidy is commonly found in plants. This process occurs through the hybridisation of haploid gametes from related species. However, there is a risk that this results in a sterile hybrid if (some of) the homeologous chromosomes from both species are too divergent and consequently unable to synapse during meiosis. The plant can escape this fate if genome doubling follows the hybridisation, such that the chromosomes of each species can pair with their respective copies, resulting in plants containing two distinct 'diploid' genome sets (e.g. allotetraploid cotton; Snustad et al., 1997; Kellogg, 2003; Wendel, 2000).

Complete or partial chromosomal duplication

Through nondisjunction, the duplication event can also be limited to a single chromosome (aneuploidy, polysomy). The resulting chromosome imbalance often has a clear phenotypic effect, as in the classic study on *Daturia stramonium*, where 12 different capsule morphologies were observed as the result of trisomy of one of the 12 respective chromosomes (Blakeslee, 1934), and the well-known cases in human of Down or Klinefelter syndrome (Snustad et al., 1997).
Partial polysomy is also observed. As an example, a translocation of the long arm of human chromosome 21 on to the short arm of chromosome 14, leading in the second generation to individuals having a third, partial copy of chromosome 21 attached to chromosome 14, was shown to lead to Down syndrome (Abeliovich et al., 1985).

Transposition

Although a lot is still to be discovered, accumulating evidence seems to point at an important role of transposition in gene duplication. The discovery of so-called processed pseudogenes in plants as well as animals, mRNAs that are transcribed into cDNAs by reverse transcriptase and reinserted in the genome, pointed at one possible mechanism (Marx, 1982; Lewin, 1983; Drouin and Dover, 1987).

Due to the absence of upstream regulatory sequences after reinsertion, the inaccuracy of the reverse transcription process and the possible insertion at a genomic region that is not adequate for its proper expression, the most probable fate of these inserts is pseudogenisation (Graur and Li, 2000). If, however, these cDNAs are inserted through homologous recombination with the original gene, the resulting exonless gene might be expressed by the original promoter and function correctly (Fink, 1987). Alternatively, if the gene is inserted near a functional promoter of another gene, the transcript can acquire a function in a new transcriptional niche, possibly leading to a selective advantage (McCarrey and Thomas 1987; Brosius, 2003). In addition, reverse transcription and insertion of semi-processed genes carrying upstream regulatory regions was also shown to result in functional genes (Soares et al., 1985). Several examples of retrogenes that are expressed and functional have been described (Brosius, 1999 and references therein).
The actual insertion of a gene inside a retrotransposon and its subsequent replicative transposition might also allow the duplication of a gene. A retrotransposon containing a partial open reading frame of a plasma membrane proton ATPase gene has been observed in maize (Jin and Bennetzen, 1994), while a Spm/En-like transposon containing a complete and expressed MADS-box gene has also been described (Montag et al., 1996). However, although theoretically possible, the insertion of a complete gene or genomic region inside a retrotransposon has not been observed yet.

In silico analysis and characterisation of gene families

To reconstruct the evolutionary history of a gene family, investigate the functional divergence of genes and formulate hypotheses on the impact of the expansion of the family on the evolution of the organism as a whole, a correct and exhaustive characterisation of the family is necessary. Thanks to the (future) completion of the sequencing of many plant genomes, this is becoming possible. The annotation and delineation of gene families constitutes an important step towards the functional characterisation of the family. This section focuses on the methods available to do this and the lessons learned from the *Arabidopsis* genome. A general overview of the complete process is given in figure 1.

Detecting putative family members

The most widely used method to detect and characterise all members of a gene family in a particular genome is by using sequence similarity programs such as BLAST (Altschul et al., 1997) to find regions similar to known members of the family. This is preferably done at the protein level (TBLASTN), as family members are in many cases too divergent to be detected at the nucleotide level. Unfortunately, the detection of family members is sometimes hampered by the presence of introns, breaking up conserved regions of the gene. When, however, a first annotation of the genome has been performed (as is generally the case), the set of predicted proteins can be searched using BLASTP, thereby avoiding the problem of introns.

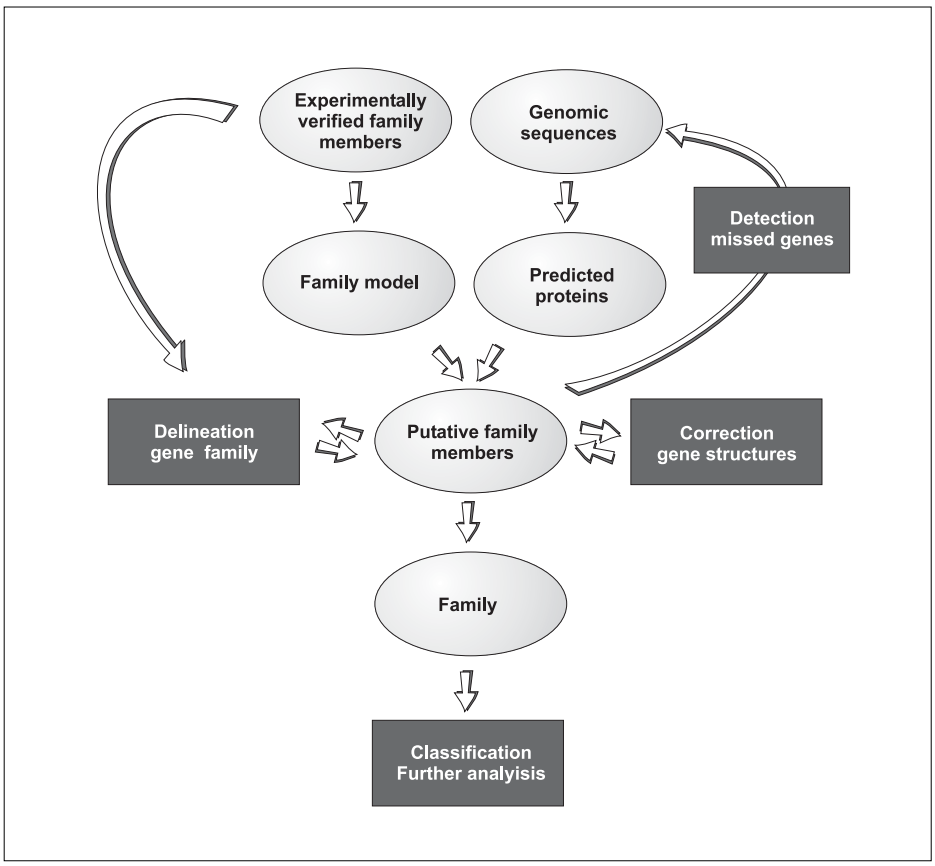

Figure 1. Overview of the gene family annotation process

Profile-based methods such as HMMER (Eddy, 1998) or PSI-BLAST (Altschul et al., 1997) provide a more sensitive alternative. Profiles are position-specific models of sequence composition within the family, generally based on Hidden Markov Models (Krogh, 1998), and are built using a large number of representative members of the family. Preferably one would like to build such a profile only based on sequences of certified family members, as this matrix constitutes the “blueprint” of the family of interest. With this matrix, one can search all predicted proteins for putative members of a family. The use of profiles is already well established in protein family characterisation, as is the case in the PFAM (Bateman et al., 2002) or PROSITE (Falquet et al., 2002) databases.

Whether BLAST or a profile-based method is used to scan the predicted proteome, one should always keep in mind that gene prediction is not error-free, and that some genes might have been missed in the annotation (see further), resulting in their absence in the dataset. Therefore, in order to detect these missed genes, the abovementioned homology search against the raw genome sequence itself remains a recommended safety precaution.

Delineation of the family

One of the hardest problems to resolve in gene family annotation comes up when a researcher is confronted with a long list of BLAST or HMMER hits, decreasing in quality when going further down the list, and has to decide which genes still belong to the family, and which ones do not.
The problem resides mainly in the definition of a gene family. When a gene family is defined as a group of homologous proteins (i.e. descending from a common ancestral gene), the statistical score (E-value) given by these programs is an important measurement for homology. An E-value of 10^{-6} is a sure indication for homology (W. Pearson, pers. comm.). However, in many cases, the gene family is defined as a subset of a larger "superfamily" of genes, where the different "subfamilies" are homologous, but more distantly related. Most of the time, the main interest of the researcher resides in this subfamily, which is more likely to contain functionally related genes. For example, within the large superfamily of phosphatases, one could only focus on the Ser/Thr phosphatase subfamily.
Fortunately, there are alternative approaches to delineate the borders of a gene family. First of all, the presence of known functionally important residues, domains or structures can be an important threshold to decide whether a gene belongs to a family or not (Kosarev et al., 2002).
This approach can be particularly useful when gene family members are very distantly related and only show similarity at the structural (secondary or tertiary) level. Second, the above-mentioned E-value score can give further indications to distinguish potential family members from false positives, or - in the case of large superfamilies - genes of other subfamilies. A clear "drop" in the E-value score can, for example in the case of a profile approach, be indicative that sequences below this threshold do not fulfil the family model as well as those above. However, this approach can potentially lead to wrong conclusions due to incomplete or biased sampling of the family for the profile. A third method is based on the phylogenetic analysis of the (super)family. A tree containing all known homologs and more distantly related members of distinct, well-known families within the same superfamily can be used to decide whether a protein belongs to the investigated family or not. As a rule of thumb, genes that significantly cluster together with experimentally certified family members can be considered to be part of the family. Of course, the building and interpretation of these trees is not always straightforward and has to be done with great care.

Structural annotation and improvement of existing annotation

In earlier stages of sequencing projects, the results of automated annotation are often not available. In addition, if available, this automated annotation is not flawless: especially the first releases of sequenced genomes still contain numerous missed or wrongly predicted genes, as was, for example, shown for *Saccaromyces cerevisiae* (Blandin et al., 2000), *Mycoplasma pneumoniae* (Dandekar et al., 2000), *Drosophila melanogaster* (Andrews et al., 2000; Gopal et al., 2001), *Caenorhabditis elegans* (Reboul et al., 2001) and *Arabidopsis thaliana* (Terryn et al.,1999; Haas et al., 2002). For this reason, the annotation or re-annotation of the gene structure of putative family members is a re-occurring theme for a researcher interested in a particular family.

The correct annotation of a genes' structure is generally founded on (the combination of) two approaches: one based on information from within the genome under investigation ('intrinsic' gene prediction) and another based on external sources, such as EST, cDNA or protein sequences ('extrinsic' gene prediction) (Mathé et al., 2002). Intrinsic gene prediction aims at predicting gene structures based on features of known genes within the same genome, such as compositional biases or codon usage (so-called 'content sensors') and/or signals such as splice sites and start or stop codons (signal sensors) (Mathé et al., 2002). In order to do this, intrinsic gene prediction software has to be 'trained' on a set of reliable gene structures for each genome separately, as software trained on or developed for one organism often produces inferior results when applied on another (Pavy et al., 1999; Pertea and Salzberg, 2002). Furthermore, as each method has its own strengths and weaknesses, a combination of several programs generally gives the best results (Pertea and Salzberg, 2002). In practice, one does not always have the opportunity to train software for the genome under study, as this is generally a specialised task and because the necessary tools are not always publicly available. And although several on-line prediction servers are available, which have been trained on several different organisms (for an overview, see Mathé et al., 2002), these tools are not available (yet) for all genomes currently being sequenced. In these cases, one would have to resort to software trained on a related organism, but results should be interpreted with caution.

Next to the intrinsic gene prediction, extrinsic methods provide a valuable way to determine, confirm or correct gene structures. First of all, the alignment of cognate ESTs and full-length cDNAs to the genomic region using programs such as Sim4 (Florea et al., 1998) allow to determine correct exon-intron borders. In general, only high-quality matching sequences (% identity > 95%) are taken into account and results are preferably manually inspected for wrong assignments and cross-matches between closely related family members. In addition, alignment of homologous protein sequences can be used to detect missing exons or wrongly predicted splice sites. The use of within-family sequence conservation is particularly useful to detect prediction errors from automated gene prediction pipelines.

Furthermore, annotation software tools such as ARTEMIS (Rutherford et al., 2000) allow to compile information from different sources and can be used to decide on a final gene structure.

Finally, predicted gene structures are preferably confirmed *in vitro*. Using automated tools for the design of gene-specific primers (e.g. SPADS; Thareau et al., in press), genes can be experimentally validated by, for example, RT-PCR.

Classification

Once the members of a gene family have been collected and their gene structure has been determined, one can determine whether subclasses exist within the family. This classification allows the transfer of function through the principles of *phylogenomics* (Eisen, 1998). Generally, classification of gene families into subclasses is done on the basis of structural motifs and/or phylogenetic analysis.

The presence or absence of certain conserved domains provides a rough classification of the gene family. Although it does not provide any fine-grained insights into the evolution of the genes, it does

provide important functional indications. Furthermore, it can complement phylogenetic analysis, if resolution of the tree is impossible due to, for example, too many family members and/or too few alignable positions (De Bodt et al., 2003). However, in general, phylogenetic analysis is the preferred tool to determine subclasses.

There are at least two general groups of methods of phylogenetic inference that can be applied to sequence data. One set of methods, to which belong maximum parsimony, maximum likelihood and Bayesian approaches, uses discrete character data, while the other set of methods, the so-called pairwise distance methods, is based on the computation of overall similarity or dissimilarity between the sequences. Maximum parsimony is historically the most widely used method. Parsimony methods for inferring phylogenies (Fitch, 1971) select that tree that minimises the total tree length, being the number of nucleic acid substitutions or amino acid replacements required to explain a given set of data. In practice, for each possible topology, the ancestral sequences at each branching point are reconstructed. Subsequently, the minimum number of substitutions to explain the sequence differences over the whole tree is computed. The tree topology requiring the smallest number of substitutions is chosen as the final tree.

Maximum likelihood (Felsenstein, 1981) seeks that tree that is most consistent with a set of sequences on a statistical basis. To apply a maximum likelihood approach, a concrete model of the evolutionary process that specifies the transition probabilities from one nucleotide or amino acid to another is used. Then, for all possible trees, given this model of evolution, the probability of the set of sequences having resulted from that particular tree topology is computed. This probability constitutes the likelihood of the data given that particular tree. The topology that shows the highest likelihood is chosen as the final tree.

Only recently, Bayesian methods (Huelsenbeck et al., 2001) have been developed to infer phylogenies. These methods are also based on an explicit model of evolution but, in contrast to Maximum Likelihood methods, the posterior probability, being the probability of the tree given the data and the evolutionary model, is used to find the most probable topology.

Distance methods, the second major group of tree inferring methods, fit a tree to a matrix of pairwise evolutionary distances. For every two sequences, the distance is a single value based on the fraction of positions in which both sequences differ, corrected for multiple substitutions by applying a specific evolutionary model that makes assumptions about the nature of evolutionary changes (cfr. maximum likelihood; Graur and Li, 2000). When all the pairwise distances have been computed for a set of sequences, a tree topology can then be inferred by a variety of clustering methods, the most well-known of which is probably the neighbor-joining method (Saitou and Nei, 1987).

In practice, phylogenetic analysis of a gene family is preferably performed using - next to the genes detected in the genome under study - a broad taxonomic sampling of homologs in other species. After alignment of the protein sequences (unless the sequences are very conserved, a protein alignment is preferred over a DNA alignment), the alignment should be checked and manually improved using alignment visualisation and editing tools such as BioEdit (Hall, 1999). Furthermore, nonalignable regions should be removed because, being a source of noise, they can seriously jeopardise the reliability of the obtained tree.

In addition, it is advisable to use multiple tree construction methodologies (see above) and compare the results. To infer the reliability of nodes, random sampling tests, such as the bootstrap, are used. Bootstrap (Felsenstein, 1985) is a technique in which a pseudo-alignment is sampled from the original alignment by picking columns at random. Because the sampling is done with replacement, the same columns can be selected several times, resulting in a pseudo-alignment in which some columns are overrepresented, while others are not. From this alignment a tree is constructed. By repeating this process multiple times (generally 100-1000 bootstrap replicates are performed), the reliability of nodes in the tree can be estimated by the number of times this node is found in the set of trees.
Finally, one should be very careful with interpretation of the tree, as nodes can indicate speciation as well as duplication events, and take processes like gene loss (or the absence of the gene from the databases) into account when drawing conclusions.

Functional annotation

Several *in silico* methods exist to learn more about the function of newly detected family members. For example, one can predict the subcellular localisation of the encoded protein using programs such as TargetP (Emanuelsson et al., 2000). Furthermore, many tools exist to detect posttranslational modifications such as myristoylation (Maurer-Stroh et al., 2002), glycosylation (Gupta and Brunak, 2002) and phosphorylation (Blom et al., 1999). Information about protein domains can also provide functional clues. Webservers such as Interpro (Mulder et al, 2003) provide the possibility to scan a protein sequence for known domains, while programs such as MEME (Bailey and Elkan, 1994) can detect additional, conserved domains in the family. In addition, an increasing amount of functional data becomes publicly available. For example, the massive EST sequencing efforts and the growing amount of publicly available data from micro-array experiments resources provide a valuable initial source of information on the expression of genes (Schultz et al., 2002).
By combining information from various sources with literature data, one has a first glance at the function of the different family members. Therefore, this first bioinformatics-driven analysis is of great importance for further functional studies. The *in silico* integration of knowledge gained from different sources allows the researcher to pinpoint target genes within the family, before engaging in expensive and time-consuming experimental research. By focussing on a small subset of candidate genes, the chance of success is increased, while, at the same time, time and resources are spared.

Gene duplication, source of biological novelty

As the functional divergence of genes within families is frequently observed, the process of gene duplication and evolution of new function has, since long, been of considerable interest. The duplication of genes is becoming widely accepted as one of the driving forces behind the increasing phenotypic complexity during the evolution of eukaryotes, as it provides new, raw material on which evolution can work (see box; Aburomia et al., 2003; Meyer and Van de Peer, 2003).

Indeed, since duplicated genes are redundant, one of the copies is, at least in theory, freed from functional constraint, and can therefore evolve a new function. Ohno (1970) predicted that mutations in the second copy are selectively neutral and will either turn the gene into a non-functional pseudogene, or alternatively, turn the duplicate gene into a gene with a new function, due to a series of non-deleterious random mutations. Although intuitively appealing a theory, little evidence has been found for genes that have obtained novel functions this way. For example, the analysis of duplicated genes of the tetraploid frog *Xenopus laevis* has shown that both copies are under purifying selection, indicating that mutations in the second copy are far from neutral (Hughes, 1999).

The role of gene and genome duplication in evolution - a brief history.

In his now classic book 'Evolution by Gene Duplication', published in 1970, **Ohno** claimed that *"if evolution had been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged, while big leaps in evolution would have been impossible without the creation - through duplication - of many new gene loci with previously nonexistent functions"*. Although his theories are currently widely accepted as the foundations for current evolutionary research on gene duplications, Ohno was not the first to acknowledge the importance of duplications in evolution: in fact, since the 1930s, pioneers in evolutionary and genetical research already saw the possibilities of duplication as a source for new genetic material.

Haldane (1932,1933) was among the first to hypothesise on the advantages of duplication as an evolutionary mechanism, inspired by the phenomenon of hybrid vigour, commonly observed in alloploid plants. He wrote**:** *"Another possible mode of making rapid evolutionary jumps is by hybridisation. […] Hybridisation (where the hybrids are fertile) usually causes an epidemic of variation in the second generation which may include new and valuable types which could not have arisen within a species by slower evolution."*

Even at the single gene level, Haldane had a prophetic view on the possibilities created by gene redundancy: *"Mutation pressure must be a slow cause of evolution, but it certainly cannot be neglected when organisms are in a fairly constant environment over long periods. Among other things it will favour polyploids, and particularly allopolyploids, which possess several pairs of sets of genes, so that one gene may be altered without disadvantage, provided its functions can be performed by a gene in one of the other sets of chromosomes."*

He did, however, attribute a more important role to large-scale duplication events: *"Duplications affecting only a few genes would confer only a slight advantage. But duplication of a large section, polysomy of a whole chromosome, or polyploidy, might confer a considerable advantage, provided it caused neither unbalance nor sterility. Whether this advantage is sufficient to be of evolutionary importance is not clear, but the possibility exists."*

Interestingly, in his 1935 paper on the banding patterns of Drosophila salivary chromosomes, **Bridges** claims to have postulated similar hypotheses seventeen years earlier: *"In my first report on duplications at the 1918 meeting of the A.A.A.S., I emphasized the point that the main interest in duplications lay in their offering a method for evolutionary increase in lengths of chromosomes with identical genes which could subsequently mutate separately and diversify their effects."*

Already in 1938, **Serebrowsky** correctly hypothesised, when analysing the *scute* and *achaete* loci of *Drosophila*, that they have originated by gene duplication and subsequent division of multiple functions of the ancestral gene, a phenomenon which decades later will be termed subfunctionalisation: *"This principle of loss of duplicate functions by one of the homologues in the process of genic evolution is considered by us as an important (though not the single) explication of a great number of phenomena discovered by genetics. It should result in a specialisation of genes, when each then fulfils only one function which is strictly limited and important for the life of the organism."*

After further confirmation of the importance of duplication in evolution by **Gulick** (1944) and **Beadle** (1945), **Metz** (1947) for the first time comes up with a concept which remarkably reminds of gene families, long before these were recognised by sequence analysis in the early 1970s: *"Suppose we assume that the series of similar single bands found here and there in the salivary gland chromosomes do represent multiple repeats- then what about other bands in the chromosomes which resemble these particular ones? Are they homologs also?*

(continued on the next page)

They could readily have been separated from the series through inversions of larger segments of the chromosomes. From this it is only a step to the grouping of all the bands of a chromosome into a few classes on the basis of their morphological similarity and implying that al those in each class are homologous- with the result that we would only have a few kinds of genes and have many representatives of each kind, with minor grades of difference within the classes. Thus we could reach almost any height of speculation."

In an excellent paper in 1951, **Stephens** looks back at past hypotheses and reformulates them in a more contemporary context: *"As long as the gene was considered as an abstract unit of inheritance, the possibilities of mutation were limited only by the imagination of the theoretical geneticist. But if the gene owes its properties to a specific surface structure it follows that a mutation implies a loss or deformation of that structure and consequent loss or impairment of the original function – with or without the concomitant acquisition of a new function. From the evolutionary point of view this would mean that mutation per se could not provide an unlimited source of variation; at best it could only replace a finite number of functions by an equal number of new ones and at worst it could result in a net loss in the number of functions.*

From a priori reasoning it is difficult to regard such a mechanism (in which a new function could be attained only at the price of discarding an old one) as an efficient method of affecting evolutionary progress from the simple to the complex.

One might expect (still on a priori grounds) that a mechanism in which new functions could be added and the old ones retained would have considerable selective advantage. Within the bounds of the theory, the only likely manner of achieving this "improvement" would be by increasing the number of genetic loci, either by the synthesis of new loci by nongenic material or by the duplication and subsequent differentiation of existing loci. Theoretically this would make possible the retention of existing functions by genes at one locus leaving the other free to develop new functions. Further, since one locus could retain its original function, the other would initially be subject to a reduced selection pressure."

But let us come back to **Ohno**, as his 1970 book contains more than only a tribute to the evolutionary importance of duplication. He also lays the foundations for the later theory of subfunctionalisation at the regulatory level: *"Nonconcordant duplication involving only one of a group of functionally interrelated gene loci becomes permissible if the incorporation of two former alleles of that locus into the genome is quickly followed by the development of the differential genetic regulatory mechanism. As this genetic regulatory mechanism permits only one or the other former allele to engage in transcriptional activity in any given somatic cell type of an individual, the original one-to-one gene dosage relationship is effectively restored among all functionally related genes in spite of discordant duplication which had affected only one locus."* This, together with his views on polyploidy and the evolution of complexity in vertebrates, makes the wide acclaim of his work highly justified.

Therefore, several alternative models for gene evolution after duplication events have been proposed, such as subfunctionalisation, the partitioning of ancestral functions between duplicates, both at the protein or at the regulatory level (Serebrowski, 1938; Li, 1980; Piatigorsky and Wistow, 1991; Hughes, 1994; 1999; Force et al., 1999; Stoltzfus, 1999; Wagner, 2002) and the retention of redundancy in multidomain proteins (Gibson and Spring, 1998) or by gene conversion (Graur and Li, 2000). In addition, new function can arise in duplicated genes by directed evolution under processes such as positive Darwinian selection (Hughes, 1999).

To investigate the divergence of duplicate genes *in silico*, many new methods have been developed over the past few years. These will be discussed in further detail in chapter 7.

References

Abeliovich, D., Katz, M., Karplus, M., and Carmi, R. (1985). A de novo translocation, 14q21q, with a michrochromosome-14p21p. Am Jour Med Genet **22**, 29-33.

Aburomia, R., Khaner, O., and Sidow, A. (2003). Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. J Struct Funct Genom **3**, 45-52.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389-3402.

Andrews, J., Bouffard, G.G., Cheadle, C., Lu, J., Becker, K.G., and Oliver, B. (2000). Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res **10,** 2030-2043.

The *Arabidopsis* Genome Initative (AGI) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2,** 28-36.

Baird, W.V., and Meagher, R.B. (1987). A complex gene superfamily encodes actin in petunia. Embo J **6,** 3223-3231.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic Acids Res **30,** 276-280.

Beadle, G.W. (1945). Biochemical genetics. Chem Rev **37,** 15-96.

Blakeslee, A.F. (1934). New Jimson Weeds from old chromosomes. Jour Hered **25,** 81-108.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12,** 1093-1101.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., Llorente, B., Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Perrin, A., Potier, S., Souciet, J., Talla, E., Toffano-Nioche, C., Wesolowski-Louvel, M., Marck, C., and Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. FEBS Lett **487,** 31-36.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol **294,** 1351-1362.

Bridges, C.B. (1935). Salivary chromosome maps. J Hered **26,** 60-64.

Bridges, C.B. (1936). The Bar "gene" A duplication. Science **83,** 210-211.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene **238,** 115-134.

Brosius, J. (2003). Gene duplication and other evolutionary strategies: from the RNA world to the future. J Struct Funct Genom **3,** 1-17.

Brown, D.D., and Dawid, I.B. (1968). Specific gene amplification in oocytes. Oocyte nuclei contain extrachromosomal replicas of the genes for ribosomal RNA. Science **160,** 272-280.

Cronn, R.C., Zhao, X., Paterson, A.H., and Wendel, J.F. (1996). Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. J Mol Evol **42,** 685-705.

Dandekar, T., Huynen, M., Regula, J.T., Ueberle, B., Zimmermann, C.U., Andrade, M.A., Doerks, T., Sanchez-Pulido, L., Snel, B., Suyama, M., Yuan, Y.P., Herrmann, R., and Bork, P. (2000). Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucleic Acids Res **28,** 3278-3288.

Dayhoff, M.O. (1974). Computer analysis of protein sequences. Fed Proc **33,** 2314-2316.

De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G., and Van De Peer, Y. (2003). Genomewide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. J Mol Evol **56,** 573-586.

Dixon, D.P., Lapthorn, A., and Edwards, R. (2002). Plant glutathione transferases. Genome Biol **3,** reviews3004.1-3004.10.

Drouin, G., and Dover, G.A. (1987). A plant processed pseudogene. Nature **328,** 557-558.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Eisen, J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8,** 163-167.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N- terminal amino acid sequence. J Mol Biol **300,** 1005-1016.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A. (2002). The PROSITE database, its status in 2002. Nucleic Acids Res **30,** 235-238.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17,** 368-376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39,** 783-791.

Fink, G.R. (1987). Pseudogenes in yeast? Cell **49,** 5-6.

Fitch, W.M. (1971). Toward defining the course of evolution: minimal change for a specific tree topology. Syst Zool **20,** 406-416.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res **8,** 967-974.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151,** 1531-1545.

Friedman, R., and Hughes, A.L. (2003). The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol **20,** 154-161.

Fryxell, K.J. (1996). The coevolution of gene family trees. Trends Genet **12,** 364-369.

Gaut, B.S., and Doebley, J.F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A **94,** 6809-6814.

Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet **14,** 46-49; discussion 49-50.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa L.* ssp. *japonica*). Science **296,** 92-100.

Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A., Aytekin-Kurban, G., Bekiranov, S., Fajardo, J.E., Eswar, N., Sanchez, R., Sali, A., and Gaasterland, T. (2001). Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. Nat Genet **27,** 337-340.

Graur, D., and Li, W.-H. (1999). Fundamentals of Molecular Evolution. (Sunderland, Massachusetts: Sinauer).

Gu, X., Wang, Y., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet **31,** 205-209.

Gulick, A. (1944). The chemical formulation of gene structure and gene action. Adv Enzymol **4,** 1-39.

Gupta, R., and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput, 310-322.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol **3,** research0029.1-0029.12.

Haldane, J.B.S. (1932). The causes of evolution. (London: Longmans Green & Co).

Haldane, J.B.S. (1933). The part played by recurrent mutation in evolution. Am Nat **67,** 5-19.

Hall, T.A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser**,** 95-98.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294,** 2310-2314.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci **256,** 119-124.

Hughes, A.L. (1999). Adaptive evolution of genes and genomes. (New York: Oxford University Press).

Jin, Y.K., and Bennetzen, J.L. (1994). Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. Plant Cell **6,** 1177-1186.

Kellogg, E.A. (2003). What happens to genes in duplicated genomes. Proc Natl Acad Sci U S A **100,** 4369-4371.

Kosarev, P., Mayer, K.F., and Hardtke, C.S. (2002). Evaluation and classification of RING-finger domains encoded by the *Arabidopsis* genome. Genome Biol **3,** research0016.1-research0016.12.

Krogh, A. (1998). An introduction to hidden Markov models for biological sequences. In Computational Methods in Molecular Biology, S.L. Salzberg, D.B. Searls, and S. Kasif, eds (Amsterdam: Elsevier), pp. 45-63.

Leitch, I.J., and Bennett, M.D. (1997). Polyploidy in angiosperms. Trends Plant Sci **2,** 470-476.

Lewin, R. (1983). How mammalian RNA returns to its genome. Science **219,** 1052-1054.

Li, W.H. (1980). Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95,** 237-258.

Long, E.O., and Dawid, I.B. (1980). Repeated genes in eukaryotes. Annu Rev Biochem **49,** 727-764.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Marx, J.L. (1982). Is RNA copied into DNA by mammalian cells? Science **216,** 969-970.

Masterson, J. (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264,** 421-424.

Mathé, C., Sagot, M.F., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res **30,** 4103-4117.

Maurer-Stroh, S., Eisenhaber, B., and Eisenhaber, F. (2002). N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J Mol Biol **317,** 541-557.

McCarrey, J.R., and Thomas, K. (1987). Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature **326,** 501-505.

Meagher, R.B., McKinney, E.C., and Kandasamy, M.K. (2000). The significance of diversity in the plant actin gene family. In Actin: a dynamic framework for multiple plant cell functions, C.J. Staiger, F. Baluska, D. Volkmann, and P.W. Barlow, eds (Dordrecht: Kluwer Academic Publishers), pp. 3-27.

Metz, C.W. (1947). Duplication of chromosome parts as a factor in evolution. Am Nat **81,** 81-103.

Meyer, A., and Van de Peer, Y. (2003). Natural selection merely modified while redundancy created' - Susumu Ohno's idea of the evolutionary importance of gene and genome duplications. J Struct Funct Genom **3,** vii-ix.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell **15,** 809-834.

Montag, K., Salamini, F., and Thompson, R.D. (1996). The ZEM2 family of maize MADS box genes possess features of transposable elements. Maydica **41,** 241-254.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R., and Zdobnov, E.M. (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res **31,** 315-318.

Ohno, S. (1970). Evolution by Gene Duplication. (Berlin;Heidelberg;New York: Springer-Verlag).

Ohta, T. (1990). How gene families evolve. Theor Popul Biol **37,** 213-219.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. (1999). Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics **15,** 887-899.

Pertea, M., and Salzberg, S.L. (2002). Computational gene finding in plants. Plant Mol Biol **48,** 39-48.

Piatigorsky, J., and Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. Science **252,** 1078-1079.

Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., Lee, H., Hitti, J., Doucette-Stamm, L., Hartley, J.L., Temple, G.F., Brasch, M.A., Vandenhaute, J., Lamesch, P.E., Hill, D.E., and Vidal, M. (2001). Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nat Genet **27,** 332-336.

Roelofs, J., and Van Haastert, P.J. (2002). Deducing the origin of soluble adenylyl cyclase, a gene lost in multiple lineages. Mol Biol Evol **19,** 2239-2246.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16,** 944-945.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4,** 406-425.

Schultz, C.J., Rumsewicz, M.P., Johnson, K.L., Jones, B.J., Gaspar, Y.M., and Bacic, A. (2002). Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case. Plant Physiol **129,** 1448-1463.

Serebrowsky, A.S. (1938). Genes *scute* and *achaete* in *Drosophila melanogaster* and a hypothesis of gene divergency. Compt Rend Acad Sci URSS **14,** 77-81.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van De Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **99,** 13627-13632.

Snustad, P., Simmons, M.J., and Jenkins, J.B. (1997). Pinciples of genetics. (New York: John Wiley & Sons, Inc.).

Soares, M.B., Schon, E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J., and Efstratiadis, A. (1985). RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol **5,** 2090-2103.

Stephens, S.G. (1951). Possible significance of duplication in evolution. Adv Genet **4,** 247-265.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J Mol Evol **49,** 169-181.

Sturtevant, A.H. (1925). The effects of unequal crossing over at the Bar locus in *Drosophila*. Genetics **10,** 117-147.

Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van Den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Van Montagu, M., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445,** 237-245.

Thareau, V., Dehais, P., Serizet, C., Hilson, P., Rouzé, P., and Aubourg, S. Automatic design of gene-specific sequence tags for genomewide functional studies. Bioinformatics **(in press)**.

Verica, J.A., and He, Z.H. (2002). The cell wall-associated kinase (WAK) and WAK-like kinase gene family. Plant Physiol **129,** 455-459.

Wagner, A. (2002). Assymetric functional divergence of duplicate genes in yeast. Mol Biol Evol **19,** 1760-1768.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol **42,** 225-249.

Wullschleger, S.D., Jansson, S., and Taylor, G. (2002). Genomics and forest biology: *Populus* emerges as the perennial favorite. Plant Cell **14,** 2651-2655.

Zhang, J., Zhang, Y.P., and Rosenberg, H.F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet **30,** 411-415.

Zuckerkandl, E. (1975). The appearance of new structures and functions in proteins during evolution. J Mol Evol **7,** 1-57.

[Chapter 2]

Family-wise expert annotation of *Arabidopsis* genes: the GeneFarm project

Jeroen Raes , Sébastien Aubourg[1], Patrice Déhais[2] and Pierre Rouzé

Département associé de l'INRA, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[1] current address: Unité de Recherche en Génomique Végétale, INRA, 2, Rue Gaston Crémieux - CP 5708, 91000 Evry Cedex, France
[2] current address: INRA-AGENA, Chemin de Borde-Rouge - Auzeville - BP 27, 31326 Castanet-Tolosan cedex, France

Genes annotated in this project were deposited in the GeneFarm database (Aubourg et al., unpublished).

The Fam-o-tator annotation procedure is protected by a registration at the "Agence pour la Protection des Programmes" and was assigned the Inter Deposit Digital Number IDDN.FR.001.100023.000.R.P.2002.000.31235.

Introduction

The completion of the first plant genome sequence of *Arabidopsis thaliana* was (and still is) of great importance to plant research (AGI, 2000). It allowed a first glimpse at the complete pool of genes that constitutes a plant and provided a solid foundation for functional research in all aspects of (plant) biology. Almost immediately after the sequencing process, a first automatic annotation of the genomic fragments was performed by the five large sequencing consortia, providing a valuable resource for the plant community. Unfortunately, this annotation was done in a non-homogeneous way: the different consortia used different strategies and tools of different quality (Aubourg and Rouzé, 2001). Oddly, Genemark.hmm, the best performing program at that time (Pavy et al., 1999), was not used at all. Later, only two out of five consortia adapted their strategy and included this tool in their analysis (Theologis et al., 2000, Tabata et al., 2000). In addition, further analysis showed that the automatic annotation was far from flawless: for example, the manual annotation of a 400-kb contig showed discrepancies with the automated annotation for about 80% of genes (Terryn et al., 1999). Later, the analysis of 5000 full-length transcripts showed that 35% of the corresponding predicted gene structures had to be corrected, while for 5% of transcripts the genes had been missed by the prediction software (Haas et al., 2002). Large-scale EST sequencing projects also discovered about 5% new, previously not predicted genes in a set of 14831 cDNA clones (Seki et al., 2002).

The functional annotation of the *Arabidopsis* genome was also based on an automated routine, using similarity to other protein sequences found in the public databases (without verifying the correct annotation of the latter), combined with domain analysis using resources such as Prosite (Falquet et al., 2002). This approach provided rapidly a first functional annotation of the genome and was therefore of considerable interest to the scientific world. It did, however, have its limitations. For example, neither available expression data nor functional data found in the literature was attributed to the genes. In addition, the finer functional differences between members of a family could not be assigned using this automated approach.

It was in this respect that the Génoplante GeneFarm project was started in 1999 by Pierre Rouzé and Sébastien Aubourg. The goal of this project consisted of providing a homogeneous, high-quality, traceable, family-wise in-depth annotation of *Arabidopsis* genes and proteins. To achieve this goal, a network of manual annotators was set up, each having expertise in one or several processes or gene families. This network consists of about 15 laboratories, each with their different scope and expertise, both in molecular biology as in bioinformatics. Recently, the Swiss Institute for Bioinformatics (SIB) joined the project to incorporate the annotation into the renowned SWISS-PROT database (Boeckmann et al., 2003), allowing a maximal public diffusion of the projects' results toward the scientific community. The annotation is done in a family-wise way to facilitate the work and to fully benefit from the annotators' knowledge of his domain. The homogeneity and quality of the annotation is assured by the use of a standardised annotation protocol (see further), together with a web-based annotation interface containing several ontology restrictions and automated control mechanisms (e.g. on consistency of entered data).

This interface also ensures the tractability of annotation. For each gene feature entered, the source of this information has to be given, allowing the end-user to assess the value of each feature and clearly distinguish predicted from proven data. Finally, the aim of this project is to provide, for each gene or family, as much additional information as possible. This information consists of knowledge on expression, posttranscriptional and –translational modifications, biochemical and biological function, subcellular localisation etc., coming from literature. For each of these features, the source (literature reference) and nature of the experiment (e.g. Northern blot, RT-PCR, etc.) has to be given. All this information is stored in a dedicated database, named GeneFarm, designed for this project (Aubourg et al., unpublished).
As a consequence of the strict and demanding manual annotation procedure, this database constitutes a reliable resource for exhaustive, in-depth annotation of *Arabidopsis* gene families. The current status of the database consists of 1700 genes from 70 families, which should increase to 5000 genes in the next three years.

Results

Development of an annotation protocol for manual family-wise annotation

To ensure a homogeneous annotation throughout the GeneFarm project, the use of different tools (based on different principles and of variable quality) by each annotator had to be avoided. In addition, many of the partners in the project had limited knowledge of annotation and the bioinformatics tools available for this task. For this reason, a standardised minimum annotation protocol was developed covering all aspects of *in silico* structural and functional annotation (see figure 1). The protocol was designed to ensure the exhaustive, in-depth annotation of each family using the best performing tools at that time, which were either publicly available or provided to the annotators via the dedicated Génoplante-info server.

Development of Fam-o-tator, a semi-automated gene family structural annotation tool

Although the manual annotation procedure, as described above, is aimed at producing high-quality annotation, it remains, especially for large gene families, a slow and tedious task. In this respect, we tried to develop a semi-automated method to detect all members of a gene family in *Arabidopsis thaliana,* combined with a high quality gene model prediction, given a representative set of genes (e.g. of experimental origin) and a set of genomic (e.g. BAC) sequences. This method relieves the user of tedious, repetitive and human error-sensitive tasks such as data management and input/ output file reformatting (as the majority of the procedures are fully automated using perl and csh scripts), while, at the same time, it allows the user to keep full control in the steps that are error-prone when fully automated. The final result of this routine consists of the gene structure, position and mRNA/protein sequence of all the family members.

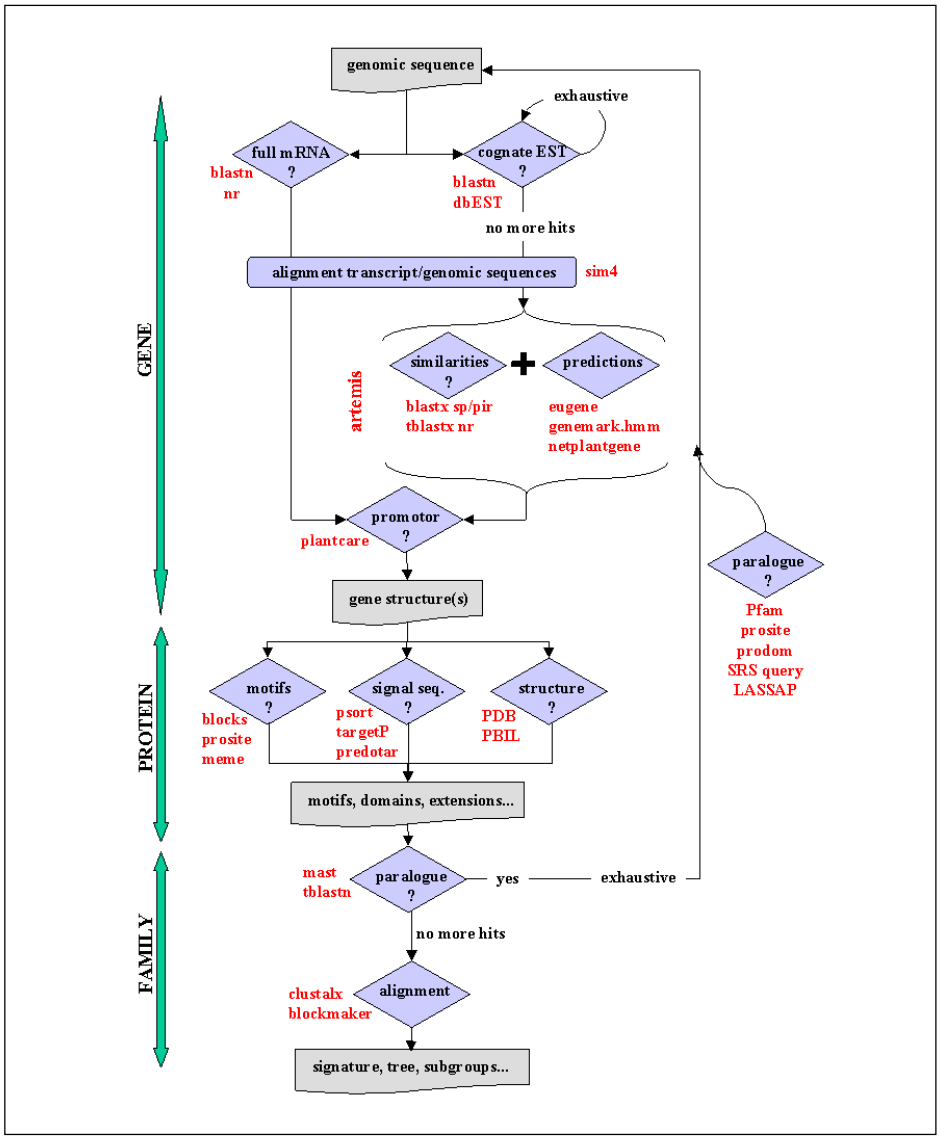

Figure 1. Annotation protocol designed for the GeneFarm project, as found on the GeneFarm website.

In addition, the automated visualisation of gene structure, protein alignment and domain representation facilitate the validation of the results. This way, the rigorousness of manual annotation is combined with the speed and ease of automatisation. An overview of the procedure used by the program is given in figure 2.

The method consists of three main procedures: first, rapid detection and genomic dataset reduction, followed by high-quality, family-specific gene model prediction in candidate regions and finally verification of the results.

In the first stage, the goal is to locate the regions of interest. This way, the genomic dataset on which prediction will be done is restricted to candidate regions in order to reduce computation time and the number of false positives. These regions are detected by BLASTing a set of representative, certified members of the family against the genomic sequences.

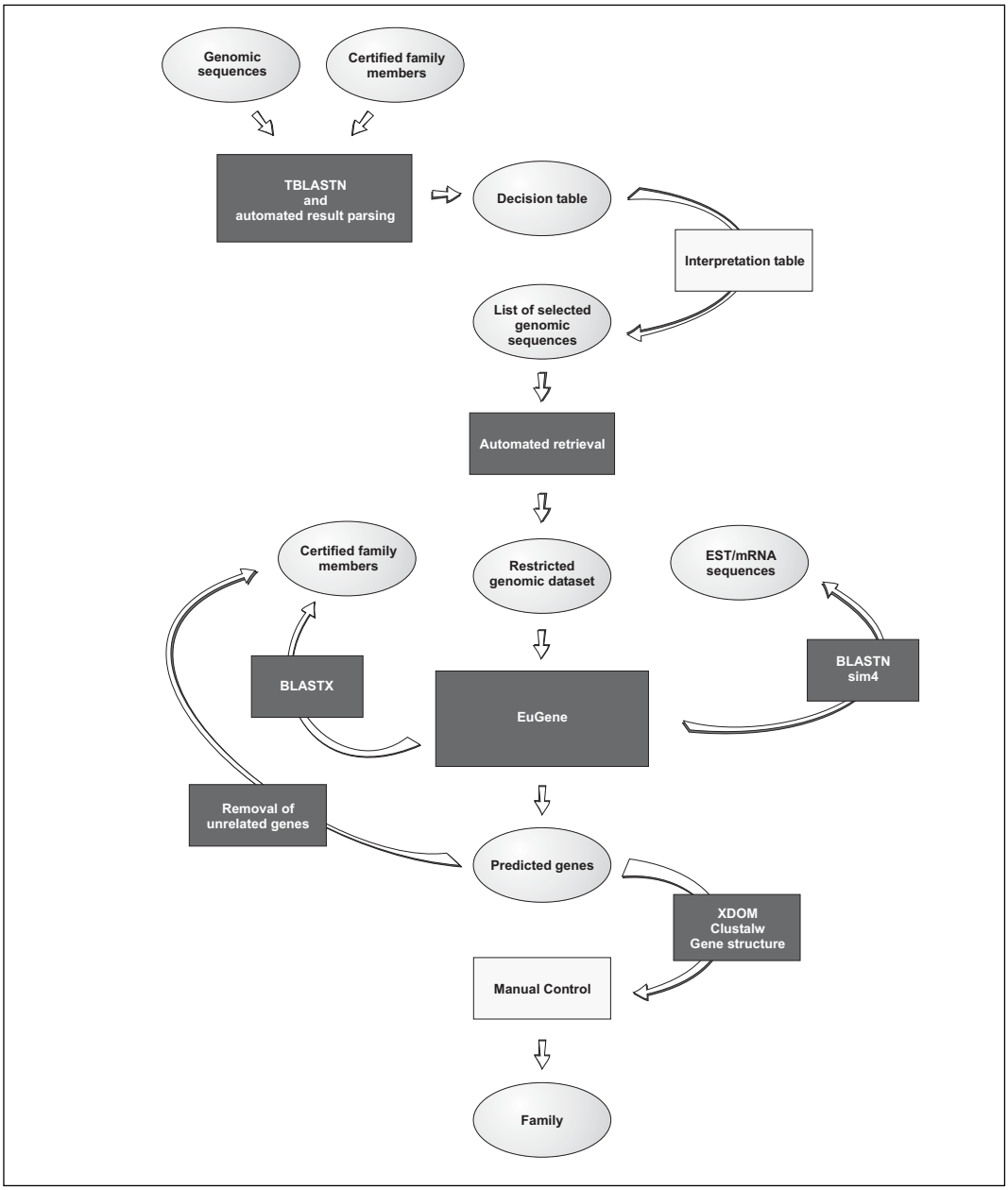

Figure 2. Overview of the Fam-o-tator procedure. Dark boxes represent automated procedures, while white boxes indicate manual user interventions.

To avoid the pitfalls of E-value based decisions and to combine the BLAST results from the different representatives, a table is produced, summarising the different BLAST outputs, to permit a user-assisted delineation of the family (Figure 3). Each line of this table consists of a genomic sequence name and a list of representatives that have “detected” this sequence in the BLAST result. The advantage of this method is that at a certain point, a clear drop-off in number of detecting representatives is distinguishable. This point appears to be a good border between false positives and true family members.

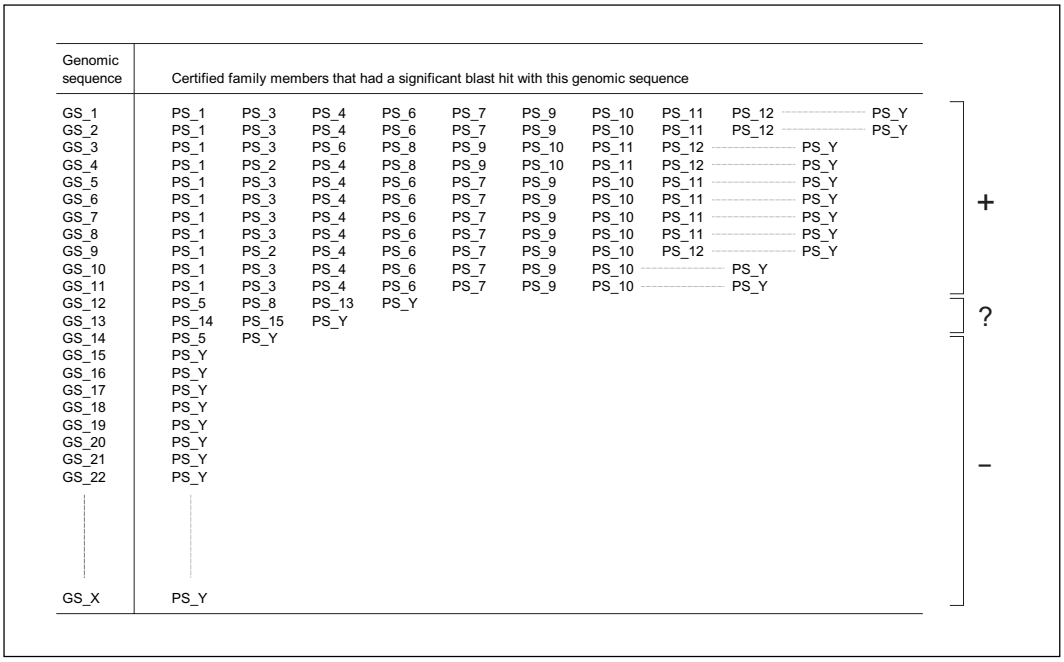

Figure 3. Decision table provided in the Fam-o-tator procedure. The first column contains the list of genomic sequences (GS) which were picked up by at least one protein sequence (PS) of a representative family member, sorted by the number of family members that pointed at a particular genomic sequence (which are shown in the next columns). This representation allows the user to decide which genomic sequences should be included in the restricted dataset for the next step of the procedure. Sequences marked by "+" are very likely candidates while those marked by "-" are false hits. Sequences in the zone marked by "?" are unsure and need to be validated manually.

The organisation of the BLAST processes and parsing of results in the table is automated using several csh and perl scripts. It must be noted that the table does not impose the cut-off on the user. Its sole purpose is to provide a guiding tool on where to draw the border, a decision which the user – being familiar with the investigated family – takes manually. This restricted genomic set, consisting of sequences which contain at least one member of the family, forms the basis of the second stage of the method: the prediction.

All the chosen genomic sequences are analysed using EuGene (Schiex et al., 2001), a gene modelling program which is currently the best available for *Arabidopsis* (Thomas Schiex, Cathérine Mathé, unpublished data). EuGene combines several sources of information using a graph-based method. These different sources include splice site prediction by NetPlantGene (Hebsgaard et al., 1996) and SplicePredictor (Brendel and Kleffe, 1998), translation start prediction with NetStart (Pedersen and Nielsen, 1997) and finally alignment of ESTs, full-length cDNA sequences and proteins. This final feature is exploited by our routine to improve the prediction in a family-specific way. By feeding the algorithm with the protein sequences of only the set of certified members of the family, we make sure that the prediction is not compromised by false annotation in the databases. By taking into account only the protein sequence of certified family members as well as EST and mRNA data for the gene structure prediction itself, we achieve a family-specific prediction of very high quality.

After automated parsing of results and extraction of sequences, this step results in a set of predicted proteins per genomic sequence, both containing "true" members of the family as well as false positives. Using a simple BLASTP filter, the obvious false positives are removed automatically. The result, both family members as well as "border cases", is subjected to the last step: the verification of the results. In this step, the user can manually assess the quality of the prediction using different representations generated by the program. An XDOM (Gouzy et al., 1997) visualisation allows the verification of the presence of conserved domains, while a CLUSTALW (Thompson et al., 1994) protein alignment gives a more detailed view of the family members. Finally, a graphical overview of all the predicted gene structure is provided in HTML format and can be visualised using any common web-browser.

Semi-automated annotation of the MYB transcription factor family in Arabidopsis thaliana

The family of MYB proteins is a group of functionally diverse transcription factors, which are found in the animal as well as in the plant kingdom. In animal genomes, the number of family members is limited: for example, in human, 10 distinct members have been estimated, of which 3 have been thoroughly analysed experimentally (A-, B- and c-MYB) (Rosinski and Atchley, 1998). In contrast, it is believed to be one of the largest transcription factor families in plants, containing more than 80 members in maize, more than 100 members in *Arabidopsis* and at least 40 in *Petunia* (Avila et al., 1993; Romero et al., 1998; Rabinowicz et al., 1999).
MYB proteins are characterized by a 50 amino acid motif, coding for a helix-turn-helix structure, which is proven to bind DNA in a sequence-specific manner (Martin and Paz-Ares, 1997). Three distinct types of MYB proteins have been described so far, which are classified by the number of occurrences of this motif in the protein.
In 1982, an oncogene v-MYB was detected in the avian myeloblastosis virus. This led to the discovery of its human progenitor, c-MYB, which has become the so-called prototype of the family (Klempnauer et al., 1982). This protein contains three (imperfect) repeats of the MYB domain. The class of genes with this structure has consequently been named 3R or R1R2R3, describing the three repeats. It has been detected in vertebrates and invertebrates, as well as in fungi and all major plant lineages (Lipsick, 1996; Rosinski and Atchley, 1998; Kranz et al., 2000). In animals, the 3R subfamily has been found to be implicated in cell proliferation control, apoptosis prevention and commitment to development (Romero et al., 1998 and references therein). Interestingly, the plant 3R genes have been linked to cell cycle control via the regulation of cyclin B genes (Ito, 2000).
A few examples of MYBs containing only one repeat have been reported, which show similarity to either the R1/R2 or the R3 repeat. They seem to be present in all fungi, plants and animals, although further research into this subfamily will be necessary to clarify this further (Bilaud et al., 1996; Kirik and Baumlein, 1996; Lipsick, 1996; Feldbrügge et al., 1997). On the functional level, these genes are involved in very diverse processes, such as circadian clock regulation, light dependent activation, epidermal cell differentiation and telomeric binding (Jin and Martin, 1999).

The predominant group of MYBs in plants, however, has only the two last repeats of the 3R MYBs. As such, it is commonly called the R2R3 subfamily. The huge diversification of this family in plants strongly contrasts with the complete absence of this group in animal lineages, which might indicate that the R2R3 type evolved to serve plant-specific functional needs (Romero et al., 1998). Phylogenetic analysis shows that this type originated by loss of the R1 repeat from an ancestral 3R protein (Rosinski and Atchley, 1998). Furthermore, this class can be divided in three further subfamilies: two smaller ones called A, which contains genes closely related to the c-MYB gene in humans, and B, and one larger (C), containing the majority of R2R3-MYBs (Romero et al., 1998). Another analysis classified this family into 22 subgroups, based on phylogenetic relationships as well as small conserved motifs C-terminal of the MYB repeats (Kranz et al., 1998).

The R2R3-MYBs are implemented in various processes. First of all, they are involved in secondary metabolism, in particular in the regulation of the phenylpropanoid and flavonoid pathway. Furthermore, they are linked to cellular morphogenesis: they are responsible for the conical shape of petal epidermal cells and trichome (hair cell) differentiation in some parts of the leaf and stem. Finally, they have been found to be implicated in hormone response during seed development and germination, in particular in gibberellic (GA) and abscisic (ABA) acid-based induction (Martin and Paz-Ares, 1997, and references therein).

In the framework of the GeneFarm project, we detected and exhaustively annotated all members of the MYB family of transcription factors in the *Arabidopsis* genome. In a first stage (before the assembly of the complete genome), the Fam-o-tator procedure was applied to all publicly available BAC sequences from chromosomes 1, 3 and 5 and nonredundant fragments of the - at that time - already assembled chromosomes 2 and 4. A set of representative, experimentally verified MYB genes from all three known classes was compiled based on literature and database searches (Table 1).

Table 1. Experimentally verified genes used as representative set. References are given to either the paper in which the gene was characterised or the EMBL accession number of the full-length cDNA-sequence.

Name	Type	Reference/full length transcript
CDC5	1R	Hirayama and Shinozaki, 1996
mybL2	1R	Kirik and Baumlein, 1996
CCA1	1R	Wang et al., 1997
CPC	1R	Wada et al., 1997
LHY	1R	Schaffer et al., 1998
MYB3	R2R3	AF062859
MYB4	R2R3	Jin et al., 2000
MYB6	R2R3	Li and Parish, 1995
MYB7	R2R3	Li and Parish, 1995
MYB12	R2R3	AF062864
MYB15	R2R3	Y14207
MYB23	R2R3	Z68158
MYB30	R2R3	Daniel et al., 1999
MYB31	R2R3	Quaedvlieg et al., 1996
MYB36	R2R3	AF062878
MYB44	R2R3	Kirik et al., 1998
MYB51	R2R3	Z95774
MYB59	R2R3	AF062894
MYB60	R2R3	AF062895
MYB68	R2R3	AF062901
MYB71	R2R3	U62743
MYB75	R2R3	Borevitz et al., 2000
MYB77	R2R3	Z54137
MYB86	R2R3	AB005889
MYB94	R2R3	AF062918
MYB102	R2R3	Quaedvlieg et al., 1996
MYB3R-2	3R	Braun and Grotewold, 1999
MYB3R-3	3R	AY034964
MYB3R-1	3R	Braun and Grotewold, 1999

This analysis resulted in the detection of 113 R2R3-type MYB genes, four 3R MYBs and five 1R MYBs. Upon closer inspection and comparison with a study describing a manual annotation of MYB transcription factor genes in Arabidopsis (Stracke et al., 2001), 13 more R2R3, 1 extra 3R and a 4R gene were detected.

The main reason for the inability of the Fam-o-tator routine to predict these additional genes was found to be their presence at BAC extremities, which reduced the efficiency of the gene prediction program. These genes were further annotated manually (using the by then available assembled chromosomes) and all annotations were transferred to the GeneFarm database. A complete overview of the family is given in table 2. The evolutionary relationships between the members of the (largest) R2R3 subfamily are depicted in figure 4.

Table 2. Overview of the MYB family in *Arabidopsis thaliana*

Gene	AGI	Strand	Gene	AGI	Strand	Gene	AGI	Strand
1R-MYB								
CDC5	At1g09770	+	MYB41	At4g28110	-	MYB89	At5g39700	-
AtmybL2	At1g71030	-	MYB42	At4g12350	+	MYB90	At1g66390	+
CPC	At2g46410	-	MYB43	At5g16600	+	MYB91	At2g37630	-
CCA1	At2g46830	+	MYB44	At5g67300	+	MYB92	At5g10280	+
LHY	At1g01060	-	MYB45	At3g48920	-	MYB93	At1g34670	+
			MYB46	At5g12870	-	MYB94	At3g47600	-
R2R3-MYB			MYB47	At1g18710	-	MYB95	At1g74430	+
MYB0	At3g27920	-	MYB48	At3g46130	+	MYB96	At5g62470	-
MYB1	At3g09230	+	MYB49	At5g54230	-	MYB97	At4g26930	+
MYB2	At2g47190	+	MYB50	At1g57560	+	MYB98	At4g18770	+
MYB3	At1g22640	-	MYB51	At1g18570	-	MYB99	At5g62320	-
MYB4	At4g38620	+	MYB52	At1g17950	+	MYB100	At2g25230	-
MYB5	At3g13540	+	MYB53	At5g65230	+	MYB101	At2g32460	-
MYB6	At4g09460	+	MYB54	At1g73410	-	MYB102	At4g21440	-
MYB7	At2g16720	-	MYB55	At4g01680	-	MYB103	At1g63910	+
MYB8	At1g35515	-	MYB56	At5g17800	+	MYB104	At2g26950	-
MYB9	At5g16770	+	MYB57	At3g01530	-	MYB105	At1g69560	-
MYB10	At3g12820	-	MYB58	At1g16490	-	MYB106	At3g01140	+
MYB11	At3g62610	+	MYB59	At5g59780	-	MYB107	At3g02940	-
MYB12	At2g47460	+	MYB60	At1g08810	-	MYB108	At3g06490	-
MYB13	At1g06180	+	MYB61	At1g09540	+	MYB109	At3g55730	-
MYB14	At2g31180	-	MYB62	At1g68320	-	MYB110	At3g29020	-
MYB15	At3g23250	+	MYB63	At1g79180	-	MYB111	At5g49330	-
MYB16	At5g15310	+	MYB64	At5g11050	+	MYB112	At1g48000	+
MYB17	At3g61250	+	MYB65	At3g11440	+	MYB113	At1g66370	+
MYB18	At4g25560	+	MYB66	At5g14750	-	MYB114	At1g66380	+
MYB19	At5g52260	+	MYB67	At3g12720	-	MYB115	At5g40360	+
MYB20	At1g66230	+	MYB68	At5g65790	+	MYB116	At1g25340	+
MYB21	At3g27810	+	MYB69	At4g33450	-	MYB117	At1g26780	+
MYB22	At5g40430	-	MYB70	At2g23290	-	MYB118	At3g27780	-
MYB23	At5g40330	+	MYB71	At3g24310	-	MYB119	At5g58850	+
MYB24	At5g40350	-	MYB72	At1g56160	-	MYB120	At5g55020	-
MYB25	At2g39880	-	MYB73	At4g37260	+	MYB121	At3g30210	+
MYB26	At3g13890	-	MYB74	At4g05100	+	MYB122	At1g74080	-
MYB27	At3g53200	-	MYB75	At1g56650	+	MYB123	At5g35550	+
MYB28	At5g61420	-	MYB76	At5g07700	+	MYB124	At1g14350	+
MYB29	At5g07690	+	MYB77	At3g50060	-	MYB125	At3g60460	-
MYB30	At3g28910	+	MYB78	At5g49620	-			
MYB31	At1g74650	+	MYB79	At4g13480	+	*3R-MYB*		
MYB32	At4g34990	-	MYB80	At5g56110	+	MYB3R-1	AT4g32730	+
MYB33	At5g06100	+	MYB81	At2g26960	-	MYB3R-2	At4g00540	-
MYB34	At5g60890	+	MYB82	At5g52600	-	MYB3R-3	AT3g09370	+
MYB35	At3g28470	-	MYB83	At3g08500	-	MYB3R-4	AT5g11510	+
MYB36	At5g57620	+	MYB84	At3g49690	+	MYB3R-5	At5g02320	+
MYB37	At5g23000	+	MYB85	At4g22680	-			
MYB38	At2g36890	+	MYB86	At5g26660	+	*4R-MYB*		
MYB39	At4g17780	-	MYB87	At4g37780	-	MYB4R1	At3g18100	+
MYB40	At5g14340	+	MYB88	At2g02820	-			

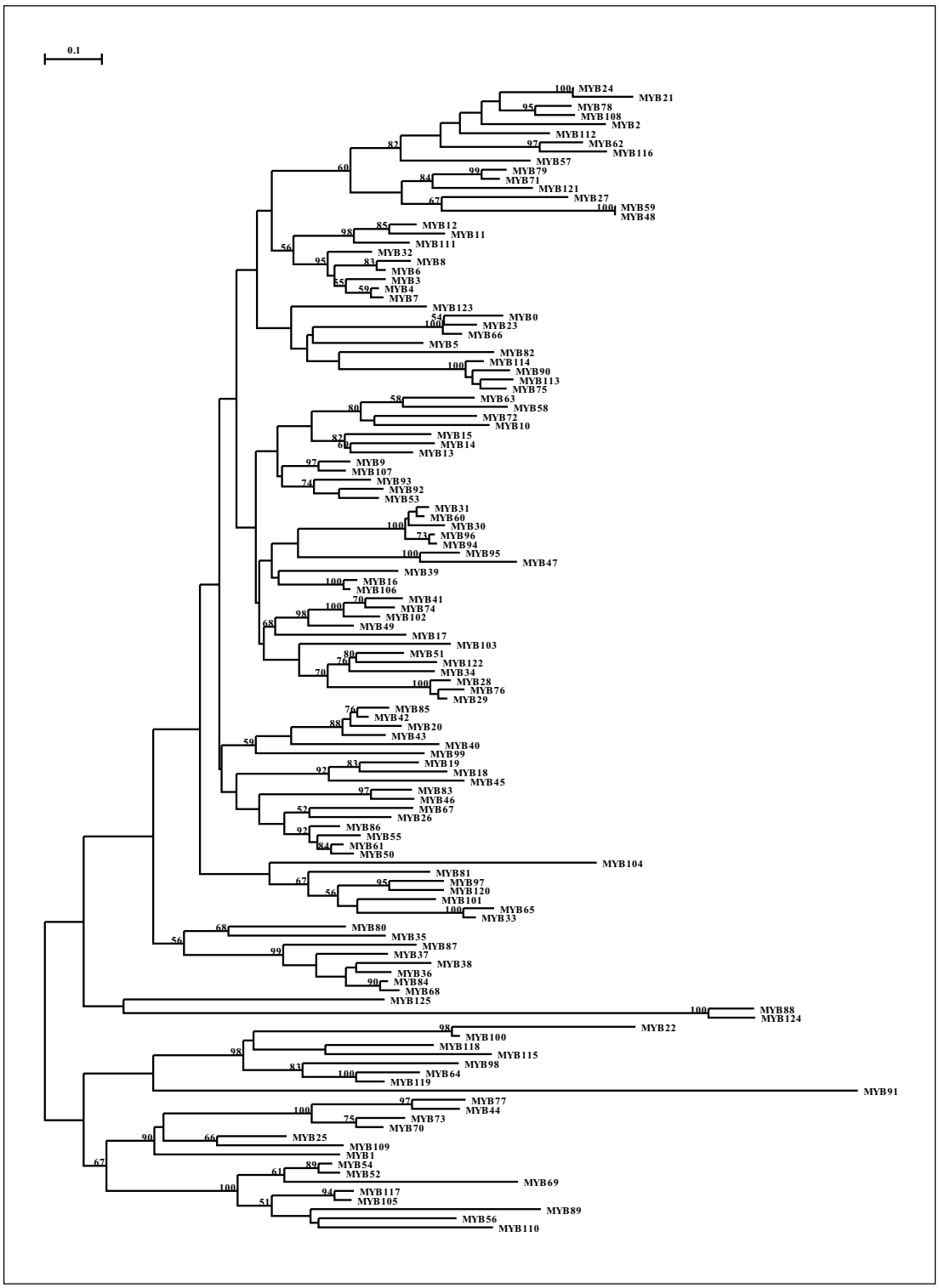

Figure 2. Neighbor-joining tree of the R2R3-MYB family, based on Poisson-corrected evolutionary distances. Significance of nodes was tested using 500 bootstrap replicates. Only bootstrap values >50 are shown.

Conclusions

In the framework of the GeneFarm *Arabidopsis* re-annotation project, we have tried to conceive a rigorous methodology for manual annotation, as well as design a semi-automated approach to speed up this process without risk. Our experience is that fully automated pipelines, as used by the *Arabidopsis* Genome Initiative, can't reach the quality level provided by manual annotation. The use of semi-automated methods, however, allows to speed up the manual approach, without having to sacrifice the added value of human control, as was shown in the annotation of the MYB family of transcription factors.

References

The Arabidopsis Genome Initiative (AGI) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

Aubourg, S., and Rouzé, P. (2001). Genome annotation. Plant Physiol Biochem **39,** 181-193.

Avila, J., Nieto, C., Canas, L., Benito, M.J., and Paz-Ares, J. (1993). *Petunia hybrida* genes related to the maize regulatory C1 gene and to animal myb proto-oncogenes. Plant J **3,** 553-562.

Bilaud, T., Koering, C.E., Binet-Brasselet, E., Ancelin, K., Pollice, A., Gasser, S.M., and Gilson, E. (1996). The telobox, a Myb-related telomeric DNA binding motif found in proteins from yeast, plants and human. Nucleic Acids Res **24,** 1294-1303.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31,** 365-370.

Brendel, V., and Kleffe, J. (1998). Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucleic Acids Res **26,** 4748-4757.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A. (2002). The PROSITE database, its status in 2002. Nucleic Acids Res **30,** 235-238.

Feldbrugge, M., Sprenger, M., Hahlbrock, K., and Weisshaar, B. (1997). PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit. Plant J **11,** 1079-1093.

Gouzy, J., Eugene, P., Greene, E.A., Kahn, D., and Corpet, F. (1997). XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. Comput Appl Biosci **13,** 601-608.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol **3,** research0029.1-0029.12.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. (1996). Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res **24,** 3439-3452.

Ito, M. (2000). Factors controlling cyclin B expression. Plant Mol Biol **43,** 677-690.

Jin, H., and Martin, C. (1999). Multifunctionality and diversity within the plant MYB-gene family. Plant Mol Biol **41,** 577-585.

Kirik, V., and Baumlein, H. (1996). A novel leaf-specific myb-related protein with a single binding repeat. Gene **183,** 109-113.

Klempnauer, K.H., Gonda, T.J., and Bishop, J.M. (1982). Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. Cell **31,** 453-463.

Kranz, H., Scholz, K., and Weisshaar, B. (2000). c-MYB oncogene-like genes encoding three MYB repeats occur in all major plant lineages. Plant J **21,** 231-235.

Kranz, H.D., Denekamp, M., Greco, R., Jin, H., Leyva, A., Meissner, R.C., Petroni, K., Urzainqui, A., Bevan, M., Martin, C., Smeekens, S., Tonelli, C., Paz-Ares, J., and Weisshaar, B. (1998). Towards functional characterisation of the members of the R2R3-MYB gene family from *Arabidopsis thaliana*. Plant J **16,** 263-276.

Martin, C., and Paz-Ares, J. (1997). MYB transcription factors in plants. Trends Genet **13,** 67-73.

Mount, D.W. (2001). Bioinformatics: sequence and genome analysis. (New York: Cold Spring Harbor Laboratory Press).

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. (1999). Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics **15,** 887-899.

Pedersen, A.G., and Nielsen, H. (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. Proc Int Conf Intell Syst Mol Biol **5,** 226-233.

Rabinowicz, P.D., Braun, E.L., Wolfe, A.D., Bowen, B., and Grotewold, E. (1999). Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. Genetics **153,** 427-444.

Romero, I., Fuertes, A., Benito, M.J., Malpica, J.M., Leyva, A., and Paz-Ares, J. (1998). More than 80 R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. Plant J **14,** 273-284.

Rosinski, J.A., and Atchley, W.R. (1998). Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. J Mol Evol **46,** 74-83.

Schiex, T., Moisan, A., and Rouzé, P. (2001). EuGène: An eukaryotic gene finder that combines several sources of evidence. In Computational Biology: Selected Papers (Lecture Notes in Computer Science, Vol. 2066), O. Gascuel and M.-F. Sagot, eds (Berlin: Springer-Verlag), pp. 111-125.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A., and Shinozaki, K. (2002). Functional annotation of a full-length Arabidopsis cDNA collection. Science **296,** 141-145.

Stracke, R., Werber, M., and Weisshaar, B. (2001). The R2R3-MYB gene family in *Arabidopsis thaliana*. Curr Opin Plant Biol **4,** 447-456.

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

Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van Den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Van Montagu, M., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445,** 237-245.

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

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22,** 4673-4680.

[Chapter 3]

Genome-wide characterization of the lignification toolbox in *Arabidopsis*

Jeroen Raes[†], Antje Rohde[†], Jørgen Holst Christensen, Yves Van de Peer, and Wout Boerjan[*]

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[†] *The two authors contributed equally to this work.*
[*] *Author for correspondence (e-mail wout.boerjan@psb.ugent.be; fax: +32 9 331 3809)*

Published in: Plant Physiology, in press.

Abstract

Lignin, one of the most abundant terrestrial biopolymers, is indispensable for plant structure and defense. With the availability of the full genome sequence, large collections of insertion mutants and functional genomics tools, *Arabidopsis thaliana* constitutes a perfect model system to profoundly unravel the monolignol biosynthesis pathway. In a genome-wide bioinformatics survey of the *Arabidopsis* genome, 34 candidate genes were annotated that encode genes homologous to the ten presently known enzymes of the monolignol biosynthesis pathway, 11 of which have not been described before. By combining evolutionary analysis of these gene families with in silico promoter analysis and expression data (from a reverse-transcription polymerase chain reaction analysis on an extensive tissue panel, mining of expressed sequence tags from publicly available resources, and assembling expression data from literature), 12 genes could be pinpointed as the most likely candidates for a role in vascular lignification. Furthermore, a possible link was detected between the presence of the AC-regulatory promoter element and the biosynthesis of G lignin during vascular development. Together, these data describe the full complement of monolignol biosynthesis genes in *Arabidopsis* and serve as a basis for further functional studies.

Introduction

Lignin is an aromatic heteropolymer that is mainly present in secondary thickened plant cells, where it provides rigidity and impermeability to the cell walls. In addition, lignin deposition may be induced upon wounding and infection to protect plant tissues against invading pathogens. Lignin is a highly heterogeneous, three-dimensional polymer that is composed of different phenylpropanoids, predominantly the monolignols *p*-coumaryl, coniferyl and sinapyl alcohols that differ in their degree of methoxylation (Figure 1). When these monolignols are incorporated into lignin, they are called *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) units, respectively. In addition to the three monolignols, other phenylpropanoids, such as hydroxycinnamyl aldehydes, hydroxycinnamyl acetates, hydroxycinnamyl *p*-hydroxybenzoates, hydroxycinammyl *p*-coumarates, and hydroxycinnamate esters, are also present in the polymer (Boerjan et al., 2003). Considerable variation exists in lignin composition between taxa, cell types, and developmental and environmental conditions; for example, lignin of gymnosperms is mainly built by H and G units, whereas angiosperm lignin additionally incorporates S units in large amounts. Perturbation of particular steps in monoligol biosynthesis results in the incorporation of pathway intermediates into the polymer (Boerjan et al., 2003; Ralph et al., 2001).

Because lignin is considered a negative factor in a number of economically and environmentally important processes, such as chemical pulping and fodder digestibility by ruminants, there is considerable interest in understanding the biochemical pathway that leads to the synthesis of the three monolignols. Genetic engineering of this pathway has already resulted in improved wood quality for chemical pulping (O'Connell et al., 2002; Pilate et al., 2002) and improved fodder digestibility (Guo et al., 2001).

Over the last decade, there has been a tremendous effort in cloning new genes involved in the monolignol biosynthetic pathway, and in tackling the enzyme kinetics of the corresponding proteins as well as the in vivo role these enzymes play in controlling the amount and composition of lignin to be deposited in the cell wall (Anterola and Lewis, 2002; Humphreys and Chapple, 2002; Boerjan et al., 2003). As a consequence, the monolignol biosynthetic pathway has virtually been rewritten. However, the exact route toward the monolignols is still a matter of debate (Figure 1).

Although in vitro enzymatic assays and transgenic plants have contributed extensively to our understanding of the in vivo role of the enzymes, there are several important limitations that confound a detailed understanding of the pathway. First, multiple copies exist for many of the genes in the genome, with different spatio-temporal expression patterns. The use of gene silencing to unravel the function of these genes inevitably risks the simultaneous silencing of several or all members of the gene family. Hence, the observed phenotype is not necessarily linked to the function of a single gene. Secondly, down-regulation of a particular gene may lead to the accumulation of pathway intermediates, which in turn may affect the expression of other genes of the pathway at the transcriptional or enzyme activity level (Mavandad et al., 1990; Blount et al., 2000; Anterola et al., 2002; Boerjan et al., 2003), and may as well cause alterations in other biosynthetic pathways or in plant development (Hu et al., 1999; Jones et al., 2001). Third, several enzymes of the monolignol biosynthesis are positioned at the outer face of the endoplasmic reticulum

(ER), where they have been proposed to participate in metabolic channeling of the pathway intermediates (Chapple, 1998; Winkel-Shirley, 1999). Inevitably, down-regulation of a single enzyme may perturb the entire metabolic complex with a phenotype that cannot readily be explained as a result.

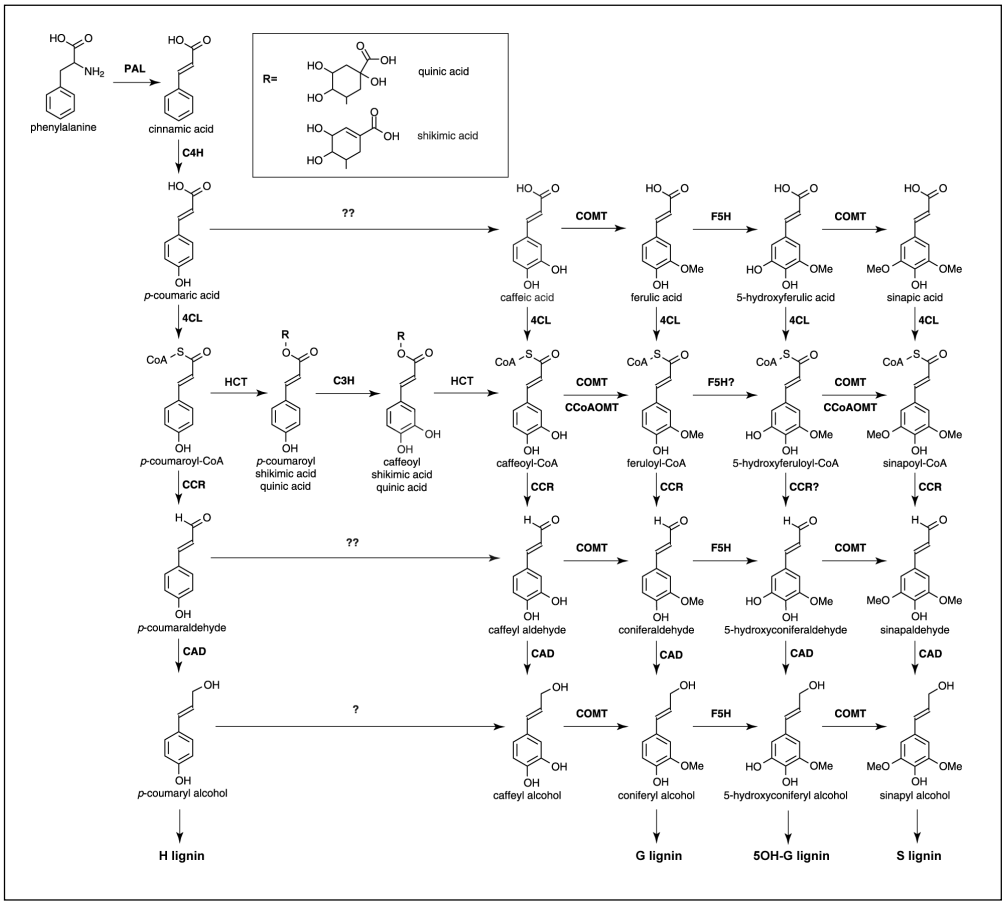

Figure 1. The monolignol biosynthetic pathway.
All the enzymatic reactions presented in the pathway have been demonstrated at least in vitro. The currently most favored route to S and G monolignols starts with the deamination of phenylalanine to cinnamic acid, catalyzed by phenylalanine ammonia lyase (PAL). Subsequently, cinnamic acid is hydroxylated by cinnamic acid 4-hydroxylase (C4H) to *p*-coumaric acid, which is esterified by 4-coumarate:CoA-ligase (4CL) to the corresponding coenzyme A-thioester *p*-coumaroyl-CoA. Then, *p*-coumaroyl-CoA is transesterified to its quinate and/or shikimate ester by hydroxycinnamoyl-CoA:quinate/shikimate hydroxycinnamoyltransferase (HCT), followed by the hydroxylation at the C3 position by coumarate 3-hydroxylase (C3H), and the conversion back to the caffeoyl-CoA thioester by HCT. The C3-hydroxyl group of caffeoyl-CoA is methylated by caffeoyl-CoA-*O*-methyltansferase (CCoAOMT) to feruloyl-CoA. The reduction of feruloyl-CoA by cinnamoyl-CoA reductase (CCR) is the first step in the monolignol-specific branch of the lignin biosynthesis pathway, and results in the synthesis of coniferaldehyde. For the synthesis of G lignin, coniferaldehyde is reduced to coniferyl alcohol by cinnamyl alcohol dehydrogenase (CAD). The synthesis of S units demands the further substitution of the aromatic C5 position. The 5-hydroxylation of coniferaldehyde is carried out by ferulate 5-hydroxylase (F5H; Cald5H), followed by the methylation by caffeic acid *O*-methyltransferase (COMT; AldOMT). Both the 5-hydroxylation and methylation may occur at the alcohol level, as well. The reduction of sinapaldehyde has been postulated to be carried out by sinapyl alcohol dehydrogenase. Because of the variety in isoenzymes and kinetic properties, alternative routes through the metabolic pathway may exist. A question mark after an enzyme name means that the substrate has not been tested yet with this enzyme. For reactions with a single question mark direct conversion has been detected, but the respective enzyme is unknown, whereas for those with a double question mark no direct conversion has been detected.

These limitations can only be dealt with in plant species, such as *Arabidopsis*, for which the genome sequence and efficient reverse genetics tools are available (*Arabidopsis* Genome Initiative, 2000). Furthermore, the advent of the genome-wide micro-arrays will make it possible to study the transcriptional differences that are the consequence of single gene perturbations, and will allow the often pleiotropic phenotype of particular mutants at the molecular level to be explained.
As a first step toward studying the role of individual family members by insertion mutagenesis and microarrays, we have undertaken a bioinformatics approach to identify in *Arabidopsis*, all the gene family members of all monolignol biosynthesis genes known today. In many cases, only a subset of a given gene family has been characterized previously, leading to an important bias in the range of sequence data available in public databases. The fact that for some genes, a plethora of sequences from various organisms exists, whereas for others only (predicted) *Arabidopsis* sequences can be found, has to be attributed to homology-based gene isolation in the past. Consequently, more distant members of a family might not be discovered when, for example, primers are designed on only a few members of the family.
Using the complete genome sequence and sensitive computational approaches, we have detected 34 candidate monolignol biosynthesis genes in the *Arabidopsis* genome, including more distantly related family members. Second, we have analyzed the expression of all 34 genes throughout development and compiled data from all expression studies published so far on these genes, including information extracted from various expressed sequence tag (EST) databases. Subsequently, the combination of phylogenetic analyses with these expression studies and promoter sequence analyses of the individual family members has allowed us to select 12 genes as the most likely candidates to be involved in the developmental lignification in vascular tissues. Importantly, the promoter comparisons revealed a possible link between G lignin biosynthesis and the presence of the AC element that is correlated with a strong xylem expression.

Methods

Annotation

For each of the 10 enzymes of the monolignol biosynthetic pathway, the corresponding genes were annotated in four steps: (i) experimentally certified family members were collected from a variety of species and a family-specific profile was created; (ii) an *Arabidopsis* protein database was scanned with this profile; (iii) true family members were selected; and (iv) prediction on the selected genes was improved with information from different sources, such as cDNA and EST sequences and within-family sequence similarity.
More specifically, based on literature and public database searches, a set of experimentally certified family members from different plant species was compiled for each enzyme. From a CLUSTALW protein alignment of these sequences, a hidden Markov model-based profile was created using the HMMER package (Thompson et al., 1994; Eddy, 1998).

The use of a profile-based approach was preferred over a BLAST-based method because of its capacity to detect remote family members by taking into account the known sequence diversity within the family. This profile was used to scan an *Arabidopsis* protein database, which was constructed through a Genemark.hmm prediction (Lukashin and Borodovsky, 1998) on the complete *Arabidopsis* genome sequence (version 180101, downloaded from the MIPS ftp site, at ftp://ftpmips.gsf.de/cress/). In a second scan, the complete genome sequence was searched with TBLASTN to detect genes that were not or wrongly annotated and would have been missed by using the protein database.

To delineate the gene family, several factors were taken into account. First, only HMMER hits with an E-value score below the default cut-off value (E=10.0) were considered. Second, the E-value score gives indications to distinguish potential family members from false positives, or –in the case of large superfamilies- genes of other subfamilies. In most cases, a clear "drop" in the E-value score could be detected, indicating that sequences below this threshold did not fulfil the family model as well as those above. This approach can potentially lead to wrong conclusions because of incomplete or biased sampling of the family. For this reason, a third method was applied, based on a phylogenetic analysis of the (super)family. The GenBank sequence databases (Benson et al., 2003) as well as the literature were searched for homologs and more distantly related members of distinct, well-known families within the same superfamily. These genes, together with the previously detected gene family members, were subjected to phylogenetic analyses to get an overview of the relations within the complete (super)family. The resulting tree was used to decide whether a protein belonged to the investigated family or not. As a rule, genes that clustered together with experimentally certified family members were considered to be part of the family. (Groups of) genes that did neither belong to the family nor cluster with other known families within the superfamily were considered as "likes", when they formed a sister group to the family investigated. These genes were not analyzed in further detail, but were included for the sake of completeness, because their function might be biochemically related to that of the monolignol biosynthesis genes. Details of the methodology used in the phylogenetic analyses are described further below.

For the family members selected through these three criteria, the automatic annotation was improved by using information from different sources. First, the public databases were searched using BLASTN (Altschul et al., 1997) for ESTs and full-length cDNAs. Only high-quality matching sequences (% identity > 95%) were considered, and manually inspected for wrong assignments and cross-matches between closely related family members. These transcripts were aligned to the genomic region using Sim4 to verify intron-exon borders (Florea et al., 1998). Second, the deduced protein sequences were aligned with the other family members to detect prediction errors (for example, missed exons). Third, predictions for candidate genes were verified with an alternative gene prediction tool called EuGene (specificity = 0.63, sensitivity = 0.74 at the gene level; available at www.inra.fr/bia/T/EuGene/ ; Schiex et al., 2001). This information was compiled with ARTEMIS (Rutherford et al., 2000) and was used to decide on a final gene structure.

For the annotation of the C4H, C3H and F5H families, a substantial amount of information from the P450 databases (at http://www.biobase.dk/P450/p450.shtml and http://drnelson.utmem.edu/CytochromeP450.html) was used to improve the annotation.

Prediction of myristoylation sites was done with the algorithm of Maurer-Stroh et al. (2002), available at http://mendel.imp.univie.ac.at/myristate/. Small Perl scripts were written to detect putative C-terminal farnesylation and geranylgeranylation sites (CaaX, CCXX, XCXC, and XXCC with a, aliphatic, C, cysteine, and X, any amino acid; Randall and Crowell, 1999; Nambara and McCourt, 1999; Thompson and Okuyama, 2000). Signals for subcellular localization were predicted with the TargetP server (http://www.cbs.dtu.dk/services/TargetP/; specificity cut-off of >0.90; Emanuelsson et al., 2000).
The annotation results were submitted to the TAIR and MIPS databases for public access and are also accessible at http://www.psb.ugent.be/bioinformatics/lignin/.

Phylogenetic analysis and mapping of genes onto duplicated blocks

The nonredundant protein database was scanned for homologous sequences using BLASTp (Altschul et al., 1997) and the results were inspected manually. Sequences were aligned with CLUSTALW v.1.84 (Thompson et al., 1994) and alignments were improved manually. Trees were constructed on conserved positions of the alignment with the neighbor-joining algorithm, as implemented in TREECON (Van de Peer and De Wachter, 1997), and by maximum-likelihood analysis (quartet puzzling) with TREE-PUZZLE (Schmidt et al., 2002). Alignments were edited and reformatted with ForCon (Raes and Van de Peer, 1999) and BioEdit (Hall, 1999). Statistical significance of nodes in the neighbor-joining approach was tested by using 500 bootstrap replicates. Duplicated blocks (i.e., large regions of colinearity) in the *Arabidopsis* genome have been described previously (Simillion et al., 2002).

Promoter analysis

Both strands of upstream regions (1,000 bp before the ATG codon or the distance between the previous gene and the ATG) as well as first and second introns of the genes were analyzed for regulatory elements with MatInspector (Quandt et al., 1995). To avoid false positives, we opted for a conservative approach with very strict parameters (core similarity = 0.9; matrix similarity = 0.9). Furthermore, 1,000 random intergenic regions uniformly distributed throughout the *Arabidopsis* genome were searched with these parameters to have a rough estimate of the random occurrence of the motifs.
A list of potentially interesting motifs was compiled on the basis of the following three criteria: the motif had to be (i) experimentally characterized, (ii) implicated in transcriptional regulation of known genes in the monolignol biosynthesis pathway, and/or (iii) involved in elicitor, wound, or pathogen response. The motifs (and their respective calculated random occurrences in the *Arabidopsis* genome) that passed these criteria were: for *Arabidopsis*: GCC box (1/73,000 bp; Rushton et al., 2002), jasmonate- and ethylene-responsive element (1/1,239,000 bp; Rushton et al., 2002), W box (1/2,300 bp; Rushton et al., 2002; withdrawn from analysis because of its high random occurrence), and S box (1/24,000 bp; Rushton et al., 2002); for parsley (FP56; not detected in the random set; Neustaedter et al., 1999) and E box (1/31,000 bp; Grimmig and Matern, 1997);

for pea AT-rich sequence (1/26,000 bp; Seki et al., 1996); for tobacco: salicylic acid-responsive element (1/18,000 bp; Shah et al., 1996) and hypersensitive-response element (1/92,000 bp; Pontier et al., 2001). Furthermore, the joint presence of the *Arabidopsis* OBP-1 binding site (1/38,000 bp; Chen et al., 1996) with an As-1 box (not detected in the random set; Krawczyk et al., 2002), or a common bean H box (1/7700 bp; Lindsay et al., 2002; also considered without G box) with a G box (1/3300 bp; Loake et al., 1992; only considered in conjunction with an H box), respectively, were tested. For the AC I and AC II elements, one unifying profile was built from all experimentally confirmed AC I and AC II elements from different species, in order to increase sensitivity (see supplemental data). The following AC elements were used: eucalyptus AC I (Lacombe et al., 2000), common bean AC I and II (Hatton et al., 1995), parsley AC II (Hauffe et al., 1993), and an AC II element (CTCACCAACCCCCAC) from the poplar *gPtCCoAOMT1* promoter (Chen et al., 2000; C. Chen et al., in preparation). The occurrence of an AC element at random using this matrix was once per 37,000 bp. In addition, the A box, suggested to work in conjunction with AC elements in parsley was included, even though not experimentally verified (1/11,000 bp; Logemann et al., 1995). Motifs used were retrieved from or submitted to the PlantCARE database (Lescot et al., 2001; http://oberon.fvms.ugent.be:8080/PlantCARE/).

Experimental verification of annotation and expression study

Plant material

Expression analysis was carried out in *Arabidopsis thaliana* (L.) Heynh. ecotype Columbia plants. Seeds were surface sterilized and placed on MS medium supplemented with 10 g L^{-1} sucrose. After the seeds had undergone a cold treatment for homogenous germination (overnight at 4°C), they were exposed to 20°C, 50 μmol m^{-2} sec^{-1} light intensity, 70% relative humidity, under a 16-h light/8-h dark cycle. Fourteen days after germination, plants were transferred to soil and cultivated in a greenhouse. Conditions were as follows: 23°C, 50 ìmol m^{-2} sec^{-1} light intensity at plant level (MBFR/U 400 W incandescent lamps; Philips, Eindhoven, The Netherlands), 40% relative humidity, and a 16-h light/8-h dark cycle, without shielding from incident day light. Material was harvested from a number of plants (within brackets) and pooled: seedling leaves and roots of 14-day-old in vitro plants (n=100); rosette leaves, flowers, and green siliques of 7-week-old plants (n=50); and inflorescence stems at 1, 3, 5, 10, 15, and 20 cm length (n=20 for 1, 3, and 5 cm; n=10 for all later stages). At 20 cm, the stems were fully grown.

RNA extraction, primer design, and reverse transcription PCRs

Total RNA was extracted with a LiCl method according to Goormachtig et al. (1995) and digested with DNase I to eliminate residual genomic contamination. Subsequently, 5 μg total RNA were reverse-transcribed into double-stranded cDNA (cDNA Synthesis System Plus, Amersham Biosciences, Little Chalfont, UK). Primers were designed either with the SPADS program that selects specific primers for a particular gene from the *Arabidopsis* genome (21 genes; available at

http://www.psb.ugent.be/databases/SPADS/) or manually (*PAL-1*, *PAL-2*, *PAL-3*, *PAL-4*, *C4H*, *4CL-2*, *4CL-3*, *HCT*, *CCoAOMT-1*, *COMT*, *F5H-1*, *CAD-7*, and *CAD-8*). Primers were designed to span at least one intron for reliable distinction of amplification from cDNA (except for *C3H-2* and *C3H-3* that are single exon genes). Products ranged from 272 bp to 1191 bp for the cDNA (for a complete list of primers and amplification products, see supplemental data). Prior to the expression analysis, primers were tested on genomic DNA and random cDNA to verify correct amplification products. In RT-PCR experiments, 25-μl reaction buffer supplied with the *Taq* polymerase and 50 ng of each primer contained a modified nucleotide mix: 200 pmol of dCTP, dTTP, and dGTP, whereas dATP was reduced to 20 pmol. To each reaction, 0.1 μl of ^{33}P-labeled dATP (10 mCi/ml, 2,500 Ci/mmol) was added, resulting in a hot-to-cold dATP ratio of 1:2,500. Products were separated on 3% or 4.5% polyacrylamide gels and visualized on dried gels through autoradiography. To increase the reliability of the assays, the PCR reaction was run with at least two template concentrations (1 μl 1:10 diluted cDNA, 1 μl undiluted cDNA). The amount of amplified cDNA was categorized as low (+/-), moderate (+), or high (++). These expression categories for a particular gene apply only for comparison of different tissues, but not between genes because of the different PCR dynamics of shorter or longer amplification products.

EST Analysis

Data on size and nature of EST libraries was obtained from http://www.ncbi.nlm.nih.gov/UniLib/, http://www.ncbi.nlm.nih.gov/Entrez/ and additionally for the RIKEN *Arabidopsis* full-length cDNA clones (RAFL) from Seki et al. (2002). A total of 160,776 *Arabidopsis* ESTs were grouped into 11 categories: whole plant (35,544; 22.1%), aboveground organs (17,934; 11.2%), seedling (3,207; 2.0%), roots (20,332; 12.6%), flowers (6,814; 4.2%), inflorescence stem (1,384; 0.9%), siliques and seeds (25,043; 15.6%), pathogen infection (2,366; 1.5%), wounded leaves (707; 0.4%), various stresses (44,007; 27.4%), and yet unclassified ESTs (542; 0.3%). Stress ESTs are from subtracted, normalized as well as non-subtracted, non-normalized libraries. The whole-plant category includes whole plants, whole rosettes and cell suspensions as starting material. Aboveground organs include, next to libraries that are described as such, libraries from mixed aboveground sources, such as whole inflorescences. Each EST was assigned to one class only. Although inevitably arbitrary and subjective, this classification was done to create clarity and to allow an easier interpretation of the results. Full details on classes and a complete list of ESTs found for each gene is available as supplemental data and at http://www.psb.ugent.be/bioinformatics/lignin/.

Results

We searched the complete *Arabidopsis* genome for members of the gene families currently known to be involved in monolignol biosynthesis. A semi-automatic structural annotation was performed using prediction results, experimental data, and information from homologous sequences (see Methods). A total of 34 candidate monolignol biosynthesis genes were annotated, of which 11 had, to our knowledge, never been described before. Additionally, 27 closely related superfamily members ("likes") were identified in this process. Besides annotation and evolutionary analysis of the gene families, putative promoter elements that drive expression during lignification, in pathogen and wound response and after induction by stress-related hormones, as well as potential subcellular localization signals are presented. To get a first insight into whether all these genes are indeed expressed and, more importantly, whether their expression pattern correlates with developmental lignification, their expression was analyzed in a set of tissues and for six developmental stages of inflorescence stem known to contain a high portion of lignifying cells. These data were compared with previous expression data from *Arabidopsis* and with information extracted from public EST databases. Together, these data describe the full complement of monolignol biosynthesis genes in *Arabidopsis* and serve as a basis for further functional studies.

Phenylalanine ammonia-lyase (PAL)

Phenylalanine ammonia-lyase (PAL, EC 4.3.1.5) is the first enzyme of the general phenylpropanoid pathway and catalyzes the non-oxidative deamination of phenylalanine to *trans*-cinnamic acid and NH_3 (Figure 1). PAL mediates the influx from primary metabolism into the phenylpropanoid pathway and becomes rate limiting when its activity is reduced below a threshold of 20-25% in transgenic tobacco (Bate et al., 1994; Sewalt et al., 1997).

By using the annotation method (see Methods), four genes encoding PAL proteins were detected in the *Arabidopsis* genome, three of which have been described previously (Ohl et al., 1990; Wanner et al., 1995). The phylogenetic analysis of *PAL* genes from various species provided no evidence for different classes in the *PAL* gene family (Figure 2), although *PAL-1* is most closely related to *PAL-2*, and *PAL-3* always clusters together with *PAL-4* (data not shown).

PAL-1, *PAL-2*, *PAL-3*, and *PAL-4* are situated on chromosome 2, 3, 5, and 3, respectively (Figure 12). The four *PAL* genes are part of two duplicated regions in the *Arabidopsis* genome, which originated during a complete genome duplication, approximately 75 million years ago (Simillion et al., 2002; Raes et al., 2003). The duplication that created the two *PAL* groups (*PAL-1* and *PAL-2*; *PAL-3* and *PAL-4*) in *Arabidopsis* occurred before this date and has been postulated to have predated the monocot-dicot split (Wanner et al., 1995), but the latter is not confirmed by our phylogenetic tree (Figure 2).

PAL-1 and *PAL-2* are not only structurally very similar, but they also share common promoter elements and a similar expression pattern (Table 1). mRNAs from both genes are most abundant in roots and stems, where the expression increases during the later stages of development (Table 1; Wanner et al., 1995).

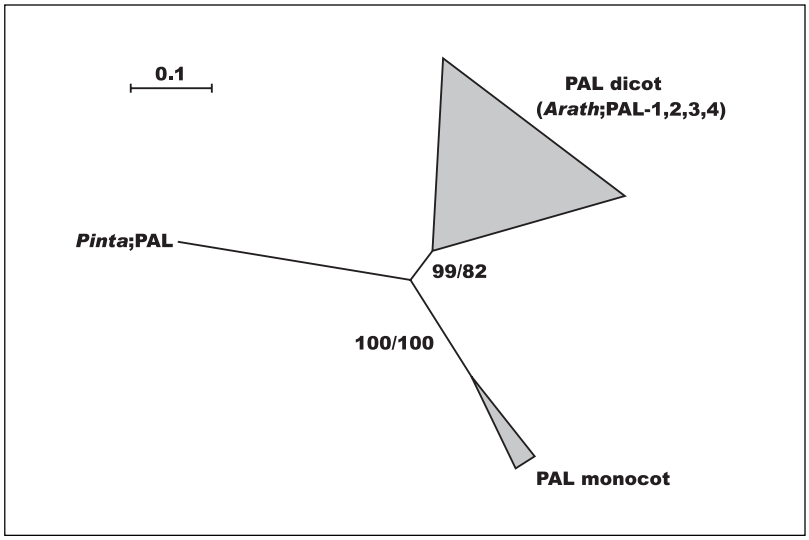

Figure 2. Neighbor-joining tree of the *PAL* family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per nucleic acid. Clusters of sequences are represented as triangles with a height equal to the average distance separating the terminal nodes from the deepest branching point in the cluster, and a base proportional to the number of sequences composing it. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: dicots: *Populus* (169453, 485808, 1109640), *Glycine* (18376), *Trifolium* (437711), *Citrus* (4808125, 4808127, 1276902), *Rubus* (7208613, 7208615), *Camellia* (662270), *Petroselinum* (534892), *Nicotiana* (170349); *Digitalis* (2631994), *Lactuca* (18001006); monocot: *Oryza* (20280, 871493), and gymnosperm: *Pinus*: 1143311. Abbreviations: *Arath*, *Arabidopsis thaliana*; *Pinta*, *Pinus taeda*.

Analyses of the fusion between *AtPAL1* and α-glucuronidase (*GUS*) revealed that the expression is located in the vascular tissues (Ohl et al., 1990; Leyva et al., 1995). Besides *PAL-1* and *PAL-2*, also *PAL-4* is highly expressed in root and stem tissue, as shown by our reverse transcription (RT)-PCR expression analysis and by the high number of ESTs (Table 1). Additionally, *PAL-2* and *PAL-4* are abundantly expressed in the seed, as judged from the EST data (Table 1). Although all four genes are almost ubiquitously expressed in the tissues investigated in this study, *PAL-3* seems to be generally expressed at a lower level (Table 1; Wanner et al., 1995; Mizutani et al., 1997; Ruegger et al., 1999).

PAL-1 was one of the first plant defense genes identified and its involvement in pathogen infection and abiotic stress has been studied. *PAL-1* as well as *PAL-2* expression is induced by pathogens (Wanner et al., 1993; Leyva et al., 1995; Mauch-Mani and Slusarenko, 1996; Ehlting et al., 1999) and by wounding (Ohl et al., 1990; Lee et al., 1997; Mizutani et al., 1997), whereas *PAL-3* expression decreases in response to wounding (Mizutani et al., 1997). Among the ESTs derived from diverse stresses, *PAL-1* and *PAL-2* are clearly the most important stress-responsive family members with 20 out of 41 ESTs and 17 out of 50 ESTs in total, respectively, even taking into account the relative database sizes (Table 1).

Table 1. Expression characteristics of the *PAL* gene family in *Arabidopsis*.
Data from *Arabidopsis* literature, ESTs, and our experimental RT-PCR are given in the table as +/- for low, + for moderate, ++ for high, and – for decreasing expression. Because of different PCR dynamics of fragments of different size and separate RNA gel blots, data can be compared only among the different tissues, but not between genes or experiments. In case of chimeric promoter-GUS constructs, only those of *Arabidopsis* promoters analyzed in *Arabidopsis* were included. Data from GUS and immunohistochemistry were included whenever available. Shaded fields without a number indicate that the tissue/condition was studied, but no expression detected. ESTs are given in absolute (EST) as well as in relative (EST rel) numbers to account for the different sizes of EST classes and to estimate overrepresentation of ESTs in a particular condition. To this end, the number of ESTs for a particular gene in a given class was divided by the total number of ESTs in this class and multiplied by 100,000 to yield a comparable relative number in ESTs/100,000 ESTs (rounded to the nearest whole number). See Methods for the full description of classification and total numbers in the different classes. Shaded fields without a number indicate that no ESTs were found in the tissue or condition. Abbreviations: ER, endoplasmic reticulum; ER-anchored, localization in the ER membrane through the membrane anchor of P450 enzymes; mRNA, RNA gel blots; A, A box; AC, AC-unified element; AT, AT-rich element; E, E box; G+H, G box in conjunction with H box; GCC, GCC box; H, H box; S, S box; SARE, salicylic acid-responsive element. When an element occurs more than once in a particular promoter, the number is given within parentheses after the respective element. Promoter elements searched for, but not found in any of the 34 genes involved in monolignol biosynthesis are: As-1 box in conjunction with an OBP-1 binding site, the jasmonate- and ethylene-responsive element (JERE), the FP56, and the hypersensitivity-response element (HSRE). See Methods for the respective random occurrences of the elements in the *Arabidopsis* genome.

Gene	other names	AGI number	signals for localization	expression																					regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Arath;PAL-1	*PAL1* [a]	At2g37040		RT-PCR	+				+	+		+		+	+	++	++	++	++						AC	
				EST (41)*	2		1	1	13			2											1	20	S	
				EST rel (25.5)*	62		3	6	64			8											42	45	H	
				mRNA[a]															+		++		+a1			
				mRNA[b]																		+b1				
				mRNA[c]																			+c1			
				mRNA[d]															+				+d1			
				mRNA[e]	+				++	+	+															
				mRNA[g]																	+					
				mRNA[h]					+	+	+	+	+						+		++					
				mRNA[i]																		+i1				
				mRNA[j]	+	+																				
				AtPAL1::GUS [a]	+a2				+a3	+a4	+a5								+a6		+a6		+a1 6			
				AtPAL1::GUS [d]									+d2									+d3	+d4			
				AtPAL1::GUS [f]						+f1												+f2				
Arath;PAL-2	*PAL2* [e]	At3g53260		RT-PCR	++				+	+	+	+		+	+	+	++	++	++						AC (2)	
				EST (50)			2	2	17			12												17	S	
				EST rel (31.1)			6	11	84			48												39		
				mRNA[b]																		+b1				
				mRNA[e]					++	+	+															
				mRNA[g]																	+					
				mRNA[h]					+	+	+	+	+						+/-		+/-					
				mRNA[j]	+	+																				
				RT-PCR[k]	+k1																					
Arath;PAL-3	*PAL3* [e]	At5g04230		RT-PCR	+				+	+	+	+/-		+	+/-	+	+	+/-	+						E	
				EST (1)*																					GCC	
				EST rel (0.6)																						
				mRNA[e]																						
				mRNA[g]																	+					
				mRNA[h]						+	+/-	+	+						+		-					
				mRNA[j]																						
				RT-PCR[e]	+				+							+										
Arath;PAL-4		At3g10340		RT-PCR	+				+	+	+	+		+	+	+	++	++	++						A	
				EST (28)			2	1	5			16												4	G+H	
				EST rel (17.4)			6	6	25			64												9		

[a] Ohl *et al*., 1990
[a1] induction by $HgCl_2$
[a2] in all tissues except the root tip and the shoot apical meristem
[a3] except the root tip, strong in vascular tissue
[a4] in vascular tissue
[a5] in sepals, anthers and carpels, not in petals, very strong in pollen
[a6] *GUS* transcript
[b] Wanner *et al*., 1993
[b1] *Pseudomonas* infection
[c] Deikman and Hammer, 1995
[c1] cytokinin induction
[d] Leyva *et al*., 1995
[d1] low temperature
[d2] very strong GUS activity in protoxylem cells
[d3] *GUS* transcript, upon *Pseudomonas* infection
[d4] at low temperatures, GUS activity in the cortical cells (photosynthetically active) of the inflorescence stem
[e] Wanner *et al*., 1995
[f] Mauch-Mani and Slusarenko, 1996
[f1] in vascular tissue
[f2] *Peronospora* infection
[g] Lee *et al*., 1997
[h] Mizutani *et al*., 1997
[i] Ehlting *et al*., 1999
[i1] *Peronospora* infection
[j] Ruegger *et al*., 1999
[k] Jin *et al*., 2000
[k1] in seedling leaves
[*] 1 EST is unclassified

In accordance with the expression pattern, the promoters of *PAL-1* and *PAL-2* contain well-conserved AC elements that specify vascular expression of phenylpropanoid genes (Table 1; Ohl et al., 1990; Wanner et al., 1995; Hatton et al., 1995; Hauffe et al., 1993; Lacombe et al., 2000; C. Chen et al., in preparation). An A box, proposed to work in conjunction with the AC elements in the parsley *PAL-1* and *PAL-4* genes (Logemann et al., 1995), was not detected in the *Arabidopsis PAL-1* and *PAL-2* promoters (Table 1). *PAL-4* contains an A box, but lacks an AC element. Interestingly, an H box and a G box were found in the *PAL-4* promoter. This combination of *cis* elements was shown to be sufficient for the feed-forward induction of the chalcone synthase (*CHS*) promoter by *p*-coumaric acid in bean (Loake et al., 1992; Lindsay et al., 2002). This observation may indicate that *PAL-4* is regulated by the reaction product of C4H.
Additionally, a number of regulatory elements, shown to be involved in promoter responsiveness to elicitors, wounding, and pathogen infection, were found (Table 1). An S box was detected in the promoters of *PAL-1* and *PAL-2*, an E box and a GCC box in that of *PAL-3*, and an H box in those of *PAL-1* and *PAL-4*. In synthetic reporter-gene constructs, the S box and the GCC box conferred elicitation-dependent transient expression in parsley protoplasts and, in *Arabidopsis*, expression upon wounding and infection with different pathogens (Rushton et al., 2002). The E box was shown to be involved in elicitation and basal expression in parsley protoplasts (Grimmig and Matern, 1997), and the H box to be the binding sequence for the KAP-2 transcription factor, which is responsible for induced expression of the bean *CHS15* gene after elicitation (Yu et al., 1993; Lindsay et al., 2002).
In conclusion, all *PAL* genes are expressed in the inflorescence stem, a tissue with a high portion of lignifying cells. However, the presence of an AC element qualifies *PAL-1* and *PAL-2* as the most likely candidates to be involved in monolignol biosynthesis in the vascular lignifying cells. In accordance, the corresponding mutants show defects in lignin formation (A. Rohde et al., in preparation). However, *PAL-4* remains a very interesting candidate as well, because of its increasing expression during stem development.

trans-Cinnamate 4-hydroxylase (C4H)

trans-Cinnamate 4-hydroxylase (C4H, E.C. 1.14.13.11) controls the conversion of cinnamate into *p*-coumarate (Figure 1). C4H (CYP73A5) belongs to the cytochrome P450-dependent monooxgenases, like the two other hydroxylases in the pathway (C3H, F5H). So far, only one *C4H* gene has been described in *Arabidopsis* (Bell-Lelong et al., 1997; Mizutani et al., 1997; Urban et al., 1997). Although multiple family members have been detected in other plants (Betz et al., 2001, and references therein), we could not find any evidence for additional *CYP73* genes in *Arabidopsis*. The discovery of more divergent members of this family in common bean (*Phaseolus vulgaris*), Valencia orange (*Citrus sinensis)*, ice plant (*Mesembryantheum crystallinum*), and maize (*Zea mays*), has led to the hypothesis that two classes of *C4H* genes exist in plants (Nedelkina et al., 1999; Betz et al., 2001), which is confirmed by our phylogenetic analysis (Figure 3).

Furthermore, the tree topology indicates that the origin of these two classes has predated the divergence of gymnosperms and angiosperms, suggesting that class II members must have existed at some time in evolution for most plant lineages. The *C4H* gene detected on chromosome 2 in *Arabidopsis* belongs to class I, whereas a class II homolog was most probably lost during its evolution.

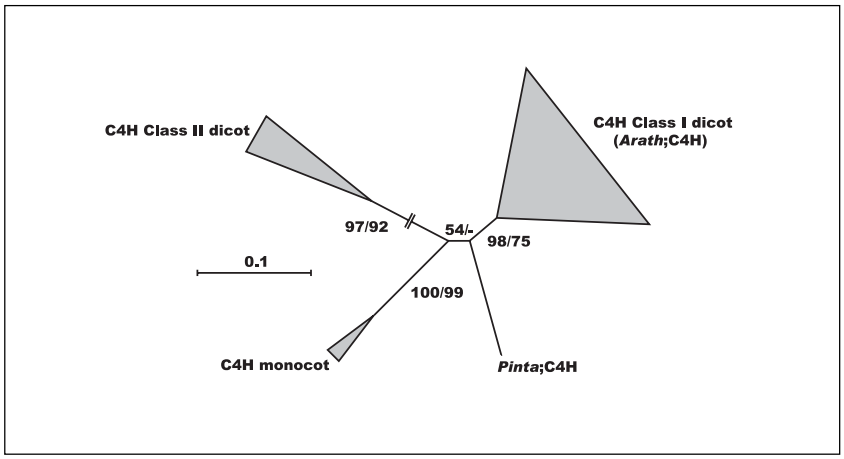

Figure 3. Neighbor-joining tree of the C4H family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: Class I dicots: *Populus* (12276037, 3915089, 3915096), *Gossypium* (9965899, 9965897), *Petroselinum* (3915088), *Ruta* (13548653), *Citrus* (8572559, 14210375), *Catharanthus* (1351206), *Lithospermum* (16555879, 16555877), *Capsicum* (3603454, 12003968), *Zinnia* (3915112), *Helianthus* (417863), *Glycine* (3915111), *Phaseolum* (586082), *Glycyrrhiza* (3915095), *Cicer* (14917048), *Medicago* (586081), *Pisum* (3915077, 9957081); Class II dicots: *Mesembryanthemum* (4206116), *Citrus* (7650489), *Phaseolus* (7430650), *Nicotiana* (14423323, 14423325); monocots: *Triticum* (10442761), *Sorghum* (14192803); gymnosperm: *Pinus*: 4566493. Abbreviations: *Arath*, *Arabidopsis thaliana*; *Pinta*, *Pinus taeda.*

C4H is expressed in all tissues and upon exposure to light, wounding, and fungal infection (Table 2). A strong expression in roots and inflorescence stems has been reported repeatedly (Table 2; Bell-Lelong et al., 1997; Meyer et al., 1998; Nair et al., 2002). In 12 consecutive samples of *Arabidopsis* inflorescence stems, *C4H* expression increased strongly from the 6th sample from the top toward the base, where the lignification process is more active (Meyer et al., 1998).

In our RT-PCR experiment, *C4H* expression increased as well during the later stages of stem development (Table 2). Activity of AtC4H::GUS coincides in the inflorescence stem and in leaves with vascular cells, but in roots the promoter is active in all cells. Consequently, *AtC4H::GUS* expression is strongest in roots (Bell-Lelong et al., 1997; Nair et al., 2002). A strong *C4H* expression is also found in siliques and seeds, where it could be involved in the production of sinapate esters (Chapple et al., 1994). The *C4H* promoter contains an H box, which might be responsible for induction of *C4H* expression after elicitation.

Table 2. Expression characteristics of the *C4H* gene in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																					regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																										
Arath;C4H	CYP73A5	At2g30490	ER-anchored	RT-PCR	+				+	+		+		+	+	+	+	++	++						H	
	REF3 [i]			EST (29)			6		2		1	5											1	14		
				EST rel (18.0)			17		10		15	20											32	42		
				mRNA[a]	+	+/-			++	+	+	+	++							+		++				
				mRNA[b]																		+				
				mRNA[c]					+	+	+	+	+							+		++				
				mRNA[d]									++													
				mRNA[e]																			+e1			
				mRNA[f]	+	+																				
				mRNA[h]					++	+	+	++	+													
				RT-PCR [g]	+g1																-g1					
				AtC4H::GUS [a]	+a1				+a2	+a3	+a4	+a5	+a6									+				
				AtC4H::GUS [h]	+				+h1	+h2	+h3	+h24	+h2									+				

[a] Bell-Lelong *et al*., 1997
[a1] GUS in primary leaves, cotyledons, strongest in root
[a2] highest expression in roots
[a3] restricted to veins in mature leaves
[a4] weak throughout the flower, including vasculature of sepals, with stronger staining at the stigma
[a5] GUS stronger in older siliques than in younger ones
[a6] restricted to xylem
[b] Lee *et al*., 1997
[c] Mizutani *et al*., 1997
[d] Meyer *et al*., 1998
[e] Ehlting *et al*., 1999
[e1] *Peronospora* infection
[f] Ruegger *et al*., 1999
[g] Jin *et al*., 2000
[g1] in seedling leaves
[h] Nair *et al*., 2002
[h1] expression highest in roots, all cell types
[h2] in the vascular tissue of stem, petiole, leaf, and silique wall
[h3] overall staining in the flower, unlike *C3H::GUS*
[h4] strong in seed, unlike *C3H::GUS*
[i] C. Chapple, personal communication

By TargetP (Emanuelsson et al., 2000), the C4H protein is predicted to contain an ER targeting peptide. However, this peptide coincides with the membrane-anchor region of P450 enzymes, whose features are a stretch of hydrophobic amino acids, followed by small region rich in basic amino acids and a hinge region of the conserved (P/I)PGPx(G/P)xP sequence (Chapple, 1998). In contrast to the class I, the class II C4H proteins lack the conserved (P/I)PGPx(G/P)xP sequence of the hinge region for membrane anchoring. Albeit first demonstrated in orange (Betz et al., 2001), our analysis of all proteins included in the phylogenetic analysis (Figure 3) shows that this hinge region as well as the basic amino acid region is divergent in all class II C4H proteins, when compared to those of class I. Although the function of these class II C4H proteins is unclear at the moment, the shared degeneration of this crucial region could be an important clue in discovering their function.

4-Coumarate:coenzyme A ligase (4CL)

4-Coumarate:coenzyme A (CoA) ligase (4CL; EC 6.2.1.12) catalyzes the formation of Coenzyme A esters of *p*-coumaric acid, caffeic acid, ferulic acid, 5-hydroxyferulic acid, and sinapic acid (Figure 1; Lee et al., 1997; Hu et al., 1998; Lindermayr et al., 2002). The reaction involves an adenylate intermediate and, therefore, 4CL shares conserved motifs with other adenylate-forming enzymes, such as peptide synthetases, luciferases, and other CoA ligases. The plethora of potential substrates may explain why there are many 4CL isoenzymes in most plants.

In addition to the different substrate specificities, the genes typically have a distinct spatio-temporal expression pattern (Lewis and Yamamoto, 1990; Hu et al., 1998; Harding et al., 2002).
The 4CL family has recently been subdivided into classes I and II, with the majority of well-characterized 4CL proteins found in class I (Ehlting et al., 1999; Cukovic et al., 2001). We detected four *4CL* and nine *4CL*-like genes in the *Arabidopsis* genome. Phylogenetic analysis of the predicted proteins together with characterized 4CL proteins, as well as luciferases, acetate, and fatty acid CoA-ligases (data not shown), confirms that 4CL proteins fall into two classes (Figure 4).

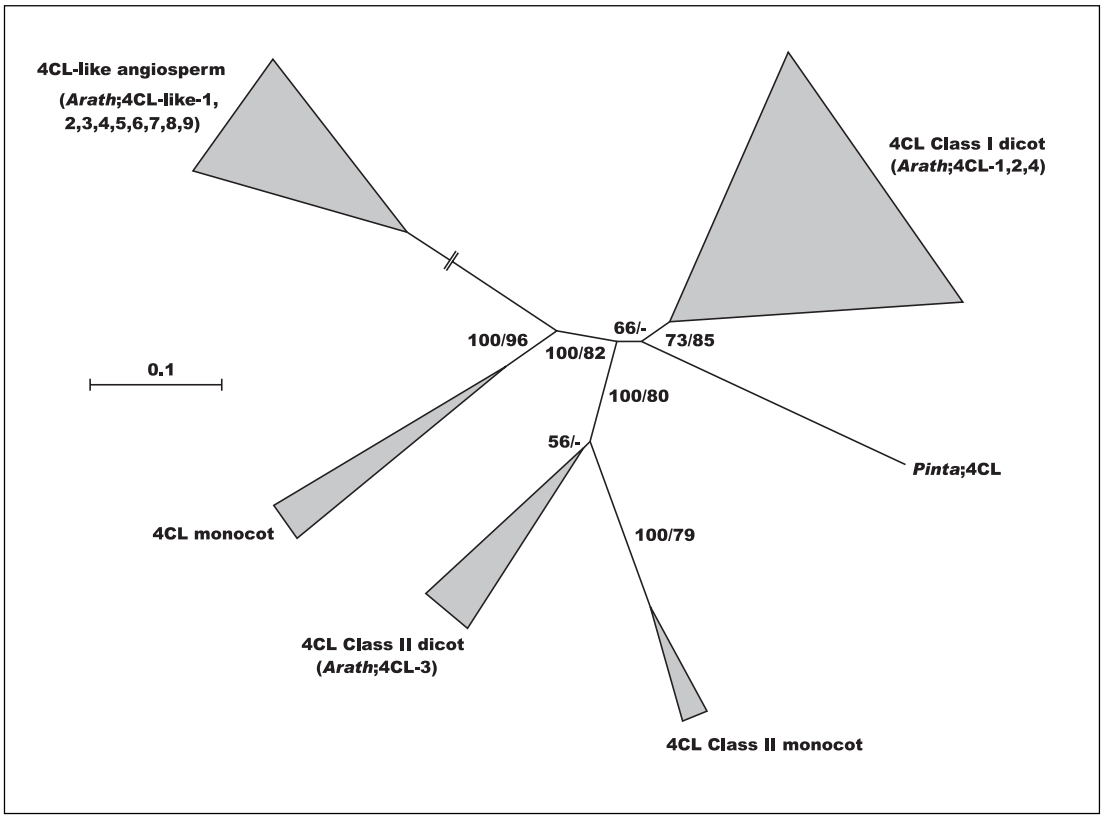

Figure 4. Consensus of two Neighbor-joining trees of the 4CL and 4CL-like proteins, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: Class I dicots: *Solanum* (398963, 398965, 5163399), *Capsicum* (12003966), *Nicotiana* (12229631, 7428495, 12229632), *Lithospermum* (1117778), *Petroselinum* (112800, 112801), *Rubus* (9651915, 9651917), *Populus* (7437854, 7437855, 14289344, 18032806, 7437852, 15636677), *Amorpha* (17063848); gymnosperm: *Pinus* 4CL: 7437872; Class II monocots: *Lolium* (7188335), *Oryza* (12229650); Class II dicots: *Lithospermum* (9988455), *Glycine* (18266852), *Populus* (7437853, 14289346); monocots: *Oryza* (112802), *Lolium* (7188337, 7188339); 4CL-like: *Oryza* (12039389). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Pinta*, *Pinus taeda*.

Three of the putative *Arabidopsis* proteins belong to class I (4CL-1, 4CL-2, and 4CL-4) and 4CL-3 to class II; the remaining nine are classified as 4CL-like, because they do not correspond to any of the 4CL or other enzyme classes mentioned above. Three *4CL* genes have already been described (Lee et al., 1995, 1997; Ehlting et al., 1999; Stuible et al., 2000), whereas, to our knowledge, *4CL-4* is described for the first time in this study.

4CL-1 and *4CL-3* are on chromosome 1 and *4CL-2* and *4CL-4* on chromosome 3 (Figure 12). Whereas the position of *4CL-4* next to *4CL-2* on chromosome 3 could suggest that both genes arose through a recent tandem duplication and, therefore, may be functionally redundant, in the phylogenetic tree *4CL-1* and *4CL-2* are more closely related to each other than to *4CL-4* (data not shown). Additionally, the analysis of duplicated regions in the *Arabidopsis* genome revealed that the *4CL* class I genes are part of a duplicated segment originating from the complete genome duplication 75 million years ago (Figure 12). A possible explanation for these observations is that an ancestor of these three genes (*4CL-1*, *4CL-2*, and *4CL-4*) was duplicated in tandem, after which the resulting two genes were duplicated “en bloc” by the genome duplication, followed by the loss of one of the genes on chromosome 1.

Table 3. Expression characteristics of the *4CL* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																				regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																									
Arath;4CL-1	*4CL1* [a]	At1g51680		RT-PCR	+				+	+	+/-	+		+	+	+	++	++	++						AC	
				EST (8)	1		2					1											4			
				EST rel (5.0)	31		6					4											9			
				mRNA[a]	+				+	+/-	+/-				++	++	++		++			+	+[a1]			
				mRNA[b]																			+			
				mRNA[c]					+	+	+	+	++							+			+			
				mRNA[d]	+				++	+/-	+/-	+	++									++	++[d1]	++[d2]		
				mRNA[e]	+	+																				
				RT-PCR [f]	+[f1]																f1,2					
Arath;4CL-2	*4CL2* [d]	At3g21240		RT-PCR	+				+	+		+		+/-	+	+	++	++	++						AC (2)	AC*
				EST (13)			2		6		1	2											2			
				EST rel (8.1)			6		30		15	8											4			
				mRNA[d]	+				++	+/-	+/-	+	+/-									+	++[d1]	+[d2]		
Arath;4CL-4		At3g21230		RT-PCR					+	+								+	+						AT	AC*
				EST (2)																			2		H	
				EST rel (1.2)																			4			
Class II																									
Arath;4CL-3	*4CL3* [d]	At1g65060		RT-PCR					+/-	+/-	+/-	+				+	+/-	+						H	
				EST (8)			1		1		1	2											3			
				EST rel (5.0)			3		5		15	8											7			
				RT-PCR [d]					+/-	+	++	+	+/-									d3	d3	++[d2]		
				RT-PCR [f]	+[f1]																+[f1]					
Class 4CL-likes																									
Arath;4CL-like-1		At1g20510																							
Arath;4CL-like-2		At1g20500																							
Arath;4CL-like-3		At1g20490																							
Arath;4CL-like-4		At1g20480																							
Arath;4CL-like-5		At1g62940																							
Arath;4CL-like-6		At4g19010																							
Arath;4CL-like-7		At4g05160																							
Arath;4CL-like-8		At5g63380																							
Arath;4CL-like-9		At5g38120																							

[a] Lee *et al.*, 1995
[a1] *Pseudomonas* infection
[b] Lee *et al.*, 1997
[c] Mizutani *et al.*, 1997
[d] Ehlting *et al.*,1999
[d1] *Peronospora* infection
[d2] UV irradiation
[d3] *4CL-3* expression not affected by *Peronospora* or wounding
[e] Ruegger *et al.*, 1999
[f] Jin *et al.*, 2000
[f1] in seedling leaves
[f2] *4CL-1* transcript unaffected by sucrose
[*] in the first intron

In soybean, a single amino acid deletion determines whether or not 4CL can use sinapic acid as a substrate (Lindermayr et al., 2003), a function lacking for 4CL-1, 4CL-2, and 4CL-3 in *Arabidopsis* (Ehlting et al., 1999). Interestingly, *4CL-4* shows a deletion in the region where the single amino acid deletion of soybean resides, suggesting that this gene may have acquired an altered substrate specificity toward sinapic acid after duplication.

Our expression analysis showed that *4CLs* are expressed in most investigated tissues (Table 3). With its expression in leaves, root and mature stem, *4CL-4* has the most restricted expression (Table 3). The latter observation is supported by the smallest number of ESTs found for *4CL-4* among the *4CLs*. *4CL-1* and *4CL-2* are expressed throughout inflorescence stem development and expression increases during the later stages (Table 3; Lee et al., 1995; Mizutani et al., 1997, Ehlting et al., 1999). On the contrary, *4CL-3* and *4CL-4* are expressed only during the later stages of inflorescence stem development (Table 3). These 4CL proteins may use other substrates that typically occur in more mature tissues. The expression of *4CL-3* differs clearly from the class I *4CL* genes (Table 3; Ehlting et al., 1999). *4CL-1* and *4CL-2* expression is affected by wounding and *Peronospora* infection, while *4CL-3* is unaffected by both. In addition, the highest *4CL-3* expression was found in the flower (Ehlting et al., 1999).In accordance with the expression analysis, the promoters of both *4CL-1* and *4CL-2* contain AC elements, which have been shown to correlate with a strong vascular expression. Furthermore, the promoter analysis identified an AT-rich sequence motif in the *4CL-4* promoter and an H box in the *4CL-3* and *4CL-4* promoters, hinting to a role in particular stress responses (Rushton et al., 2002; Seki et al., 1996).

In conclusion, 4CL-1 and 4CL-2 are the best candidates for a function in monolignol biosynthesis during developmental lignification, as suggested previously by Ehlting et al. (1999). Their expression correlates with tissues containing a high portion of lignifying cells and AC elements are present in their promoters. To the contrary, 4CL-3 (class II) was suggested to channel activated *p*-coumarate to CHS and subsequently to the flavonoid biosynthesis (Ehlting et al., 1999). 4CL-4 (class I), although expressed more specifically or at a lower level, might have yet another substrate specificity, possibly including sinapic acid.

Hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase (HCT)

Hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase (HCT) belongs to a large family of acyltransferases that are involved in the biosynthesis of diverse secondary metabolites. Only recently, the first HCT has been purified from tobacco stems and the corresponding gene cloned (Hoffmann et al., 2003). In tobacco, HCT catalyzes the conversion of *p*-coumaroyl-CoA and caffeoyl-CoA to the corresponding shikimate or quinate esters (Figure 1). These shikimate and quinate esters, themselves being important intermediates in the phenylpropanoid pathway, have recently been shown to be good substrates for C3H (Kühnl et al., 1987; Schoch et al., 2001; Franke et al., 2002a, 2002b; Nair et al., 2002). Moreover, HCT catalyzes also the reverse *trans*-esterification (Hoffmann et al., 2003). Therefore, HCT might play a critical role up- and downstream of C3H. For the *Arabidopsis* HCT homolog, a biochemical activity similar to that of the tobacco HCT has been shown (Hoffmann et al., 2003).

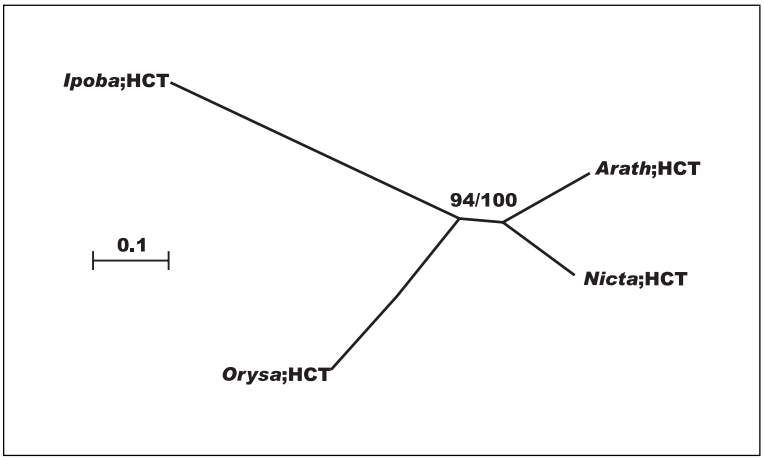

Figure 5. Neighbor-joining tree of the HCT family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: *Ipomoea* (6469032), *Oryza* (21740518), and *Nicotiana* (GenBank entry not available - see Hoffmann et al., 2002). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Ipoba*, *Ipomoea batatas*; *Nicta*, *Nicotiana tabacum*; *Orysa*, *Oryza sativa*.

Here, no more members of the HCT family were detected in the *Arabidopsis* genome, although a small number of homologs have been found in other species (Figure 5). With only two characterized members, the delineation of this family is not straightforward. For this reason and the apparent well-conserved nature of the family (~60% identity between monocot and dicot members; data not shown), no more distantly related genes were included. *HCT* lies on chromosome 5 (Figure 12). The expression analysis shows that *HCT* is expressed in all tissues investigated, but strongly in the inflorescence stem with an increase during the later stages of development (Table 4).

Table 4. Expression characteristics of the *HCT* gene in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																			regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Arath;HCT		At5g48930		RT-PCR	+				+/-	+	+/-	+		+	++	++	++	++	++						AC	
				EST (57)*			7	13	11		1	8											15		G+H	
				EST rel (35.5)			20	73	54		15	32											34			

* 2 ESTs are unclassified

The promoter contains an AC element. The high and ubiquituous expression is confirmed by the second highest number of ESTs found for the 10 gene families analyzed (Table 4). Interestingly, the combined presence of an H and a G box was observed, as for *PAL-4* and *F5H-2*, suggesting transcriptional regulation by the pathway intermediate *p*-coumaric acid (Loake et al., 1992).

p-Coumarate 3-hydroxylase (C3H)

p-Coumarate 3-hydroxylase (C3H) was originally named after its suspected function in C3-hydroxylation of *p*-coumaric acid, but recently CYP98A3 (C3H-1) was shown to preferentially convert the shikimate and quinate esters of *p*-coumaric acid into the corresponding caffeic acid conjugates, whereas *p*-coumaric acid and *p*-coumaroyl-CoA were no substrates of this enzyme (Figure 1; Schoch et al., 2001; Franke et al., 2002b; Nair et al., 2002).
We detected three *C3H* genes in the *Arabidopsis* genome, which all belong to the CYP98 class of the P450 enzymes. Only a few enzymes of this class could be found from other species for phylogenetic analysis (Figure 6). *Arabidopsis* C3H-1 clusters with all known C3Hs in other species, whereas C3H-2 and C3H-3 (CYP98A8 and CYP98A9, respectively) probably constitute a different class that diverged before the gymnosperm-angiosperm split (Figure 6). *C3H-1* is located on chromosome 2, whereas *C3H-2* and *C3H-3* are probably the result of a tandem duplication on chromosome 1 (Figure 12).

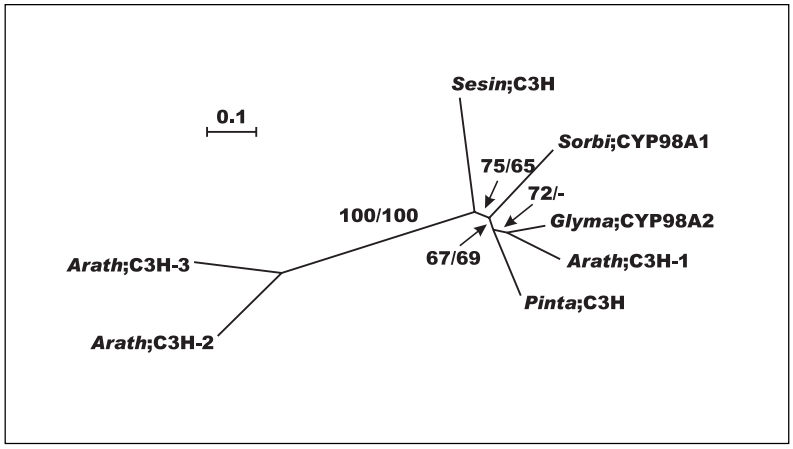

Figure 6. Neighbor-joining tree of the C3H family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: *Sesamum* (17978831), *Sorghum* (5915857), *Pinus* (17978651), and *Glycine* (5915858). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Glyma*, *Glycine max*; *Sesin*, *Sesamum indicum*; *Sorbi*, *Sorghum bicolor*; *Pinta*, *Pinus taeda*.

Our expression analysis shows that *C3H-1* is expressed in all tissues investigated, an observation that is supported by ESTs from various tissues (Table 5). This ubiquitous expression is in accordance with previous studies that detected the highest expression in the vascular tissues of stem and root, the expression in root being only moderate in our RT-PCR analysis (Table 5; Schoch et al., 2001; Franke et al., 2002b; Nair et al., 2002). On the contrary, *C3H-2* and *C3H-3* are expressed only during particular stages of inflorescence stem development: *C3H-2* is expressed in older stems and *C3H-3* in young developing stems (Table 5). The fact that only one EST is found for *C3H-2* and none for *C3H-3* suggests that they are either conditionally regulated or expressed at low levels.

Table 5. Expression characteristics of the *C3H* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																					regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																										
Arath;C3H-1	CYP98A3 *REF8* [b]	At2g40890	ER-anchored	RT-PCR	+				+/-	+	+/-	+		+	+	+	+	+	+						AC	
				EST (36)			5	3	5			7										2		14		
				EST rel (22.4)			14	17	25			28										283		32		
				mRNA[a]					+	+/-	+/-	+	++									+				
				mRNA[b]	+				++	+	+	+	++													
				mRNA[c]					++	+	+/-	++	+													
				AtC3H::GUS [c]	+				++[c1]	+[c2]	+[c2]	+[c3]	+[c2]									+				
				C3H[a]					+[a1]				+[a2]													
				C3H[c]					+[c4]				+[c5]													
Class II																										
Arath;C3H-2 *	CYP98A8	At1g74540		RT-PCR					+		+					+/-	+/-	+/-	+/-						A	
				EST (1)			1																			
				EST rel (0.6)			3																			
Arath;C3H-3 *	CYP98A9	At1g74550		RT-PCR	+				+	+	+			+	+											
				EST (0)																						
				EST rel (0)																						

[a] Schoch *et al*., 2001
[a1] immunolocalization using a polyclonal anti-CYP98A3 antibody; mainly in differentiating xylem, also in secondary phloem in the cortical zone of mature root
[a2] immunolocalization using a polyclonal anti-CYP98A3 antibody; very strong in differentiating xylem
[b] Franke *et al*., 2002a
[c] Nair *et al*., 2002
[c1] expression highest in roots, expressed in stele and endodermis; not expressed in root apical meristem, epidermis and cortex
[c2] in the vascular tissue of stem, petiole, leaf, petal, sepal, anther, stigma
[c3] in the vascular tissue of the silique wall, not in seed
[c4] immunolocalization using a polyclonal anti-CYP98A3 antibody; in stele
[c5] immunolocalization using a polyclonal anti-CYP98A3 antibody; in meta- and protoxylem cells in the young stem, strongest in lignified interfascicular fibers and xylem vessels of older stem
* *C3H-2* and *C3H-3* are single exon genes

The promoter analysis reveals a well-conserved AC element in the promoter of *C3H-1*, in agreement with its vascular expression detected by the GUS reporter system. An A box is found in the promoter region of *C3H-2*.

Analysis of the N-terminus by TargetP predicts the C3H-1 protein to contain an ER targeting peptide, but it overlaps, as for C4H, with the membrane-anchor region of P450 enzymes. The C3H-1 protein has previously been localized in the membrane fraction in yeast (Franke et al., 2002b). In contrast to C3H-1, the sequences of C3H-2 and C3H-3 is divergent in both the stretch of basic amino acids and the hinge region. Because these regions are necessary for the correct insertion of the enzyme in the membrane (Chapple, 1998), the degeneration of this region suggests they are not membrane-anchored proteins.

In conclusion, C3H-1 is involved in the monolignol pathway, as is functionally demonstrated with the reduced epidermal fluorescence (*ref8*) mutant (Franke et al., 2002a, 2002b). Not much is known about the developmental and the stress- or elicitor response of the other two *C3H* genes. C3H-2 and C3H-3 do not hydroxylate shikimate and quinate esters of *p*-coumaric acid, but their activity toward other substrates remains to be investigated (Schoch et al., 2001).

Caffeoyl-CoA 3-O-methyltransferase (CCoAOMT)

Caffeoyl-CoA 3-*O*-methyltransferase (CCoAOMT, EC 2.1.1.104) catalyzes the methylation of caffeoyl-CoA to feruloyl-CoA (in vitro and in vivo) and 5-hydroxferuloyl-CoA to sinapoyl-CoA (at least in vitro) and is, together with COMT, responsible for the methylation of the monolignol precursors (Figure 1; Ye et al., 1994; Zhong et al., 1998; Pinçon et al., 2001).
Seven putative members of the *CCoAOMT* gene family were detected in the *Arabidopsis* genome (Figure 7). Plant *CCoAOMT* genes fall into two classes: class I contains the *Arabidopsis CCoAOMT-1* gene together with the majority of experimentally characterized *CCoAOMT* genes (e.g. Zhong et al., 1998; Meyermans et al., 2000), whereas class II consists of six *Arabidopsis* genes and a few sequences from other species. The latter class does not closely resemble most of the certified *CCoAOMT* genes, but contains an experimentally characterized chickweed (*Stellaria longipes*) *CCoAOMT* able to methylate caffeoyl-CoA (Zhang and Chinnappa, 1997).

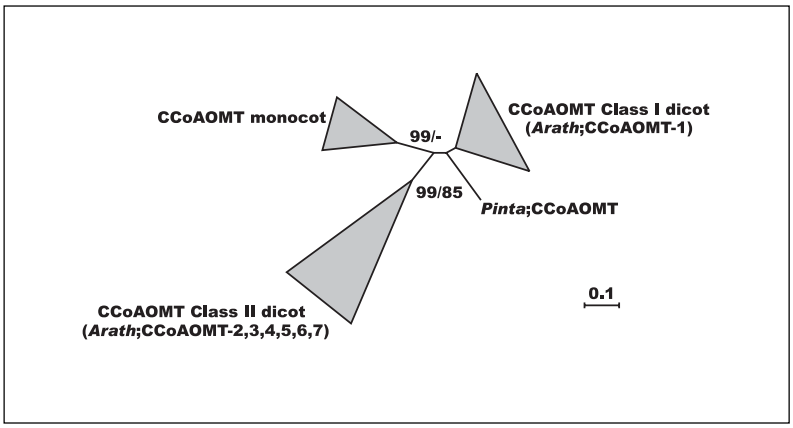

Figure 7. Neighbor-joining tree of the *CCoAOMT* family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per nucleic acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: Class I dicots: *Populus* (2960355, 857577, 13249170, 2960357), *Zinnia* (533120), *Petroselinum* (169648), *Nicotiana* (2511736), *Citrus* (6561880), *Vitis* (1000518), *Eucalyptus* (5739372, 1934858); gymnosperm: *Pinus CCoAOMT* (4104458); Class II dicots: *Stellaria* (438896), *Populus* (1785476); and monocots: *Zea* (5101869, 5101867), *Oryza* (5091496, 5257255 [three genes]). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Pinta*, *Pinus taeda*.

CCoAOMT-1 and *CCoAOMT-7* are found on chromosome 4 and *CCoAOMT-3* and *CCoAOMT-4* in tandem on chromosome 3 (Figure 12). Upstream of this tandem, another gene with sequence similarity to *CCoAOMT-3* has been found that is heavily truncated at the 3' end. With only a small region of 35 amino acids conserved with *CCoAOMT-3*, but not with any other *CCoAOMT*, it is probably a remnant of a duplication event. Curiously, this gene is expressed (data not shown).
CCoAOMT-2, *CCoAOMT-5*, and *CCoAOMT-6* are situated in a large internal block duplication on chromosome 1 that originated during the complete genome duplication 75 million years ago (Figure 12).

CCoAOMT-5 and *CCoAOMT-6* resemble each other more than either resembles *CCoAOMT-2* (data not shown). Therefore, the most probable evolutionary scenario is that an ancestor of these three genes had duplicated, yielding *CCoAOMT-2* and an ancestor of *CCoAOMT-5* and *CCoAOMT-6* during the genome duplication and later on, *CCoAOMT-5* and *CCoAOMT-6* originated through tandem duplication.

CCoAOMT-1 is expressed in all tissues investigated and has by far the highest number of ESTs (Table 6). Moreover, the *CCoAOMT-1* gene has two AC elements in its promoter. *CCoAOMT-1* is highly expressed in the basal portion of the inflorescence as compared to the apical portion (Goujon et al., 2003b). Of the class II genes, *CCoAOMT-5* and *CCoAOMT-7* are expressed in all tissues, but only the expression of *CCoAOMT-7* increases during the later stages of inflorescence stem development. Furthermore, *CCoAOMT-4* and *CCoAOMT-5* are also expressed at all stages of inflorescence stem development. Others, such as *CCoAOMT-2*, *CCoAOMT-3* and *CCoAOMT-6* are expressed later in this process (Table 6). Few ESTs have been found for most genes of class II (Table 6).

Table 6. Expression characteristics of the *CCoAOMT* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																				regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																									
Arath;CCoAOMT-1	*CCoAOMT* [c]	At4g34050		RT-PCR	++				++	++	+	++		+	++	++	++	++	++	++					AC (2)	
				EST(45)	1		4		19		1	9	1										1	9	H	
				EST rel (28.0)	31		11		93		15	72	36										42	21		
				mRNA[c]	+/-					+/-			+[c1]													
Class II																									
Arath;CCoAOMT-2		At1g24735		RT-PCR	+				+	+		+		+/-		+	+	+	+						H	
				EST(0)																						
				EST rel (0)																						
Arath;CCoAOMT-3		At3g61990	ER	RT-PCR	+				+	+							+	+	+						S	
				EST(6)			3																	3		
				EST rel (3.7)			9																	7		
Arath;CCoAOMT-4		At3g62000		RT-PCR	+					+				+/-	+/-	+	+	+	+							
				EST(2)			1																	1		
				EST rel (1.2)			3																	2		
Arath;CCoAOMT-5		At1g67990		RT-PCR	+/-				+	+	+	+		+	+	+	+	+	+							
				EST(1)							1															
				EST rel (0.6)							15															
Arath;CCoAOMT-6	*CCoAOMT* [a]	At1g67980		RT-PCR						+		+					+/-	+/-	+							
				EST(2)			2																			
				EST rel (1.2)			6																			
				RT-PCR [b]	+[b1]																+[b1]					
Arath;CCoAOMT-7		At4g26220		RT-PCR	++				+	++	+/-	+		+	+	+	++	++	++							
				EST(4)								3												1		
				EST rel (2.5)								12												2		

[a] Zou and Taylor, 1994
[b] Jin *et al.*, 2000
[b1] in seedling leaves
[c] Goujon *et al.*, 2003a
[c1] highly expressed in the basal portion as compared to the apical portion of the inflorescence stem

None of the *Arabidopsis CCoAOMT* genes has yet been characterized for pathogen or elicitor induction. However, *CCoAOMT* genes of other species were shown to be responsive to these treatments (e.g. Pakusch et al., 1991; Chen et al., 2000). Promoter elements involved in stress-induced expression were identified in *CCoAOMT-1* and *CCoAOMT-2* (H box), and in *CCoAOMT-3* (S box) (Table 6). CCoAOMT-3, with an extended N-terminal sequence shared by no other CCoAOMT, is predicted to contain an ER targeting peptide, indicating that this protein is secreted or functional at or in the ER membrane.
Based on its clustering in class I, its expression characteristics and level, and the presence of two AC elements in its promoter, *CCoAOMT-1* is the main candidate gene to be involved in the monolignol pathway during developmental lignification.

Cinnamoyl-CoA reductase (CCR)

Cinnamoyl-CoA reductase (CCR; E.C.1.2.1.44) catalyzes the conversion of cinnamoyl-CoA esters to their respective cinnamaldehydes and is the first enzyme of the monolignol-specific part of the lignin biosynthetic pathway (Figure 1). The two previously described *CCR* genes and five new *CCR*-like genes were found (Figure 8; Lauvergeat et al., 2001; Jones et al., 2001). The latter do not cluster with any other gene family in *Arabidopsis*, but there are no indications that they are genuine *CCR* genes.

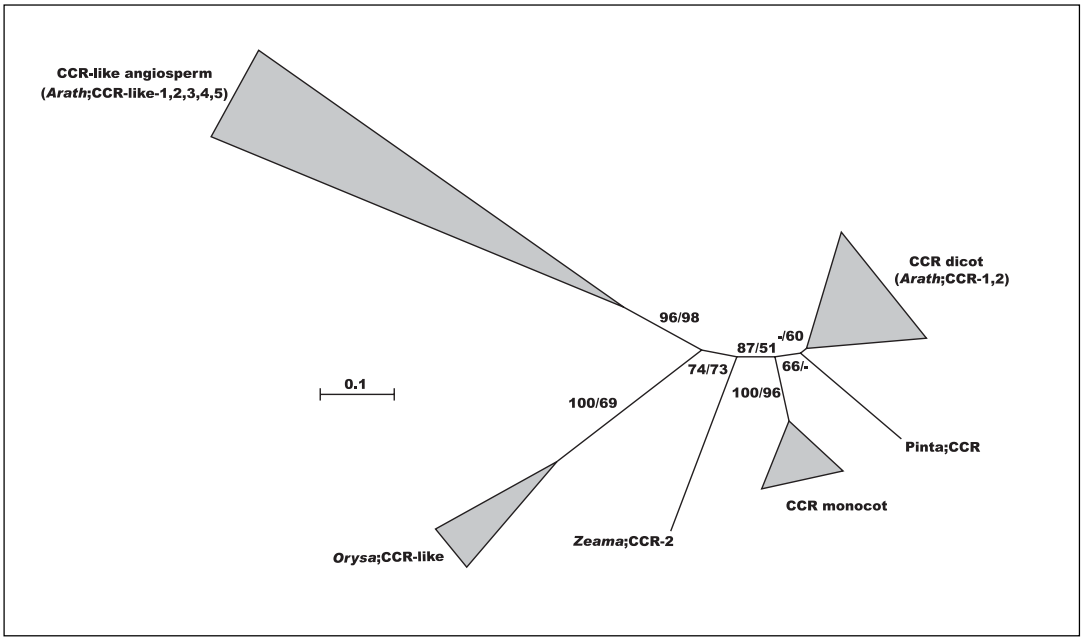

Figure 8. Neighbor-joining tree of the CCR family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: CCR dicots: *Eucalyptus* (7431407, 7431408, 10304406), *Populus* (7239228, 2960364, 9998901); CCR monocots: *Lolium* (9964087), *Saccharum* (3341511, 17978549), *Zea* (7431410, 3242328); gymnosperm: *Pinus* CCR (17978649); *Zea* CCR-2 (3668115); *Oryza* CCR-like (13486725, 13486726, 18307514). CCR-like angiosperm: *Oryza* (15624051). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Orysa*, *Oryza sativa*; *Pinta*, *Pinus taeda*; *Zeama, Zea mays*.

Members of another closely related family, the VR-ERE-like aldehyde reductases that have a high affinity for 3-substituted benzaldehydes (Guillén et al., 1998) were not withheld as putative *CCR*-like genes. *CCR-1* and *CCR-2* are both located on chromosome 1 in a duplicated block that arose through the complete genome duplication 75 million years ago (Figure 12).

CCR-1 is highly expressed in all tissues examined, whereas *CCR-2* in all tissues but flowers and the earliest stage of inflorescence development (Table 7). *CCR-1* has previously been found to be strongly expressed in the stem (Lauvergeat et al., 2001; Goujon et al., 2003a). Although *CCR-2* was hardly detected in stem by RNA gel blots (Lauvergeat et al. 2001), the more sensitive RT-PCR clearly detects *CCR-2* expression in the inflorescence stem (Table 7). In contrast to *CCR-1*, *CCR-2* expression increases with age during inflorescence stem development and may, thus, correlate with more lignified tissues (Table 7). Corresponding with the differences in expression levels of *CCR-1* and *CCR-2*, almost 10 fold more ESTs are found for *CCR-1* than for *CCR-2* (Table 7).

Table 7. Expression characteristics of the *CCR* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																			regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1 cm	3 cm	5 cm	10 cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																									
Arath;CCR-1	*CCR1* [ab]	At1g15950		RT-PCR	++				++	++	+	++		++	++	++	++	++	++						AC	
	IRX4 [a]			EST (43)			6	8	8		2	8	1										2	8		
				EST rel (26.7)			17	45	39		29	72	32										85	18		
				mRNA[b]						+	++		++										+[b1]			
				mRNA[c]						+			+[c1]													
Arath;CCR-2	*CCR2* [ab]	At1g80820		RT-PCR	+				+	+		+			+/-	+/-	+	+	+							
				EST (4)					3														1			
				EST rel (2.5)					15														42			
				mRNA[b]																			+[b2]	b3		
Class CCR-likes																									
Arath;CCR-like-1		At1g76470																							
Arath;CCR-like-2		At2g02400																							
Arath;CCR-like-3		At2g33590																							
Arath;CCR-like-4		At2g33600																							
Arath;CCR-like-5		At5g58490																							

[a] Jones *et al.*, 2001
[b] Lauvergeat *et al.*, 2001
[b1] *Xanthomonas* infection
[b2] induced by *Xanthomonas* infection and salicylic acid
[b3] not induced by methyl jasmonate or ethylene
[c] Goujon *et al.*, 2003b
[c1] moderately expressed in the basal part of the inflorescence stem, highly expressed in the apical part of the inflorescence stem

Both genes are induced by *Xanthomonas* infection and ESTs linked with stress and pathogen infection have been detected (Lauvergeat et al., 2001; Table 7). The promoter of *CCR-1* contains a well-conserved AC element, conform with its function in lignification and the strong stem expression (Lauvergeat et al., 2001; Table 7).

In conclusion, *CCR-1* and *CCR-2* are expressed during both developmental lignification and pathogen response, as documented by our expression analysis and ESTs (Table 7). CCR-1 may be preferentially correlated with developmental lignification and CCR-2 with pathogen response. The role of CCR-1 in lignification has clearly been established through the irregular xylem (*irx4*) mutant characterization (Jones et al., 2001).

Both CCR-1 and CCR-2 use feruloyl-CoA and sinapoyl-CoA, but CCR-1 is 5 fold more efficient than CCR-2 (Lauvergeat et al., 2001). Although CCR-2 seems to be implicated in stress and elicitor response, the expression results do not exclude a (minor) role for CCR-2 in developmental lignification. It must be noted, however, that CCR-2 does not complement the *irx4* mutant (Jones et al., 2001).

Ferulate 5-hydroxylase (F5H)

Ferulate 5-hydroxylase (F5H), also called coniferaldehyde 5-hydroxylase (Cald5H), is a cytochrome P450-dependent monooxygenase (CYP84) that is required for the production of syringyl lignin. Originally, it had been thought to convert ferulic acid to 5-hydroxyferulic acid, in a syringyl-specific branch of the monolignol pathway. However, because the enzyme has a 1,000-fold greater affinity to coniferaldehyde and coniferyl alcohol than to ferulic acid, it was assigned to be responsible for the 5-hydroxylation of coniferaldehyde and/or coniferyl alcohol (Figure 1; Humphreys et al., 1999; Li et al., 2000; Humphreys and Chapple, 2002). Thus, F5H introduces the final hydroxyl group at the C5 of the aromatic ring necessary to generate the methoxy group typical of syringyl monomers. The *Arabidopsis* genome harbors two *F5H* homologs, both belonging to the *CYP84* family of the P450 monooxygenases. *F5H-1* (*CYP84A1*) has been characterized in *Arabidopsis*, as well as its homologs of Liquidambar and Brassica (Meyer et al., 1996; Osakabe et al., 1999; Nair et al., 2000), whereas *F5H-2* (*CYP84A4*), a more divergent member of the *CYP84* family, is described for the first time in this study. So far, no genes that closely resemble *F5H-2* have been detected in other plants, although the phylogeny indicates that the two proteins found in *Arabidopsis* diverged before the divergence of the different Rosidae subfamilies (Figure 9).

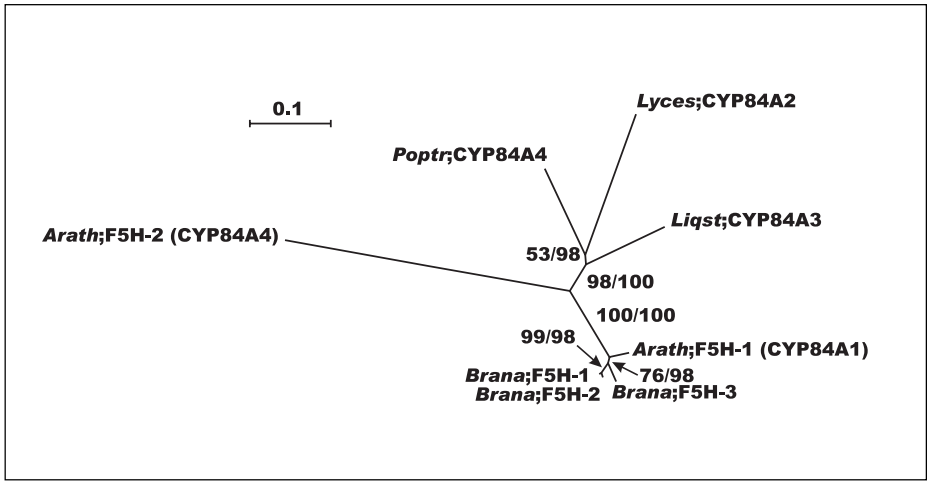

Figure 9. Neighbor-joining tree of the F5H family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: *Populus* CYP84A4 (6688937), *Lycopersicon* CYP84A2 (5002354), *Liquidambar* CYP84A3 (5731998), and *Brassica* F5H-1, F5H-2, and F5H-3 (10197650, 10197652, 10197654). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Brana*, *Brassica napus*; *Liqst*, *Liquidambar styraciflua*; *Lyces*, *Lycopersicon esculentum*; *Poptr*, *Populus trichocarpa*.

F5H-1 resides on chromosome 4, whereas *F5H-2* is located on chromosome 5, within the borders of the duplicated block that was linked to the expansion of the PAL family (Figure 12). However, we did not detect a copy of *F5H-2* on chromosome 3, indicating that this hypothetical copy of *F5H-2* was either lost or that *F5H-2* arose on its current position after the genome duplication event.
Our expression analysis revealed *F5H-1* expression in all tissues and an increasing expression during inflorescence development (Table 8), in accordance with results of earlier studies (Meyer et al., 1998; Ruegger et al., 1999, Goujon et al., 2003b). In 12 independent samples of inflorescence stems, *F5H-1* was expressed from the 9th sample from the top on, in contrast with *C4H*, that was already present from the 6th internode (Meyer et al., 1998). This correlation of *F5H-1* expression with later development is linked with the increase in S monomer production with increasing age of the inflorescence stem (Meyer et al., 1998). In contrast to *F5H-1*, *F5H-2* had the strongest expression in the early stages of inflorescence development (Table 8).

Table 8. Expression characteristics of the *F5H* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																				regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Arath;F5H-1	CYP84A1 *FAH* [a]	At4g36220	ER-anchored	RT-PCR	++				+	++	+	+		+	+	+	++	++	++						H	
				EST (2)*			1																			
				EST rel (1.2)			3																			
				mRNA[b]									+													
				mRNA[c]	+	+			++	+	+	+	++													
				mRNA[e]					+/-	+	+/-	++	++													
				mRNA[f]	+/-					+			+[f1]													
				RT-PCR[d]	+[d1]																					
Arath;F5H-2	CYP84A4	At5g04330	ER-anchored	RT-PCR	+				+/-		+/-			+	+	+	+	+/-	+/-						G+H	
				EST (0)																						
				EST rel (0)																						

[a] Chapple *et al.*, 1992
[b] Meyer *et al.*, 1998
[c] Ruegger *et al.*, 1999
[d] Jin *et al.*, 2000
[d1] in seedling leaves
[e] Nair *et al.*, 2002
[f] Goujon *et al.*, 2003a
[f1] highly expressed in the basal portion as compared to the apical portion of the inflorescence stem
[*] 1 EST is unclassified

In addition, *F5H-1* was expressed also in several other tissues, but mainly in young and senescent leaves and in roots (Meyer et al. 1996; Ruegger et al., 1999). Only two ESTs were found for *F5H-1* and none for *F5H-2* (Table 8).
In the promoter analysis, for both genes an H box was found and for *F5H-2* also a G box, suggesting that both genes may be inducible and that *F5H-2* may be regulated by *p*-coumarate (Loake et al., 1992; Lindsay et al., 2002). Moreover, F5H-1and F5H-2 contain a fully conserved membrane-anchor region. Additionally, F5H-2 is predicted to contain an ER targeting peptidethat, however, coincides with the region of the membrane anchor of P450 enzymes.
Remarkably, no AC element was detected for either *F5H* gene, although *F5H-1* had been shown to be involved in lignification through the analysis of the *fah1* mutant (Chapple et al., 1992).

Caffeic acid O-methyltransferase (COMT)

Caffeic acid *O*-methyltransferase (COMT; E.C. 2.1.1.68) was originally postulated to be a bifunctional enzyme methylating caffeic acid and 5-hydroxyferulic acid. However, in vitro studies revealed a much higher affinity of COMT for caffeyl aldehyde, 5-hydroxy coniferaldehyde, and 5-hydroxyconiferyl alcohol, which led to its alternative name 5-hydroxyconiferaldehyde *O*-methyltransferase (AldOMT; Osakabe et al., 1999; Li et al., 2000, Chen et al., 2001; Guo et al., 2001; Parvathi et al., 2001). These observations and the marked reduction of syringyl lignin in COMT-downregulated transgenic plants led to the new position of COMT in the pathway. Thus, the predominant role of COMT is the methylation of 5-hydroxy coniferaldehyde and/or 5-hydroxyconiferyl alcohol to sinapaldehyde and/or sinapyl alcohol, respectively (Figure 1).

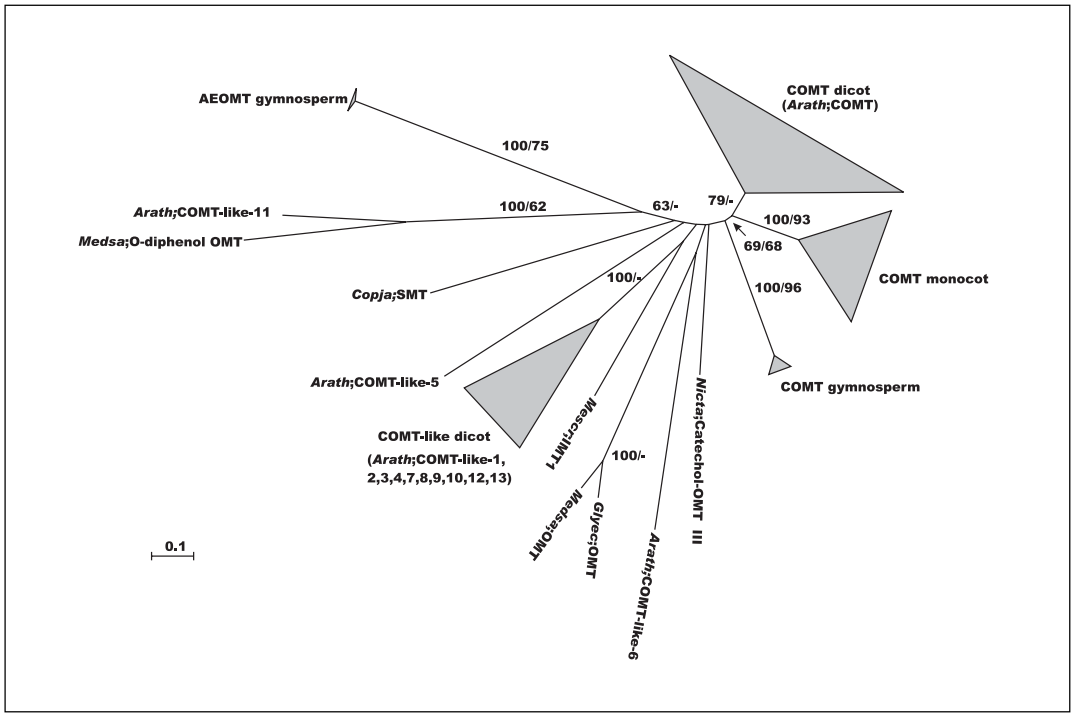

Figure 10. Neighbor-joining tree of the COMT family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: COMT dicots: *Populus* (7528266, 762870, 231757, 444327, 7332271, 7447887, 762872), *Stylosanthes* (1582580), *Medicago* (116908), *Prunus* (3913295), *Fragaria* (6760443), *Liquidambar* (5732000), *Chrysosplenium* (1184041, 567077), *Vitis* (7271883), *Capsicum* (3421382, 7488967, 12003964), *Nicotiana* (480082, 480083), *Eucalyptus* (1169009, 5739365), *Clarkia* (2832224, 3913289), *Mesembryanthemum* (7447880), *Thalictrum* (4808522, 4808524, 4808526, 4808528, 4808530), *Catharanthus* (18025321), *Ocimum* (5031492, 5031494), *Zinnia* (642952); COMT monocots: *Lolium* (4104220, 4104222, 4104224, 2388664), *Sorghum* (18033964), *Saccharum* (3341509), *Zea* (729135), *Festuca* (14578611, 14578613, 14578615, 14578617); COMT gymnosperms: *Pinus* (15524083), *Picea* (COMT-C7, COMT-C16; Michael H. Walter, personal communication); *Nicotiana* Catechol-OMT III (542050); *Glycyrrhiza* OMT (1669591), *Medicago* OMT (7447884), *Mesembryanthemum* IMT1 (1170555), *Coptis* SMT (758580), *Medicago O*-diphenol OMT (6688808); AEOMT gymnosperm: *Pinus* (7447883, 1777386, 4574324). Abbreviations: *Arath*, *Arabidopsis thaliana*; *Copja*, *Coptis japonica*; *Glyec*, *Glycyrrhiza echinata*; *Medsa*, *Medicago sativa*; *Mescr*, *Mesembryanthemum crystallinum*; *Nicta*, *Nicotiana tabacum*.

We detected only one *COMT* gene in the *Arabidopsis* genome. COMT was first described in *Arabidopsis* as interacting with a 14-3-3 protein (Zhang et al., 1997). Furthermore, 13 proteins similar to COMT were detected that clustered in-between the functionally characterized COMT clade and the cluster containing the hydroxycinnamic acid/hydroxycinnamoyl-CoA ester *O*-methyltransferase protein (AEOMT; Li et al., 1997, 1999), i.e., among proteins that have been shown to use a wide variety of substrates (Maxwell et al., 1993; Pellegrini et al., 1993; Takeshita et al., 1995; Vernon and Bohnert, 1992). Because the role of AEOMT in the monolignol pathway is still a matter of debate (Anterola et al., 2002) and other *COMT* candidate genes of conifers clustered much more closely to the known COMTs, it is unclear whether these 13 genes play any role in the monolignol pathway. Therefore, these genes were classified as *COMT*-likes. By consequence, only one class of COMTs exists in plants (Figure 10; Maury et al., 1999).

The *COMT* gene is located on chromosome 5, whereas *COMT*-like genes are found on chromosomes 1, 3, and 5. Interestingly, 10 out of 13 *COMT*-like genes are present on chromosome 1, originating from a combination of genome duplication and multiple tandem duplications (Figure 12). Our RT-PCR data show that *COMT* is expressed in all tissues investigated and the numerous ESTs point toward a generally high and ubiquitous expression (Table 9).

Table 9. Expression characteristics of the *COMT* gene in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																				regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Arath;COMT	*OMT1* [a]	At5g54160	myristoylation	RT-PCR	++				+	++	+	+		+	++	++	++	++	++							SARE*
				EST (99)			11	16	28		2	22										2	18			
				EST rel (61.6)			31	89	138		29	88										85	41			
				mRNA[a]	+				+		+		++													
				mRNA[b]	+	+																				
				mRNA[d]						+/-			+[d1]													
				RT-PCR[c]	+[c1]																					
				AtCOMT1::GUS [d]	+[d2]					+[d3]	+[d4]	+[d5]	+[d6]													
Class COMT-likes																									
Arath;COMT-like-1		At1g21100																							
Arath;COMT-like-2		At1g21110																							
Arath;COMT-like-3		At1g21120																							
Arath;COMT-like-4		At1g21130																							
Arath;COMT-like-5		At1g33030																							
Arath;COMT-like-6		At1g51990																							
Arath;COMT-like-7		At1g63140																							
Arath;COMT-like-8		At1g76790																							
Arath;COMT-like-9		At1g77520																							
Arath;COMT-like-10		At1g77530																							
Arath;COMT-like-11		At3g53140																							
Arath;COMT-like-12		At5g37170																							
Arath;COMT-like-13		At5g53810																							

[a] Zhang *et al.*, 1997
[b] Ruegger *et al.*, 1999
[c] Jin *et al.*, 2000
[c1] in seedling leaves
[d] Goujon *et al.*, 2003a
[d1] moderately expressed in the basal part of the inflorescence stem, highly expressed in the apical part of the inflorescence stem
[d2] constitutive in 3d-old-seedlings, very high in vascular tissues in 12d-old tissues
[d3] basal GUS activity in leaf blade of young leaves, in vascular tissues of mature leaves
[d4] only in the sepal veins
[d5] only in the lignified ends of silique
[d6] very high in xylem, differentiating fibers and mature phloem
[*] in the first intron

Ninety-nine *COMT* ESTs, with a fifth being stress related, is almost twice the number found for any other gene in this analysis. *COMT* expression is particularly high in the inflorescence stem, confirming previous expression analyses (Table 9; Zhang et al., 1997; Goujon et al., 2003b). Correspondingly, *COMT::GUS* expression occurs in xylem, differentiating fibers, and mature phloem (Goujon et al., 2003b).
Unlike many other monolignol biosynthesis genes, *COMT* has no AC elements in its promoter. In fact, to the best of our knowledge, AC elements have never been reported in *COMT* promoters of other plants either. In a search of other available *COMT* promoters (tobacco [*Nicotiana tabacum*; AX037003] and lotus [*Lotus japonicus*; AP004939]), no AC elements were found.
Interestingly, the COMT protein might be myristoylated. The N-terminal MGSTAETQLTPVQVTDDE sequence was identified as a "twilight zone" myristoylation signal, which corresponds both with truly myristoylated proteins as well as with false positives (Maurer-Stroh et al., 2002). Myristoylation is generally associated with cell membrane anchoring or, as recently shown for an *Arabidopsis* protein kinase, ER attachment (Lu and Hrabak, 2002). Pending the experimental verification of this observation, the putative localization of the COMT protein indicates a new research avenue in the field of monolignol channeling and export.

Cinnamyl alcohol dehydrogenase (CAD)

Cinnamyl alcohol dehydrogenase (CAD, EC 1.1.1.195) catalyzes the last step in the monolignol biosynthesis, i.e., the reduction of cinnamyl aldehydes into their corresponding alcohols (Figure 1). CAD reduces various aldehydes, present in different cell types or during different stages of development. Besides the function in developmentally regulated lignification, a number of *CAD* genes has been characterized for their response to plant pathogens (Kiedrowski et al., 1992; Galliano et al., 1993).
Here, nine putative *CAD* genes were detected in the *Arabidopsis* genome, of which eight have been described before (Table 10; Tavares et al., 2000). Our phylogenetic analysis reveals that eight of the CAD proteins fall into three classes, whereas CAD-9 is more divergent (Figure 11).
CAD-2 and CAD-6, belonging to the class I CADs, closely resemble CAD proteins that have been characterized for their involvement in lignification in other species. Although these genes are the most likely candidates for "true" *CAD* orthologs in *Arabidopsis*, they have not yet been studied. The topology of the tree indicates furthermore that the class I "true" CAD clade diverged from the other CADs before the angiosperm-gymnosperm split (Figure 11).
Class II CADs (CAD-3, CAD-4, and CAD-5) cluster with a number of alcohol dehydrogenases with diverse substrate preferences, such as the poplar sinapyl alcohol dehydrogenase (SAD; Li et al., 2001), the celery (*Apium graveolens*) mannitol dehydrogenase (MTD; Williamson et al., 1995), and the parsley ELI3/CAD proteins (Kiedrowski et al., 1992; Logemann et al., 1997). *CAD-4* (*AtELI3-1*) and *CAD-5* (*AtELI3-2*) have previously been identified as responsive to elicitor treatments and *Pseudomonas* infection (Kiedrowski et al., 1992). Moreover, CAD-5 has a substrate specificity distinct from "true" CADs, MTD, and aromatic alcohol:NADP+ oxidoreductase and was, therefore, named benzyl alcohol dehydrogenase (BAD; Somssich et al., 1996).

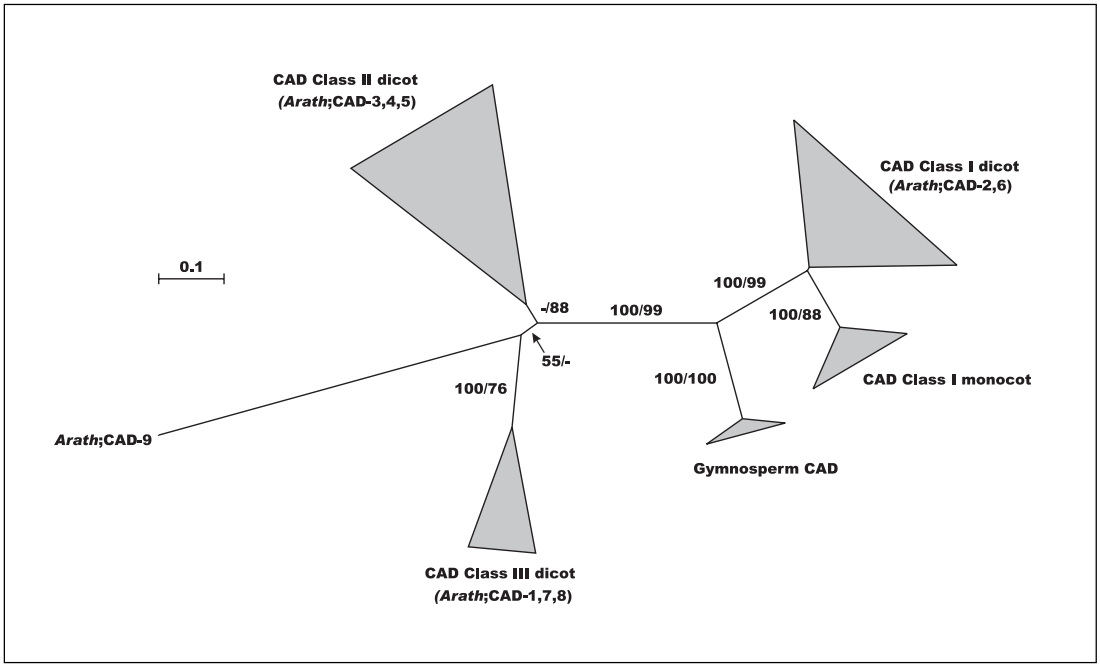

Figure 11. Neighbor-joining tree of the CAD family, inferred from Kimura corrected evolutionary distances. Bootstrap values (NJ/ML) above 50% are shown at the internodes. The scale measures evolutionary distance in substitutions per amino acid. Clusters of sequences are represented as described in Figure 2. Species and GenBank Identifier numbers of non-*Arabidopsis* sequences included in this tree are: Class I dicots: *Populus* (421814, 1168734, 9998899, 7239226), *Nicotiana* (231676, 231675), *Medicago* (399168), *Aralia* (1168727), *Zinnia* (1944403), *Eucalyptus* (1705554, 10281656, 399165, 10719920, 3913185). Class I monocots: *Saccharum* (10719916), *Zea* (3913182, 7430938), *Lolium* (3913181), *Festuca* (15428276, 15428278, 15428280, 15428282). Gymnosperm CAD: *Picea* (584872, 10719915), *Pinus* (107623, 3334135, 1168733, 3372645); Class II dicots: *Stylosanthes* (3913194), *Apium* (12643507), *Petroselinum* (1168732), *Lycopersicon* (8099340, 7430935), *Mesembryanthemum* (10720090), *Fragaria* (10720093, 13507210), *Populus* (14279694); Class III dicots: *Stylosanthus* (3913193), *Medicago* (10720088). Abbreviations: *Arath*, *Arabidopsis thaliana*.

CAD-5 nevertheless shares a striking sequence similarity with MTD (Williamson et al., 1995). Class III CADs (CAD-1, CAD-7, and CAD-8) cluster in a group with an alcohol dehydrogenase from alfalfa (*Medicago sativum*), which is able to catalyze the reduction of cinnamaldehyde, sinapaldehyde, and coniferaldehyde, but also several aliphatic aldehydes and various substituted benzaldehydes (Brill et al., 1999). Being very divergent from class I "true" CADs, this class also represents a group of multisubstrate alcohol dehydrogenases. One last protein, CAD-9, does not cluster with any known protein or with one of the classes mentioned above.

CAD-1, *CAD-6*, *CAD-3*, *CAD-4*, and *CAD-5* are located on chromosome 4, the latter three in tandem. *CAD-2* is situated on chromosome 3, *CAD-7* and *CAD-8* on chromosome 2, and *CAD-9* on chromosome 1 (Figure 12). The three genes of class III (*CAD-1*, *CAD-7*, and *CAD-8*) originated during the complete genome duplication, leading to *CAD-1* and an ancestor of *CAD-7* and *CAD-8*, followed by a duplication of the latter (Figure 12).

All *CADs* of classes I and III, and *CAD-9* are expressed in all stages of inflorescence stem development (Table 10). In accordance, *CAD-6* was shown to be highly expressed in the basal portion when compared with the apical portion of the inflorescence stem (Goujon et al., 2003b).

Table 10. Expression characteristics of the *CAD* gene family in *Arabidopsis*.
See Table 1 for the full explanation of table and abbreviations.

Gene	other names	AGI number	signals for localization	expression																				regulatory elements	
				method	seedling	etiolated seedling	whole plant	aboveground organs	roots	leaves	flower	silique, seed	inflorescence stem	1cm	3cm	5cm	10cm	15 cm	fully grown	light	sucrose	wounded leaves	pathogen infection	various stresses	upstream region	introns
Class I																									
Arath; CAD-2	*LCAD-C* [f]	At3g19450		RT-PCR	++				+	++		+		+	+	+	++	++	++						
				EST (33)			6	4	11		2	2												8	
				EST rel (20.5)			17	22	54		29	8												18	
Arath; CAD-6	*LCAD-D* [f]	At4g34230		RT-PCR	+				+	++	+/-	+		+	+	+	++	++	++						AC
				EST (23)	1		6		5			5										1	5		A
				EST rel (14.3)	31		17		25			20										42	11		
				mRNA [g]						+/-			+[g1]												
Class II																									
Arath; CAD-3	*L-CAD-A* [f]	At4g37970		RT-PCR	++				+	++				+	+	+	++	++	++						
				EST (1)																			1		
				EST rel (0.6)																			2		
				RT-PCR [f]	+				+	+	+[f1]		+												
Arath; CAD-4	*LCAD-B* [f]	At4g37980		RT-PCR	+				+	+	+	+		+		+	+	+	+						
	ELI3-1 [a,f2]			EST (26)	1		11	5			1												8		
				EST rel (16.2)	31		31	28			15												18		
				mRNA[ab]																		+[a1]			
Arath; CAD-5	*ELI3-2* [a]	At4g37990		RT-PCR	+/-					+	+/-	+					+	+	+/-						AC
	BAD [d]			EST (2)			1																1		
				EST rel (1.2)			3																2		
				mRNA[a]																		+[a1]			
Class III																									
Arath; CAD-1	*CAD1* [c]	At4g39330		RT-PCR	++				+	+		+		+	+	+	+	+	+						
				EST (32)	1		6	4			1	10											10		
				EST rel (19.9)	31		17	22			15	40											23		
				RT-PCR [e]	+[e1]																				
Arath; CAD-7	*LCAD-E* [f]	At2g21730		RT-PCR*	+				+/-	+	+/-			+	+	+	+	+	++						
				EST (0)																					
				EST rel (0)																					
Arath; CAD-8	*LCAD-F* [f]	At2g21890		RT-PCR*	+				+/-	+	+/-			+	+	+	+	+	++						
				EST (0)																					
				EST rel (0)																					
Arath; CAD-9		At1g72680		RT-PCR	++					++	+/-	+		+	+	+	+	+	+						E
				EST (9)			1					5											3		
				EST rel (5.6)			3					20											7		

[a] Kiedrowski *et al.*, 1993
[a1] *Pseudomonas* infection
[b] Leyva *et al.*, 1995
[c] Somers *et al.*, 1995
[d] Somssich *et al.*, 1996
[e] Jin *et al.*, 2000
[e1] in seedling leaves
[f] Tavares *et al.*, 2000
[f1] not in pollen
[f2] *ELI3-1* (X67816) is a recombinant clone of *ELI3-2* and *LCAD-B*
[g] Goujon *et al.*, 2003a
[g1] highly expressed in the basal portion as compared to the apical portion of the inflorescence stem
[.] coding sequences and 3'UTR cannot be distinguished by RT-PCR

Expression of most *CAD* genes is documented by ESTs, except for *CAD-7* and *CAD-8* (Table 10), which are nevertheless expressed, as indicated in the RT-PCRs (Table 10). It should be noted that *CAD-7* and *CAD-8* arose during a recent duplication event (described in detail by Tavares et al., 2000) and could not be distinguished in the RT-PCR analysis because of their high sequence similarity: 98%, 95%, and 94% identity in the coding regions, introns, and putative 3' untranslated regions, respectively.

The promoter analysis revealed that *CAD-6* from class I and *CAD-5* from class II contain AC elements (Table 10). Additionally, an A box was detected in the *CAD-6* promoter. The fact that only one gene in the pathway contains both an AC element and an A box casts doubt on the previous assumption that an A box works in conjunction with AC elements (Logemann et al., 1995). Furthermore, only one elicitation-related element (E box) was identified in *CAD-9* (Table 10).
Based on the fact that they cluster with other well-characterized "true" *CAD* genes in the phylogenetic tree, *CAD-2* and *CAD-6* are the most likely candidates for the monolignol pathway in *Arabidopsis*. Of these two, only *CAD-6* has an AC element. The function of class II and class III *CAD* genes remains less clear. However, *CAD-3*, *CAD-4*, and *CAD-5* of class II are the closest homologs of the poplar *SAD* (Li et al., 2001). Possibly, one or all of these proteins show a preference for sinapyl alcohol or sinapaldehyde turning them into S branch-specific enzymes.

Discussion

As a first step in the functional analysis of monolignol biosynthesis genes, we searched the complete *Arabidopsis* genome for members of the gene families currently known to be involved in monolignol biosynthesis and found 34 candidate genes (Tables 1-10). Eleven of these genes have, to our knowledge, not been described before. The gene annotation was complemented with, on the one hand, an exhaustive compilation of previous — most of the time fragmented — expression data, and on the other hand, an expression analysis of all 34 genes in an array of tissues providing us for the first time with an overall picture of gene expression (Tables 1-10). Moreover, five genes had not been picked up by an EST before, which provides the first expression data for these genes. Together, these data will serve as a compendium for further functional studies of these genes in *Arabidopsis*.

Fourteen monolignol biosynthesis genes are highly expressed in the inflorescence stem

Lignification is a process that occurs predominantly in cells of the vascular tissue, found in almost all organs, but most abundantly in stems and roots. In addition, flowers, seeds, and siliques accumulate significant amounts of other phenylpropanoid-derived compounds, such as sinapate esters and flavonoids (Chapple et al., 1994; Chen and McClure, 2000; Ruegger and Chapple, 2001).
All 34 genes, annotated from the *Arabidopsis* genome sequence for their potential involvement in monolignol biosynthesis, are expressed at some stage of inflorescence stem development, a tissue with a prominent portion of lignifying cells (Tables 1-10; Dharmawardhana et al., 1992). Of these genes, 23 are expressed throughout stem development (Table 11). Furthermore, of 11 of these 23 genes, the expression increases during the later stages of inflorescence stem development, when lignification is more prominent (Dharmawardhana et al., 1992). Additionally, six genes, curiously all upstream of C3H in the general part of the phenylpropanoid biosynthesis, have the highest expression in stem as compared to other tissues (Table 11).

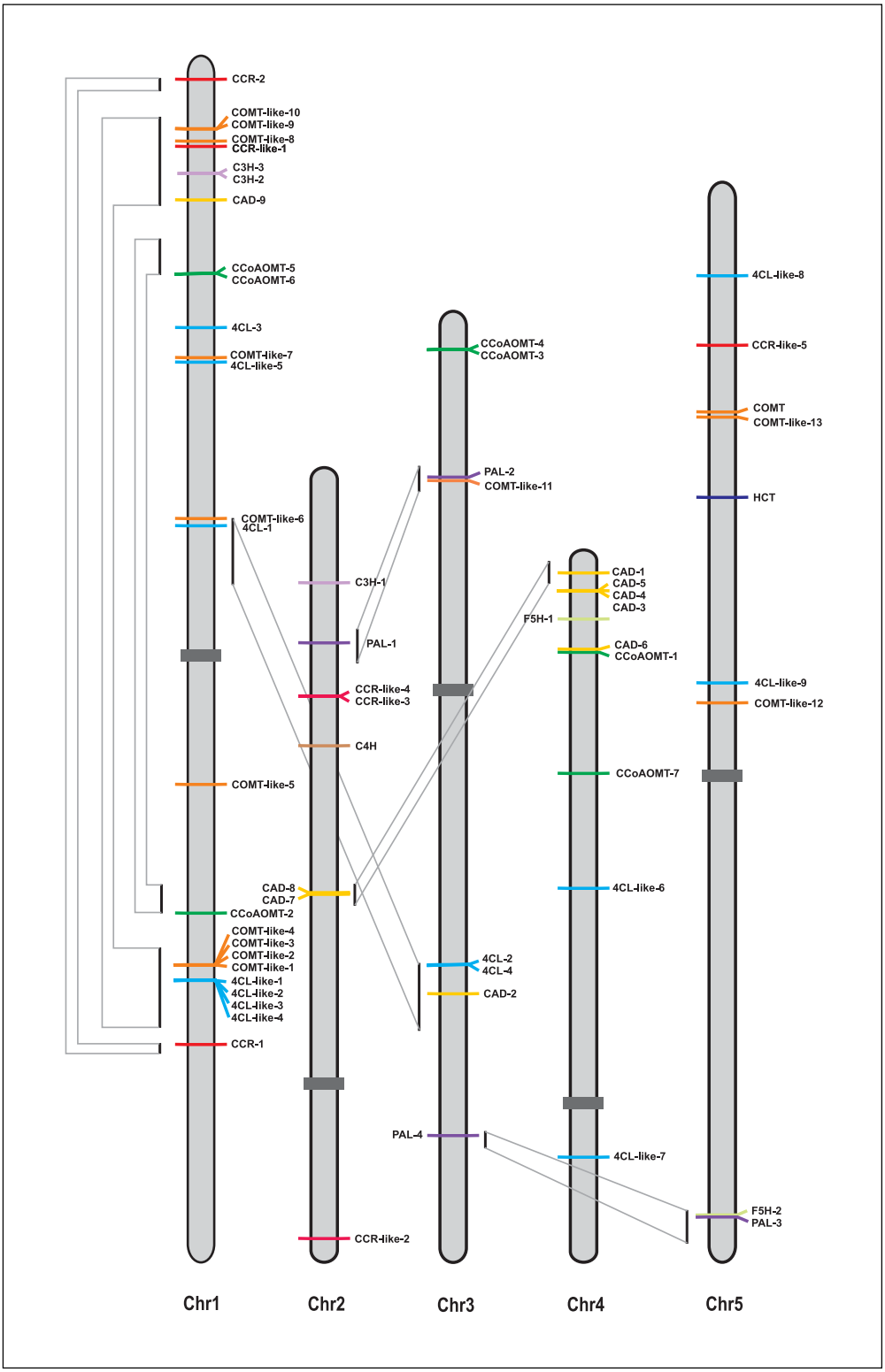

Figure 12. Chromosomal position of monolignol biosynthesis genes. Linked vertical bars represent large duplicated regions in which genes of this study have been retained after duplication. Horizontal gray bars indicate the position of centromeric regions.

Table 11. Summary of expression characteristics and occurrence of AC elements in monolignol biosynthesis genes. Characteristics are listed for each gene: the corresponding mutants with lignification-related phenotypes, the clustering with certified proteins of other species in the phylogenetic analysis, a high and constitutive expression in the inflorescence stem being eventually higher than in other tissues (as determined by our RT-PCR analysis), ESTs of the different relevant categories in total numbers, and the occurrence of AC elements. Genes in boldface have been characterized for the first time. Genes marked with asterisks are associated to a membrane or predicted to be ER targeted. The position of AC elements found with stringent parameters is given in bp from ATG and the strand within brackets. Shaded fields indicate an overrepresentation of ESTs in this particular tissue or conditions as compared to the presence of this gene in all ESTs. Overrepresentation was judged by comparing the relative occurrence of a gene in all ESTs with that of the same gene in a particular tissue or condition. When less than three ESTs were detected in a particular tissue or condition, no overrepresentation was calculated.

gene family	genes with AC element													genes without AC element										
		corresponding mutants	phylogenetic clustering	constitutively in stem	higher in stem	EST total	EST inflorescence stem	EST aboveground organs	EST root	EST seed	EST stress, wound, pathogen	AC element (+/-) strand (core:0.9;matrix 0.9)		corresponding mutants	phylogenetic clustering	constitutively in stem	higher in stem	EST total	EST inflorescence stem	EST aboveground organs	EST root	EST seed	EST stress, wound, pathogen
PAL	*PAL-1*	*pal1*[a]	x	x	x	41		1	13	2	21	483 (+)		*PAL-3*		x	x		1					
	PAL-2	*pal2*[a]	x	x		50		2	17	12	17	246 (+), 495 (+)		***PAL-4***		x	x	x	28		1	5	16	4
C4H														*C4H**	*ref3* [b]	x	x	x	29			2	5	15
4CL	*4CL-1*		x	x	x	8				1	4	159 (+)		*4CL-3*					8			1	2	3
	4CL-2		x	x	x	13			6	2	2	124 (+), 233 (+)		***4CL-4***		x			2					2
HCT	*HCT*		x	x	x	57		13	11	8	15	132 (-)												
C3H	*C3H-1**	*ref8*	x	x		36		5	5	7	16	145 (+)		***C3H-2***					2					
														C3H-3					0					
CCoAOMT	*CCoAOMT-1*		x	x		45	1		19	9	10	174 (+), 651 (+)		***CCoAOMT-2***					0					
														CCoAOMT-3*					6					3
														CCoAOMT-4			x		2					1
														CCoAOMT-5					1					
														CCoAOMT-6					2					
														CCoAOMT-7			x		4				3	1
CCR	*CCR-1*	*irx4*	x	x		43	1	8	8	8	10	269 (+)		*CCR-2*		x			4			3		1
F5H														*F5H-1**	*fah1*	x	x		2					
														F5H-2*			x		0					
COMT														*COMT**	*comt*	x	x		99		16	28	22	20
CAD	*CAD-5*					2					1	256 (-)		*CAD-1*			x		32		4		10	10
	CAD-6		x	x		23			5	5	6	515 (+)		*CAD-2*		x	x		33		4	11	2	8
														CAD-3			x		1					1
														CAD-4					26		5			8
														CAD-7			x		0					
														CAD-8			x		0					
														CAD-9			x		9				5	3

[a] Rohde *et al.*, in preparation
[b] C. Chapple, personal communication

Whereas none of the 34 genes is exclusively expressed in stem or root, their expression level, as estimated from the EST data, is highest in those EST classes known to correlate to some extent with lignification, namely root and aboveground organs (Table 11). A strong expression of monolignol biosynthesis genes in stems and roots is documented in numerous publications (Tables 1-10, and references therein).

Possibly, lignification cDNAs are relatively highly represented in root libraries because of the absence of other very abundant processes, such as photosynthesis, or, as could be concluded from *AtC4H::GUS* analysis (Nair et al., 2002), the phenylpropanoid pathway in roots is active in more cells than the vascular ones to generate compounds not destined for lignification.

A high expression level in lignifying tissues (RT-PCR and EST), and the phylogenetic classification in groups with functionally characterized proteins of other species, were used as the first two criteria to delineate those family members that are the most likely to be involved in monolignol biosynthesis during developmental lignification (Table 11). These criteria are fulfilled for 14 genes: *PAL-1*, *PAL-2*, *PAL-4*, *C4H*, *4CL-1*, *4CL-2*, *HCT*, *C3H-1*, *CCoAOMT-1*, *CCR-1*, *F5H-1*, *COMT*, *CAD-2*, and *CAD-6*. Of these 14, seven genes have been already certified for their involvement in monolignol biosynthesis through the characterization of the corresponding mutants: *PAL-1* (*pal1*), *PAL-2* (*pal2*), *C4H* (*ref3*), *C3H-1* (*ref8*), *CCR-1* (*irx4*), *F5H-1* (*fah1*), and *COMT* (*comt1*) (Chapple et al., 1992; Jones et al., 2001; Franke et al., 2002a, 2002b; Goujon et al., 2003b; C. Chapple, personal communication; A. Rohde et al., in preparation). Except for *CAD-6*, these 14 genes have the highest expression level in their respective gene families, as judged from the number of ESTs (Table 11). In conclusion, this set of 14 genes is through their expression and phylogeny eligible for being involved in the developmental monolignol biosynthesis in *Arabidopsis*.

AC Elements sign-post a number of G-branch monolignol biosynthesis genes

AC elements, originally identified in the promoters of the parsley *PAL1* gene, the bean *PAL2* and *PAL3* genes and parsley *4CL-1* gene (Cramer et al., 1989; Lois et al., 1989; Hauffe et al., 1991; Leyva et al., 1992), are thought to enhance the expression of genes in xylem and at the same time to prevent their expression in the adjacent phloem and cortical cells. A number of functional studies have proven the importance of AC elements for vascular expression (Hauffe et al., 1993; Hatton et al., 1995; Lacombe et al., 2000; C. Chen et al., in preparation). Because the deletion of the AC element results, within the vascular tissue, in derepression of phloem expression, it has been suggested that a (possibly phloem-specific) repressor is normally bound to the AC element preventing expression in cells other than xylem cells. In contrast, in xylem cells, the repressor would be released to give rise to typically high expression levels (Hauffe et al., 1993; Hatton et al., 1995). A number of MYB and other transcription factors bind to AC elements resulting in *trans*-activation of the respective promoters (e.g., Sablowski et al., 1995; Séguin et al., 1997; Sugimoto et al., 2000). Overexpression of specific MYB factors leads to lignin-related phenotypes (Tamagnone et al., 1998; Borevitz et al., 2000). Of course, AC element-controlled genes, recruited into monolignol biosynthesis in vascular cells would retain the capacity to participate outside these cells in other processes. In fact, many AC element-containing genes have complex expression patterns, emphasizing that the AC element is only one component that regulates the activity of their promoters (Tables 1-10).

Given the importance of AC elements in specifying vascular expression, the presence of an AC element in the promoters of the 34 annotated genes has been examined. In the past, most AC elements were identified by consensus sequences built from both experimentally verified AC elements as well as those detected by sequence similarity. Often on top of such a consensus, a number of mismatches were allowed. Moreover, AC elements were often subdivided into ACI and ACII boxes, despite the fact that they align perfectly and were shown to be functionally redundant with respect to vascular expression (see supplemental data; Hatton et al., 1995).

In view of the limited knowledge on the specific binding of AC elements by transcription factors in vivo, we built one unifying matrix for element identification based on the five experimentally verified and delineated elements (see supplemental data) with very stringent parameters in the search. In addition, such a matrix approach accounts for relative probabilities of bases at a particular position, whereas the consensus sequences only describe the presence or absence of one or more base(s) at a position. To illustrate the power of matrix versus consensus approach, the statistical significance of both methods was evaluated on 1,000 random intergenic regions distributed uniformly throughout the *Arabidopsis* genome. The consensus approach used, for example, by Wanner et al. (1995), has a probability to find an AC element by chance of once every 1,200 bp, whereas with our approach it is once every 37,000 bp. With these parameters, some of the AC elements that had previously been identified based on similarity to a consensus were not detected, such as the AC element in the *PAL-3*, *C4H*, and *4CL-3* promoters (Wanner et al., 1995; Mizutani et al., 1997; Ehlting et al., 1999). Note that the elements in these promoters have not been verified experimentally. By searching with the matrix approach all 29,787 *Arabidopsis* genes predicted with EuGene (Schiex et al., 2001; C. Serizet et al., in preparation), AC elements on either DNA strand were found in 780 promoters (2.6%). In the set of 34 monolignol biosynthesis genes, 10 out of 34 promoters have AC elements (29%; Table 11): 8 on the positive and 2 on the negative strands.

Seven gene families have at least one family member with an AC element in their promoter (Table 11). Genes with an AC element do not simply correspond with genes that are highly expressed as estimated from the number of ESTs (Table 11). Rather, AC elements coincide with those gene family members that were assigned to be involved in developmental lignification based on expression and phylogeny (see above): of these 14 genes, nine contain an AC element on the positive strand (Table 11). Thus, within their respective gene families the following genes are extra-qualified for playing a role in developmental lignification in vascular tissues: *PAL-1*, *PAL-2*, *4CL-1*, *4CL-2*, *HCT*, *C3H-1*, *CCoAOMT-1*, *CCR-1*, and *CAD-6*. *CAD-5* has an AC element, but did not cluster with the true *CAD* clade in the phylogenetic tree (Figure 11). In contrast, no AC elements were found in the gene families *C4H*, *F5H*, and *COMT*. Of these three gene families, *C4H* and *COMT* are single genes that, contrary to multigene families, may have acquired a more relaxed promoter organization compatible with expression in a broader range of cells and conditions. Maybe these genes contain more degenerated AC elements that were not picked up under the stringent search parameters used. The *F5H* family consists of two genes that are not functionally redundant, because F5H-2 fails to compensate for the loss of F5H-1 in the *fah1* mutant (Meyer et al., 1998). In this rationale, *F5H-1* probably has to be considered as a single gene as well. However, this hypothesis, explaining why *C4H*, *F5H* and *COMT* promoters lack an AC element, is not in agreement with *HCT*, which is a single gene as well, but has an AC element, albeit on the negative strand. Interestingly, only two of the 13 AC elements detected in the promoters are on the negative strand, potentially indicating that they are not functional.

A tantalizing alternative hypothesis starts out from the notion that all AC element-containing monolignol biosynthesis genes code for enzymes acting in the G-branch of the pathway (Figure 1, Table 11). None of the 14 other promoter elements analyzed, including stress- and elicitor-responsive elements, could be linked in a similar meaningful way to particular groups of genes (Tables 1-10),

underscoring how important the presence of AC elements may be for a common regulation of G-branch genes. A separate regulation of S-branch genes is a valid option to explain why the latter lack AC elements, given the spatio-temporal differences in deposition of S and G lignin (Dharmawardhana et al., 1992; Dixon et al., 2001; Donaldson, 2001; Jones et al., 2001). Young tissues accumulate preferentially G lignin, whereas the content of S lignin increases with tissue maturity (Meyer et al., 1998). At the level of individual cells, G-branch enzymes are involved in the lignin deposition during earlier stages of cell wall formation than S-branch enzymes (Terashima et al., 1986). Within the vascular tissue, xylem vessels contain G lignin, whereas fibers and parenchyma cells contain a mixture of G and S lignin, with the latter predominating in fibers (Donaldson, 2001; and references therein). Maybe the profound induction of G-branch enzymes suffices to achieve the required extra-production of G monolignols for secondary cell wall formation, typical of a lignifying xylem vessel cell. This suggestion is in line with the previously proposed mode of action of AC elements within the vascular tissue: AC elements drive high expression in xylem vessels, whereas in phloem (consisting primarily of fibers) they repress it (Hauffe et al., 1993; Hatton et al., 1995). The only G-branch gene family lacking a member with AC element is the single *C4H*. However, *C4H* might be regulated separately. Its transcriptional regulation was shown to be distinct from other monolignol genes in *Arabidopsis* as well as in pine (*Pinus taeda*; Jin et al., 2000; Anterola et al., 2002).

If this scenario were true, AC elements correlate with a strong expression of G-lignin genes. Furthermore, COMT and F5H would have been recruited specifically into the S branch during the evolution of angiosperms, because no S lignin is made in gymnosperms. As a consequence, a putatively S-specific alcohol dehydrogenase, as identified in poplar (Li et al., 2001), might also exist in *Arabidopsis*. The class III CAD proteins clusters with a CAD of alfalfa that reduces, among others, sinapaldehyde (Brill et al., 1999). However, the class II CAD genes, *CAD-3*, *CAD-4*, and *CAD-5* are the closest homologs of the poplar SAD, with respect to sequence. These *Arabidopsis CAD* genes do not contain AC elements, except for *CAD-5* that would need to be excluded as a putative *SAD* homolog. Mixed substrate assays, as applied for the identification of the poplar SAD (Li et al., 2001), will be very informative to clarify this question.

Putative membrane localization of six enzymes

Growing evidence suggests that cytochrome P450 enzymes provide membrane anchors in the ER for assembling multienzyme complexes involved in metabolic channeling within the phenylpropanoid pathway (Chapple, 1998; Rasmussen and Dixon, 1999; Wagner and Hrazdina, 1984; Winkel-Shirley, 1999). Metabolic channeling has been reported from phenylalanine to *p*-coumarate with a possible association of PAL and C4H on microsomal membranes (Czichi and Kindl, 1977; Wagner and Hrazdina, 1984; Rasmussen and Dixon, 1999). Among the three P450 enzyme families of the pathway (C4H, C3H, and F5H), C4H, C3H-1, and F5H-2 are predicted by TargetP to hold an ER targeting peptide. However, the predicted peptide coincides with the N-terminal hydrophobic helix, which is part of the membrane-anchoring region common to all P450 proteins (Chapple, 1998).

Because membrane-anchoring regions, such as those present in P450, are known to cause false predictions by TargetP, it is very unlikely that the P450 enzymes would contain a cleavable signal peptide (G. von Heijne, personal communication). C3H-1 has previously been localized to the microsomal fraction when expressed heterologously in yeast (Franke et al., 2002b). A *c-myc*-tagged C4H was found exclusively in the microsomal fraction of tobacco and the poplar C4H fused to GFP was shown to be ER-localized in transgenic *Arabidopsis* (Ro et al., 2001; Achnine et al., 2002). In conclusion, C4H, C3H-1, F5H-1, and F5H-2 have a well-conserved membrane-anchoring region, in agreement with their proposed localization in the ER membrane. C3H-2 and C3H-3, are not predicted to contain an ER targeting peptide and do not comply to the amino acid features of the membrane anchor.
In addition, CCoAOMT-3 contains also a putative ER targeting signal, but no evidence exists for membrane association, implying a surprisingly vacuolar or extracellular localization of this enzyme. Sinapoylglucose:malate sinapoyltransferase (SMT) and sinapoylglucose:choline sinapoyltransferase (SCT) involved in modification of sinapoylglucose have been identified as proteins with an ER targeting peptide (Lehfeldt et al., 2000; Shirley et al., 2001). These enzymes were suggested to be localized in the vacuole based on previous studies showing SMT activity in vacuoles (Strack and Sharma, 1985). Whether CCoAOMT-3 shares this localization needs experimental verification.
Finally, a putative myristoylation site was detected in the COMT, possibly involved in membrane anchoring. In agreement with this finding, a fraction of COMT from alfalfa stem was shown to be associated with the microsomal membranes, and channeling by COMT and F5H was suggested from coniferaldehyde to sinapaldehyde in the S-branch of the monolignol pathway (Guo et al., 2002). This observation was interpreted as a tight coupling of COMT with the membrane-anchored F5H (Guo et al., 2002). However, our data do not exclude that COMT itself could be anchored into the membrane by myristoylation.

Monolignol biosynthesis gene families show a large diversity in size, sequence similarity, and functional spectrum

The number of candidate monolignol biosynthesis genes found in the *Arabidopsis* genome varies greatly among the gene family studied: from single genes (*COMT*, *HCT*, and *C4H*) to medium-size (*F5H*, *C3H*, *PAL*, and *CCR*), and large (*4CL*, *CAD*, and *CCoAOMT*) gene families. For some families, clear classes were revealed by the phylogenetic analysis of all plant members (*C4H*, *C3H*, *4CL*, *CAD*, and *CCoAOMT*), whereas other families were represented by one class only (*COMT* and *CCR*). A complex history of gene duplications caused the expansion and diversification of the respective gene families. Interestingly, the polyploidy event, which happened 75 million years ago, as indicated by the presence of large blocks of genes that duplicated at that time (Figure 12; Simillion et al., 2002; Raes et al., 2003), did not create new classes within any of the investigated families. Mapping of the genome duplication on the respective phylogenetic trees shows that, in all cases, this event together with several small-scale duplications, was responsible only for a greater within-class diversity (Figures 2-11). Classes must have originated at an earlier time in evolution, i.e., before 75 million years ago and not necessarily all at the same time.

By definition, the complete genome duplication created a full, redundant, double set of monolignol biosynthesis genes. In some gene families (*PAL*, *CAD*, *CCR*, *4CL*, and *CCoAOMT)*, duplicates that originated through this event have been retained, whereas they were lost in others (*COMT*, *HCT*, and *C4H*). The mechanisms and reasons of gene conservation and loss after duplication are still unclear and various theories exist (Prince and Prickett, 2002). Some of the duplicated genes in this study that were retained have evolved different expression patterns after the genome duplication (*CCR-1* and *CCR-2*, *PAL-3* and *PAL-4*, and *CAD-1* and *CAD-7*/*CAD-8*), whereas others show no clear difference in expression (*PAL-1* and *PAL-2*). In some cases, the influence of the genome duplication is blurred by more ancient (*4CL-1*, *4CL-2*, and *4CL-4*) or recent (*CCoAOMT-2*, *CCoAOMT-5*, and *CCoAOMT-6*) tandem duplications. The observed differences in expression of the *CCR* and *CAD* families might point to a functional divergence of these genes, a process called subfunctionalization. The duplicates still exert the same biochemical function, but in different spatio-temporal “niches” in the organism (Piatigorsky and Wistow, 1991; Hughes, 1994; Force et al., 1999). A putative example of subfunctionalization after gene duplication is found in the C3H family, where *C3H-2* and *C3H-3* show a mutually exclusive expression pattern during stem development (Table 5). However, although subfunctionalization may be a possible reason why the duplicated genes are retained, only further functional studies will reveal the full consequences of these gene and genome duplications within the monolignol biosynthetic pathway and provide some more hints on the evolutionary constraints acting on these families.

In conclusion, the genome-wide analysis of monolignol biosynthesis genes, as presented here, provides the foundation of the next steps in unravelling the monolignol pathway. The combination of reverse genetics with transcript and metabolite profiling analyses of the respective mutants will profoundly enlarge our understanding of this pathway and its relation with plant development.

Acknowledgements

The authors would like to thank Vincent Thareau for the gene-specific primer design and Pierre Rouzé for helpful discussions, Gunnar von Heijne for help in the interpretation of the TargetP results, Clint Chapple, Lise Jouanin, Michael Walter, and Carine Serizet for sharing unpublished results, Vanessa Hostyn for excellent technical assistance, Stephane Rombauts and Cedric Simillion for help with promoter analysis and *Arabidopsis* duplicated regions, and Martine de Cock for help in preparing the manuscript. Part of this work was supported by funds from the European Commission programs EDEN (QLK5-CT-2001-00443) and COPOL (QLK5-2000-01493). A.R. is a postdoctoral fellow of the Fund for Scientific Research (Flanders).

References

Achnine, L., Rasmussen, S., Blancaflor, E., and Dixon, R.A. (2002). Metabolic channeling at the entry point into the phenylpropanoid pathway: physical association between L-phenylalanine ammonia-lyase and cinnamate 4-hydroxylase. In Abstract presented at the XXI International Conference on Polyphenols, Marrakech (Morocco), September 9-12, 2002.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25,** 3389-3402.

Anterola, A.M., and Lewis, N.G. (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry **61,** 221-294.

Anterola, A.M., Jeon, J.-H., Davin, L.B., and Lewis, N.G. (2002). Transcriptional control of monolignol biosynthesis in *Pinus taeda*. Factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. J. Biol. Chem. **277,** 18272-18280.

Bate, N.J., Orr, J., Ni, W., Meromi, A., Nadler-Hassar, T., Doerner, P.W., Dixon, R.A., Lamb, C.J., and Elkind, Y. (1994). Quantitative relationship between phenylalanine ammonia-lyase levels and phenylpropanoid accumulation in transgenic tobacco identifies a rate-determining step in natural product synthesis. Proc. Natl. Acad. Sci. USA **91,** 7608-7612.

Bell-Lelong, D.A., Cusumano, J.C., Meyer, K., and Chapple, C. (1997). Cinnamate-4-hydroxylase expression in *Arabidopsis*. Plant Physiol. **113,** 729-738.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2003). GenBank. Nucleic Acids Res. **31,** 23-27.

Betz, C., McCollum, T.G., and Mayer, R.T. (2001). Differential expression of two cinnamate 4-hydroxylase genes in 'Valencia' orange (*Citrus sinensis* Osbeck). Plant Mol. Biol. **46,** 741-748.

Blount, J.W., Korth, K.L., Masoud, S.A., Rasmussen, S., Lamb, C., and Dixon, R.A. (2000). Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop at the entry point into the phenylpropanoid pathway. Plant Physiol. **122,** 107-116.

Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin biosynthesis. Annu. Rev. Plant Biol. **54,** in press.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A., and Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell **12,** 2383-2393.

Brill, E.M., Abrahams, S., Hayes, C.M., Jenkins, C.L.D., and Watson, J.M. (1999). Molecular characterisation and expression of a wound-inducible cDNA encoding a novel cinnamyl-alcohol dehydrogenase enzyme in lucerne (*Medicago sativa* L.). Plant Mol. Biol. **41,** 279-291.

Chaffey, N., Cholewa, E., Regan, S., and Sundberg, B. (2002). Secondary xylem development in *Arabidopsis*: a model for wood formation. Physiol. Plant. **114,** 594-600.

Chapple, C. (1998). Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. Annu. Rev. Plant Physiol. Plant Mol. Biol. **49,** 311-343.

Chapple, C.C.S., Vogt, T., Ellis, B.E., and Somerville, C.R. (1992). An *Arabidopsis* mutant defective in the general phenylpropanoid pathway. Plant Cell **4,** 1413-1424.

Chapple, C.C.S., Shirley, B.W., Zook, M., Hammerschmidt, R., and Somerville, S.C. (1994). Secondary metabolism in *Arabidopsis*. In *Arabidopsis*, E.M. Meyerowitz and C.R. Somerville, eds (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), pp. 989-1030.

Chen, M., and McClure, J.W. (2000). Altered lignin composition in phenylalanine ammonia-lyase-inhibited radish seedlings: implications for seed-derived sinapoyl esters as lignin precursors. Phytochemistry **53,** 365-370.

Chen, W., Chao, G., and Singh, K.B. (1996). The promoter of a H_2O_2-inducible, *Arabidopsis* glutathione S-transferase gene contains closely linked OBF- and OBP1-binding sites. Plant J. **10,** 955-966.

Chen, F., Kota, P., Blount, J.W., and Dixon, R.A. (2001). Chemical syntheses of caffeoyl and 5-OH coniferyl aldehydes and alcohols and determination of lignin *O*-methyltransferase activities in dicot and monocot species. Phytochemistry **58,** 1035-1042.

Chen, C., Meyermans, H., Burggraeve, B., De Rycke, R.M., Inoue, K., De Vleesschauwer, V., Steenackers, M., Van Montagu, M.C., Engler, G.J., and Boerjan, W.A. (2000). Cell-specific and conditional expression of caffeoyl-CoA *O*-methyltransferase in poplar. Plant Physiol. **123,** 853-867.

Cramer, C.L., Edwards, K., Dron, M., Liang, X., Dildine, S.L., Bolwell, G.P., Dixon, R.A., Lamb, C.J., and Schuch, W. (1989). Phenylalanine ammonia-lyase gene organization and structure. Plant Mol. Biol. **12,** 367-383.

Cukovic, D., Ehlting, J., VanZiffle, J.A., and Douglas, C.J. (2001). Structure and evolution of 4-coumarate:coenzyme A ligase (*4CL*) gene families. Biol. Chem. **382,** 645-654.

Czichi, U., and Kindl, H. (1977). Phenylalanine ammonia-lyase and cinnamic acid hydroxylase as assembled consecutive enzymes on microsomal membranes of cucumber cotyledons: Co-operation and subcellular distribution. Planta **134,** 133-143.

Deikman, J., and Hammer, P.E. (1995). Induction of anthocyanin accumulation by cytokinins in *Arabidopsis thaliana*. Plant Physiol. **108,** 47-57.

Dharmawardhana, D.P., Ellis, B.E., and Carlson, J.E. (1992). Characterization of vascular lignification in *Arabidopsis thaliana*. Can. J. Bot. **70,** 2238-2244.

Dixon, R.A., Chen, F., Guo, D., and Parvathi, K. (2001). The biosynthesis of monolignols: a "metabolic grid", or independant pathways to guaiacyl and syringyl units? Phytochemistry **57,** 1069-1084.

Donaldson, L.A. (2001). Lignification and lignin topochemistry - an ultrastructural view. Phytochemistry **57,** 859-873.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Ehlting, J., Büttner, D., Wang, Q., Douglas, C.J., Somssich, I.E., and Kombrink, E. (1999). Three 4-coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. Plant J. **19,** 9-20.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300,** 1005-1016.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8,** 967-974.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151,** 1531-1545.

Franke, R., Hemm, M.R., Denault, J.W., Ruegger, M.O., Humphreys, J.M., and Chapple, C. (2002). Changes in secondary metabolism and deposition of an unusual lignin in the *ref8* mutant of *Arabidopsis*. Plant J. **30,** 47-59.

Franke, R., Humphreys, J.M., Hemm, M.R., Denault, J.W., Ruegger, M.O., Cusumano, J.C., and Chapple, C. (2002). The *Arabidopsis REF8* gene encodes the 3-hydroxylase of phenylpropanoid metabolism. Plant J. **30,** 33-45.

Galliano, H., Cabané, M., Eckerskorn, C., Lottspeich, F., Sandermann, H.J., and Ernst, D. (1993). Molecular cloning, sequence analysis and elicitor-/ozone-induced accumulation of cinnamyl alcohol dehydrogenase from Norway spruce (*Picea abies* L.). Plant Mol. Biol. **23,** 145-156.

Goormachtig, S., Valerio-Lepiniec, M., Szczyglowski, K., Van Montagu, M., Holsters, M., and de Bruijn, F.J. (1995). Use of differential display to identify novel *Sesbania rostrata* genes enhanced by *Azorhizobium caulinodans* infection. Mol. Plant-Microbe Interact. **8,** 816-824.

Goujon, T., Ferret, V., Mila, I., Pollet, B., Ruel, K., Burlat, V., Joseleau, J.-P., Barrière, Y., Lapierre, C., and Jouanin, L. (2003). Down-regulation of *AtCCR1* gene in *Arabidopsis thaliana*: effects on phenotype, lignins and cell wall degradability. Planta, in press.

Goujon, T., Sibout, R., Pollet, B., Maba, B., Nussaume, L., Bechtold, N., Lu, F., Ralph, J., Mila, I., Barrière, Y., Lapierre, C., and Jouanin, L. (2003). A new *Arabidopsis thaliana* mutant deficient in the expression of *O*-methyltransferase 1: impact on lignins and on sinapoyl esters. Plant Mol. Biol., in press.

Grimmig, B., and Matern, U. (1997). Structure of the parsley caffeoyl-CoA *O*-methyltransferase gene, harbouring a novel elicitor responsive *cis*-acting element. Plant Mol. Biol. **33,** 323-341.

Guillén, P., Guis, M., Martínez-Reina, G., Colrat, S., Dalmayrac, S., Deswarte, C., Bouzayen, M., Roustan, J.-P., Fallot, J., Pech, J.-C., and Latché, A. (1998). A novel NADPH-dependent aldehyde reductase gene from *Vigna radiata* confers resistance to the grapevine fungal toxin eutypine. Plant J. **16,** 335-343.

Guo, D., Chen, F., and Dixon, R.A. (2002). Monolignol biosynthesis in microsomal preparations from lignifying stems of alfalfa (*Medicago sativa* L.). Phytochemistry **61,** 657-667.

Guo, D., Chen, F., Wheeler, J., Winder, J., Selman, S., Peterson, M., and Dixon, R.A. (2001). Improvement of in-rumen digestibility of alfalfa forage by genetic manipulation of lignin O-methyltransferases. Transgenic Res. **10,** 457-464.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41,** 95-98.

Harding, S.A., Leshkevich, J., Chiang, V.L., and Tsai, C.-J. (2002). Differential substrate inhibition couples kinetically distinct 4-coumarate:coenzyme A ligases with spatially distinct metabolic roles in quaking aspen. Plant Physiol. **128,** 428-438.

Hatton, D., Sablowski, R., Yung, M.-H., Smith, C., Schuch, W., and Bevan, M. (1995). Two classes of *cis* sequences contribute to tissue-specific expression of a *PAL2* promoter in transgenic tobacco. Plant J. **7,** 859-876.

Hauffe, K.D., Lee, S.P., Subramaniam, R., and Douglas, C.J. (1993). Combinatorial interactions between positive and negative *cis*-acting elements control spatial patterns of *4CL-1* expression in transgenic tobacco. Plant J. **4,** 235-253.

Hauffe, K.D., Paszkowski, U., Schulze-Lefert, P., Hahlbrock, K., Dangl, J.L., and Douglas, C.J. (1991). A parsley 4CL-1 promoter fragment specifies complex expression patterns in transgenic tobacco. Plant Cell **3,** 435-443.

Hoffmann, L., Maury, S., Martz, F., Geoffroy, P., and Legrand, M. (2002). Purification, cloning and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. J. Biol. Chem. **278,** 95-103.

Hu, W.-J., Kawaoka, A., Tsai, C.-J., Lung, J., Osakabe, K., Ebinuma, H., and Chiang, V.L. (1998). Compartmentalized expression of two structurally and functionally distinct 4-coumarate:CoA ligase genes in aspen (*Populus tremuloides*). Proc. Natl. Acad. Sci. USA **95,** 5407-5412.

Hu, W.-J., Harding, S.A., Lung, J., Popko, J.L., Ralph, J., Stokke, D.D., Tsai, C.-J., and Chiang, V.L. (1999). Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnol. **17,** 808-812.

Hughes, L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B **256,** 119-124.

Humphreys, J.M., and Chapple, C. (2002). Rewriting the lignin roadmap. Curr. Opin. Plant Biol. **5,** 224-229.

Humphreys, J.M., Hemm, M.R., and Chapple, C. (1999). New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proc. Natl. Acad. Sci. USA **96,** 10045-10050.

Initiative, T.A.G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., and Martin, C. (2000). Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. EMBO J. **19,** 6150-6161.

Jones, L., Ennos, A.R., and Turner, S.R. (2001). Cloning and characterization of *irregular xylem4* (*irx4*): a severely lignin-deficient mutant of *Arabidopsis*. Plant J. **26,** 205-216.

Kawai, S., Mori, A.S., T., Kajita, S., Katayama, Y., and Morohoshi, N. (1996). Isolation and analysis of cinnamic acid 4-hydroxylase homologous genes from a hybrid aspen, *Populus kitakamiensis*. Biosci. Biotech. Biochem. **60,** 1586-1597.

Kiedrowski, S., Kawalleck, P., Hahlbrock, K., Somssich, I.E., and Dangl, J.L. (1992). Rapid activation of a novel plant defense gene is strictly dependent on the *Arabidopsis RPM1* disease resistance locus. EMBO J. **11,** 4677-4684.

Krawczyk, S., Thurow, C., Niggeweg, R., and Gatz, C. (2002). Analysis of the spacing between the two palindromes of *activation sequence-1* with respect to binding to different TGA factors and transcriptional activation potential. Nucleic Acids Res. **30,** 775-781.

Kühnl, T., Koch, U., Heller, W., and Wellmann, E. (1987). Chlorogenic acid biosynthesis: characterization of a light-induced microsomal 5-*O*-(4-coumaroyl)-D-quinate/shikimate 3'-hydroxylase from carrot (*Daucus carota* L.) cell suspension cultures. Arch. Biochem. Biophys. **258,** 226-232.

Lacombe, E., Van Doorsselaere, J., Boerjan, W., Boudet, A.M., and Grima-Pettenati, J. (2000). Characterization of *cis*-elements required for vascular expression of the *Cinnamoyl CoA Reductase* gene and for protein-DNA complex formation. Plant J. **23,** 663-676.

Lamb, C. (1977). trans-cinnamic acid as a mediator of the light-stimulated increase in hydroxycinnamoyl:CoA-quinate hydroxycinnmoyl transferase. FEBS Lett. **75,** 37-40.

Lauvergeat, V., Lacomme, C., Lacombe, E., Lasserre, E., Roby, D., and Grima-Pettenati, J. (2001). Two cinnamoyl-CoA reductase (CCR) genes from *Arabidopsis thaliana* are differentially expressed during development and in response to infection with pathogenic bacteria. Phytochemistry **57,** 1187-1195.

Lee, D., Meyer, K., Chapple, C., and Douglas, C.J. (1997). Antisense suppression of 4-coumarate:coenzyme A ligase activity in *Arabidopsis* leads to altered lignin subunit composition. Plant Cell **9,** 1985-1998.

Lee, D., Ellard, M., Wanner, L.A., Davis, K.R., and Douglas, C.J. (1995). The *Arabidopsis thaliana* 4-coumarate:CoA ligase (*4CL*) gene: stress and developmentally regulated expression and nucleotide sequence of its cDNA. Plant Mol. Biol. **28,** 871-884.

Lehfeldt, C., Shirley, A.M., Meyer, K., Ruegger, M.O., Cusumano, J.C., Viitanen, P.V., Strack, D., and Chapple, C. (2000). Cloning of the *SNG1* gene of *Arabidopsis* reveals a role for a serine carboxypeptidase-like protein as an acyltransferase in secondary metabolism. Plant Cell **12,** 1295-1306.

Lescot, M., Déhais, P., Moreau, Y., Van de Peer, Y., Rouzé, P., and Rombauts, S. (2002). PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. Nucleic Acids Res. **30,** 325-327.

Lewis, N.G., and Yamamoto, E. (1990). Lignin: occurrence, biogenesis, and biodegradation. Ann. Rev. Plant Physiol. Plant Mol. Biol. **41,** 455-496.

Leyva, A., Jarillo, J.A., Salinas, J., and Martinez-Zapater, J.M. (1995). Low temperature induces the accumulation of *phenylalanine ammonia-lyase* and *chalcone synthase* mRNAs of *Arabidopsis thaliana* in a light-dependent manner. Plant Physiol. **108,** 39-46.

Leyva, A., Liang, X., Pintor-Toro, J.A., Dixon, R.A., and Lamb, C.J. (1992). *cis*-Element combinations determine phenylalanine ammonia-lyase gene tissue-specific expression patterns. Plant Cell **4,** 263-271.

Li, L., Osakabe, Y., Joshi, C.P., and Chiang, V.L. (1999). Secondary xylem-specific expression of caffeoyl-coenzyme A 3-*O*-methyltransferase plays an important role in the methylation pathway associated with lignin biosynthesis in loblolly pine. Plant Mol. Biol. **40,** 555-565.

Li, L., Popko, J.L., Umezawa, T., and Chiang, V.L. (2000). 5-Hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosynthesis in angiosperms. J. Biol. Chem. **275,** 6537-6545.

Li, L., Cheng, X.F., Leshkevich, J., Umezawa, T., Harding, S.A., and Chiang, V.L. (2001). The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. Plant Cell **13,** 1567-1585.

Li, L., Popko, J.L., Zhang, X.-H., Osakabe, K., Tsai, C.-J., Joshi, C.P., and Chiang, V.L. (1997). A novel multifunctional *O*-methyltransferase implicated in a dual methylation pathway associated with lignin biosynthesis in loblolly pine. Proc. Natl. Acad. Sci. USA **94,** 5461-5466.

Lindermayr, C., Fliegmann, J., and Ebel, J. (2003). Deletion of a single amino acid residue from different 4-coumarate-CoA ligases from soybean results in the generation of new substrate specificities. J. Biol. Chem. **278,** 2781-2786.

Lindermayr, C., Möllers, B., Fliegmann, J., Uhlmann, A., Lottspeich, F., Meimberg, H., and Ebel, J. (2002). Divergent members of a soybean (*Glycine max* L.) 4-coumarate:coenzyme A ligase gene family. Primary structures, catalytic properties, and differential expression. Eur. J. Biochem. **269,** 1304-1315.

Lindsay, W.P., McAlister, F.M., Zhu, Q., He, X.-Z., Dröge-Laser, W., Hedrick, S., Doerner, P., Lamb, C., and Dixon, R.A. (2002). KAP-2, a protein that binds to the H-box in a bean chalcone synthase promoter, is a novel plant transcription factor with sequence identity to the large subunit of human Ku autoantigen. Plant Mol. Biol. **49,** 503-514.

Loake, G.J., Faktor, O., Lamb, C.J., and Dixon, R.A. (1992). Combination of H-box [CCTACC$(N)_7$CT] and G-box [CACGTG] cis elements is necessary for feed-forward stimulation of a chalcone synthase promoter by the phenylpropanoid-pathway intermediate *p*-coumaric acid. Proc. Natl. Acad. Sci. USA **89,** 9230-9234.

Logemann, E., Parniske, M., and Hahlbrock, K. (1995). Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. Proc. Natl. Acad. Sci. USA **92,** 5905-5909.

Logemann, E., Reinold, S., Somssich, I.E., and Hahlbrock, K. (1997). A novel type of pathogen defense-related cinnamyl alcohol dehydrogenase. Biol. Chem. **378,** 909-913.

Lois, R., Dietrich, A., Hahlbrock, K., and Schulz, W. (1989). A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive *cis*-acting elements. EMBO J. **8,** 1641-1648.

Lu, S.X., and Hrabak, E.M. (2002). An *Arabidopsis* calcium-dependent protein kinase is associated with the endoplasmic reticulum. Plant Physiol. **128,** 1008-1021.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26,** 1107-1115.

Mauch-Mani, B., and Slusarenko, A.J. (1996). Production of salicylic acid precursors is a major function of phenylalanine ammonia-lyase in the resistance of *Arabidopsis* to *Peronospora parasitica*. Plant Cell **8,** 203-212.

Maurer-Stroh, S., Eisenhaber, B., and Eisenhaber, F. (2002). N-terminal *N*-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J. Mol. Biol. **317,** 541-557.

Maury, S., Geoffroy, P., and Legrand, M. (1999). Tobacco *O*-methyltransferases involved in phenylpropanoid metabolism. The different caffeoyl-coenzyme A/5-hydroxyferuloyl-coenzyme A 3/5-*O*-methyltransferase and caffeic acid/5-hydroxyferulic acid 3/5-*O*-methyltransferase classes have distinct substrate specificities and expression patterns. Plant Physiol. **121,** 215-223.

Mavandad, M., Edwards, R., Liang, X., Lamb, C.J., and Dixon, R.A. (1990). Effects of *trans*-cinnamic acid on expression of the bean phenylalanine ammonia-lyase gene family. Plant Physiol. **94,** 671-680.

Maxwell, C.A., Harrison, M.J., and Dixon, R.A. (1993). Molecular characterization and expression of alfalfa isoliquiritigenin 2'-*O*-methyltransferase, an enzyme specifically involved in the biosynthesis of an inducer of *Rhizobium meliloti* nodulation genes. Plant J. **4,** 971-981.

Meyer, K., Cusumano, J.C., Somerville, C., and Chapple, C.C.S. (1996). Ferulate-5-hydroxylase from *Arabidopsis thaliana* defines a new family of cytochrome P450-dependent monooxygenases. Proc. Natl. Acad. Sci. USA **93,** 6869-6874.

Meyer, K., Shirley, A.M., Cusumano, J.C., Bell-Lelong, D.A., and Chapple, C. (1998). Lignin monomer composition is determined by the expression of a cytochrome P450-dependent monooxygenase in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **95,** 6619-6623.

Meyermans, H., Morreel, K., Lapierre, C., Pollet, B., De Bruyn, A., Busson, R., Herdewijn, P., Devreese, B., Van Beeumen, J., Marita, J.M., Ralph, J., Chen, C., Burggraeve, B., Van Montagu, M., Messens, E., and Boerjan, W. (2000). Modification in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A *O*-methyltransferase, an enzyme involved in lignin biosynthesis. J. Biol. Chem. **275,** 36899-36909.

Mizutani, M., Ohta, S., and Sato, R. (1997). Isolation of a cDNA and a genomic clone encoding cinnamate 4-hydroxylase from *Arabidopsis* and its expression manner in planta. Plant Physiol. **113,** 755-763.

Nair, R.B., Joy, R.W.I., Kurylo, E., Shi, X., Schnaider, J., Datla, R.S.S., Keller, W.A., and Selvaraj, G. (2000). Identification of a CYP84 family of cytochrome P450-dependent mono-oxygenase genes in *Brassica napus* and perturbation of their expression for engineering sinapine reduction in the seeds. Plant Physiol. **123,** 1623-1634.

Nair, R.B., Xia, Q., Kartha, C.J., Kurylo, E., Hirji, R.N., Datla, R., and Selvaraj, G. (2002). *Arabidopsis* CYP98A3 mediating aromatic 3-hydroxylation. Developmental regulation of the gene, and expression in yeast. Plant Physiol. **130,** 210-220.

Nambara, E., and McCourt, P. (1999). Protein farnesylation: a greasy tale. Curr. Opin. Plant Biol. **2,** 388-392.

Nedelkina, S., Jupe, S.C., Blee, K.A., Schalk, M., Werck-Reichert, D., and Bolwell, G.P. (1999). Novel characteristics and regulation of a divergent cinnamate 4-hydroxylase (CYP73A15) from French bean: engineering expression in yeast. Plant Mol. Biol. **39,** 1079-1090.

Neustaedter, D., Lee, S.P., and Douglas, C.J. (1999). A novel parsley *4CL cis*-element is required for developmentally regulated expression and protein-DNA complex formation. Plant J. **18,** 77-88.

O'Connell, A., Holt, K., Piquemal, J., Grima-Pettenati, J., Boudet, A., Pollet, B., Lapierre, C., Petit-Conil, M., Schuch, W., and Halpin, C. (2002). Improved paper pulp from plants with suppressed cinnamoyl-CoA reductase or cinnamyl alcohol dehydrogenase. Transgenic Res. **11,** 495-503.

Ohl, S., Hedrick, S.A., Chory, J., and Lamb, C.J. (1990). Functional properties of a phenylalanine ammonia-lyase promoter from *Arabidopsis*. Plant Cell **2,** 837-848.

Osakabe, K., Tsao, C.C., Li, L., Popko, J.L., Umezawa, T., Carraway, D.T., Smeltzer, R.H., Joshi, C.P., and Chiang, V.L. (1999). Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. Proc. Natl. Acad. Sci. USA **96,** 8955-8960.

Pakusch, A.-E., Matern, U., and Schiltz, E. (1991). Elicitor-inducible caffeoyl-coenzyme A 3-*O*-methyltransferase from *Petroselinum crispum* cell suspensions. Purification, partial sequence, and antigenicity. Plant Physiol. **95,** 137-143.

Parvathi, K., Chen, F., Guo, D., Blount, D.W., and Dixon, R.A. (2001). Substrate preferences of *O*-methyltransferases in alfalfa suggest new pathways for 3-*O*-methylation of monolignols. Plant J. **25,** 193-202.

Pellegrini, L., Geoffroy, P., Fritig, B., and Legrand, M. (1993). Molecular cloning and expression of a new class of ortho-diphenol-*O*-methyltransferases induced in tobacco (*Nicotiana tabacum* L.) leaves by infection or elicitor treatment. Plant Physiol. **103,** 509-517.

Piatigorsky, J., and Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. Science **252,** 1078-1079.

Pilate, G., Guiney, E., Holt, K., Petit-Conil, M., Lapierre, C., Leplé, J.-C., Pollet, B., Mila, I., Webster, E.A., Marstorp, H.G., Hopkins, D.W., Jouanin, L., Boerjan, W., Schuch, W., Cornu, D., and Halpin, C. (2002). Field and pulping performances of transgenic trees with altered lignification. Nature Biotechnol. **20,** 607-612.

Pinçon, G., Maury, S., Hoffmann, L., Geoffroy, P., Lapierre, C., Pollet, B., and Legrand, M. (2001). Repression of *O*-methyltransferase genes in transgenic tobacco affects lignin synthesis and plant growth. Phytochemistry **57,** 1167-1176.

Pinçon, G., Chabannes, M., Lapierre, C., Pollet, B., Ruel, K., Joseleau, J.-P., Boudet, A.M., and Legrand, M. (2001). Simultaneous down-regulation of caffeic/5-hydroxy ferulic acid-*O*-methyltransferase I and cinnamoyl-coenzyme A reductase in the progeny from a cross between tobacco lines homozygous for each transgene. Consequences for plant development and lignin synthesis. Plant Physiol. **126,** 145-155.

Pontier, D., Balagué, C., Bezombes-Marion, I., Tronchet, M., Deslandes, L., and Roby, D. (2001). Identification of a novel pathogen-responsive element in the promoter of the tobacco gene *HSR203J*, a molecular marker of the hypersensitive response. Plant J. **26,** 495-507.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nature Rev. Genet. **3,** 827-837.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MtaInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23 **4878-4884,** JR+.

Raes, J., and Van de Peer, Y. (1999). ForCon: a software tool for the conversion of sequence alignments. EMBnet.news **6,** (http://www.ebi.ac.uk/embnet.news/vol6_1).

Raes, J., Vandepoele, K., Saeys, Y., Simillion, C., and Van de Peer, Y. (2003). Investigating ancient duplication events in the *Arabidopsis* genome. J. Struct. Funct. Genom. **3,** 117-129.

Ralph, J., Lapierre, C., Marita, J.M., Kim, H., Lu, F., Hatfield, R.D., Ralph, S., Chapple, C., Franke, R., Hemm, M.R., Van Doorsselaere, J., Sederoff, R.R., O'Malley, D.M., Scott, J.T., Mackay, J.J., Yahiaoui, N., Boudet, A.-M., Pean, M., Pilate, G., Jouanin, L., and Boerjan, W. (2001). Elucidation of new structures in lignins of CAD- and COMT-deficient plants by NMR. Phytochemistry **57,** 993-1003.

Randall, S.K., and Crowell, D.N. (1999). Protein isoprenylation in plants. Crit. Rev. Biochem. Mol. Biol. **34,** 325-338.

Rasmussen, S., and Dixon, R.A. (1999). Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. Plant Cell **11,** 1537-1551.

Ro, D.K., Mah, N., Ellis, B.E., and Douglas, C.J. (2001). Functional characterization and subcellular localization of poplar (*Populus trichocarpa* (*Populus deltoides*) cinnamate 4-hydroxylase. Plant Physiol. **126,** 317-329.

Ruegger, M., Meyer, K., Cusumano, J.C., and Chapple, C. (1999). Regulation of ferulate-5-hydroxylase expression in *Arabidopsis* in the context of sinapate ester biosynthesis. Plant Physiol. **119,** 101-110.

Rushton, P.J., Reinstädler, A., Lipka, V., Lippok, B., and Somssich, I.E. (2002). Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. Plant Cell **14,** 749-762.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16,** 944-945.

Sablowski, R.W.M., Baulcombe, D.C., and Bevan, M. (1995). Expression of a flower-specific Myb protein in leaf cells using a viral vector causes ectopic activation of a target promoter. Proc. Natl. Acad. Sci. USA **92,** 6901-6905.

Sakai, T., Takahashi, Y., and Nagata, T. (1996). Analysis of the promoter of the auxin-inducible gene, *parC*, of tobacco. Plant Cell Physiol. **37,** 906-913.

Schiex, T., Moisan, A., and Rouzé, P. (2001). EUGÈNE: an eukaryotic gene finder that combines several sources of evidence. Lecture Notes in Computer Science **2066,** 111-125.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18,** 502-504.

Schoch, G., Goepfert, S., Morant, M., Hehn, A., Meyer, D., Ullmann, P., and Werck-Reichhart, D. (2001). CYP98A3 from *Arabidopsis thaliana* is a 3'-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. J. Biol. Chem. **276,** 36566-36574.

Séguin, A., Laible, G., Leyva, A., Dixon, R.A., and Lamb, C.J. (1997). Characterization of a gene encoding a DNA-binding protein that interacts *in vitro* with vascular specific *cis* elements of the phenylalanine ammonia-lyase promoter. Plant Mol. Biol. **35,** 281-291.

Seki, H., Ichinose, Y., Kato, H., Shiraishi, T., and Yamada, T. (1996). Analysis of *cis*-regulatory elements involved in the activation of a member of chalcone synthase gene family (*PsChs1*) in pea. Plant Mol. Biol. **31,** 479-491.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishiii, Y., Arakawa, T., Shibata, K., Shinagawa, A., and Shinozaki, K. (2002). Functional annotation of a full-length *Arabidopsis* cDNA collection. Science **296,** 141-145.

Sewalt, V.J.H., Ni, W., Blount, J.W., Jung, H.G., Masoud, S.A., Howles, P.A., Lamb, C., and Dixon, R.A. (1997). Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of L-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase. Plant Physiol. **115,** 41-50.

Shah, J., and Klessig, D.F. (1996). Identification of a salicylic acid-responsive element in the promoter of the tobacco pathogenesis-related ß-1,3-glucanase gene, *PR-2d*. Plant J. **10,** 1089-1101.

Shirley, A.M., McMichael, C.M., and Chapple, C. (2001). The *sng2* mutant of *Arabidopsis* is defective in the gene encoding the serine carboxypeptidase-like protein sinapoylglucose:choline sinapoyltransferase. Plant J. **28,** 83-94.

Simillion, C., Vandepoele, K., Van Montagu, M., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99,** 13627-13632.

Somers, D.A., Nourse, J.P., Manners, J.M., Abrahams, S., and Watson, J.M. (1995). A gene encoding a cinnamyl alcohol dehydrogenase homolog in *Arabidopsis thaliana*. Plant Physiol. **108,** 1309-1310.

Somssich, I.E., Wernert, P., Kiedrowski, S., and Hahlbrock, K. (1996). *Arabidopsis thaliana* defense-related protein ELI3 is an aromatic alcohol:$NADP^+$ oxidoreductase. Proc. Natl. Acad. Sci. USA **93,** 14199-14203.

Strack, D., and Sharma, V. (1985). Vacuolar localization of the enzymatic synthesis of hydroxycinnamic acid esters of malic acid in protoplasts from *Raphanus sativus* leaves. Physiol. Plant. **65,** 45-50.

Stuible, H.-P., Büttner, D., Ehlting, J., Hahlbrock, K., and Kombrink, E. (2000). Mutational analysis of 4-coumarate:CoA ligase identifies functionally important amino acids and verifies its close relationship to other adenylate-forming enzymes. FEBS Lett. **467,** 117-122.

Sugimoto, K., Takeda, S., and Hirochika, H. (2000). MYB-related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the tobacco retrotransposon *Tto1* and defense-related genes. Plant Cell **12,** 2511-2527.

Takeshita, N., Fujiwara, H., Mimura, H., Fitchen, J.H., Yamada, Y., and Sato, F. (1995). Molecular cloning and characterization of *S*-adenosyl-L-methionine:scoulerine-9-*O*-methyltransferase from cultured cells of *Coptis japonica*. Plant Cell Physiol. **36,** 29-36.

Tamagnone, L., Merida, A., Parr, A., Mackay, S., Culianez-Macia, F.A., Roberts, K., and Martin, C. (1998). The AmMYB308 and AmMYB330 transcription factors from Antirrhinum regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. Plant Cell **10,** 135-154.

Tavares, R., Aubourg, S., Lecharny, A., and Kreis, M. (2000). Organization and structural evolution of four multigene families in *Arabidopsis thaliana*: AtLCAD, AtLGT, AtMYST and AtHD-GL2. Plant Mol. Biol. **42,** 703-717.

Terashima, N., Fukushima, K., and Tsuchiya, S. (1986). Heterogeneity in formation of lignin. VII. An autoradiographic study on the formation of guaiacyl and syringyl lignin in poplar. J. Wood Chem. Technol. **6,** 495-504.

Terashima, N., Fukushima, K., and Takabe, K. (1986). Heterogeneity in formation of lignin. VIII. An autoradiographic study on the formation of guaiacyl and syringyl lignin in *Magnolia Kobus* DC. Holzforschung **40, Supp.,** 101-105.

Thompson, G.A.J., and Okuyama, H. (2000). Lipid-linked proteins in plants. Prog. Lipid Res. **39,** 19-39.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22,** 4673-4680.

Urban, P., Mignotte, C., Kazmaier, M., Delorme, F., and Pompon, D. (1997). Cloning, yeast expression, and characterization of the coupling of two distantly related *Arabidopsis thaliana* NADPH-cytochrome P450 reductases with P450 CYP73A5. J. Biol. Chem. **272,** 19176-19186.

Van de Peer, Y., and De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. **10,** 569-570.

Vandepoele, K., Simillion, C., and Van de Peer, Y. (2002). Detecting the undetectable: Uncovering duplicated segments in *Arabidopsis* through rice. Trends Genet. **18,** 606-608.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res. **12,** 1792-1801.

Vandepoele, K., Raes, J., De Veylder, L., Rouzé, P., Rombauts, S., and Inzé, D. (2002). Genome-wide analysis of core cell cycle genes in *Arabidopsis*. Plant Cell **14,** 903-916.

Vernon, D.M., and Bohnert, H.J. (1992). A novel methyl transferase induced by osmotic stress in the facultative halophyte *Mesembryanthemum crystallinum*. EMBO J. **11,** 2077-2085.

Wagner, G.J., and Hrazdina, G. (1984). Endoplasmic reticulum as a site of phenylpropanoid and flavonoid metabolism in *Hippeastrum*. Plant Physiol. **74,** 901-906.

Wanner, L.A., Mittal, S., and Davis, K.R. (1993). Recognition of the avirulence gene *avrB* from *Pseudomonas syringae* pv. *glycinea* by *Arabidopsis thaliana*. Mol. Plant-Microbe Interact. **6,** 582-591.

Wanner, L.A., Li, G., Ware, D., Somssich, I.E., and Davis, K.R. (1995). The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana*. Plant Mol. Biol. **27,** 327-338.

Whitbred, J.M., and Schuler, M.A. (2000). Molecular characterization of *CYP73A9* and *CYP82A1* P450 genes involved in plant defense in pea. Plant Physiol. **124,** 47-58.

Williamson, J.D., Stoop, J.M.H., Massel, M.O., Conkling, M.A., and Pharr, D.M. (1995). Sequence analysis of a mannitol dehydrogenase cDNA from plants reveals a function for the pathogenesis-related protein ELI3. Proc. Natl. Acad. Sci. USA **92,** 7148-7152.

Winkel-Shirley, B. (1999). Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. Physiol. Plant. **107,** 142-149.

Yamazaki, S., Sato, K., Suhara, K., Sakaguchi, M., Mihara, K., and Omura, T. (1993). Importance of the proline-rich region following signal-anchor sequence in the formation of correct conformation of microsomal cytochrome P-450s. J. Biochem. **114,** 652-657.

Ye, Z.-H., Kneusel, R.E., Matern, U., and Varner, J.E. (1994). An alternative methylation pathway in lignin biosynthesis in *Zinnia*. Plant Cell **6,** 1427-1439.

Yu, L.M., Lamb, C.J., and Dixon, R.A. (1993). Purification and biochemical characterization of proteins which bind to the H-box *cis*-element implicated in transcriptional activation of plant defense genes. Plant J. **3,** 805-816.

Zhang, X.-H., and Chinnappa, C.C. (1997). Molecular characterization of a cDNA encoding caffeoyl-coenzyme A 3-*O*-methyltransferase of *Stellaria longipes*. J. Biosci. **22,** 161-175.

Zhang, H., Wang, J., and Goodman, H.M. (1997). An *Arabidopsis* gene encoding a putative 14-3-3-interacting protein, caffeic acid/5-hydroxyferulic acid *O*-methyltransferase. Biochim. Biophys. Acta **1353,** 199-202.

Zhong, R., Morrison, W.H.I., Negrel, J., and Ye, Z.-H. (1998). Dual methylation pathways in lignin biosynthesis. Plant Cell **10,** 2033-2046.

Zou, J., and Taylor, D.C. (1994). Isolation of an *Arabidopsis thaliana* cDNA homologous to parsley (*Petroselinum crispum*) *S*-adenosyl-L-methionine:*trans*-caffeoyl-coenzyme A 3-*O*-methyltransferase, an enzyme involved in disease resistance. Plant Physiol. Biochem. **32,** 423-427.

[Chapter 4]

Genome-wide structural annotation and evolutionary analysis of the type I MADS-box genes in plants

Stefanie De Bodt[1], Jeroen Raes[1], Kobe Florquin[1], Stephane Rombauts[1], Pierre Rouzé[1,2], Günter Theissen[3] and Yves Van de Peer[1*]

[1] Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium,
[2] Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium,
[3] University of Jena, Lehrstuhl for Genetics, Philosophenweg 12, D-07743 Jena, Germany
[*] *Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*

Published in: Journal of Molecular Evolution **56**, 573-586 (2003)

Abstract

The type I MADS-box genes constitute a largely unexplored subfamily of the extensively studied MADS-box gene family, well known for its role in flower development. Genes of the type I MADS-box subfamily possess the characteristic MADS-box but are distinguished from type II MADS-box genes by the absence of the keratin-like box. In this *in silico* study, we have structurally annotated all 47 members of the type I MADS-box gene family in *Arabidopsis thaliana* and exerted a thorough analysis of the C-terminal regions of the translated proteins. On the basis of conserved motifs in the C-terminal region, we could classify the gene family into three main groups, two of which could be further subdivided. Phylogenetic trees were inferred in order to study the evolutionary relationships within this large MADS-box gene subfamily. These suggest for plant type I genes a dynamic of evolution that is significantly different from both the mode of animal type I (SRF) and plant type II (MIKC-type) gene phylogeny. The presence of conserved motifs in the majority of these genes, the identification of *Oryza sativa* MADS-box type I homologs, and the detection of expressed sequence tags for *Arabidopsis thaliana* and other plant type I genes suggest that these genes are indeed of functional importance to plants. It is therefore the more intriguing that, from an experimental point of view, almost nothing is known about the function of these MADS-box type I genes.

Introduction

The MADS-box gene family encodes a family of transcription factors involved in diverse aspects of plant development, and has been designated by an acronym (Schwarz-Sommer et al. 1990) after a few of its earliest members, namely ***M****CM1* found in yeast (Passmore et al. 1988), ***A****GAMOUS* in *Arabidopsis thaliana* (Yanofsky et al. 1990), ***D****EFICIENS* in *Antirrhinum majus* (Sommer et al. 1990; Schwarz-Sommer et al. 1992) and ***S****RF* in human (Norman et al. 1988). All MADS-box genes encode a strongly conserved MADS domain - found in the N-terminal region, that is responsible for DNA binding to $CC(A/T)_6GG$ boxes in the regulatory region of their target genes (Shore and Sharrocks 1995). Recent analyses have shown that this large gene family can be divided into two major lineages, named type I and type II (Alvarez-Buylla et al. 2000a). Since both type I and type II genes are found in plants, animals and fungi, both types of MADS-box genes are assumed to have originated by duplication before the divergence of these kingdoms. Based on the structure of the MADS domain, type I and type II genes are also referred to as MADS SRF-like and MADS MEF2-like genes, respectively (Alvarez-Buylla et al. 2000a).

In animals, type I genes are involved in response to growth factors while type II genes are involved in muscle development (Norman et al. 1988; Yu et al. 1992). Besides the highly conserved MADS domain, animal type I (SRF-like) and type II (MEF2-like) genes contain an additionally conserved region, the SAM and MEF2 domain, respectively (Shore and Sharrocks 1995, Riechmann and Meyerowitz 1997; Alvarez-Buylla et al. 2000a). The same is true for Fungi.

Plant type II MADS-box genes possess a strongly conserved MEF2-like MADS box, followed by a weakly conserved I (intervening) box, a K (keratin-like) box, and a C box and are therefore termed the MIKC-type (short: MIKC) genes (Münster et al. 1997). The moderately conserved K domain has been shown to be important for protein-protein interactions and probably forms a coiled-coil structure. The poorly conserved carboxyl-terminal (C) region may function as a trans-activation domain (Riechmann and Meyerowitz 1997). Plant type II MADS-box genes have been extensively studied during the last decade and are best known for their role in flower development (see e.g. Riechmann and Meyerowitz 1997; Pelaz et al. 2000; Theissen et al. 2000; Ng and Yanofsky 2001; Theissen 2001; Theissen and Saedler 2001). Besides this role, MADS-box genes also have an important function in the development of other plant organs such as fruit (Liljegren et al. 1998, 2000), roots (Zhang and Forde 2000; Alvarez-Buylla et al. 2000b; Burgeff et al. 2002) and ovules (Angenent and Colombo 1996). The type II MADS-box transcription factors provide an excellent genetic toolkit to study the evolution of plant development. Alterations in the expression of genes coding for transcriptional regulators, such as MADS-box genes, are emerging as a major source of the diversity and change that underlie evolution and can be linked to changes in plant body plan or the generation of evolutionary novelties (Riechmann et al. 2000; Theissen 2001).

Unlike the type II MADS-box genes in plants, the type I subfamily has remained largely unexplored. Plant type I MADS-domain proteins are characterized by an SRF-like MADS domain but the C-terminal region of these genes is still not well defined and is of variable length. Furthermore, type I genes are characterized by the absence of the well-defined K box.

Based on phylogenetic tree inference, Alvarez-Buylla et al. (2000a) concluded that this K box arose in plant type II genes after the divergence of plants and animals and fungi. Hitherto, only a few members of this subfamily have been identified by 'in silico' prediction in *Arabidopsis thaliana* whereas their function remains completely unknown (Alvarez-Buylla et al. 2000a). The recent discovery of this new subfamily of MADS-box genes in *Arabidopsis thaliana* and the lack of knowledge about their function urges upon the full characterization of this gene family in *Arabidopsis thaliana* and the identification of homologs in other plants. Moreover, further analysis of the type I MADS-box gene family may be very important in understanding the origin and evolution of the whole MADS-box gene family. In this respect, we have analysed the size and the structural characteristics of the type I subfamily in *Arabidopsis thaliana* and have identified the first type I MADS-box genes in *Oryza sativa*. The completion of the *Arabidopsis thaliana* genome sequence (Arabidopsis Genome Initiative 2000) allows investigation of the full complement of MADS box type I genes in this model plant. The structural annotation of the gene family was done in a semi-automated way, combining high throughput gene prediction with a manual control step. By using this approach we tried to combine speed with accuracy because future research on these sequences depends on the correctness of their annotation. Additionally, we performed a phylogenetic analysis of the type I subfamily of MADS-box genes in order to study the evolutionary relationships between the newly annotated genes.

Methods

Structural annotation of type I MADS-box genes

The annotation of the type I MADS-box gene family in *Arabidopsis thaliana* was based on homology searches with the conserved part of the genes of the family. Hence, the MADS domain of the type I MADS-domain proteins identified by Alvarez-Buylla et al. (2000) was used as a query sequence in BLAST (tblastn using default parameters) searches (Altschul et al. 1990) against the sequences of the *Arabidopis* genome. The E-value cut-off was initially set at 1e-10, where hits with higher E-value were selected manually, taking into account the conserved, possibly functionally important residues in the MADS-domain. The genomic sequences containing putative type I MADS-box genes were subjected to gene prediction using GeneMark.hmm (Lukashin and Borodovsky 1998). A manual control step of the annotation involved the inspection of the exon-intron structure and the multiple alignment of the MADS-domain protein sequences using Artemis (Rutherford et al. 2000) and BioEdit (Hall 1999). Based on similarity with close relatives of the gene family, wrongly predicted exon borders and over- or underprediction of exons were detected and corrected. To identify more distantly related proteins, we also constructed a HMMer profile (Eddy 1998) based on the already predicted and manually corrected genes. This profile was used to search a non-redundant database containing a collection of *Arabidopsis thaliana* proteins found through prediction with GeneMark.hmm (Lukashin and Borodovsky 1998) on the *Arabidopsis thaliana* genome (genome version of January 18 2001 (v180101), and downloaded from the MIPS ftp-site at ftp://ftpmips.gsf.de/cress/).

These gene predictions were then again manually checked. Additionally, we searched for type I MADS-domain proteins in *Oryza sativa*. Based on the multiple sequence alignment of the *Arabidopsis thaliana* type I MADS-domain proteins, a HMMer profile (Eddy 1998) was built to search a rice protein database for type I MADS-domain proteins. This database contained 24,305 rice proteins predicted with GeneMark.hmm (Lukashin and Borodovsky 1998) on rice BAC sequences from the Rice Genome Project covering approximately 29% of the rice genome (*Oryza sativa* spp. *japonica*; Sasaki and Burr 2000; http://rgp.dna.affrc.go.jp/). Furthermore, we screened the draft sequence of *Oryza sativa* spp. *indica* (Yu et al. 2002) for putative type I MADS-box genes using BLAST, with other type I genes as query sequences.
Duplicated blocks (i.e. large regions of colinearity) in the *Arabidopsis thaliana* genome were detected and dated as described earlier (Raes et al. 2002; Vandepoele et al. 2002).

Structural analysis of the C-terminal region

All type I MADS-box genes possess the strongly conserved MADS-box. However, the C-terminal region of these genes is much less conserved and has a variable length. We performed a motif search on all type I MADS-domain protein sequences using MEME (Multiple Expectation Minimization for Motif Elicitation) version 3.0 (Bailey and Elkan 1994). Based on the conserved motifs found by MEME, the type I MADS-box gene family was further subdivided into smaller subgroups after which these subgroups were realigned, now taking into account additional sites that could be proven to belong to shared and conserved motifs.
A HMMer profile (Eddy 1998) was built from the different motifs identified by MEME (see Results). These profiles were scanned against our in-house *Arabidopsis thaliana* protein database (see Structural annotation of type I MADS-domain genes) and the MIPS protein database to search for other proteins that contain similar motifs. The InterPro database (release 4.0, Nov. 2001, Apweiler et al. 2001) was also checked for the presence of the C-terminal motifs.
In order to make sure that no type II MADS-domain proteins have been included in our data set, all sequences were analyzed for the presence of the type II specific K-domain using InterPro searches (release 4.0, Nov. 2001, Apweiler et al. 2001) and Multicoil (Wolf et al. 1997) for coiled-coil prediction based on the presence of heptat-repeat signature motifs (abcdefg in which a and d are hydrophobic residues and are pointing to the core of the coiled-coil and b, d, e, f and g are hydrophylic residues) in the sequences (Lupas 1996).

Phylogenetic Analysis of Type I MADS-domain Proteins

The complete alignment of all type I MADS-domain proteins was edited and reformatted for phylogenetic analysis using BioEdit (Hall 1999) and ForCon (Raes and Van de Peer 1999) resulting in an alignment of the conserved residues (MADS-domain + residues of shared motifs). Neighbor-Joining (Saitou and Nei 1987) trees were constructed using TREECON (Van de Peer and De Wachter 1997) based on Poisson-corrected distances. To assess support for the inferred relationships, 500 bootstrap samples (Felsenstein 1985) were generated.

Maximum likelihood trees were constructed for type I MADS-box genes (see below) using TREE-PUZZLE 5.0 (Strimmer and von Haeseler 1996; Schmidt et al. 2002) and PAML (Yang 2000). In TREE-PUZZLE, the mutation probability matrix of Müller and Vingron (2000) was used whereas the number of puzzling steps was set to 20,000. Bootstrapped maximum parsimony trees for class M and class N genes were constructed with PAUP* (Swofford 1998). Predicted sequences and multiple alignments are available from our website at http://www.psb.ugent.be/bioinformatics/MADS/.

Results

Structural annotation and phylogenetic analysis

Based on a genome wide analysis, we identified 47 type I MADS-box genes in the genome of *Arabidopsis thaliana*, of which 14 correspond to genes previously described by Alvarez-Buylla et al. (2000a) and of which 33 are new (see Table 1). Additionally, we discovered the presence of a new group of MADS-like genes. These genes are different from type I (and also type II) MADS-box genes due to a highly divergent N-terminal region of the MADS-box. Furthermore, although most of these genes are overall strongly conserved, they do not possess the C-terminal conserved regions characteristic for type I (or type II) genes. For these reasons, we did not include these genes (listed in Table 2) in our analyses.

Table 1. *Arabidopsis thaliana* and *Oryza sativa* type I MADS-box genes

Locus name	Gene	Accession number[c]	Start	Stop	Length	Strand	Chr.	EST	Class
At1g28460		AC010155	35082	35630	182	-	1		M
At1g28450		AC010155_2	37337	37894	185	+	1		M
At1g60880		AC018908_2	24777	25352	191	-	1		M
At1g60920		AC018908_1	6660	7265	201	+	1		M
At3g04100		AC016829	84782	85405	207	+	3		M
At1g01530	AGL28	Y12776	6766	7788	247	+	1		M
At1g65360	AGL23	AC004512_2	47399	48213	226	+	1		M
At2g24840		AC006585	25227	25859	210	+	2		M
At5g60440		AB011483	26829	28020	299	+	5		M
At4g36590	AGL40	AL161589	121429	123079	243	-	4		M
At5g38620		AB005231	463826	464875	349	-	5		M
At5g49420		AB023034	34638	36134	402	-	5		M
At2g34440	AGL29	AC004077	16781	17299	172	+	2		M
At1g48150		AC023673	497767	498738	323	+	1		M
At5g27130	AGL39	AF007271	71901	75618	435	-	5		M
At1g47760		AC012463	70240	70948	184	-	1		M
At3g66656		AC036106	29224	29760	178	-	3		M
At4g14530		AL161539	46973	47658	213	-	4		M
At5g49490		AB023033	10587	11330	247	+	5		M
At5g04640		AL162875	89521	90489	322	+	5		M
		Os_AP003951_1	28733	29365	633	-	6		M
		Os_AP003951_2	50199	50771	572	-	6		M
		Os_AP003627	102168	102794	627	+	1		M
		Os_AP004093	72268	73128	861	+	2		M
		Os_Contig2417	5705	9967	210	+			M
		Os_Contig4095	1453	2109	218	+			M
		Os_Contig4276	6289	6921	210	+			M
		Os_Contig28459	1540	2078	141	-			M
		Os_Contig18609	465	1049	194	+			M
At5g26580/At5g26575[b]		AF058914	4471	5508	304	+	5		N
At5g26630/At5g26625[b]		AF058914_2	40425	47737	315	-	5		N
At5g26650/At5g26645[b]		AF058914_3	53688	54794	327	-	5	X	N
At1g65330		AC004512_1	32543	33382	279	-	1		N

Table 1 (continued).

Locus name	Gene	Accession number[c]	Start	Stop	Length	Strand	Chr.	EST	Class
At1g65300		AC004512_3	21003	21827	278	+	1		N
At3g05860		AC012393	56275	57224	260	-	3		N
At2g28700		AC007184	3732	4502	256	-	2		N
At5g27960		AC007627	64277	65368	363	-	5		N
At5g48670		AB015468	59946	60911	321	-	5		N
At1g31630		AC074360_2	59951	60970	339	+	1		N
At1g31640		AC074360_1	55322	56806	464	+	1		N
At2g40210		AC018721	40137	42106	402	-	2	X	N
At2g26880	AGL41	AC005168	51386	52188	260	-	2		N
		Os_AP002070_1	57225	57947	240	+	1		N
		Os_AP002070_2	71207	72127	306	+	1		N
		Os_Contig28311	1167	1580	138	-			N
		Os_Contig603	3973	4776	267	+			N
		Os_Contig23118	1479	1904	141	+			N
		Os_Contig18573	850	1479	209	+			N
		Os_Contig119850	52	667	205	-			N
		Os_Contig31610	1805	2215	136	-			N
		Os_Contig18149	1420	2035	205	-			N
At2g03060[a]	AGL30	AC004138	81852	83914	364	+	2		O
At1g31140		AC004793	29171	30813	211	+	1		O
At1g22590		AC006551	24033	24524	163	-	1	X	O
At1g77950		AC009243	50294	52808	244	+	1	X	O
At1g72350		AC016529	73282	73956	224	+	1		O
At1g17310		AC026479	2189	2827	212	-	1		O
At5g26950	AGL26	AF007270	84574	85554	292	-	5		O
At2g26320	AGL33	AC004484	66595	68743	209	-	2		O
At1g18750[a]		AC011809	60126	62604	440	+	1		O
At1g22130		AC069252	2812402	2814274	335	-	1		O
At1g77980		AC009243_2	58477	60332	303	-	1		O
At1g69540		AC073178	88592	90480	359	-	1		O
At5g06500		AP002543	7047	7775	728	+	5		O
At5g58890	AGL43	AB016885	33758	34642	294	+	5		O
At5g55690		AB009050	40372	41205	277	-	5		O
		Os_AP000616	39576	42209	855	+	6		O
		Os_AP003104	53129	54943	1815	+	1		O
		Os_AP003331_1	86944	88167	1224	+	1		O
		Os_AP003331_2	89653	90947	975	+	1		O
		Os_AP003380	8256	9365	1110	-	1		O
		Os_AP003436	171451	172890	1440	-	1		O
		Os_AP003763	127881	128597	279	+	6		O
		Os_AP003742	63104	64429	645	-	7		O
		Os_AP004322	4659	10836	477	+	6		O
		Os_AP003331_3	95818	98653	1188	+	1		O
		Os_Contig19550	853	1324	375	+			O
		Os_Contig52002	790	?	?	+			O
		Os_Contig20368	?	405	?	-			O
		Os_Contig45237	1180	?	?	+			O
		Os_Contig11428	5081	?	?	+			O
		Os_Contig32902	2555	?	?	+			
		Os_Contig2175	12725	?	?	-			Unassigned[d]
		Os_Contig5668	?	5842	?	-			
		Os_Contig21589	?	2624	?	-			

[a] Locus names of genes in MIPS (Schoof et al. 2002) that differ in their structural annotation with those presented here.
[b] Genes on BAC AF058914 have different locus names in MIPS and TIGR (http://www.tigr.org) respectively.
[c] Rice genes are in bold.
[d] These genes could not be classified unambiguously because the prediction was incomplete (see text for details).

Figure 1 shows the distribution of the type I MADS-box genes on the different chromosomes. Seven genes could be linked to block duplications, namely both the gene pairs AC016529 and AC026479, and the gene pairs AC009243_2 and AC069252, which are all located in an internally duplicated block that contains 172 duplicated genes on chromosome 1 (Raes et al. 2002; Simillion et al. 2002). Additionally, genes AC012393 and AF058914_2 (and its neighbour AF058914_3) belong to a smaller block of 13 genes duplicated between chromosomes 3 and 5 (Fig. 1). The largest block has been dated 69 ± 17 MYA, while the smaller block duplication was dated 78 ± 29 MYA, which implies that they could have both originated during the same complete genome duplication event, estimated to have occurred at around that time (Lynch and Conery 2000; Raes et al. 2002; Simillion et al. 2002).

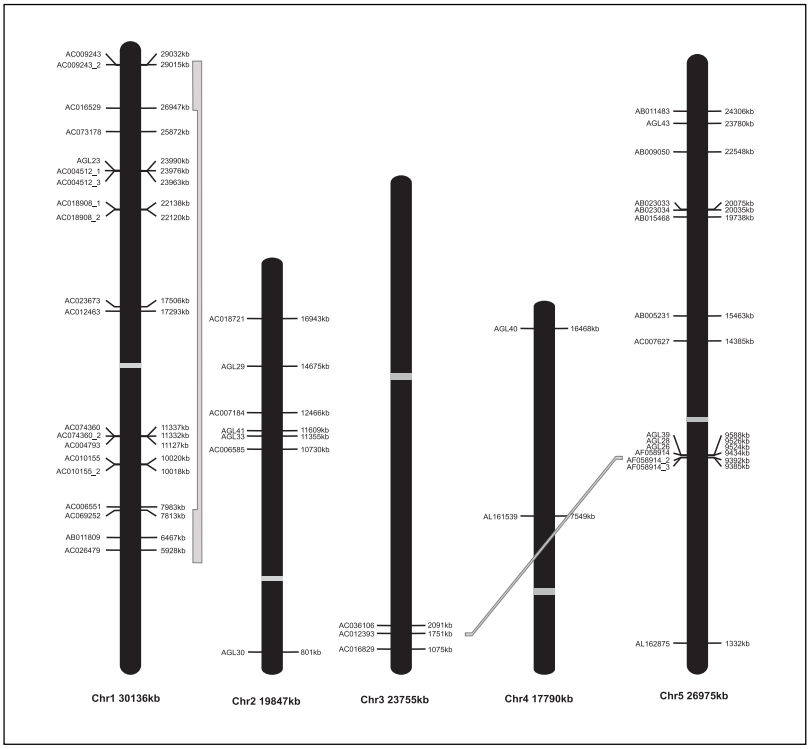

Figure 1. Chromosomal localization of the type I MADS-box genes in *Arabidopsis thaliana*. Grey bands denote duplicated blocks (see text for details).

Figure 2a shows the distribution of the number of exons found in type I MADS-box genes. As can be observed, the majority of the type I genes consist of only one or two exons, which is quite different from type II MADS-box genes, where most genes consist of 7 exons (Fig. 2b).

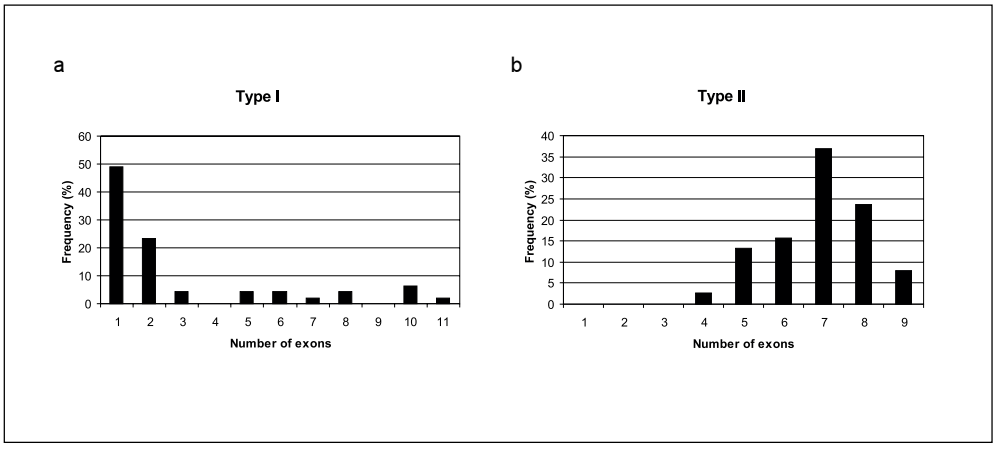

Figure 2. Distribution of the number of exons in the type I (a) and type II (b) MADS-box gene family.

In addition to the *Arabidopsis thaliana* type I genes 16 rice type I MADS-box genes were annotated on BAC sequences of the rice consortium (Sasaki and Burr 2000). Preliminary analysis of the draft sequence of rice resulted in the additional identification of 19 putative type I MADS-box genes. Six other genes were found through BLAST searches on the rice draft sequence but could not be ascribed unequivocally to the type I subfamily. Further analysis and manual annotation of these rice genes will be necessary to decide whether these are type I or type II genes. Furthermore, to improve gene prediction in rice, an assembly of the contigs of the draft sequence will be necessary because many MADS-box genes are located at the end of the contigs. We also searched the publicly available databases for type I MADS-box genes of other plants, but could not find any other type I homologs. It should be noted that the sequencing and annotation of other plant sequences is still ongoing which will probably result in the detection of many more type I MADS-domain proteins in the near future.

Table 2. List of MADS-like genes in *Arabidopsis thaliana*

Locus name	Accession number BAC
At5g27090	AF170760
At5g27070	AF170670
At5g27580	AC007478
At5g26950	AF007270
At4g11250	AL096882
At5g65330	AB011479
At5g40220	AB010699
At5g39750	AB016876
At5g38740	AB011478
At5g40120	AB010699
At5g39810	AB016876
At5g41200	AB010072
At3g18650	AB026654
At5g27050	AF170670
At1g60040	AC005966
At1g59810	AC007258

The construction of reliable phylogenetic trees of the complete type I subfamily of MADS-domain proteins is very difficult due to the small size (60 amino acids) of the conserved MADS-domain. Trees constructed on such a low number of residues often turn out to be unreliable and poorly supported by statistical analyses. As can be seen in Figure 3, very few nodes are well supported and no conclusion can be drawn about possible subclasses present in the type I MADS-box gene family. Therefore, we applied alternative approaches to resolve the phylogeny of the gene family (see also Methods).

Detailed structural analysis using MEME (Bailey and Elkan 1994) enabled us to discover several conserved motifs in the C-terminal region of the type I MADS-domain proteins (summarized in Figs. 4 and 5). Two main distinct classes of type I MADS-domain proteins, which we designate class M and class N, can be identified, each of which can be further subdivided. Class M possesses three types of genes, viz. type I M1 genes that are characterized by motifs 1 2, and 3; type I M2 genes characterized by motifs 1 and 3, and type I M3 genes which only contain motif 1 (Fig. 4).

Class N possesses three types of genes, viz. type I N1 genes that are characterized by motifs 4, 5, 6, 7, sometimes 8, and 9; type I N2 genes that possess motifs 4 and 5 and have a degenerated form of motif 6, and finally type I N3 genes that only contain motifs 4 and 5 (Fig. 5). Next to class M and class N genes, there is a third class O of genes that do not possess the same conservation in the C-terminal region as the proteins in the other classes. Thus, although specific motifs could be identified for class M and N genes, it was not possible to find any conserved motif for the proteins that we classified as belonging to class O. It should be noted that type I MADS-box genes of rice have been found for all three classes (see Fig. 1).

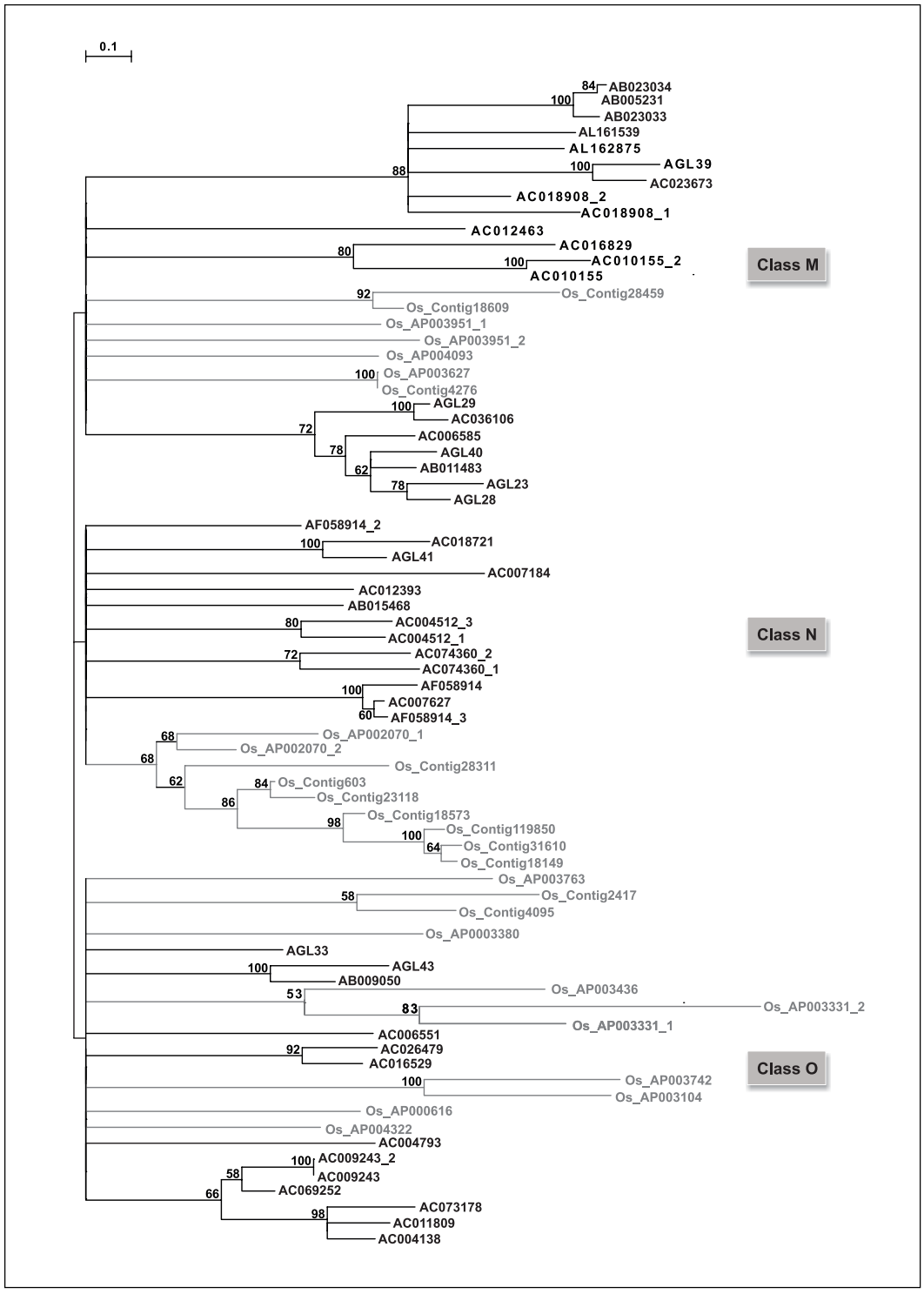

Figure 3. Phylogenetic distance tree of all type I MADS-box proteins identified in *Arabidopsis thaliana* and *Oryza sativa*. Tree construction was based on only 47 conserved residues in the MADS domain. Five hundred bootstrap samples (Felsenstein 1985) were taken and branches are drawn as unresolved when supported by less than 50%. Based on the presence or absence of C-terminal motifs, genes were ascribed to class M, N or O (see text for more details). Rice proteins are in indicated in grey. The scale indicates 0.1 substitutions per site.

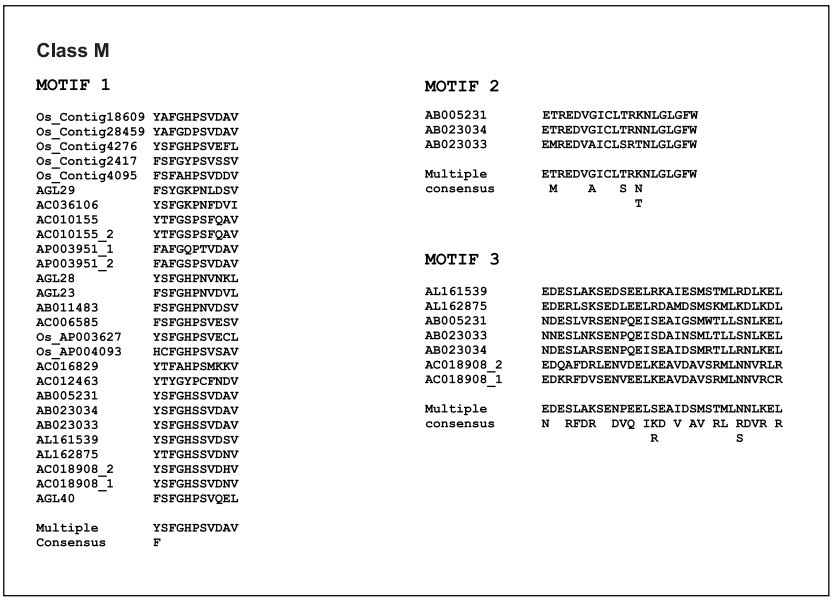

Figure 4. Conserved motifs in the C-terminal region of class M proteins of the type I MADS-box gene family found by MEME (Bailey and Elkan 1994). Rice genes are preceded by the prefix Os. Multiple consensus sequences are in bold. The multilevel consensus sequence is calculated from the motif position-specific probability matrix computed by MEME. For each column of the motif, the amino acid residues are sorted in decreasing order by the probability with which they are expected to occur at a certain position of the motif. The most probable amino acid is put on top. Only amino acids with probabilities of 0.2 or higher at that position in the motif are printed.

Classification of the type I MADS-box genes into classes M and N on the basis of the presence of certain conserved motifs allowed alignment of longer regions of the type I MADS-box genes. Therefore, a phylogenetic tree was constructed for genes belonging to class M from an alignment of 76 conserved residues, including the MADS domain and motif 1 (shared between all the genes belonging to class M), whereas a second tree for class N genes was constructed from an alignment of 116 conserved residues, based on the MADS domain and the motifs 4 and 5. These trees are shown in Figures 6 and 7, respectively. Both trees were artificially rooted based on the presence or absence of certain motifs.

As expected, in general there is a clear correlation between the tree topology and the structural characteristics of a group of proteins. In other words, proteins with the same C-terminal motif composition seem to be more closely related. In a few cases, remnants of common ancestry can be found, but the conservation was too low to be picked up by MEME. For example, genes of type I N2 do not contain motif 6 according to MEME, but some residues of the consensus sequence of this motif can still be recognized in these proteins. Therefore, these motifs are represented by dashed boxes (Figs. 6 and 7).

The trees shown in Figures 6 and 7 are neighbor-joining trees (Saitou and Nei 1987) based on Poisson corrected distances computed with TREECON (Van de Peer and De Wachter 1997).

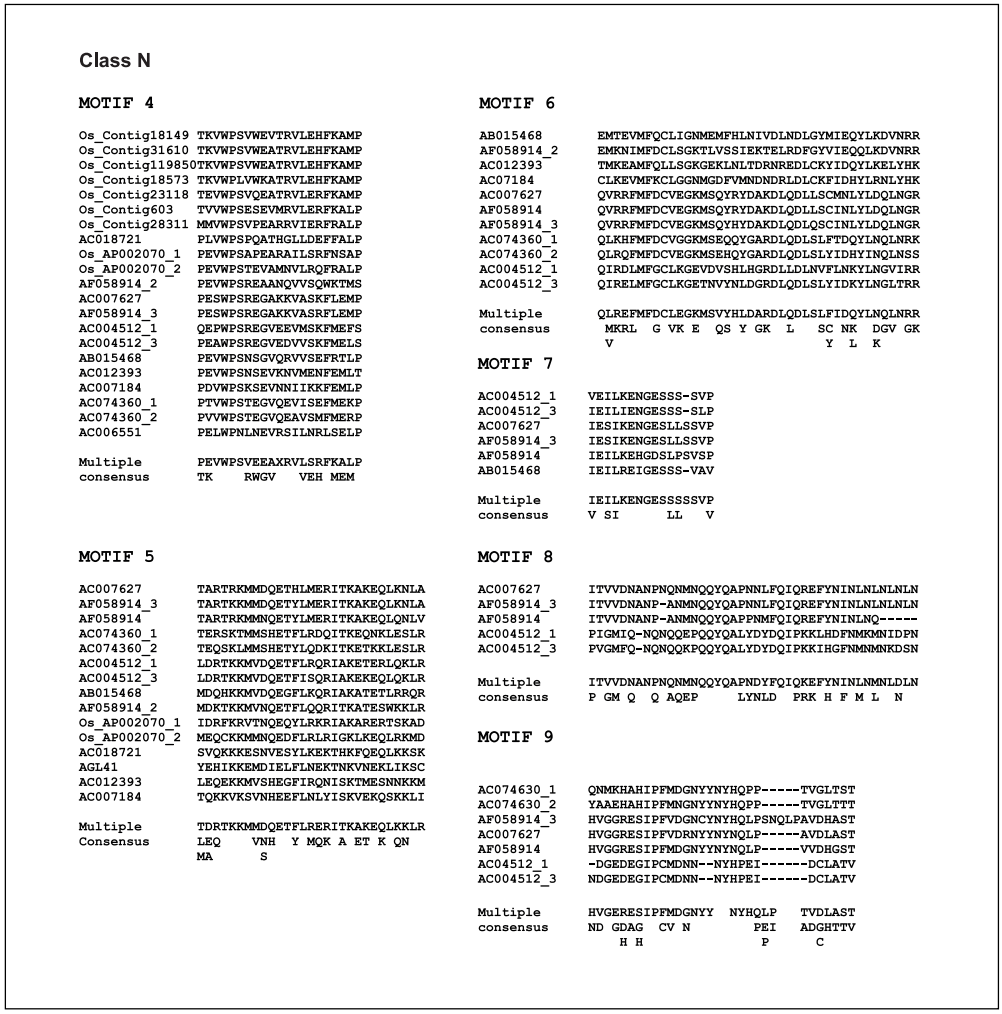

Figure 5. Conserved motifs in the C-terminal region of class N proteins of the type I MADS-box gene family found by MEME. Interpretation is as in Figure 4.

Overall, maximum likelihood trees and maximum parsimony trees gave similar results and differences were only observed for non-supported nodes. As expected, the resolution of the trees seems to be correlated with the number of residues that could be taken into account for tree inference.

The tree of class M genes, shown in Fig. 6 and based on 76 alignment positions, is still not very well resolved, apart from one subgroup of sequences that also contain additional conserved motifs (Type I M1 and Type I M2). Although strong conclusions cannot be drawn regarding the rice genes, due to the uncertainty of most branching orders, it seems that none of the rice genes is specifically related with any of the *Arabidopsis thaliana* genes. This is also observed in the tree of the N genes, based on 116 alignment positions, where the rice genes clearly form a monophyletic group, which is well supported by bootstrap analysis and with different methods of tree construction (Fig. 7).

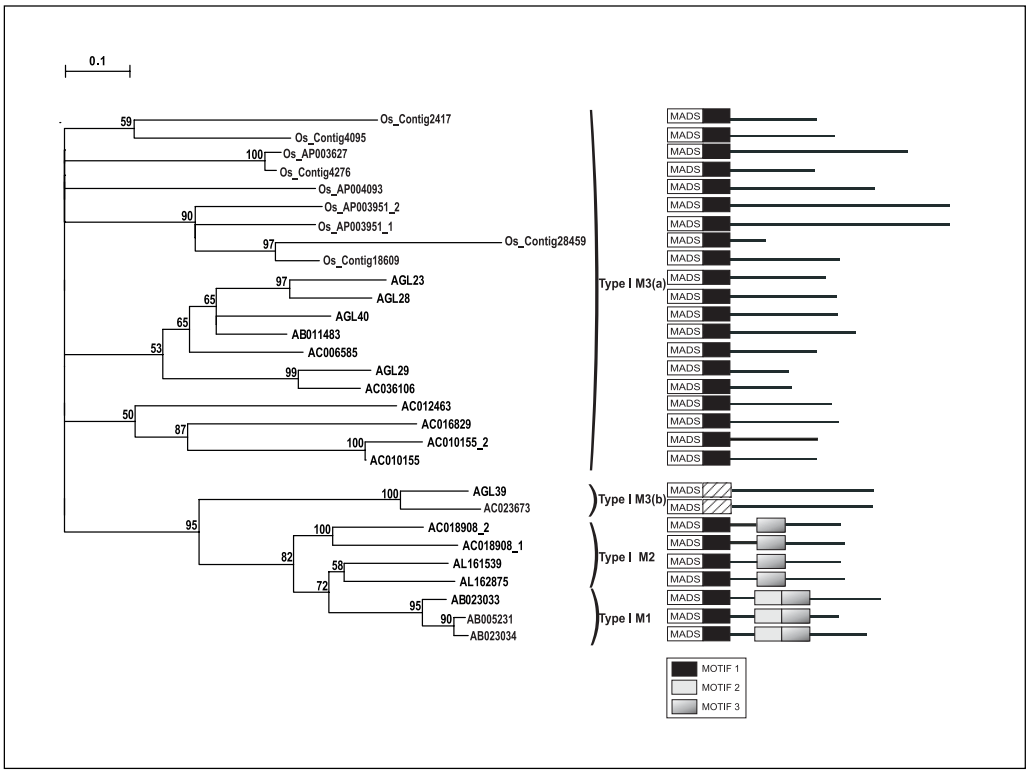

Figure 6. Pairwise distance tree of the type I MADS-box genes belonging to class M (see text for details), inferred from a sequence alignment including sites of to the MADS domain and motif 1. The motif composition of each gene is denoted by a black line (representing the length of the sequence) and coloured boxes. A dashed box denotes a degenerated form of the motif. Rice genes are preceded by the prefix Os. Interpretation of the scale is as in Figure 3.

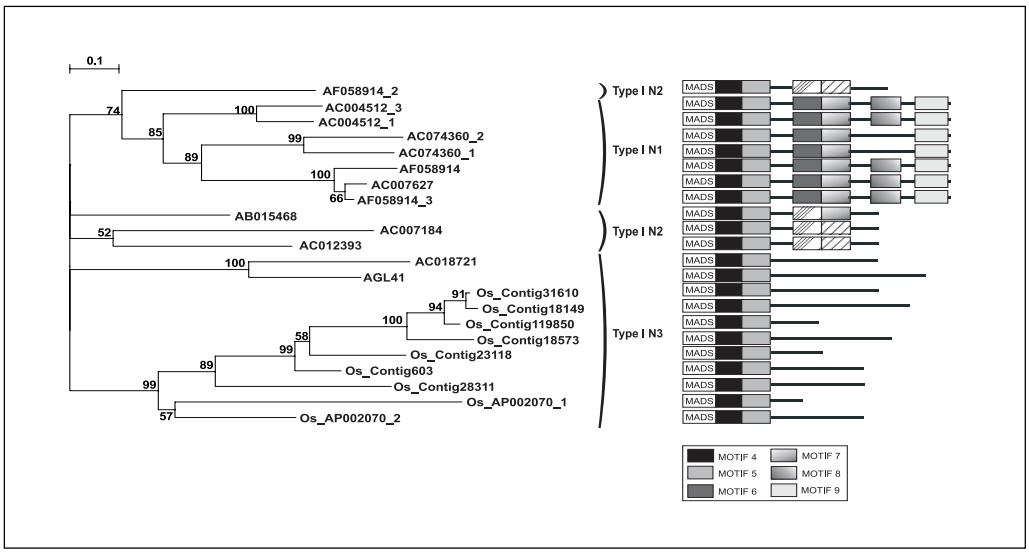

Figure 7. Pairwise distance tree of the type I MADS-box genes belonging to class N, inferred from a sequence alignment including sites of the MADS domain, and motifs 4 and 5. Interpretation is as in Figure 6. Interpretation of the scale is as in Figure 3.

Functional annotation

In order to assign a putative function to the type I MADS-box genes, we analyzed the C-terminal part of these genes in more detail. Genes that encode transcription factors often contain a transcription-activating domain. Three types of trans-activation domains are described in the literature: they are either rich in acidic residues, in proline residues, or in glutamine residues, but have low overall conservation on the primary structure level (Latchman 1998). Type I M1 and Type I M2 proteins contain an acidic region in their characteristic motif 3. Class N proteins all contain a proline-rich region, approximately starting from position 160. This region shows low conservation on the primary sequence level and does not correlate with any particular C-terminal motif designated by MEME. However, as stated before, the abundance of prolines in this region might possibly refer to the trans-activation domain of these proteins (Latchman 1998). However, apart from these putative trans-activation domains, little can be said about the C-terminal region. For example, no similarity could be found between the profiles inferred from the conserved motifs and any previously described motifs or domains (InterPro release 4.0, Nov. 2001; Apweiler et al. 2001).

Table 3. List of ESTs found for type I MADS-box genes in different plant species

EST	Gene	Plant species	Expression[a]
AV558219	AF058914_3	*Arabidopsis thaliana*	Organ: green siliques
AV823886	AC018721	*Arabidopsis thaliana*	Developmental stage: in various developmental stages from germination to mature seeds Treatment: dehydration and cold
AV787106	AC018721	*Arabidopsis thaliana*	Idem
AV787440	AC018721	*Arabidopsis thaliana*	Idem
AV788503	AC018721	*Arabidopsis thaliana*	Idem
AV784963	AC018721	*Arabidopsis thaliana*	Idem
AU238686	AC006551	*Arabidopsis thaliana*	Treatment: cold
Z37169	AC006551	*Arabidopsis thaliana*	Tissue type: green shoots
F13558	AC006551	*Arabidopsis thaliana*	Tissue type: green shoots
AU236968	AC009243	*Arabidopsis thaliana*	Organ: flowers and siliques
AV556667	AF058914	*Arabidopsis thaliana*	Organ: green siliques
BE610209		*Glycine max*	Tissue type: immature seed coats of greenhouse-grown plants
BE823841		*Glycine max*	from cDNA libraries from various tissues and stages of development of soybean that represent 2,639 sequences from immature cotyledons, 1,770 from immature seed coats, 3,938 from flowers, and 869 from young pods
AW508033		*Glycine max*	from a cDNA library that was constructed from mRNA isolated from immature cotyledons of greenhouse grown plants
BE054256		*Gossypium arboreum*	Tissue type: Fibers isolated from bolls harvested 7-10 dpa
BE999756		*Medicago truncatula*	Tissue type: senescent root nodules Developmental stage: mixture of effective nodules from 40 day old plants harvested 36 hours post shoot removal and nodules collected from 2-month-old plants at mid-pod stage
AW029842		*Lycopersicon esculentum*	Tissue type: callus Developmental stage: 25-40 days old
BI929334		*Lycopersicon esculentum*	Tissue type: flower Developmental stage: 3 to 8-mm buds
BG139571		*Lycopersicon pennellii*	Tissue type: pollen Developmental stage: pollen collected from open flowers
BJ247094		*Triticum aestivum*	Tissue type: spike at flowering date Developmental stage: Feekes' scale 10.5.1
BJ248139		*Triticum aestivum*	Idem
BJ218990		*Triticum aestivum*	Tissue type: spike at meiosis Developmental stage: Feekes' scale 9
BG525865		*Stevia rebaudiana*	Tissue type: leaf Developmental stage: field grown, mid-size
AW010840		*Pinus taeda*	Organ: shoot tips
BE643398		*Ceratopteris richardii*	Tissue type: gametophyte; cell type: spore Developmental stage: 20 hours after germination initiation
BJ184681		*Physcomitrella patens*	Tissue type: mixture of chloronemata, caulonemata, and malformed buds

[a] Expression details (e.g. tissue or organ, condition) are as described in the EMBL entries.

In order to get more information on the expression of type I MADS-box genes, and their possible functional annotation, we screened *Arabidopsis thaliana* ESTs, rice ESTs and an EST collection containing all publicly available ESTs from diverse plant species. However, the number of ESTs corresponding to type I MADS-box genes of *Arabidopsis thaliana* was extremely small (see Table 3), in particular in comparison with ESTs for type II genes where per gene, on average 4 to 5 ESTs could be identified. We found one EST (C99890) for type I gene AGL39 (type I M3(b)), which had also been identified previously by Alvarez-Buylla et al. (2000a) and ESTs for four other *Arabidopsis thaliana* genes. Some ESTs from other plant species could be found that were long enough to demonstrate unambiguously that they are ESTs from type I MADS-box genes (Table 3). ESTs of type I MADS-box genes are found in diverse plant species such as *Glycine max*, *Lycopersicon esculentum*, *Triticum aestivum*, and even in *Ceratopteris richardii* (a fern) and *Physcomitrella patens* (a moss).

Discussion

Detailed structural and evolutionary analysis of the type I subfamily of MADS-box genes suggest that these genes are indeed of functional importance in plants. The type I subfamily possesses 47 members which is more than the number of members of the very well studied type II subfamily (unpublished results). Moreover, in a first preliminary analysis, already 33 type I genes are identified in *Oryza sativa* spp. *japonica* on BAC sequences of the rice consortium (December 2001) and on the draft sequence of *Oryza sativa* spp*. indica*. Furthermore, *Arabidopsis thaliana* and rice type I proteins still have conserved common motifs in their C-terminal region (rice genes are present in the type I M3(a) and type I N3 classes). This conservation is most likely due to functional constraints on the C-terminal region, although the overall functional constraint within the type I genes has probably been lower than that within the type II genes. This is, amongst other things, supported by the higher evolutionary distances between type I MADS-box genes (Alvarez-Buylla et al. 2000a; our own observations).

Unfortunately, based on *in silico* analyses, we cannot assign a putative function to the type I MADS-domain proteins. The small number of ESTs found for type I MADS-box genes of different plant species can probably be attributed to the fact that most of the type I genes have a very low expression level, or that the genes are expressed under very specific conditions that are not yet monitored in EST-sequencing projects. Strikingly, nearly half of type I genes are intronless (Fig. 2). This gene structure could possibly be interpreted as a result of the evolutionary history of the type I genes through reverse transcription, with the possibility that many of them are inactive pseudogenes. However, it should be noted that gene AC006551, for which we found three ESTs, consists of only one exon, which argues that, at least some of these genes, are expressed and functional and not pseudogenes as put forward by Ng and Yanofsky (2001). In maize, transposon-like elements have been identified that have recently hijacked AGAMOUS-like (type II) MADS-boxes and distributed them through the maize genome (Fischer et al. 1995; Montag et al. 1995; Montag et al. 1996).

In order to investigate whether this could have been the case for the *Arabidopsis thaliana* type I genes, we looked for characteristic transposon-like elements in the flanking and coding regions of the type I genes.

To this end, we searched for similarity with known (retro)transposons and with proteins involved in their activity such as pol, gag, RT, etc. (Bennetzen 2000). However, no evidence for the presence of transposable elements could be found in our analyses.

As stated previously, all the type I MADS-box class N rice genes form a well supported monophyletic grouping, while a monophyletic origin of the rice class M genes can also not be ruled out on the basis of tree inference. If true, and provided that the root in Figs. 6 and 7 is placed correctly, this would suggest that the expansion of both the *Arabidopsis thaliana* and rice class M and N type I MADS-box genes (nothing can be said about genes from class O) occurred after the divergence of these two plants, somewhere between 150 and 200 MYA (Wikstrom et al. 2001). This is in clear contrast with observations in MADS type II phylogenies, according to which the last common ancestor of extant gymnosperms and angiosperms already contained at least seven different MIKC-type MADS-box genes (Becker et al. 2000). If type I MADS-box genes were present in the most recent common ancestor of plants, animals, and fungi, as suggested by Alvarez-Buylla et al. (2000a), and our observations are correct, this would imply that type I MADS-box genes may have remained low-copy (or even single-copy) for many hundreds of millions of years until the most recent common ancestor of *Arabidopsis thaliana* and rice, and then started to multiply independently, giving rise to high gene numbers in both *Arabidopsis thaliana* and rice. This seems highly unrealistic, given the evolutionary history of type II MADS-box genes (Becker et al. 2000; Krogan and Ashton 2000; Theissen et al. 2001). An alternative explanation could be that the type I genes from animals and plants are not monophyletic, i.e. that they originated two times independently in plants and animals, and, at least for plants, much more recently than previously suggested. In line with this, the type I genes from animals (SRF-like genes) have a structure which is significantly different from that of plant type I genes, and obvious sequence similarity between both gene types is restricted to the MADS-domain anyway (Alvarez-Buylla et al. 2000). Animal type I genes have an evolutionary history which is different from that of plant type I genes: while the gene number of the latter increased dramatically in the lineages that led to extant *Arabidopsis thaliana* and rice (this work), SRF seems to have remained a single copy gene throughout the more than 500 million years of animal evolution, and represents the evolutionary most conserved subfamily of MADS-box genes (Escalante and Sastre 1998; Hoffmann and Kroiher 2001; Scheffer et al. 1997). As already stated by Alvarez-Buylla et al. (2000), the type I MADS-box clade in plants is defined by only one putative synapomorphy while some synapomorphies are shared by all but one or a few sequences; this cannot be considered as strong proof for a monophyletic origin of type I MADS-box genes. On the other hand, it is possible that there are orthologous type I genes in *Arabidopsis thaliana* and rice, but that phylogeny reconstruction, due to the limited number of phylogenetically informative sites, is unable to correctly identify them. Probably, the identification of type I genes from other plants will be necessary to clarify this. This however not possible yet due to the limited amount of genomic data from other plant species.

Hopefully, as previously suggested by Riechmann and Ratcliffe (2000), *in silico* studies, about the annotation and classification of specific gene families, such as the one described here, can guide future experimental work and enhance the functional characterization of genes.

Acknowledgements

The authors want to thank Klaas Vandepoele and Cedric Simillion for technical help. S.D. and K.F. are indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) for a predoctoral fellowship. Annotated sequences have been submitted to the MAtDB (Schoof et al. 2002) and the TAIR (Huala et al. 2001) databases. Supplementary data will be available on http://www.psb.ugent.be/bioinformatics/MADS/.

Note added in proof

After acceptance, novel MADS-box genes were identifed in *Physcomitrella* [Henschel K, Kofuji R, Hasebe M, Saedler H, Munster T, Theissen G (2002) Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. Mol Biol Evol **19**, 801–814]. By including the MIKC* (type II) genes (PPM3, PPM4, PPMADS2, and PPMADS3) in our analysis, some of the Arabidopsis genes that we denoted as being of type I clustered with the *Physcomitrella* genes. Although these *Arabidopsis* genes did not seem to possess a conserved K-box (the reason why they were included), a relic of this box could be identifed through comparison with the very degenerated K-box found in *Physcomitrella*. Therefore, some of the genes (i.e., AC011809, AC073178, AC004138, AC069252, AC009243, AC009243_2, and AC004484; Fig. 1) should probably be classifed as type II rather than type I genes in our study.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389-3402.

Alvarez-Buylla, E.R., Liljegren, S.J., Pelaz, S., Gold, S.E., Burgeff, C., Ditta, G.S., Vergara-Silva, F., and Yanofsky, M.F. (2000). MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. Plant J **24,** 457-466.

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci U S A **97,** 5328-5333.

Angenent, G.C., and Colombo, L. (1996). Molecular control of ovule development. Trends Plant Sci **1,** 228-232.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., and Zdobnov, E.M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res **29,** 37-40.

The *Arabidopsis* Genome Initiative(AGI) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2,** 28-36.

Becker, A., Winter, K.U., Meyer, B., Saedler, H., and Theissen, G. (2000). MADS-Box gene diversity in seed plants 300 million years ago. Mol Biol Evol **17,** 1425-1434.

Bennetzen, J.L. (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol Biol **42,** 251-269.

Burgeff, C., Liljegren, S.J., Tapia-Lopez, R., Yanofsky, M.F., and Alvarez-Buylla, E.R. (2002). MADS-box gene expression in lateral primordia, meristems and differentiated tissues of *Arabidopsis thaliana* roots. Planta **214,** 365-372.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Escalante, R., and Sastre, L. (1998). A Serum Response Factor homolog is required for spore differentiation in *Dictyostelium*. Development **125,** 3801-3808.

Felsenstein, F. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution **39,** 1901-1911.

Fischer, A., Baum, N., Saedler, H., and Theissen, G. (1995). Chromosomal mapping of the MADS-box multigene family in *Zea mays* reveals dispersed distribution of allelic genes as well as transposed copies. Nucleic Acids Res **23,** 1901-1911.

Hall, T.A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser**,** 95-98.

Hoffmann, U., and Kroiher, M. (2001). A possible role for the cnidarian homologue of serum response factor in decision making by undifferentiated cells. Dev Biol **236,** 304-315.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., and Rhee, S.Y. (2001). The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res **29,** 102-105..

Krogan, N.T., and Ashton, N.W. (2000). Ancestry of plant MADS-box genes revealed by bryophyte (*Physcomitrella patens*) homologues. New Phytol **147,** 505-517.

Latchman, D.S. (1998). Eukaryotic transcription factors. (Academic Press, California).

Liljegren, S.J., Ferrándiz, C., Alvarez-Buylla, E.R., Pelaz, S., and Yanofsky, M.F. (1998). Arabidopsis MADS-box genes involved in fruit dehiscence. Flowering News Letter **25,** 9-19.

Liljegren, S.J., Ditta, G.S., Eshed, Y., Savidge, B., Bowman, J.L., and Yanofsky, M.F. (2000). *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. Nature **404,** 766-770.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26,** 1107-1115.

Lupas, A. (1996). Coiled coils: new structures and new functions. Trends Biochem Sci **21,** 375-382.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Montag, K., Salamini, F., and Thompson, R.D. (1995). ZEMa, a member of a novel group of MADS box genes, is alternatively spliced in maize endosperm. Nucleic Acids Res **23,** 2168-2177.

Montag, K., Salamini, F., and Thompson, R.D. (1996). The ZEM2 family of maize MADS box genes possess features of transposable elements. Maydica **41,** 241-254.

Muller, T., and Vingron, M. (2000). Modeling amino acid replacement. J Comput Biol **7,** 761-776.

Munster, T., Pahnke, J., Di Rosa, A., Kim, J.T., Martin, W., Saedler, H., and Theissen, G. (1997). Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. Proc Natl Acad Sci U S A **94,** 2415-2420.

Ng, M., and Yanofsky, M.F. (2001). Function and evolution of the plant MADS-box gene family. Nat Rev Genet **2,** 186-195.

Norman, C., Runswick, M., Pollock, R., and Treisman, R. (1988). Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell **55,** 989-1003.

Passmore, S., Elble, R., and Tye, B.K. (1989). A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. Genes Dev **3,** 921-935.

Pelaz, S., Ditta, G.S., Baumann, E., Wisman, E., and Yanofsky, M.F. (2000). B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405,** 200-203.

Raes, J., and Van de Peer, Y. (1999). ForCon: A software tool for the conversion of sequence alignments. EMBNet.news **6,** (http://www.ebi.ac.uk/embnet.news/vol6_1/).

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., and Van de Peer, Y. (2003). Investigating ancient duplication events in the *Arabidopsis* genome. Journal of structural and functional genomics **3,** 117-123.

Riechmann, J.L., and Meyerowitz, E.M. (1997). MADS domain proteins in plant development. Biol Chem **378,** 1079-1101.

Riechmann, J.L., and Ratcliffe, O.J. (2000). A genomic perspective on plant transcription factors. Curr Opin Plant Biol **3,** 423-434.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16,** 944-945.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4,** 406-425.

Sasaki, T., and Burr, B. (2000). International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol **3,** 138-141.

Scheffer, U., Krasko, A., Pancer, Z., and Müller, W.E.G. (1997). High conservation of the serum response factor within Metazoa: cDNA from the sponge *Geodia cydonium*. Biol J Linn Soc **61,** 127-137.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18,** 502-504.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., and Mayer, K.F. (2002). MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30,** 91-93.

Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H., and Sommer, H. (1990). Genetic control of flower development by homeotic genes in *Antirrhinum majus*. Science **250,** 931-936.

Schwarz-Sommer, Z., Hue, I., Huijser, P., Flor, P.J., Hansen, R., Tetens, F., Lonnig, W.E., Saedler, H., and Sommer, H. (1992). Characterization of the *Antirrhinum* floral homeotic MADS-box gene deficiens: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. Embo J **11,** 251-263.

Shore, P., and Sharrocks, A.D. (1995). The MADS-box family of transcription factors. Eur J Biochem **229,** 1-13.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van De Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **99,** 13627-13632.

Sommer, H., Beltran, J.P., Huijser, P., Pape, H., Lonnig, W.E., Saedler, H., and Schwarz-Sommer, Z. (1990). Deficiens, a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus*: the protein shows homology to transcription factors. Embo J **9,** 605-613.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M.A., Tzouvara, K., and Vaughan, R. (2002). The EMBL Nucleotide Sequence Database. Nucleic Acids Res **30,** 21-26.

Strimmer, K., and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol **13,** 964-969.

Swofford, D. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). (Sunderland, Massachusetts: Sinauer Associates).

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. Curr Opin Plant Biol **4,** 75-85.

Theissen, G., and Saedler, H. (2001). Plant biology. Floral quartets. Nature **409,** 469-471.

Theissen, G., Munster, T., and Henschel, K. (2001). Why don't mosses flower? New Phytol **150,** 1-8.

Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J.T., Munster, T., Winter, K.U., and Saedler, H. (2000). A short history of MADS-box genes in plants. Plant Mol Biol **42,** 115-149.

Van de Peer, Y., and De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci **10,** 569-570.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res. **12,** 1792-1801.

Wikstrom, N., Savolainen, V., and Chase, M.W. (2001). Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci **268,** 2211-2220.

Wolf, E., Kim, P.S., and Berger, B. (1997). MultiCoil: A Program for Predicting Two- and Three-Stranded Coiled Coils. Protein Sci **6,** 1179-1189.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13,** 555-556.

Yanofsky, M.F., Ma, H., Bowman, J.L., Drews, G.N., Feldmann, K.A., and Meyerowitz, E.M. (1990). The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature **346,** 35-39.

Yu, Y.T., Breitbart, R.E., Smoot, L.B., Lee, Y., Mahdavi, V., and Nadal-Ginard, B. (1992). Human myocyte-specific enhancer factor 2 comprises a group of tissue- restricted MADS box transcription factors. Genes Dev **6,** 1783-1798.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296,** 79-92.

Zhang, H., and Forde, B.G. (2000). Regulation of *Arabidopsis* root development by nitrate availability. J Exp Bot **51,** 51-59.

[Chapter 5]

And then there were many: MADS goes genomic

Stefanie De Bodt[1], Jeroen Raes[1], Yves Van de Peer[1,*] and Günter Theissen[2]

[1] Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Gent, Belgium
[2] University of Jena, Lehrstuhl for Genetics, Philosophenweg 12, D-07743 Jena, Germany

[*] *Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*

Published in: Trends in Plant Science, in press

Abstract

During the last decade, MADS-box genes became known as key regulators in both reproductive and vegetative plant development. Today, research on MADS-box genes has entered the (post) genomic era and starts to reveal the true complexity of this large gene family. Traditional genetics and functional genomics tools are now available to elucidate the expression and function of this complex gene family on a much larger scale. Moreover, comparative analysis of the MADS-box genes in diverse flowering and non-flowering plants, boosted by bioinformatics, contributes to our understanding of how this important gene family has expanded during the evolution of land plants. Therefore, recent advances in comparative and functional genomics enable researchers to identify the full range of MADS-box gene functions and will have a significant impact on a better understanding of plant development and evolution.

Throughout plant evolution, MADS-box genes have been recruited as transcriptional regulators active in the development of diverse plant structures. Since the discovery of the first MADS-box genes more than a decade ago, biologists have made great progress in the elucidation of the role of these genes in plant development. Expression studies and mutant analyses on MADS-box genes in diverse plant species such as *Arabidopsis thaliana*, *Antirrhinum majus* and *Zea mays*, among others, revealed the crucial importance of MADS-box genes in the regulation of both reproductive (flower, seed, fruit) and vegetative (root, leaf) development (Ng and Yanofsky, 2001). Furthermore, MADS-box genes, employed in the control of floral patterning, form the ideal genetic toolkit to study the diversification of flower architecture (Theissen et al., 2000).

The MADS-box genes constitute a large gene family named after a few of its earliest members, ***M****CM1*, found in yeast (Passmore et al., 1988), ***A****GAMOUS*, in *Arabidopsis thaliana* (Yanofsky et al., 1990), ***D****EFICIENS*, in *Antirrhinum majus* (Sommer et al., 1990; Schwarz-Sommer et al., 1992), and ***S****RF*, in human (Norman et al., 1988). The gene family can be divided into two main lineages (referred to as type I and type II) both present in plants, animals and fungi, which all members possess the on average 180 nucleotides long MADS-box (Alvarez-Buylla et al., 2000). It encodes the MADS-domain of the transcription factors that is responsible for nuclear localization, DNA-binding, dimerisation and accessory factor binding (Ng and Yanofsky, 2001; Theissen et al., 2000; Immink et al., 2002). In plants, type II MADS-domain proteins, referred to as MIKC proteins, possess three additional functional domains: a well-conserved K (Keratin)-domain, responsible for dimerisation, a less conserved I (Intervening)-domain, which constitutes a key regulatory determinant for the selective formation of DNA-binding dimers, and a variable C-terminal region, which is involved in transcriptional activation or in the formation of ternary or quaternary protein complexes (Riechmann and Meyerowitz, 1997; Egea-Cortines et al., 1999; Honma and Goto, 2001) and contributes to functional specificity (Lamb and Irish, 2003). Contrary to type II genes, which have been the subject of extensive research, not much is known about the type I genes in plants. Except for the MADS-box, the type I genes share no sequence similarity with type II genes. However, some type I genes share conserved C-terminal motifs among each other (De Bodt et al., 2003; Parenicova et al., 2003). In addition, a third group of genes has been identified recently, which are referred to as MADS-like genes and which possess only half of the MADS-box or which are overall highly divergent (De Bodt et al., 2003).

In this review, we present a survey on the recent progress that has been made in the field of MADS-box gene research, especially the contribution of genomics, bioinformatics and protein-protein interaction studies to the understanding of the MADS-box gene family and the future ways for plant developmental studies in the phylogenomics and phyloproteomics era ahead.

Genetics lays the foundations

The study of plant MADS-box genes was initially prompted by their importance in flower development. Gain- and loss-of-function phenotypes generated through T-DNA, transposon- or EMS-induced mutations in MADS-box genes have uncovered the function of many of these genes in diverse aspects of this process, ranging from the determination of flowering time (e.g.

FLOWERING LOCUS C, *SUPPRESSOR OF CONSTANS1*) to the specification of floral meristem (e.g. *APETALA1, CAULIFLOWER*) and floral organ identity (e.g. *APETALA1*, *APETALA3*, *PISTILLATA*, *AGAMOUS*) (Ng and Yanofsky, 2001). As a result, for example, developmental biologists have been able to molecularly clone almost all of the genes providing the floral homeotic functions that, according to the ABC-model, act in a combinatorial way to specify floral organ identity (Coen and Meyerowitz, 1991; Weigel and Meyerowitz, 1994; Theissen, 2001). Later, more key players of the floral developmental pathway were identified leading to the extension of this model to the ABCDE and the protein-based quartet model (Theissen, 2001).
Whereas research on floral developmental genes is progressing rapidly, the functional analysis of other MADS-box genes is lagging behind. Nevertheless, MADS-box genes have also been shown to function in the control of fruit development (*SHATTERPROOF1* and *2, FRUITFULL*), seed development (e.g. *TRANSPARENT TESTA 16*) and root growth (e.g. *ARABIDOPSIS NITRATE-RESPONSIVE 1*) (Rounsley et al., 1995; Liljegren et al., 2000; Zhang and Forde, 2000; Burgeff et al., 2002; Ferrandiz et al., 2000; Nesi et al., 2002; and others).
Unfortunately, when analysing large families, such as the MADS-box gene family, one is confronted with a number of problems. First, due to the high functional redundancy found in MADS-box genes, the construction of double or even multiple mutants is often inevitable to uncover the complete spectrum of gene functions by mutant phenotype. As such studies are relatively time consuming, the prediction of functional redundancy by phylogeny reconstructions helps to minimize the effort (Liljegren et al., 2000; Riechmann et al., 2000; Pelaz et al., 2000; Smyth, 2000; Pinyopich et al., 2003). In addition, the incomplete sampling of MADS-box genes in most organisms makes it difficult to assign the correct orthologous and paralogous relationships between genes and restricts a comprehensive comparison of the gene functions (Becker and Theissen, 2003). Moreover, lineage specific gene family expansion through gene duplication has led to extant plants having established orthologous relationships between clades of paralogous genes rather than between individual genes and could have led to differences in functional divergence of these duplicated genes in different plant lineages (Theissen and Becker, in press).

Genomics reveals new roads ahead

Since the beginning of the 21st century, plant molecular biology has been flushed with a previously unseen amount of sequence data. The completion of the genome sequence of *Arabidopsis thaliana* and *Oryza sativa* now allows the investigation of the full complement of MADS-box genes in both eudicot and monocot plants (Arabidopsis Genome Initiative, 2000; Goff et al., 2002; Yu et al., 2002). The genome-wide structural annotation of the MADS-box gene family in these organisms has resulted in the discovery of more than 100 genes in *Arabidopsis* (104 genes in De Bodt et al. (2003), 107 in Parenicova et al. (2003), 105 in Kofuji et al. (2003)), and 71 genes in rice (De Bodt et al., 2003; TIGR annotation; our unpublished results). A list of MADS-box genes in selected model species can be found as supplementary material on our web site (www.psb.ugent.be/bioinformatics/MADS). It should be noted that the true number of MADS-box genes in rice might be higher than 71, since the annotation of the rice genome is far from completed.

Structural annotation of the novel type I subfamily in the *Arabidopsis* and rice genomes has resulted in the discovery of 40 (+7 MIKC*, see further) and 37 MADS-box genes, respectively. Additionally, 20 highly diverged MADS-like genes have been identified in the *Arabidopsis thaliana* genome, for which no rice homologs have been found yet (De Bodt et al., 2003; our unpublished results; supplementary material). However, in the rice genome, a number of genes can be detected that possess remnants of the MADS-box but degenerated into pseudogenes through the insertion of stop codons. In contrast, all *Arabidopsis* MADS-like genes consist of complete open reading frames. The genome-wide identification of MADS-box genes has led to new views on the evolution of the gene family. Through the ongoing *Arabidopsis* genome sequencing project, a great amount of new data became available that was first used by Alvarez-Buylla et al. (2000) to infer the phylogeny of the MADS-box gene family. Their phylogenetic analyses, comprising 45 MADS-box genes from *Arabidopsis thaliana* and representative genes from animal and fungal species, uncovered, for the first time, the existence of two MADS-box lineages (type I and type II) in plants, animals and fungi (Alvarez-Buylla et al., 2000) (Figure 1). The authors suggested that the two lineages arose through an ancestral duplication that occurred in the common ancestor of plants, animals and fungi and that the K domain, specific to plant type II genes, probably evolved in the plant lineage after its divergence from the animals and fungi.
Structural analysis of all MADS-box genes has indicated two main differences between type I and type II genes, namely the absence of the K-box in type I genes and the fact that most type I MADS-box genes are single exon genes, while type II genes consist of 7 exons, on average (Alvarez-Buylla et al., 2000; De Bodt et al., 2003).

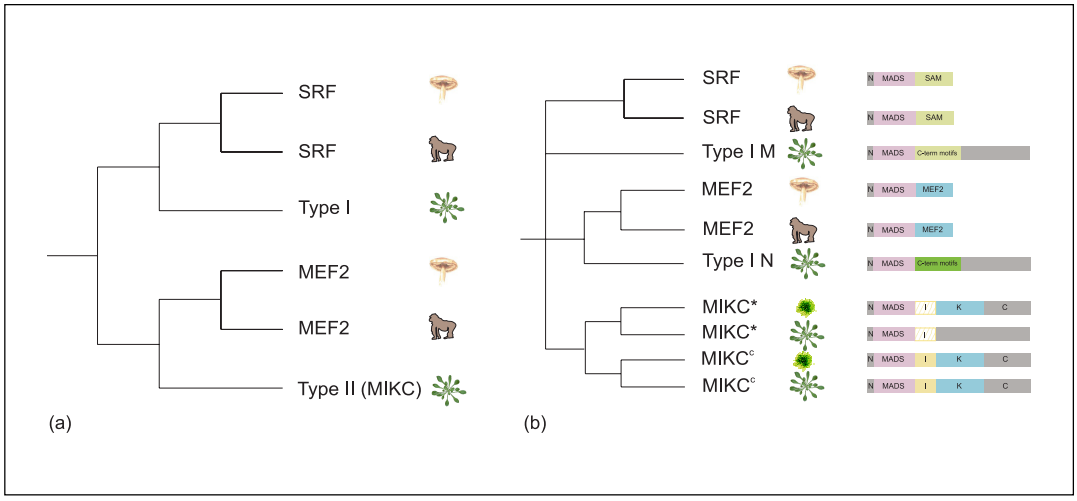

Figure 1: Evolution and structure of MADS-box genes of higher plants, mosses, animals and fungi, according to (a) Alvarez-Buylla et al. (2000) and (b) using new data and alternative approaches (see text).

Phylogeny reconstructions based on a more extensive set of MADS-domain sequences indicated that 7 *Arabidopsis* sequences, originally assigned to a subtype of type I genes, termed class O genes, might actually represent deviant type II genes, termed MIKC*-type genes (De Bodt et al., 2003).

These genes constitute a novel subtype of plant MIKC-type (type II) genes, which have been marked by an asterisk to distinguish them from the "classical" MIKC-type genes (hence also termed MIKC^c-type genes) (Henschel et al., 2002). First analyses have shown that MIKC*-type genes may be mainly expressed in pollen (Kofuji et al., 2003). In line with this, a novel MADS-box gene, closely related to the MIKC*-type genes from *Arabidopsis*, was recently identified in *Nicotiana tabacum* through its differential expression in pollen (Steiner et al., 2003; supplementary material).
Plant type I genes *sensu stricto* (i.e. without the MIKC*-type genes) have an evolutionary dynamic which is significantly different from that of both animal type I (SRF) and plant type II (MIKC) genes. For example, their evolutionary rate is much higher than that of plant type II genes (De Bodt et al., 2003). One possible explanation for this would be that the functional constraint on type I genes is lower than on type II genes, and that type I genes are therefore of less functional importance to the plant. This could be the reason why no mutant phenotype has ever been reported for a plant type I gene. On the contrary, a single or (in case of redundant genes) multiple mutant phenotype is known for 18 type II MADS-box genes from *Arabidopsis* and for many other plant type II genes, but all of them are type II genes (Becker and Theissen, 2003). The absence of mutant phenotypes for type I genes could be due to their functional redundancy with other genes, which is also shown for plant type II genes (Liljegren et al., 2000; Riechmann et al., 2000; Pelaz et al., 2000; Smyth, 2000; Pinyopich et al., 2003).
Another explanation is that plant type I genes have only subtle functions, or work only under exceptional environmental conditions. In line with this, the expression level of most plant type I genes, if any, is much lower than that of type II genes. For example, for many type I genes, expression could only be detected by RT-PCR, whereas it was impossible to detect expression of most type I genes through RNA gel blot analysis, macro-array and in situ hybridisation (Parenicova et al., 2003; Kofuji et al., 2003). In addition, some cases were found where expression was detected using a macro-array approach, while expression was not detectable via RT-PCR (e.g. AGL103, AGL34) (Parenicova et al., 2003; Kofuji et al., 2003). It is clear that a meticulous analysis (e.g. including more tissues and conditions) of the expression patterns of these genes will be needed to resolve these issues.
A final reason why no type I gene mutant phenotype is known could be that type I genes are (evolving to) pseudogenes. We and others indeed found evidence that at least one type I MADS-box gene (At5g49490) is a processed pseudogene, since a poly A-tail is found downstream of the gene (our unpublished results) (Kofuji et al., 2003). Since many type I genes consist of only a single exon, one could also presume that these genes arose through (retro)transposition (De Bodt et al., 2003; Kofuji et al., 2003). Moreover, type I genes are mainly located on chromosomes 1 and 5 (De Bodt et al., 2003; Parenicova et al., 2003; Kofuji et al., 2003) which fits this hypothesis, since several plant transposons show preferential local integration. If (retro)transposition is responsible for the origin of many (most) single exon MADS-box genes, we might expect to find repeat and known transposon-like sequences in close proximity of these genes. For some type I genes, short repeats can indeed be found 1 kb up- and downstream. However, for most type I genes, remnants of (retro)transposition can not be found, but this does not rule out (retro)transposition events early in the history of type I genes.

On the other hand, it cannot be excluded that plant type I genes represent an absolutely novel and unprecedented class of transposable elements lacking any sequence hallmarks defined before. Transposons carrying a MADS-box would not be unprecedented. In maize (*Zea mays*) and its relatives, *En*/*Spm*-like transposable elements have been identified which have captured a MADS-box and have distributed it throughout the *Zea* genome (Fischer et al., 1995; Montag et al., 1996). However, these elements contain an *AGAMOUS*-like (hence type II) MADS-box, and share no other domains with the type I genes.

Do type I genes have a function? The fact that type I genes in *Arabidopsis* and rice contain similar C-terminal motifs (De Bodt et al., 2003; Parenicova et al., 2003) suggests sequence conservation due to functional constraint, despite the high evolutionary rate of type I genes. But function does not necessarily imply a function for the host plant. The alternative could be that type I sequences, rather than being conventional genes, represent transposable elements or some other kind of "selfish" sequence elements.

Recently, it has been shown that the type I gene *PHERES1* (*AGL37*) is transiently expressed during embryo and endosperm development, and that up-regulation of *PHERES1* in Polycomb-group gene mutants such as *medea* is responsible for developmental defects, such as seed abortion, in these mutants. Moreover, *PHERES1* is obviously a direct target gene of some Polycomb-group proteins including MEDEA (Köhler et al., 2003). These findings raise the hope that a function to at least one plant type I MADS-box gene can be assigned soon.

However, no current hypothesis on plant type I genes fits all the data satisfactorily, and maybe no single hypothesis ever will, if type I genes are a phylogenetically or functionally heterogeneous class of genes. To solve the frustrating conundrum of type I genes, comprehensive and careful analysis of plant gene mutants, e.g. obtained by reverse genetic screens, will elucidate whether these sequence elements are of functional importance to the plants. To circumvent putative problems with redundancy, the generation of double or even multiple gene knock-outs (guided by phylogeny reconstructions) might prove necessary. While loss-of-function phenotypes for a number of genes will almost certainly exclude the transposon hypothesis (at least for the respective genes), the inability to identify phenotypes would be less conclusive, because lack of a recognizable phenotype does not necessarily mean that the gene has no function. In these cases, however, the defining characteristic of transposons, i.e. their ability to change their chromosomal position, might reveal the transposon character of these sequence elements. Transposition of mobile elements might be observed by Southern blot analysis or a technique called transposon display, as recently demonstrated for an active transposon family in rice (Jiang et al., 2003).

Whereas the function of type I genes largely remains a mystery, the functional importance of type II MADS-box genes has been clearly shown both through the functional characterisation of single MADS-box genes and through moderate to large scale cDNA sequencing projects in diverse plants, such as the eudicot angiosperm *Petunia hybrida* (Immink et al., 2003), the monocot *Zea mays* (Münster et al., 2002), the gymnosperms *Gnetum gnemon* (Winter et al., 1999; Becker et al., 2000), *Pinus radiata* (Mouradov et al., 1999; Mouradov et al., 1998; Walden et al., 1998), *Picea abies* (Rutledge et al., 1998), and *Ginkgo biloba* (Jager et al., 2003), the fern *Ceratopteris richardii* (Münster et al., 1997; Hasebe et al., 1998), and the moss *Physcomitrella patens* (Henschel et al., 2002).

An overview on the current status of MADS-box gene sequences from *Arabidopsis thaliana, Oryza sativa, Zea mays* and *Petunia hybrida* and their function (where known) in these plants is given in the supplementary material.

The cDNA sequencing efforts have allowed phylogenetic analyses of type II MADS-box genes, which showed that these genes can be subdivided into distinct clades, each clade comprising orthologs from different seed plants. MADS-box genes from ferns and mosses, however, could so far not be assigned to any of these clades and probably possess a more ubiquitious expression and function than their counterparts in flowering plants (Theissen et al., 2000; Henschel et al., 2002; Münster et al., 1997; Hasebe et al., 1998). Thus the study of these genes allows the correlation of the appearance of new types (clades) of developmental control genes with the origin of novel morphological structures (such as ovules/seeds and flowers) in plants (Theissen et al., 2000).

Two different approaches have been used to date the origin of the distinct clades of type II MADS-box genes and correlate the evolution of MADS-box genes with the divergence of major plant lineages. Becker et al. (2000; 2003) based their study on gene sampling and obtained estimates of 300 – 400 million years for the origins of many type II gene clades, whereas the Nam and coworkers' study (2003) used molecular clock-based dating, leading to much older age estimates. The study by Nam et al. (2003) implies that class B and class C floral homeotic gene lineages originated about 660 and 570 million years ago, respectively, i.e. before the separation of the lineages that led to mosses, ferns and seed plants. This suggests that representatives of these clades were either lost in extant mosses and ferns, or are present, but have simply not been identified. Another explanation is that type II genes in the lineage that led to extant ferns evolved at a higher rate than genes in the seed plant lineage, so that fern orthologs of seed plant genes cannot be recognized anymore. Alternatively, molecular clock estimates extrapolating from gymnosperm and angiosperm data might overestimate the ages of the clades, because type II gene evolution in the lineage that led to extant seed plants could have been much faster 300 – 400 MYA (after the fern lineage split off), and slowed down 300 MYA, after the angiosperm - gymnosperm split. If so, it would be interesting to find out which changes (e.g. in gene functions, modes of protein-protein interactions) can be correlated with these differences in evolutionary rate.

In Figure 2, a phylogenetic tree of MIKC^c genes from diverse plant lineages with their expression patterns in distinct tissues is presented. It generally corroborates the view that members of the same gene subfamilies tend to have similar expression patterns (Theissen et al., 1996), but it also demonstrates that this correlation is stricter for genes involved in flower formation than for genes mainly expressed in non-floral organs (Becker and Theissen, 2003). In some cases, lineage-specific expansions led to the occurrence of orthologous pairs of genes which possess a distinct pattern of divergence on expression level; both genes of one pair have kept the expression pattern of their ancestral gene, suggesting functional redundancy, while genes of the other pair have subdivided the expression pattern resulting in genes with a more specific functional activity (for example, expression of *AP1/CAL* from *Arabidopsis thaliana* and *PFG*/*FBP26* from *Petunia hybrida* in reproductive structures).

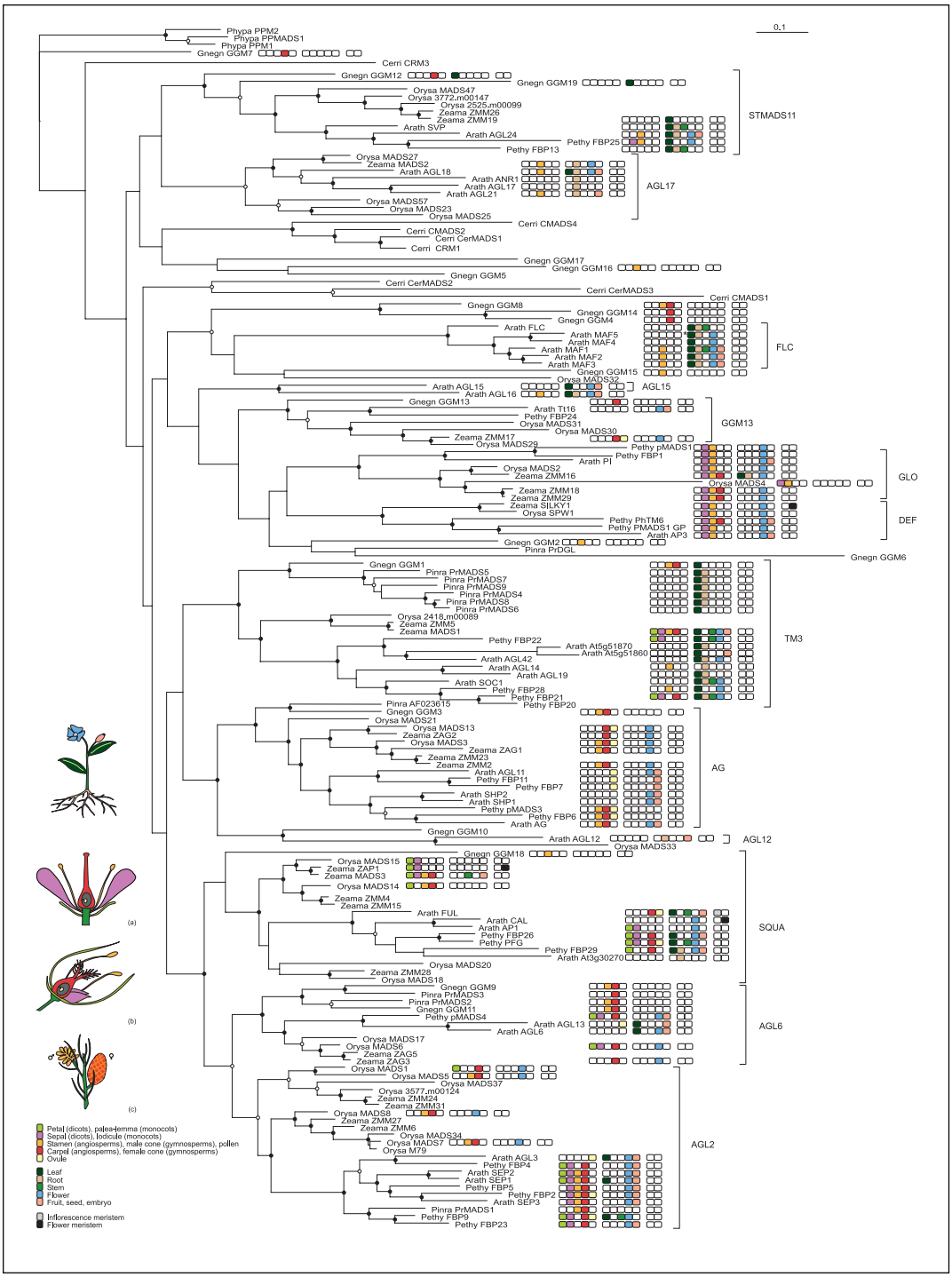

Figure 2: Phylogenetic tree and expression patterns (where known) of MIKC genes from (a) the eudicots *Arabidopsis thaliana* (Arath) and *Petunia hybrida* (Pethy), (b) the monocots *Oryza sativa* (Orysa) and *Zea mays* (Zeama), (c) the gymnosperms *Pinus radiata* (Pinra) and *Gnetum gnemon* (Gnegn), the fern *Ceratopteris richardii* (Cerri) and the moss *Physcomitrella patens* (Phypa). The phylogenetic tree is constructed using MrBayes (1,000,000 generations, 4 chains). Nodes supported by posterior probabilities higher than 70 are denoted by a black dot, posterior probabilities between 50-70 by an open circle. The scale indicates 0.1 substitutions per site. Expression patterns (tissue-specific) are extracted from literature on specific genes, on the one hand, and from the genome-wide analyses of Parenicova et al. (2003) and Kofuji et al. (2003), on the other hand. In case of conflict, preference was given to indicate a gene as being expressed when more sensitive approaches (e.g. RT-PCR) gave a positive result where others did not (e.g. macro-array, Northern). The expression of genes that could only be detected through macro-array analysis (2003) and not through other methods, is marked with an asterisk (*).

The extensive analysis of the *Arabidopsis* MIKC genes has allowed the transfer of knowledge about functions to orthologous genes from other plants through the principle of “phylogenomics” (Eisen and Wu, 2002). In particular, high functional redundancy, found through the analyses of *Arabidopsis* MADS-box genes (e.g. *SEPALLATA* and *SHATTERPROOF* genes), can be anticipated in similar studies in other organisms (Riechmann et al., 2000; Pelaz et al., 2000; Smyth, 2000; Pinyopich et al., 2003). On the other hand, the analysis of MADS-box genes in species other than *Arabidopsis* has provided us with greater insights into *Arabidopsis* genes. For example, studies in the gymnosperm *Gnetum gnemon* led to the discovery of a novel MADS-box gene subfamily with a sister-group relationship to the class B genes having members in *Gnetum gnenom* (*GGM13*), but also *Arabidopsis thaliana* (*ABS*), and *Zea mays* (*ZMM17*), among other plants (Becker et al., 2002). An independent and parallel functional characterisation of the *ABS* (*Arabidopsis B-sister*) gene (or *TRANSPARANT TESTA16*, *TT16)*, demonstrated its involvement in endothelial cell specification and in the genetic control of seed coat pigmentation (Riechmann et al., 2000). In addition, MADS cDNA sequencing in the moss *Physcomitrella patens* led to the identification of an additional class of MIKC genes, which possess a divergent I box and are referred to as MIKC*-type genes, as mentioned above (Henschel et al., 2002). So two interesting classes of MADS-box genes have first been identified in lower plants such as a moss and a gymnosperm rather than in the model plant *Arabidopsis*. Moreover, comparative analysis of the MADS-box gene family in angiosperms, and in particular, the AP3/PI clade of genes (Figure 2), which act as B class floral organ identity genes, uncovered distinct C-terminal motifs which can be correlated with their functional specificity (Lamb and Irish, 2003). At least for some genes, it has been recently shown that these specific motifs have probably arisen through one (or more) nucleotide insertions or deletions, causing translational frame-shifts, and subsequent sequence conservation (Vandenbussche et al., 2003). What is remarkable about this finding is that frame-shift mutations in C-terminal regions of duplicate genes are selected for and hence the gene has been retained together with the unchanged gene duplicate. 3' terminal frame-shift mutations might therefore represent an important novel mechanism in the functional diversification of transcription factor gene families (Vandenbussche et al., 2003).

The results of recent genome-wide studies, as those described above, urge an unambiguous definition and nomenclature for the different classes of MADS-box genes, preferably based on careful, evolutionary analyses (Table 1). Unfortunately, the phylogenetic analyses of the whole gene family in *Arabidopsis* and rice result in poorly resolved trees, mainly due to the combination of a limited number of phylogenetically informative positions in the short MADS-domain (60 amino acids), and the large number of genes (De Bodt et al., 2003). Therefore, type I MADS-box genes were first classified based on structural characteristics rather than on poorly resolved phylogenetic trees, resulting in the class M and N type I genes which can be distinguished through the presence of conserved, C-terminal motifs (De Bodt et al., 2003). Detailed phylogenetic analyses of these classes in both *Arabidopsis* and rice showed extensive expansion of the number of these genes after the divergence of monocots and eudicots (De Bodt et al., 2003). In order to reconstruct the evolution of other complex whole gene families, alternative approaches have been employed, for example by limiting the number of genes and choosing only genes from a few representative species (Bharathan et al., 1997).

Table 1. Classification of the MADS-box genes in *Arabidopsis thaliana*.

Alvarez-Buylla et al. 2000	De Bodt et al. 2003	Parenicova et al. 2003	Kofuji et al. 2003
Type I (SRF-like)	Type I M	Mα [a, d, e]	M [a, f, g]
Type I (SRF-like)	Type I N	Mγ [c, e]	M
Type I (SRF-like)	MIKC* [g, i]	Mδ [g]	MIKC*
Type I (SRF-like)	Type I O [e, f, h]	-	M
Type I (SRF-like)	MADS-like	Mβ [b, e]	M
Type II (MEF2-like)	Type II	MIKC[f]	MIKC[c, i]

[a] At1g29960 and At1g54760 are assigned to class Mα by Parenicova et al. (2003), to class M by Kofuji et al. (2003) and are not identified as MADS-box genes by De Bodt et al. (2003).
[b] At4g02240 and At5g37420 are assigned to class Mβ by Parenicova et al. (2003) and are not identified as MADS-box genes by De Bodt et al. (2003) and Kofuji et al. (2003).
[c] At2g15660 is assigned to class Mγ according to Parenicova et al. (2003) and is not identified as a MADS-box gene by De Bodt et al. (2003) and Kofuji et al. (2003).
[d] At1g46408 was identified for the first time in Parenicova et al. (2003) and belongs to class Mα.
[e] At1g72350 and At1g17310 are assigned to class Mα, At5g06500 and At1g22590 to class Mγ and At5g26950, At5g58890 and At5g55690 to class Mb by Parenicova et al. (2003), and belong to type I O according to De Bodt et al. (2003).
[f] At1g31140 is assigned to class type I O by De Bodt et al. (2003), to class M by Kofuji et al. (2003), but to class MIKC by Parenicova et al. (2003).
[g] Originally considered Type I O, but then identified as MIKC* by De Bodt et al. (2003). At2g26320 is assigned to MIKC* by De Bodt et al. (2003), to class Md by Parenicova et al. (2003), and to class M by Kofuji et al. (2003).
[h] Type I O genes sensu stricto are class O genes according to De Bodt et al. (2003), except the MIKC* genes mentioned in the same paper
[i] term MIKC[c] and MIKC* introduced by Henschel et al. (2002).

Another solution is to replace well-supported clades of genes by their ancestral sequence. The two latter approaches, applied to MADS-box genes, give a topology as depicted in Figure 1b, which clearly contrasts with the results of Alvarez-Buylla et al. (2000) (Figure 1a). Although these approaches are not able to unequivocally resolve the deeper branching order between subclades of MADS-box genes, they suggest a polyphyletic origin of different groups of type I genes. However, it remains very difficult to elucidate the evolutionary relationships between these different groups of type I genes and their animal and fungal counterparts. More extensive sampling of MADS-box genes from diverse species, including basal plants, will hopefully contribute to the reconstruction of the evolutionary history of the MADS-box gene family.
In the future, more large sequencing projects such as the floral genome project (Soltis et al., 2002) combined with high-throughput functional characterization approaches will undoubtedly enable more comprehensive comparative analyses (both functional and evolutionary) and will consequently allow us to gain deeper insights into the role of different classes of MADS-box genes (type I and II) in the evolution of the gene family and in plant development.

Functional genomics provides the tools (for high-throughput analysis)

The availability of complete genome sequences as well as large sets of expressed sequence tags (ESTs) has triggered the development of high throughput methods to functionally analyse these raw data. Oligo and cDNA micro-arrays now allow the genome-wide analysis of spatial and temporal expression patterns (Lockhart and Winzeler, 2000 ; Schulze and Downward, 2001). To gain insights into the expression of regulatory genes such as MADS-box genes, specific arrays for the profiling of these genes are being designed (Paz-Ares, 2002).

In addition, the effect of MADS-box gene perturbation can be analysed using micro-arrays, which allows the identification of the downstream genes in the developmental pathway. For example, the global identification of target genes regulated by class B floral homeotic genes *APETALA3* and *PISTILLATA* was conducted through the use of cDNA micro-arrays (Zik and Irish, 2003). Similar analyses are being conducted, although on a smaller scale, in other plants. For example, Moore and co-workers (Moore et al., 2002) are investigating the effect of tomato *ripening-inhibitor* (*rin*) and *non-riping* (*nor*) mutants on gene expression using different genomics tools.
Large-scale interaction studies such as yeast two- and three-hybrid screens and FRET (Fluorescence-Resonance-Energy-Transfer) analyses provide insights into protein-protein and RNA-protein interactions (Fields and Song 1989; Sengupta et al., 1996; Immink and Angenent, 2002). The FRET technology has been shown to be effective in the identification of dimeric complexes of MADS-domain proteins involved in flower development *in planta* (e.g. the formation of complexes consisting of organ identity MADS-box genes) (Sengupta et al., 1996; Immink and Angenent, 2002). Moreover, yeast one-hybrid experiments are used to detect protein-DNA interactions and to isolate new proteins that bind to a specific target (regulatory) element (Luo et al., 1996). These experiments can be conducted on a large scale when an extensive collection of promoters and their *cis*-acting regulatory elements is available for plants. The ChIP (Chromatin Immuno Precipitation) technology recently allowed the identification of several targets of AGL15 (Fernandez et al., 2000; Wang et al., 2002). The development of micro-arrays containing regulatory regions for all *Arabidopsis* genes will speed up the detection of candidate target genes using this approach, as has been demonstrated in yeast (Ren et al., 2000) and human (Weinmann and Farnham, 2002). In parallel with these experimental studies, *in silico* analyses of promoters can be exerted using clusters of coregulated genes or through a comparative approach using homologous genes in different organisms (Koch et al., 2001; Hong et al., 2003; Rombauts et al., 2003).
As such, a complete survey of all genes, including the largely unexplored type I MADS-box and MADS-like genes, can be compiled in an efficient way, giving a glimpse of the processes in which these genes are active and which can be used to select interesting genes for more in depth analyses on both the RNA and protein level.

Conclusion and outlook

These are exciting times in the MADS world. The availability of complete genomes and the rise of novel sophisticated technologies open up many possibilities for plant research. Thanks to the combination of comparative developmental biology and genomics, exciting new insights are being revealed in the evolution of development and the underlying regulatory mechanisms. To be most profitable, efforts should focus on plant species of evolutionary importance, for which genetic and genomic tools exist or can be developed (Pryer et al., 2002). However, the choice of adequate model systems is not self-evident, due to the large genome size and the long generation time of many phylogenetically interesting plants (e.g. gymnosperms). On the other hand, the moss *Physcomitrella patens* is an example of an especially interesting and useful plant model organism, not only because it has quite a small genome and is easy to grow, but especially because it is the

only land plant which is amenable to efficient gene targeting via homologous recombination (Rensing et al., 2002; Schaefer, 2001). We believe that only an integrative approach, combining classical genetics, functional genomics, bioinformatics, and comparative genomics will be able to unravel the evolution and functional divergence of large transcription factor families such as the MADS-box gene family. Probably, future research will even go beyond this comprehensive "phylogenomics" approach, as there is much evidence that the specificity of MADS-box gene action is conferred by combinatorial protein-protein interactions (for a review, see Messenguy and Dubois, 2003). Examples are the quartet model (Theissen, 2001) and some others, termed "The second model" and "A third model" (Jack, 2001), describing the specification of floral organ identity. It can be predicted, therefore, that future studies will focus more and more on trying to understand MADS-domain protein-protein interactions. Employing techniques such as X-ray crystallography and NMR, FRET, gel retardation assays and the yeast two-hybrid system in a phylogenetic context, "phyloproteomics" of MADS-domain transcription factors might be at the horizon.

Acknowledgements

SDB is indebted to the "Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT)" for a predoctoral fellowship. We thank three anonymous reviewers for helpful comments on an earlier version of the manuscript.

References

The *Arabidopsis* Genome Initiative(AGI) (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408,** 796-815.

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci U S A **97,** 5328-5333.

Becker, A., and Theissen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phyl Evol **In press**.

Becker, A., Winter, K.U., Meyer, B., Saedler, H., and Theissen, G. (2000). MADS-Box gene diversity in seed plants 300 million years ago. Mol Biol Evol **17,** 1425-1434.

Becker, A., Kaufmann, K., Freialdenhoven, A., Vincent, C., Li, M.A., Saedler, H., and Theissen, G. (2002). A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. Mol Genet Genomics **266,** 942-950.

Bharathan, G., Janssen, B.J., Kellogg, E.A., and Sinha, N. (1997). Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc Natl Acad Sci U S A **94,** 13749-13753.

Burgeff, C., Liljegren, S.J., Tapia-Lopez, R., Yanofsky, M.F., and Alvarez-Buylla, E.R. (2002). MADS-box gene expression in lateral primordia, meristems and differentiated tissues of Arabidopsis thaliana roots. Planta **214,** 365-372.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature **353,** 31-37.

De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G., and Van De Peer, Y. (2003). Genomewide Structural Annotation and Evolutionary Analysis of the Type I MADS-Box Genes in Plants. J Mol Evol **56,** 573-586.

Egea-Cortines, M., Saedler, H., and Sommer, H. (1999). Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus. Embo J **18,** 5370-5379.

Eisen, J.A., and Wu, M. (2002). Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol **61,** 481-487.

Fernandez, D.E., Heck, G.R., Perry, S.E., Patterson, S.E., Bleecker, A.B., and Fang, S.C. (2000). The embryo MADS domain factor AGL15 acts postembryonically. Inhibition of perianth senescence and abscission via constitutive expression. Plant Cell **12,** 183-198.

Ferrandiz, C., Liljegren, S.J., and Yanofsky, M.F. (2000). Negative regulation of the SHATTERPROOF genes by FRUITFULL during Arabidopsis fruit development. Science **289,** 436-438.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature **340,** 245-246.

Fischer, A., Baum, N., Saedler, H., and Theissen, G. (1995). Chromosomal mapping of the MADS-box multigene family in Zea mays reveals dispersed distribution of allelic genes as well as transposed copies. Nucleic Acids Res **23,** 1901-1911.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296,** 92-100.

Hasebe, M., Wen, C.K., Kato, M., and Banks, J.A. (1998). Characterization of MADS homeotic genes in the fern Ceratopteris richardii. Proc Natl Acad Sci U S A **95,** 6222-6227.

Henschel, K., Kofuji, R., Hasebe, M., Saedler, H., Munster, T., and Theissen, G. (2002). Two ancient classes of MIKC-type MADS-box genes are present in the moss Physcomitrella patens. Mol Biol Evol **19,** 801-814.

Hong, R.L., Hamaguchi, L., Busch, M.A., and Weigel, D. (2003). Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell **15,** 1296-1309.

Honma, T., and Goto, K. (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature **409,** 525-529.

Immink, R.G., and Angenent, G.C. (2002). Transcription factors do it together: the hows and whys of studying protein-protein interactions. Trends Plant Sci **7,** 531-534.

Immink, R.G., Gadella, T.W., Jr., Ferrario, S., Busscher, M., and Angenent, G.C. (2002). Analysis of MADS box protein-protein interactions in living plant cells. Proc Natl Acad Sci U S A **99,** 2416-2421.

Immink, R.G., Ferrario, S., Busscher-Lange, J., Kooiker, M., Busscher, M., and Angenent, G.C. (2003). Analysis of the petunia MADS-box transcription factor family. Mol Genet Genomics **268,** 598-606.

Jack, T. (2001). Relearning our ABCs: new twists on an old model. Trends Plant Sci **6,** 310-316.

Jager, M., Hassanin, A., Manuel, M., Guyader, H.L., and Deutsch, J. (2003). MADS-Box Genes in Ginkgo biloba and the Evolution of the AGAMOUS Family. Mol Biol Evol **20,** 842-854.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R., and Wessler, S.R. (2003). An active DNA transposon family in rice. Nature **421,** 163-167.

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

Kofui, R., Sumikawa, N., Yamasaki, M., Kondo, K., Ueda, K., Ito, M., and Hasebe, M. (2003). Evolution and divergence of MADS-box gene family based on genome wide expression analyses. Mol Biol Evol **in press**.

Kohler, C., Hennig, L., Spillane, C., Pien, S., Gruissem, W., and Grossniklaus, U. (2003). The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev **17,** 1540-1553.

Lamb, R.S., and Irish, V.F. (2003). Functional divergence within the APETALA3/PISTILLATA floral homeotic gene lineages. Proc Natl Acad Sci U S A **100,** 6558-6563.

Liljegren, S.J., Ditta, G.S., Eshed, Y., Savidge, B., Bowman, J.L., and Yanofsky, M.F. (2000). SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature **404,** 766-770.

Lockhart, D.J., and Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature **405,** 827-836.

Luo, Y., Vijaychander, S., Stile, J., and Zhu, L. (1996). Cloning and analysis of DNA-binding proteins by yeast one-hybrid and one-two-hybrid systems. Biotechniques **20,** 564-568.

Messenguy, F., and Dubois, E. (2003). Role of MADS-box proteins and their cofactors as promoter architects in combinatorial control of gene expression and cell development. Gene **in press**.

Montag, K., Salamini, F., and Thompson, R.D. (1996). The ZEM2 family of maize MADS box genes possess features of transposable elements. Maydica **41,** 241-254.

Moore, S., Vrebalov, J., Payton, P., and Giovannoni, J. (2002). Use of genomics tools to isolate key ripening genes and analyse fruit maturation in tomato. J Exp Bot **53,** 2023-2030.

Mouradov, A., Hamdorf, B., Teasdale, R.D., Kim, J.T., Winter, K.U., and Theissen, G. (1999). A DEF/GLO-like MADS-box gene from a gymnosperm: Pinus radiata contains an ortholog of angiosperm B class floral homeotic genes. Dev Genet **25,** 245-252.

Mouradov, A., Glassick, T.V., Hamdorf, B.A., Murphy, L.C., Marla, S.S., Yang, Y., and Teasdale, R.D. (1998). Family of MADS-Box genes expressed early in male and female reproductive structures of monterey pine. Plant Physiol **117,** 55-62.

Munster, T., Pahnke, J., Di Rosa, A., Kim, J.T., Martin, W., Saedler, H., and Theissen, G. (1997). Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. Proc Natl Acad Sci U S A **94,** 2415-2420.

Münster, T., Deleu, W., Wingen, L.U., Ouzunova, M., Cacharrón, J., Faigl, W., Werth, S., Kim, J.T., Saedler, H., and Theissen, G. (2002). Maize MADS-box genes galore. Maydica **47,** 287-301.

Nam, J., DePamphilis, C.W., Ma, H., and Nei, M. (2003). Antiquity and Evolution of the MADS-Box Gene Family Controlling Flower Development in Plants. Mol Biol Evol **in press**.

Nesi, N., Debeaujon, I., Jond, C., Stewart, A.J., Jenkins, G.I., Caboche, M., and Lepiniec, L. (2002). The TRANSPARENT TESTA16 locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. Plant Cell **14,** 2463-2479.

Ng, M., and Yanofsky, M.F. (2001). Function and evolution of the plant MADS-box gene family. Nat Rev Genet **2,** 186-195.

Norman, C., Runswick, M., Pollock, R., and Treisman, R. (1988). Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell **55,** 989-1003.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D.S., Favalli, C., Busscher, J., Cook, H.E., Ingram, R.M., Kater, M.M., Davies, B., Angenent, G.C., and Colombo, L. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell **15,** 1538-1551.

Passmore, S., Maine, G.T., Elble, R., Christ, C., and Tye, B.K. (1988). Saccharomyces cerevisiae protein involved in plasmid maintenance is necessary for mating of MAT alpha cells. J Mol Biol **204,** 593-606.

Paz-Ares, J. (2002). REGIA, an EU project on functional genomics of transcription factors from Arabidopsis thaliana. Comp Funct Genom **3,** 102-108.

Pelaz, S., Ditta, G.S., Baumann, E., Wisman, E., and Yanofsky, M.F. (2000). B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405,** 200-203.

Pinyopich, A., Ditta, G.S., Savidge, B., Liljegren, S.J., Baumann, E., Wisman, E., and Yanofsky, M.F. (2003). Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature **424,** 85-88.

Pryer, K.M., Schneider, H., Zimmer, E.A., and Ann Banks, J. (2002). Deciding among green plants for whole genome studies. Trends Plant Sci **7,** 550-554.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., and Young, R.A. (2000). Genome-wide location and function of DNA binding proteins. Science **290,** 2306-2309.

Rensing, S.A., Rombauts, S., Van de Peer, Y., and Reski, R. (2002). Moss transcriptome and beyond. Trends Plant Sci **7,** 535-538.

Riechmann, J.L., and Meyerowitz, E.M. (1997). MADS domain proteins in plant development. Biol Chem **378,** 1079-1101.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., and Yu, G. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290,** 2105-2110.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., and van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol **132,** 1162-1176.

Rounsley, S.D., Ditta, G.S., and Yanofsky, M.F. (1995). Diverse roles for MADS box genes in Arabidopsis development. Plant Cell **7,** 1259-1269.

Rutledge, R., Regan, S., Nicolas, O., Fobert, P., Cote, C., Bosnich, W., Kauffeldt, C., Sunohara, G., Seguin, A., and Stewart, D. (1998). Characterization of an AGAMOUS homologue from the conifer black spruce (Picea mariana) that produces floral homeotic conversions when expressed in Arabidopsis. Plant J **15,** 625-634.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays—a technology review. Nat Cell Biol **3,** E190-195.

Schwarz-Sommer, Z., Hue, I., Huijser, P., Flor, P.J., Hansen, R., Tetens, F., Lonnig, W.E., Saedler, H., and Sommer, H. (1992). Characterization of the Antirrhinum floral homeotic MADS-box gene deficiens: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. Embo J **11,** 251-263.

SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., and Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. Proc Natl Acad Sci U S A **93,** 8496-8501.

Smyth, D. (2000). A reverse trend—MADS functions revealed. Trends Plant Sci **5,** 315-317.

Soltis, D.E., Soltis, P.S., Albert, V.A., Oppenheimer, D.G., dePamphilis, C.W., Ma, H., Frohlich, M.W., and Theissen, G. (2002). Missing links: the genetic architecture of flowers [correction of flower] and floral diversification. Trends Plant Sci **7,** 22-31; dicussion 31-24.

Sommer, H., Beltran, J.P., Huijser, P., Pape, H., Lonnig, W.E., Saedler, H., and Schwarz-Sommer, Z. (1990). Deficiens, a homeotic gene involved in the control of flower morphogenesis in Antirrhinum majus: the protein shows homology to transcription factors. Embo J **9,** 605-613.

Steiner, C., Bauer, J., Amrhein, N., and Bucher, M. (2003). Two novel genes are differentially expressed during early germination of the male gametophyte of Nicotiana tabacum. Biochim Biophys Acta **1625,** 123-133.

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. Curr Opin Plant Biol **4,** 75-85.

Theissen, G., and Becker, A. (2004). The ABCs of flower development in Arabidopsis and rice. Prog Bot **in press**.

Theissen, G., Kim, J.T., and Saedler, H. (1996). Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. J Mol Evol **43,** 484-516.

Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J.T., Munster, T., Winter, K.U., and Saedler, H. (2000). A short history of MADS-box genes in plants. Plant Mol Biol **42,** 115-149.

Vandenbussche, M., Theissen, G., Van de Peer, Y., and Gerats, T. (2003). Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Res **31,** 4401-4409.

Walden, A.R., Wang, D.Y., Walter, C., and Gardner, R.C. (1998). A large family of TM3 MADS-box cDNAs in Pinus radiata includes two members with deletions of the conserved K domain. Plant Sci **138,** 167-176.

Wang, H., Tang, W., Zhu, C., and Perry, S.E. (2002). A chromatin immunoprecipitation (ChIP) approach to isolate genes regulated by AGL15, a MADS domain protein that preferentially accumulates in embryos. Plant J **32,** 831-843.

Weigel, D., and Meyerowitz, E.M. (1994). The ABCs of floral homeotic genes. Cell **78,** 203-209.

Weinmann, A.S., and Farnham, P.J. (2002). Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods **26,** 37-47.

Winter, K.U., Becker, A., Munster, T., Kim, J.T., Saedler, H., and Theissen, G. (1999). MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc Natl Acad Sci U S A **96,** 7342-7347.

Yanofsky, M.F., Ma, H., Bowman, J.L., Drews, G.N., Feldmann, K.A., and Meyerowitz, E.M. (1990). The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature **346,** 35-39.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296,** 79-92.

Zhang, H., and Forde, B.G. (2000). Regulation of Arabidopsis root development by nitrate availability. J Exp Bot **51,** 51-59.

Zik, M., and Irish, V.F. (2003). Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. Plant Cell **15,** 207-222.

[Chapter 6]

Genome-wide analysis of core cell cycle genes in *Arabidopsis*

Klaas Vandepoele[1], Jeroen Raes[1,2], Lieven De Veylder[1], Pierre Rouzé[2], Stephane Rombauts[1], and Dirk Inzé[1*]

[1] Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[2] Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Universiteit Gent, Technologiepark 927, B-9052 Ghent, Belgium

[*] *Author for correspondence (e-mail diinz@psb.ugent.be; fax: +32 9 331 3809).*

Published in: The Plant Cell **14**, 903-916 (2002)

Abstract

Cyclin-dependent kinases and cyclins master together with the help of different interacting proteins the progression through the eukaryotic cell cycle. A high-quality, homology-based annotation protocol was applied to determine all core cell cycle genes in the recently completed *Arabidopsis* genome sequence. In total, 61 genes were identified belonging to seven selected families of cell cycle regulators, for which 30 are new or corrections of the existing annotation. A new class of putative cell cycle regulators was found that probably are competitors of E2F/DP transcription factors, which mediate the G1-to-S progression. In addition, the existing nomenclature for cell cycle genes of *Arabidopsis* was updated and physical positions of all genes were compared with segmentally duplicated blocks in the genome, showing that 22 core cell cycle genes emerged through block duplications. This genome-wide analysis illustrates the complexity of the plant cell cycle machinery and provides a tool for elucidating the function of new family members in the future.

Introduction

Cell proliferation is controlled by a universally conserved molecular machinery, in which the core key players are serine/threonine kinases, known as cyclin-dependent kinases (CDKs). CDK activity is regulated in a complex manner, including phosphorylation/dephosphorylation by specific kinases/ phosphatases and the association with regulatory proteins. Although many cell cycle genes of plants have been identified in the last decade (for review, see Stals and Inzé, 2001), the correct number of CDKs, cyclins, and interacting proteins with a role in the cell cycle control is still unknown. Now that the complete sequence of the nuclear genome of *Arabidopsis* is available (The *Arabidopsis* Genome Initiative, 2000), it is possible to scan an entire plant genome for all these core cell cycle genes and determine their number, position on the chromosomes, and phylogenetic relationship. From an evolutionary point of view, this core cell cycle gene catalogue would be extremely interesting because it allows us to determine which processes are plant specific and which are conserved among all eukaryotes. Furthermore, there is a unique opportunity to unravel in future experiments the function and interactions of newly found family members of primary cell cycle regulators, thus expanding our knowledge on how cell cycle is regulated in plants.

Nevertheless, a genome-wide inventory of all core cell cycle genes is only possible when the available raw sequence data are correctly annotated. Although the genome-wide annotation of organisms sequenced by large consortia produced a huge amount of information, which, no doubt, benefits the scientific community, one has to realize that this automated high-throughput annotation is far from optimal (Devos and Valencia, 2001). For this reason, it is often not trivial to extract clear biological information out of these public databases. When high-quality annotation is needed, a supervised semi-automatic annotation may be a good compromise between quality and speed.

Annotation is generally performed in two steps: first, a structural annotation that aims at finding and characterizing biologically relevant elements within the raw sequence (such as exons and translation starts), and secondly, functional annotation, in which biological information is attributed to the gene or its elements. Unfortunately, there are some problems inherent to both.

When structural annotation is performed, the first problem occurs whenever no cDNA or expressed sequence tag (EST) information is available, which is the case for 60% of all *Arabidopsis* genes (The *Arabidopsis* Genome Initiative, 2000). Then, one has to resort to intrinsic gene prediction software, which remains limited, although a lot of improvement has been made over the last few years. Errors range from wrongly determined splice sites or start codons, over so-called spliced (one gene predicted as two) or fused (two genes predicted as one) genes, up to completely missed or nonexisting predicted genes (Rouzé et al., 1999). In addition, no general and well-defined prediction protocol is used by the different annotation centers with the generation of redundant, non-uniform, structural annotation as a result. Furthermore, clear information is lacking on methods and programs used as well as the motivation for applying a special protocol, making it impossible to trace the annotation grounds.

The problem with functional annotation is related to the difficulty to couple biological knowledge to a gene. Such a link is made generally on the basis of sequence similarity that is derived either from full-length sequence comparisons or by means of multiple alignments, patterns, and domain

searches. Of major concern is the origin of the assigned function, because transfer of low-quality or bad functional annotation propagates wrong annotations in the public databases. Even correct annotations can be erroneously disseminated: one can easily imagine the wrong transfer of a good functional assignment from a multidomain protein to a protein that only has one of the domains. This problem can be avoided by using only experimentally derived information to predict unambiguously a gene's structure and function.
Here, we applied a homology-based annotation by using experimental references to build a full catalogue with 61 core cell cycle genes of *Arabidopsis*. In total, 30 genes are new or are genes for which the previous annotation was incorrect. Based on phylogenetic analysis we updated and rationalized their nomenclature. Furthermore, relations between gene family members were correlated with large segmental duplications.

Methods

Annotation of Arabidopsis cell cycle genes

The genome version of January 18, 2001 (v180101) was downloaded from the ftp site (ftp://ftpmips.gsf.de/cress/) of the Martiensried Institute for Protein Sequences (MIPS) center (Martiensried, Germany). Regions of interest on the chromosomes were localized by the BLAST software (Altschul et al., 1997) with experimental representatives as query sequence. For the regions returned by BLAST, chromosome sequences were extracted with 15 kb upstream and downstream from the hit to prevent unreliable prediction due to border effects.
Gene prediction was done with Eugene (Schiex et al., 2001), in combination with GeneMark.hmm (Lukashin and Borodovsky, 1998), because the latter had been reported previously to give the best scores in *Arabidopsis* (Pavy et al., 1999). New analysis (C. Mathé, personal communication), however, showed that Eugene has become the best gene prediction tool for *Arabidopsis*. The Eugene program combines NetGene2 (Tolstrup et al., 1997) and SplicePredictor (Brendel and Kleffe, 1998) for splice site prediction, NetStart (Pedersen and Nielsen, 1997) for translation initiation prediction, Interpolated Markov model-based content sensors, and information from protein, EST, and cDNA matches to predict the final gene model.
The predicted candidate gene products were aligned with the experimental representatives by using CLUSTAL W (Thompson et al., 1994). On the final alignments, HMMer was used to generate profiles for each specific gene family with hidden Markov models. These profiles were then used to search for new family members (Eddy, 1998). The genome-wide non-redundant collection of *Arabidopsis* protein-encoding genes was predicted with GeneMark.hmm. Based on these predictions, we built a database of virtual transcripts (and corresponding protein database) that we designated genome-predicted transcripts (GPTs). Manual annotation was done with Artemis (Rutherford et al., 2000).

Phylogeny and nomenclature

Phylogenetic analysis was performed on more conserved positions of the alignment. Editing of the alignment and reformatting was done with BioEdit (Hall, 1999) and ForCon (Raes and Van de Peer, 1999). Similarity between proteins was based on a BLOSUM62 matrix (Henikoff and Henikoff, 1993). Trees were constructed with various distance and parsimony methods. Distance matrices were calculated based on Poisson, Kimura, or PAM correction and trees were constructed with the Neighbor-joining algorithm by means of the software packages TREECON (Van de Peer and De Wachter, 1994) and PHYLIP (Felsenstein, 1993). The latter was also used for the parsimony analysis. Bootstrap analysis with 500 replicates was performed to test the significance of nodes.

Protein structure analysis

Protein secondary structure prediction was done with PSIpred v2.0 (Jones, 1999).

Segmental duplications in the Arabidopsis genome

For the detection of large segmental duplications, duplicated blocks were identified by a method similar to that by Vision et al. (2000). Initially, protein-coded genes predicted by GeneMark.hmm (in total 26,352 present in our GPT database) were ordered according to the location on the corresponding chromosome. BLASTP was used to identify genes with high sequence similarity and all BLASTP scores were stored in a matrix to be analyzed. Initially, filtering was performed to reduce low-similarity hits (E-value < $1e^{-50}$; Friedman and Hughes, 2001), followed by a procedure to define duplicated blocks in the scoring matrix. Finally, by post-processing only blocks of appropriate size (i.e. blocks containing more than seven genes) were selected.

Results

Strategy

In order to correctly annotate all core cell cycle genes, a strategy was defined that uses as much reliable information as possible, combining experimentally derived data with the best prediction tools available for *Arabidopsis* (see “Methods”). First, experimental representatives for each family were used as bait to locate regions of interest on the different chromosomes. For these selected regions, genes were predicted and candidate genes were validated; the presence of mandatory domains in their gene products was determined by aligning them with the experimental representatives and, if necessary, the predicted gene structure was modified by using the family-related characteristics or ESTs. Still, in some cases, this approach did not allow us to conclude whether a region of interest really coded for a potential gene or whether a candidate gene was a core cell cycle gene. To clarify such situations, a more integrated analysis was performed. First, the members of every family were used to build a profile for that specific family.

By taking into account the new predicted genes for creating the profile, a more “flexible” (i.e. all diversity within a class/subclass being represented) and plant-specific profile could be established. With this new profile, novel family members were sought within a collection of genome-wide predicted *Arabidopsis* proteins. Subsequently, the predicted gene products were again validated or modified by comparing them with those of other family members in a multiple alignment. With this additional approach, we could determine clearly whether the predicted genes were similar to a certain class of cell cycle genes.
To characterize subclasses within the gene families, phylogenetic trees were generated that included reference cell cycle genes from other plants and known genes from *Arabidopsis*; by different methods and statistical analysis of nodes the significance of the derived classification was tested. Based on the position in the tree and the presence of class-specific signatures, genes were named according to the proposed nomenclature rules for cell cycle genes (Renaudin et al., 1996; Joubès et al., 2000). A complete list of core cell cycle genes in *Arabidopsis* in presented in Table 1. Additional data regarding nomenclature and gene models can be found at http://www.plantgenetics.rug.ac.be/bioinformatics/coreCC/.

Annotation and nomenclature

CDK

In yeasts one CDK is sufficient to drive cells through all cell cycle phases, whereas multicellular organisms evolved to use a family of related CDKs, all with specific functions. In plants, two major classes of CDKs have been studied so far, known as A-type and B-type CDKs. The A-type CDKs regulate both the G1-to-S and G2-to-M transitions and the B-type CDKs seem to control the G2-to-M checkpoint only (Hemerly et al., 1995; Magyar et al., 1997; Porceddu et al., 2001). In addition, the presence of C-type CDKs and CDK-activating kinases (CAKs) have been reported (Magyar et al., 1997; Umeda et al., 1998; Joubès et al., 2001). Whereas the latter were shown to regulate the activity of the A-type CDKs, the function of the C-type CDKs remains unknown. Until now, one A-type and four B-type CDKs have been described for *Arabidopsis* (Joubès et al., 2000; Boudolf et al., 2001). Furthermore, C-type CDKs and one CAK have been reported as well (Umeda et al., 1998; Lessard et al., 1999). In alfalfa, one E-type CDK has been identifed, but no counterparts had been found previously in *Arabidopsis* (Magyar et al., 1997). By the homology-based annotation method used here, we identified in total eight CDKs (one A-type, four B-type, two C-type, one E-type) and four CAKs (three D-type and one F-type).
The previously described CAK homolog of *Arabidopsis* (cak1At) differs substantially from the known rice CAK, R2 (Umeda et al., 1998; Yamaguchi et al., 1998). R2 has been suggested to be specific for monocots (Yamaguchi et al., 1998). However, with the rice sequence as experimental reference, three related sequences were identified in *Arabidopsis*, designated CDKD;1, CDKD;2 and CDKD;3 with 75%, 68% and 79% sequence similarity with R2 from rice, respectively. These genes are only distantly related to cak1At, indicating that *Arabidopsis* has two functional classes of CAK.

Table 1. Characteristics of all 61 core cell cycle genes in Arabidopsis

Gene	Chr.	Start[a]	Stop[b]	Strand	Status[c]	Features[d]	ORF name
Arath;CDKA;1	3	18,368,303	18,370,279	+	EXP	PSTAIRE	AT3g48750
Arath;CDKB1;1	3	20,355,861	20,357,226	+	EXP	PPTALRE	AT3g54180
Arath;CDKB1;2	2	16,301,446	16,302,758	+	EXP	PPTALRE	AT2g38620
Arath;CDKB2;1	1	28,430,923	28,429,129	-	EXP	PSTTLRE	AT1g76540
Arath;CDKB2;2	1	7,294,679	7,292,770	-	EXP	PPTTLRE	AT1g20930
Arath;CDKC;1	5	3,224,679	3,221,723	-	AI993037	PITAIRE	AT5g10270
Arath;CDKC;2	5	25,955,460	25,958,387	+	AV439592	PITAIRE	AT5g64960
Arath;CDKD;1	1	27,423,792	27,425,694	+	PRED	NVTALRE	AT1g73690
Arath;CDKD;2	1	24,603,461	24,605,698	+	AV554642	NFTALRE	AT1g66750
Arath;CDKD;3	1	6,206,888	6,209,316	-	AF344314	NITALRE	AT1g18040
Arath;CDKE;1	5	25,465,021	25,463,612	-	BG459367	SPTAIRE	AT5g63610
Arath;CDKF;1	4	13,494,330	13,495,958	+	EXP	none	AT4g28980
Arath;CYCA1;1	1	16,354,762	16,352,618	-	AV556475	LVEVxEEY	AT1g44110
Arath;CYCA1;2	1	28,792,710	28,790,480	-	PRED	LVEVxEEY	AT1g77390
Arath;CYCA2;1	5	8,885,657	8,887,990	+	EXP	LVEVxEEY	AT5g25380
Arath;CYCA2;2	5	3,604,472	3,601,820	-	EXP	LVEVxDDY	AT5g11300
Arath;CYCA2;3	1	5,363,054	5,365,235	+	EXPe	LVEVxEEY	AT1g15570
Arath;CYCA2;4	1	29,923,266	29,925,430	+	AV558333	LVEVxEEY	AT1g80370
Arath;CYCA3;1	5	17,293,193	17,294,681	+	PRED	LVEVxEEY	AT5g43080
Arath;CYCA3;2	1	17,022,212	17,023,757	+	AT50514	LVEVxEEY	AT1g47210
Arath;CYCA3;3	1	17,024,852	17,026,370	+	PRED	LVEVxEEY	AT1g47220
Arath;CYCA3;4	1	17,027,927	17,029,762	+	PRED	LVEVxEEY	AT1g47230
Arath;CYCB1;1	4	16,830,051	16,827,976	-	EXP	HxRF	AT4g37490
Arath;CYCB1;2	5	1,861,577	1,859,551	-	EXP	HxKF	AT5g06150
Arath;CYCB1;3	3	3,627,150	3,625,489	-	EXPf	HxKF	AT3g11520
Arath;CYCB1;4	2	11,548,850	11,552,088	+	PRED	HxKF	AT2g26760
Arath;CYCB2;1	2	7,813,050	7,815,144	+	EXP	HxKF	AT2g17620
Arath;CYCB2;2	4	16,107,598	16,109,617	+	EXP	HxKF	AT4g35620
Arath;CYCB2;3	1	7,137,288	7,135,091	-	PRED	HxKF	AT1g20610
Arath;CYCB2;4	1	28,338,772	28,336,622	-	PRED	HxKF	AT1g76310
Arath;CYCB3;1	1	5,584,476	5,582,409	-	PRED	HxKF	AT1g16330
Arath;CYCD1;1	1	26,148,702	26,150,664	+	EXP	LxCxE	AT1g70210
Arath;CYCD2;1	2	9,704,757	9,703,043	-	EXP	LxCxE	AT2g22490
Arath;CYCD3;1	4	15,563,758	15,565,156	+	EXP	LxCxE	AT4g34160
Arath;CYCD3;2	5	26,836,277	26,837,626	+	AI995751	LxCxE	AT5g67260
Arath;CYCD3;3	3	18,862,632	18,861,289	-	AV527915	LxCxE	AT3g50070
Arath;CYCD4;1	5	26,143,713	26,141,558	-	EXP	LxCxE	AT5g65420
Arath;CYCD4;2	5	3,282,347	3,280,801	+	PRED	no LxCxE	AT5g10440
Arath;CYCD5;1	4	16,885,341	16,886,338	+	AI998509	LFLCxE	AT4g37630
Arath;CYCD6;1	4	1,432,497	1,431,184	-	PRED	no LxCxE	AT4g03270
Arath;CYCD7;1	5	417,084	418,547	+	PRED	LxCxE	AT5g02110
Arath;CYCH;1	5	9,813,161	9,816,075	+	AV560893	none	AT5g27620
Arath;CKS1	2	12,060,430	12,059,793	-	EXP	none	AT2g27960
Arath;CKS2	2	12,061,999	12,061,350	-	AV553882	none	AT2g27970
Arath;DEL1	3	18,079,607	18,081,809	+	EXP	none	AT3g48160
Arath;DEL2	5	4,858,640	4,861,044	+	PRED	none	AT5g14960
Arath;DEL3	3	126,812	124,606	-	EXP	none	AT3g01330
Arath;DPa	5	544,155	844,977	-	EXP	none	AT5g02470
Arath;DPb	5	842,841	845,196	+	EXP	none	AT5g03410
Arath;E2Fa	2	15,268,582	15,271,784	+	EXP	none	AT2g36010
Arath;E2Fb	5	7,431,826	7,434,541	+	EXP	none	AT5g22220
Arath;E2Fc	1	17,356,113	17,358,730	+	EXP	none	AT1g47870
Arath;KRP1	2	10,126,806	10,125,908	-	EXP	none	AT2g23430
Arath;KRP2	3	19,096,470	19,097,325	+	EXP	none	AT3g50630
Arath;KRP3	5	19,794,310	19,792,575	-	EXP	none	AT5g48820
Arath;KRP4	2	14,022,387	14,024,238	+	EXP	none	AT2g32710
Arath;KRP5	3	9,060,905	9,061,654	+	EXP	none	AT3g24810
Arath;KRP6	3	6,617,597	6,616,567	-	EXP	none	AT3g19150
Arath;KRP7	1	18,087,625	18,086,761	-	EXP	none	AT1g49620
Arath;Rb	3	3,919,344	3,913,685	-	AF245395	none	AT3g12280
Arath;WEE1	1	673,409	676,125	+	EXP[g]	none	AT1g02970

[a] Position of start codon on the chromosome.
[b] Position of stop codon on the chromosome.
[c] Expression status of the gene: PRED, prediction; EXP, experimentally characterized; number is EST accession number.
[d] Family-specific protein signatures.
[e] EST BE528080 found for the first exon completes the structural annotation.
[f] Gene structure was determined by using partial mRNA L27224 and AV546264.
[g] Gene structure was determined by using two cDNA sequences, confirming the manual annotation.

To stress this functional difference and to have a more uniform nomenclature, cak1At was renamed as CDKF;1. The phylogenetic relationship among CDKs of *Arabidopsis* are shown in Figure 1.

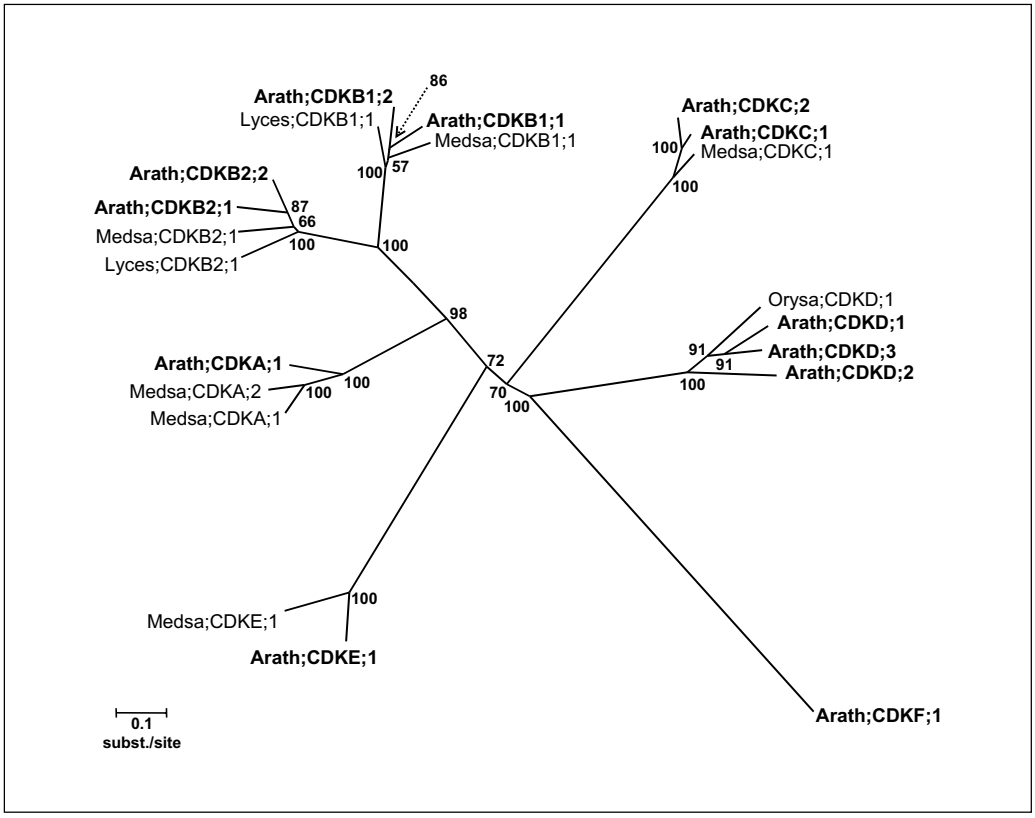

Figure 1. Unrooted Neigbor-joining tree of the A, B, C, D, E, and F class of CDKs with the Poisson correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Scales indicate evolutionary distance. Abbreviations: Arath, Arabidopsis (*Arabidopsis thaliana*); Medsa, alfalfa (*Medicago sativa*); Lyces, tomato (*Lycopersicon esculentum*); Orysa, rice (*Oryza sativa*). Reference genes are Medsa;CDKC;1 (Accession number CAA65979.1), Orysa;CDKD;1 (CAA41172.1), Medsa;CDKE;1 (CAA65981.1), Medsa;CDKA;1 (AAB41817.1), Medsa;CDKA;2 (CAA50038.1), CDKB1;1 (CAA65980.1), Lyces;CDKB1;1 (CAC15503.1), Lyces;CDKB2;1 (CAC15504.1), and Medsa;CDKB2;1 (CAA65982.1).

Cyclins

Monomeric CDKs have no kinase activity and have to associate with regulatory proteins called cyclins to be activated. Because the cyclin protein levels fluctuate in the cell cycle, cyclins are the major factors that determine the timing of CDK activation. Cyclins can be grouped into mitotic cyclins (designated A- and B-type cyclins in higher eukaryotes and CLBs in budding yeast) and G1-specific cyclins (D-type cyclins in mammals and CLNs in budding yeast). H-type cyclins regulate the activity of the CAKs. All four types of cyclins known in plants were identified mostly by analogy to their human counterparts. For *Arabidopsis*, currently four A-type, five B-type, five D-type, but no H-type, cyclins have been described (Soni et al., 1995; Renaudin et al., 1996; De Veylder et al., 1999; Swaminathan et al., 2000).

By using the known plant cyclin sequences as probes, a total of 30 cyclins could be detected in the *Arabidopsis* genome. For 19 cyclins, an EST could be found (Table 1).

Three different subclasses of plant A-type cyclins (A1, A2, and A3) have been described previously (Renaudin et al., 1996) and were all found in *Arabidopsis*, comprising 10 cyclins. Two members of A1-type members (CYCA1;1 and CYCA1;2), four A2-type (CYCA2;1, CYCA2;2, CYCA2;3, and CYCA2;4), and four A3-type genes were detected (CYCA3;1, CYCA3;2, CYCA3;3, and CYCA3;4). B-type cyclins are subdivided into two subclasses, B1 and B2. In total, *Arabidopsis* contains nine B-type cyclins, of which four belong to the B1 class (CYCB1;1; CYB1;2, CYCB1;3, and CYCB1;4) and four to the B2 class (CYCB2;1, CYCB2;1, CYCB2;3, and CYCB2;4). One gene could be attributed neither the B1 nor the B2 classes, although it clearly contained a B-type-like cyclin box in combination with the B-type specific HxKF signature. On the other hand, no B1- nor B2-like destruction box could be detected. The phylogenetic position of this gene within the B cluster depended on the number of positions used for the analysis. Because cyclin sequences are known to be saturated with substitutions (Renaudin et al., 1996), a technique was applied to construct trees on unsaturated positions only (Van de Peer et al., 2001). No support was found to designate this gene to one of the two classes of B-type cyclins (data not shown). On this basis, it seems justified to create a new subclass of cyclins, the B3-type (Figure 2).

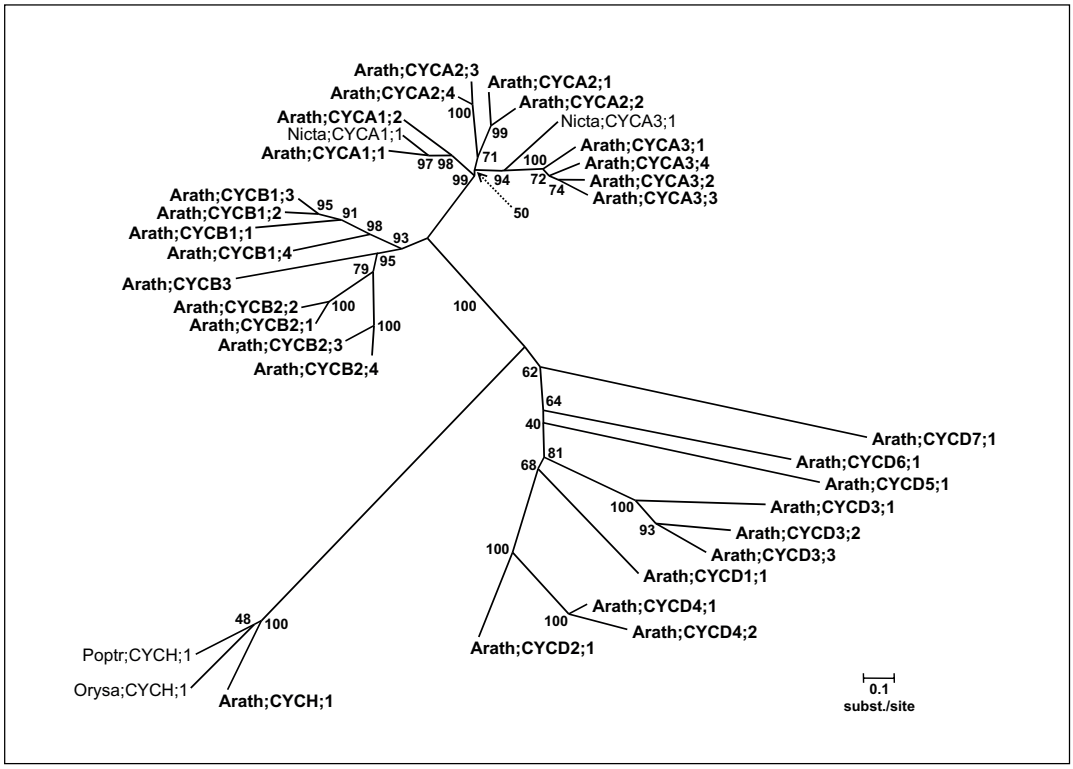

Figure 2. Unrooted Neigbor-joining tree of the A, B, D, and H subgroups of the cyclin family with Poisson correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Scales indicate evolutionary distance. Abbreviations: Arath, Arabidopsis (*Arabidopsis thaliana*); Nicta, tobacco (*Nicotiana tabacum*); Orysa, rice (*Oryza sativa*); Poptr, poplar (*Populus tremula* x *Populus tremuloides*). Reference genes are Nicta;CycA1;1 (Accession number BAA09366.1), Nicta;CycA3;1 (CAA63540.1), Poptr;cycH (AAD02871.1), and Orysa;cycH (BAB11694.1).

In addition to the five D-type cyclins already described (CYCD1;1, CYCD2;1, CYCD3;1, CYCD3;2, and CYCD4;1), five new D-type genes were detected. Based on their phylogenetic position, two were attributed to the D3 (CYCD3;3 and CYCD3;4) and one to the D4 (CYCD4;2) classes. The remaining new D-type cyclins were further subdivided into classes CYCD5, CYCD6, and CYCD7 according to their phylogenetic positions. It is remarkable that CYCD4;2 and CYCD6;1 do not contain the LxCxE retinoblastoma (Rb)-binding motif, whereas CYCD5;1 contains a divergent Rb-binding motif (FxCxE), located at the N-terminus. The biological function of cyclins lacking the conserved Rb-binding motif remains unclear. One *Arabidopsis* gene was found with high sequence similarity to cyclin H of poplar (71%) and rice (66%).
Aligning all cyclins allowed us to identify the cyclin and destruction box consensus sequences for A-, B-, D-, and H-type cyclins (Table 2).

Table 2. Consensus sequences for cyclin and destruction box in Arabidopsis cyclins.

Subclass	Cyclin box signature	Destruction box
Cyclin A1	MR-(I/V)L(I/V)DW	RAPL(G/S)(D/N)ITN
Cyclin A2	MR-(I/V)L(I/V)DW	RAVL(K/G)(D/E)(I/V)(T/S)N
Cyclin A3[a]	MR-(I/V)L(I/V)DW	RVVLGEL(P/L)N
Cyclin B1	MR-IL(I/V/F)DW	R-(A/V)LGDIGN
Cyclin B2	MR-IL(I/V/F)DW	RR(A/V)L--IN
Cyclin B3	TRGILINW	N.D.
Cyclin D1	REDSVAW	N.D.
Cyclin D2	RNQALDW	N.D.
Cyclin D3	R(E/K)(E/K)A(L/V)(D/G)W	N.D.
Cyclin D4	R(R/I)(D/Q)AL(N/G)W	N.D.
Cyclin D5	RLIAIDW	N.D.
Cyclin D6	RNQAISS	N.D.
Cyclin D7	RFHAFQW	N.D.
Cyclin H[b]	MRAFYEAK	N.D.

[a] CycA3;1: cyclin box KRGVLVDW not included in consensus, no destruction box detected.
[b] Plant cyclin H consensus for cyclin box: MR(A/V)(F/Y)YE-K (based on sequence of Arath;*CYCH*, Orysa;*CYCH* (accession number BAB11694) and cyclin H of poplar (*Populus tremula* x *Populus tremuloides*; accession number AAD02871).
N.D., not detected.

Although A- and B-type cyclin boxes are very similar, these two types of cyclins can be discriminated by their destruction boxes. For two genes within the A- and B-type cyclins (CYCA3;1 and CYCB3;1), no destruction box could be detected. In addition, these genes have a highly diverged cyclin box compared with their subclass consensus. The low overall sequence similarity within D-type cyclins is also reflected in their cyclin box.
In addition to the cyclins described above, two presumed pseudogenes were predicted, which were very similar to B-type cyclins. The precise number of pseudogenes for the seven selected families remains unclear, because the detection of pseudogenes depends on the degree of conservation still present in their gene structure and of detection by prediction tools of these degenerated structures.

CDK/cyclin interactors and regulatory proteins

CKS proteins act as docking factors that mediate the interaction of CDKs with putative substrates and regulatory proteins. Besides the already described CDK subunit gene in *Arabidopsis* (Arath;*CKS1*; De Veylder et al., 1997), a second *CKS* gene was found (Arath;*CKS2*) with sequence (83% identical and 90% similar amino acids) and gene structure (number and size of exons and introns) very similar to those of Arath;*CKS1* (Figure 3A). The two CKS gene products miss both the N- and C-terminal extension when compared with the yeast Suc1p/Cks1p homologs (De Veylder et al., 1997). Upon the occurrence of stress or the perception of antiproliferation agents, the CDK/cyclin complexes are repressed by the CDK inhibitor (CKI) proteins. In mammals, two different classes of CKIs exist (the INK4 and the Kip/Cip families), each with their own CDK-binding specificity and protein structure. Seven *CKI* genes, belonging to the group of Kip/Cip CKIs, have been described previously for *Arabidopsis*, designated *KRP1* to *KRP7* (De Veylder et al., 2001). No extra KRPs could be detected in the complete genome and no plant counterparts of the INK4 family were found as well.

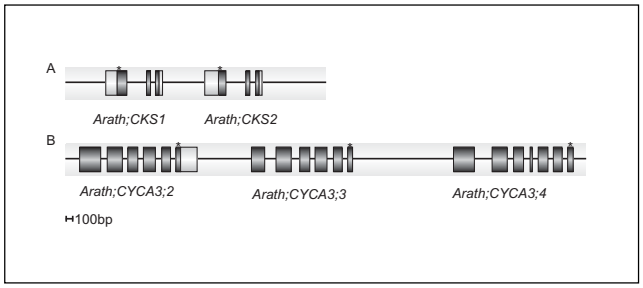

Figure 3. Gene tandem duplication of CKS and A3-type cyclin genes. Black rectangles are protein-encoding exons; white rectangle represent untranslated regions based on hits with EST or mRNA. Asterisks denote the exon with stop codon. (A) Gene structure of *CKS1* and *CKS2* on chromosome 2. The indicated chromosome region spans from 12,059 kb to 12,063 kb. (B) Gene structure of *CYCA3;2*, *CYCA3;3*, and *CYCA3;4* on chromosome 1. The indicated region spans from 17,022 kb to 17,030 kb. ESTs AT50714, AT50514, and AT37419 hit with *CYCA3;2* (data not shown).

CDK/cyclin activity is negatively regulated by phosphorylation of the CDK subunit by the WEE1 kinase and positively when the inhibitory phosphate groups are removed by the CDC25 phosphatase. A single *WEE1* gene was identified on chromosome 1. The WEE1 kinase was annotated by using two cDNA sequences that were at our disposal (L. De Veylder, unpublished results) and has the highest homology to the WEE1 kinase of maize, showing 56% similarity with the gene product of a partial mRNA (Sun et al., 1999). No CDC25 phosphatase could be identified.

Rb and E2F/DP

Rb and the E2F/DP proteins are key regulators that control the entry of DNA replication. When the E2F/DP transcription factors are bound to Rb, they are inactive, but they become active when Rb is phosphorylated by G1-specific CDK/cyclin complexes, stimulating transcription of genes needed for G1-to-S and S phase progression. Only one Rb could be identified in the *Arabidopsis* genome that was located on chromosome 3. *E2F* genes are known for tobacco, carrot, and wheat (Ramírez-Parra et al., 1999; Sekine et al., 1999; Albani et al., 2000; Magyar et al., 2000), but no *Arabidopsis* family members have been described until now, whereas two *Arabidopsis DP* genes (*DPa* and *DPb*) have been reported.

The *E2F* and *DP* genes were analyzed in a combined approach, because the sequence of both types of proteins are partially similar (22% overall similarity). In total, eight genes were detected in *Arabidopsis*. Although the sequence similarity between these eight members of the *E2F/DP* family is rather low (20% overall mean similarity), three groups had emerged based on prior experimental information (Magyar et al., 2000) and phylogenetic analysis (Figure 4).

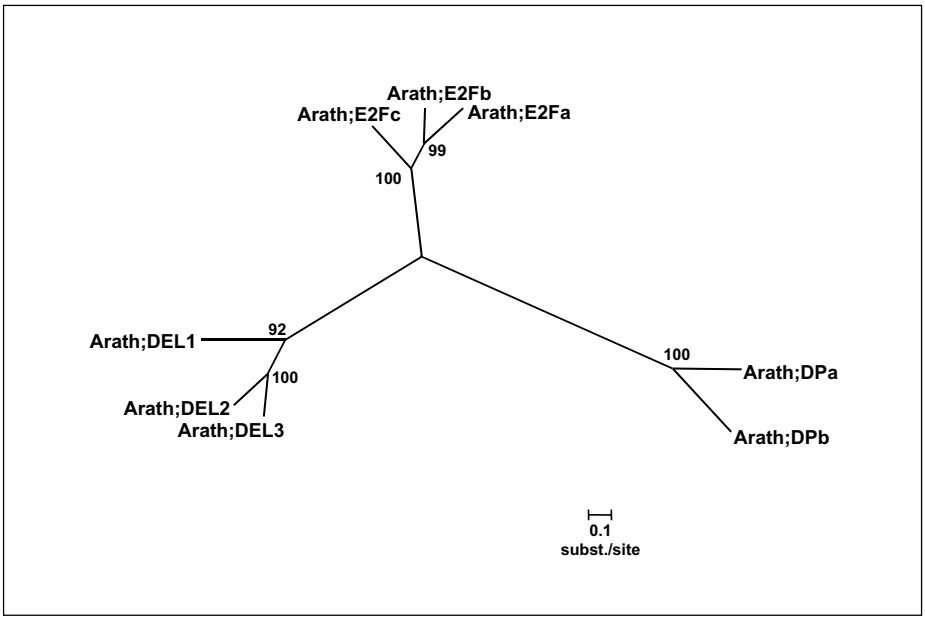

Figure 4. Unrooted Neigbor-joining tree of E2F, DP and DEL families with Poisson correction for evolutionary distance calculation. Bootstrap values of 500 bootstrap iterations are shown. Scales indicate evolutionary distance. Genes are Arath;*E2Fa* (Accession number AF242582), Arath;*E2Fb* (AD242580), Arath;*E2Fc* (AF242581), Arath;*DPa* (AJ294531), and Arath;*DPb* (AJ294532). Abbreviation: Arath, Arabidopsis.

The first group comprises the *E2F* transcription factors that are most similar to the mammalian E2F factors and were designated *E2Fa*, *E2Fb*, and *E2Fc* (46% overall similarity). The second group consists of the two already known DP factors.

The third group contains three new genes with an internal similarity of 59% and a sequence similarity with both *E2F* (21%) and *DP* genes (18%), initially indicating some kind of relation with the *E2F/DP* genes. When the boxes present in the *E2F* genes (DNA-binding, dimerization, Marked and Rb-binding box) and *DP* genes (DNA-binding and dimerization box) were compared with these three new genes, only a DNA-binding domain was found, but in duplex (Figure 5A). Both DNA-binding domains are highly similar to the *E2F* DNA-binding domain. Because of their phylogenetic position, they form a distinct class, which we designated as DP-E2F-like (DEL).

The DNA-binding domain of the *E2F* and *DP* genes have a limited across-family homology (Figure 5B), including the RRxYD DNA recognition motif (in their a3 helices), which interacts with half of the palindromic promoter-binding site (CGCGCG and CGCGCG).

Within all three *DEL* genes, the conserved DNA recognition motif RRxYD is also present in two copies. The *E2F/DP* heterodimer binds and recognizes the palindromic sequence of the binding site in an essentially symmetric arrangement (Zheng et al., 1999).

Protein secondary structure prediction for the *DEL* genes showed that the winged-helix DNA-binding motif, a fold found in the cell cycle transcription factors *E2F/DP* (three a helices and a ß sheet), is present in duplex in all these *DEL* genes. The first and second *DEL* DNA-binding domain have an overall similarity of 61% and 47% with the *E2F* DNA-binding domain, respectively. Currently, no experimental data are available about the putative function and role of the *DEL* genes in cell cycle regulation.

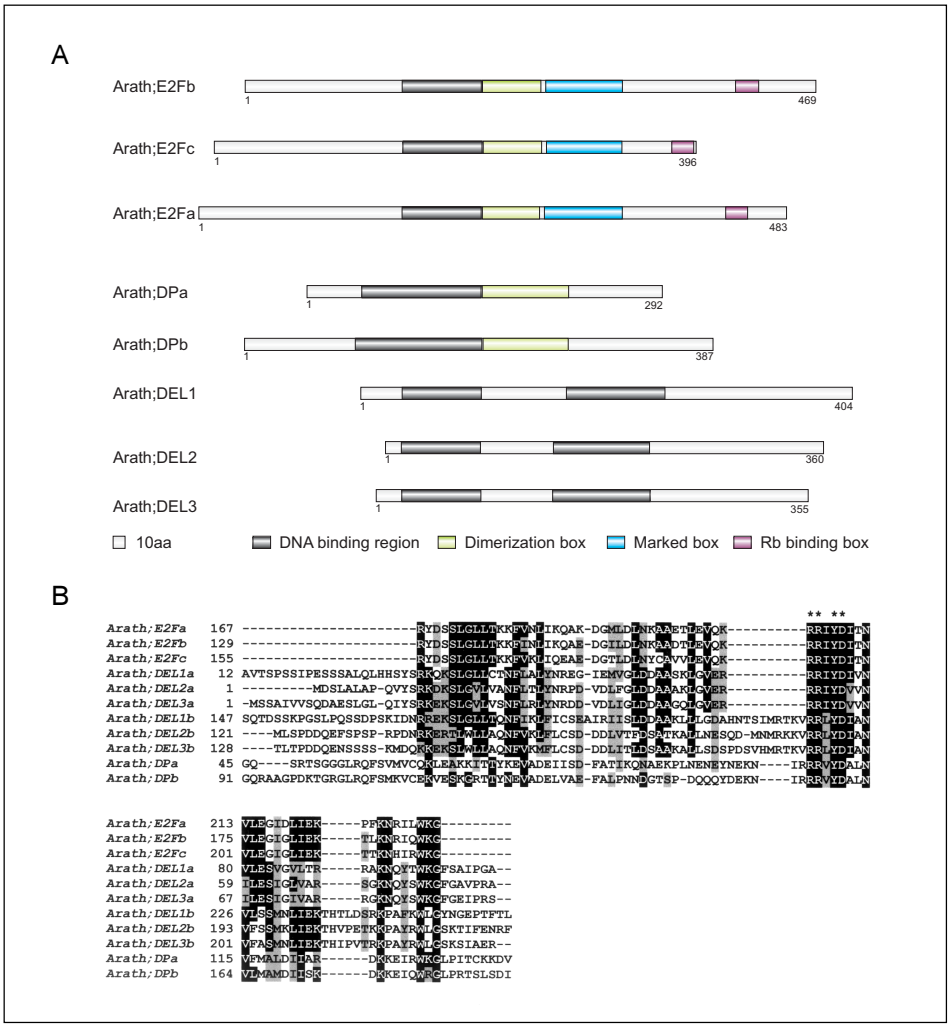

Figure 5. Structural organization of the E2F, DP, and DEL families at the protein level. Numbers indicate protein length in amino acids. (A) Schematic representation of the DNA-binding, dimerization, Marked, and Rb-binding boxes in *E2F*, *DP*, and *DEL* genes of Arabidopsis. (B) Alignment of putative DNA-binding domains of E2F, DP, and DEL proteins. All DEL proteins were split in two (parts a and b) to compare both DNA-binding motifs with those of the E2F and DP. The RRxYD DNA-binding motif is indicated by asterisks.

Gene/Genome organization

In order to find out whether the segmental or genomic duplications and the acquisition of new cell cycle regulation mechanisms are linked, we mapped all cell cycle genes on the five different chromosomes (Figure 6). Subsequently, all duplicated regions in the *Arabidopsis* genome were defined and the position of every cell cycle gene was compared with the coordinates of each duplicated block.

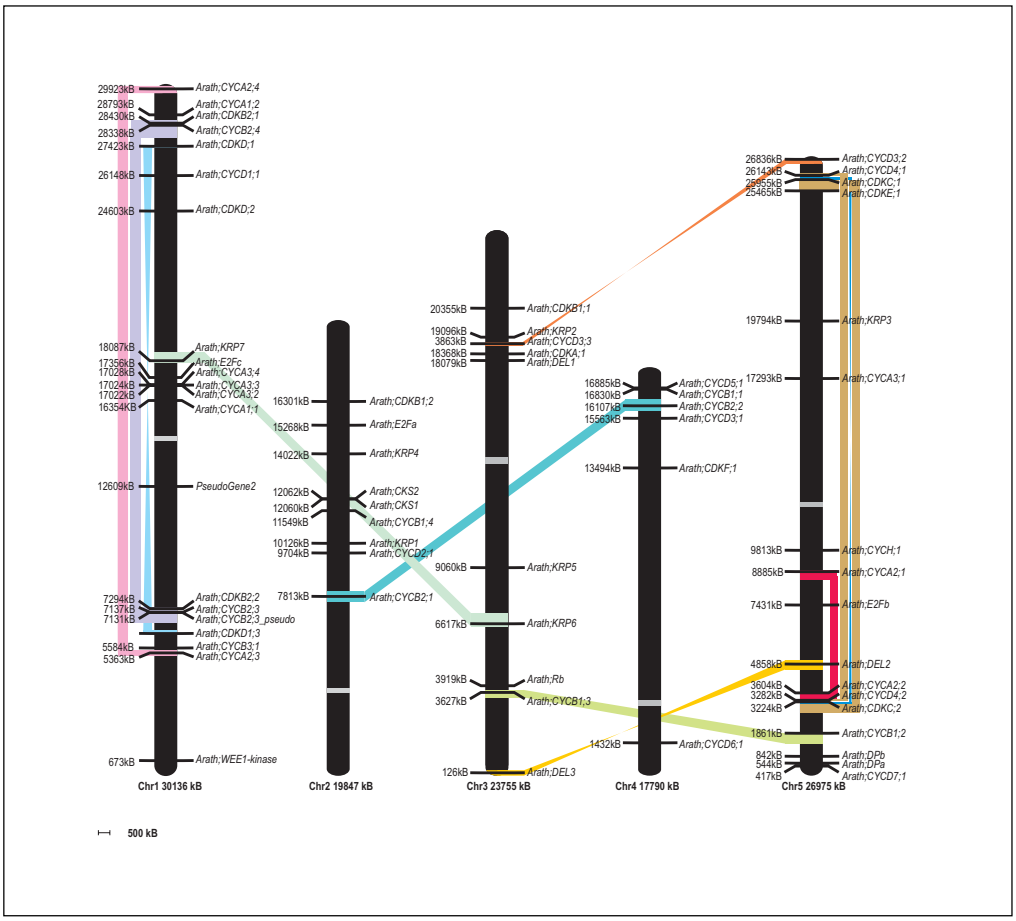

Figure 6. Physicial position of core cell cycle genes on the Arabidopsis genome. Segmental duplicated regions are only drawn when a cell cycle gene is present in a duplication event. Colored bands connect corresponding duplicated blocks. Duplicated blocks in reverse orientation are connected with twisted colored bands. Centromeres are represented as grey boxes.

Comparison of the position of A2 cyclin genes with the position of duplicated blocks in the *Arabidopsis* genome revealed that all four members are located in duplicated blocks: one internal duplication on chromosome 1 (*CYCA2;3* linked with *CYCA2;4*) and one on chromosome 5 (*CYCA2;2* linked with *CYCA2;1*). The three *CYCA3* genes were organized in tandem (*CYCA3;2*, *CYCA3;3*, and *CYCA3;4* spanning a region of less than 8 kb) and have a highly similar gene structure (number and size of exons and introns), as well as highly similar protein sequences (74.3% overall similarity). Only *CYCA3;2*

had one significant EST hit, whereas *CYCA3;4* had an additional small predicted exon (33 nucleotides) when compared with the other *CYCA3* genes that occur in the same tandem (Figure 3B).
Similar to the A2-type cyclins, all four B2-type cyclins were located within duplicated blocks: one duplicated block between chromosomes 2 and 4 (linking *CYCB2;1* and *CYCB2;2*) and one internal duplication on chromosome 1 (linking *CYCB2;3* and *CYCB2;4*).
Although in total 10 D-type cyclins were detected, only few of them were located in duplicated blocks. *CYCD3;2* and *CYCD3;3* are members of an inverted block between chromosome 5 and 3, whereas *CYCD4;1* and *CYCD4;2* are located within an internal block of chromosome 5.
The two *CKS* genes were located in a gene tandem duplication, where the stop codon of *CKS2* was separated by only 916 bp from the start codon of *CKS1* (Figure 3A).
Special attention is required for two duplication events. On chromosome 1, a large internal duplication occurred (spanning an area of approximately 4890 kb or 16% of chromosome 1) that was followed by several inversions (data not shown), leading to the formation of multiple smaller blocks, one of which contained two pairs of cell cycle genes: *CDKB2;2* linked with *CDKB2;1* and *CYCB2;3* linked with *CYCB2;4*. The *CYCB2;3* gene was present in tandem (interspersed by one gene) and the second copy was designated Arath;*CYCB2;3_pseudo*, because its gene structure was degraded and imperfect with respect to *CYCB2;3*. We conclude that this tandem duplication occurred after the segmental duplication event, because in the region linked to the duplicated block, no trace of another extra B2-like cyclin was found.
Another special, internally duplicated event was found on chromosome 5. Two duplicated blocks (Figure 6, brown blocks) were detected that connected both extremities of the chromosome. Although these blocks could be regarded as one, we clearly distinguished an invertedly duplicated block in between (Figure 6, blue block). *CYCD4;1* and *CYCD4;2* both fit nicely into the first block. *CDKC;1* and *CDKC;2* mapped in this region as well, located in the small invertedly duplicated block. It is remarkable that, although both couples of linked genes were located in duplicated blocks with different orientations, their relative positions were the same (i.e. at the bottom and the top of chromosome 5, a C-type CDK was followed by a D4-type cyclin). This configuration suggests that initially one large duplication event occurred (Figure 6; the region spanning brown and blue blocks) that was later reshuffled by inversions (and perhaps some deletions), resulting in adjacent, duplicated blocks with different orientations and sizes.

Discussion

The members of the *Arabidopsis* genome sequencing consortia use different tools to perform automated genome annotations together with similarities to ESTs and known protein sequence to refine gene models. This procedure has generated a large quantity of information on the *Arabidopsis* gene repertoire. However, the extraction of clear biological information for a particular process from these public databases is not always that trivial (for instance, the word ‘cyclin’ as query in the MIPS database returned 37 hits with 23 putative cyclin or cyclin-like hits). To solve this problem, we designed a protocol, mainly focused on high-quality homology-based annotation.

We used a combination of two selected high-quality *Arabidopsis* prediction tools (Pavy et al., 1999; Schiex et al., 2001; C. Mathé and P. Rouzé, personal communication), together with pure experimental information as reference material. A first advantage of this method is that the chance of finding new and rarely expressed genes is maximized because it is structurally characterized by tools with higher specificity and sensitivity than those used by the different consortia for generating genome annotation (Gopal et al., 2001). Secondly, focus on families with available experimental references allows comparisons with functionally well-characterized genes and diminishes the risk of propagation of wrong annotation is diminished. In addition, the use of hidden Markov profiles, which represent the complete diversity within a family, is clearly more powerful than that of a single sequence for remote-homolog detection (Karplus et al., 1998).

With this strategy, we have built a catalogue of 61 core cell cycle genes, belonging to seven selected families. From these, 30 had not been described before and for 22 of them the gene prediction provided by the *Arabidopsis* Genome Initiative was incorrect. Corrected gene models have been submitted to TAIR and can also be found at the web site http://www.plantgenetics.rug.ac.be/bioinformatics/coreCC/. These results highlight the complexity of the cell cycle regulation in *Arabidopsis*, indicating a larger variety of genes than what was currently known experimentally.

Like in mammals, plants evolved to use different classes of CDKs to regulate their cell cycle. In *Arabidopsis*, a total of six different CDK classes can be identified, designated from A through F. Although some of these CDKs have been proven to be active during specific phases of the cell cycle (Magyar et al., 1997; Porceddu et al., 2001; Sorrell et al., 2001), no functional correlation can be made with CDKs of other eukaryotes on the basis of protein sequences. For example, no clear ortologs can be identified for the mammalian G1/S-specific CDK4 and CDK6, suggesting that plants developed independently additional CDKs for more specialized functions in the cell cycle control. This hypothesis is in agreement with the observation that the cyclin-binding motifs found in the plant B-type CDKs cannot be found in any CDK of other eukaryotes.

Within the CDK family, we identified three new CAK members, being close homologs of the rice *R2* gene (Hata, 1991). These CAKs (CDKD;1, CDKD;2 and CDKD;3) differ structurally from the previously isolated *Arabidopsis* cak1At, renamed CDKF;1. The high sequence diversity (35% overall sequence similarity between D- and F-type CDKs) suggests that plants utilize two distinct classes of CAKs. When the *Arabidopsis* CDKF;1 is compared with the rice R2, both classes are functionally different: they both can complement yeast CAK mutant strains, but show a different substrate specificity; the rice R2 phosphorylates both CDKs and the carboxyl-terminal domain of the largest subunit of RNA polymerase II, whereas CDKF;1 phosphorylates CDKs only (Umeda et al., 1998; Yamaguchi et al., 1998).

The complexity of the cyclin gene family appears to be higher in plants than in mammals. Compared to human, *Arabidopsis* has approximately 14 more A- and B-type cyclins, and seven more D-type cyclins. A major part of the A-cyclins originated through large segmental duplications. For the 10 A-type cyclins, all four members of the A2-type subclass are part of duplicated blocks and three genes out of the four A3-type cyclins are organized in tandem.

Several analyses of the *Arabidopsis* genome sequence had already concluded that genes had duplicated extensively in the history of the model plant. More than 50% of the genes in *Arabidopsis* belong to a gene family with three or more members. After analyzing regions of chromosomes 2, 4 and 5, Blanc et al. (2000) estimated that more than 60% of the genome consisted of duplicated regions and suggested the possibility that *Arabidopsis* was an ancient tetraploid. In a later analysis, Vision et al. (2000) concluded that in fact several large independent duplications of chromosome segments had happened at different time points in the plants' evolution. This view was blurred by extensive deletion, inversion and translocation of genes and chromosome segments, as well as smaller and tandem gene duplications (The *Arabidopsis* Genome Initiative, 2000; Vision et al., 2000). In our analysis, we detected that 22 core cell cycle genes are part of a segmental duplication in the *Arabidopsis* genome. Whether there is functional redundancy within A- and B-type cyclins, or whether some cyclin subclasses are differently regulated (and expressed) will have to be analyzed.

In contrast to the A- and B-type cyclins, D-type cyclins lack high sequence similarity among each other, which is reflected within the phylogenetic analysis resulting in seven D-type subclasses. When compared with A- and B-type cyclins, of which some complete subclasses (A2 and B2) are located within segmentally duplicated blocks, no large duplications can be found for the D-type cyclins. Only the D3 and D4 subclasses have different members. Redundancy of the D3-type cyclins has been proposed previously as an explanation of the failure to observe mutant phenotypes, when knocking out a single D3-type cyclin (Swaminathan et al., 2000). Our analysis clearly confirms this hypothesis: the fact that two D3-type cyclins are linked via a recent segmental duplication strengthens our belief that these D3 cyclins are functionally redundant. A similar hypothesis could hold for D4-type cyclins, because two out of three are located in a duplicated block.

The much larger divergence seen for D-type cyclins when compared to A- and B-type cyclins might reflect the presumed role of D-type cyclins in integrating developmental signals and environmental cues into the cell cycle. For example, D3-type cyclins have been shown to respond to plant hormones, such as cytokinins and brassinosteroids, whereas *CYCD2* and *CYCD4* are activated earlier in G1 and react to sugar availability (for review, see Stals and Inzé, 2001). Because of the large number of various D-type cyclins with different response to developmental and environmental signals, cell division and growth in sessile plants might be more flexible than what is observed in other eukaryotes.

Whereas plants clearly share all elements needed for G1/S entry with other higher eukaryotes, they lack the typical class of E-type cyclins, known to be essential regulators of DNA replication (Duronio et al., 1996). Presumably some of the A- or D-type cyclins take over the role of the E-type cyclins. Also the lack of a consensus Rb-binding motif in some D-type cyclins suggests that some cyclins might have gained other novel functions during evolution. Alternatively, some of the core cell cycle genes might have undergone such dramatic changes during evolution that they cannot be recognized anymore as functional homologs of animal and yeast counterparts, of which the *CDC25* gene is the most likely example. Both the presence of the antagonistic WEE1 kinase and accumulating biochemical evidence point to the existence of a CDC25 phosphatase in plants (Zhang et al., 1996; Sun et al., 1999), although it could not be identified as such in the *Arabidopsis* genome.

It is surprising that mammals and plants have approximately the same number of core cell cycle genes, with the exception of the above described difference in cyclin number. Complex, multicellular organisms may need many more cell cycle genes to coordinate cell cycle progression with the diverse developmental pathways. Therefore, the pool of mammalian cell cycle genes is probably larger than expected because of the frequent occurrence of alternative splicing. For example, spliced variants of cyclin E are known, with an expression profile and substrate specificity different from that of cyclin E itself (Mumberg et al., 1997; Porter and Keyomarsi, 2000). At least five distinct *DP-2* mRNAs are synthesized in a tissue-specific fashion (Rogers et al., 1996). Depending on the splice variant, the DP family members lack a nuclear localization signal and, when associated with E2F, these different DP molecules have opposing effects on the E2F/DP activity (De la Luna et al., 1996). Furthermore, alternative splicing in humans is known for CDKs, CDC25, and CKIs (Wegener et al., 2000; Hirano et al., 2001; Herrmann and Mancini, 2001). For cell cycle genes of plants, only one case of alternative splicing has been reported (Sun et al., 1997).
E2F/DP transcription factors are characterized by the presence of both a DNA-binding and transcription activation domain. Binding of these transcription factors to the E2F/DP palindromic binding site is mediated by a small DNA recognition motif (RRxYD). By scanning the genome for E2F/DP-related proteins, a putatively novel class of cell cycle-regulating genes was identified, designated DEL. The DEL proteins have two E2F-like DNA-binding boxes, each including the RRxYD motif, but have no activation domain. By competing for the same DNA binding sites, monomeric DEL proteins could act as competitors of the E2F/DP proteins and, because they lack an activation domain, they would act as a repressor of E2F/DP-regulated genes. This mechanism would avoid G1-to-S transition, in cases where conditions are not appropriate for entry in the S phase (such as DNA damage and stress). This new class of putative cell cycle regulators seems not to be plant specific, because one homolog was found in *Caenorabditis elegans* (data not shown).
In conclusion, our genome-wide analysis demonstrated an unexpected complexity of the core cell cycle machinery in plants that is comparable with that seen in mammals. The major challenge for the future is to understand the specific role of all these individual genes in regulating cell division during plant development.

Acknowledgements

We especially thank Yvan Saeys for providing us with the necessary programs to define duplicated blocks in the *Arabidopsis* genome, Dr. Yves Van de Peer for help with the analysis of saturated positions in the cyclin alignments, Dr. Catherine Mathé for additional information about Eugene, Patrice Déhais for the programs developed to run Eugene, and Martine De Cock and Rebecca Verbanck for help in preparing the manuscript and artwork, respectively. This work was supported by grants from the Belgian Programme on Interuniversity Poles of Attraction (Prime Minister's Office, Science Policy Programming #38), the European Union (ECCO QLG2-CT1999-00454), Génoplante (project BI1999087), and CropDesign N.V. (0235). K.V. is indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie for a predoctoral fellowship and L.D.V. is a postdoctoral fellow of the Fund for Scientific Research (Flanders).

Note added in proof

The postulated function of the DEL proteins has recently been confirmed (Mariconti, L., Pellegrini, B., Cantoni, R., Stevens, R., Bergounioux, C., Cella, R., and Albani, D. [January 10, 2002] J. Biol. Chem. 10.1074/jbc.M110616200), but the gene prediction for one DEL family member (*E2Ff~DEL3*) differs from the one we present here. The gene structure we propose has been validated experimentally in our laboratory.

References

Albani, D., Mariconti, L., Ricagno, S., Pitto, L., Moroni, C., Helin, K., and Cella, R. (2000). DcE2F, a functional plant E2F-like transcriptional activator from *Daucus carota*. J. Biol. Chem. **275,** 19258-19267.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25,** 3389-3402.

The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* (London) 408, 796-815.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12,** 1093-1101.

Boudolf, V., Rombauts, S., Naudts, M., Inzé, D., and De Veylder, L. (2001). Identification of novel cyclin-dependent kinases interacting with the CKS1 protein of *Arabidopsis*. J. Exp. Bot. **52,** 1381-1382.

Brendel, V., and Kleffe, J. (1998). Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucleic Acids Res. **26,** 4748-4757.

de la Luna, S., Burden, M.J., Lee, C.-W., and La Thangue, N.B. (1996). Nuclear accumulation of the E2F heterodimer regulated by subunit composition and alternative splicing of a nuclear localization signal. J. Cell Sci. **108,** 2443-2452.

De Veylder, L., Segers, G., Glab, N., Casteels, P., Van Montagu, M., and Inzé, D. (1997). The *Arabidopsis* Cks1At protein binds to the cyclin-dependent kinases Cdc2aAt and Cdc2bAt. FEBS Lett. **412,** 446-452.

De Veylder, L., Beeckman, T., Beemster, G.T.S., Krols, L., Terras, F., Landrieu, I., Van Der Schueren, E., Maes, S., Naudts, M., and Inzé, D. (2001). Functional analysis of cyclin-dependent kinase inhibitors of *Arabidopsis*. Plant Cell **13,** 1653-1667.

De Veylder, L., De Almeida Engler, J., Burssens, S., Manevski, A., Lescure, B., Van Montagu, M., Engler, G., and Inzé, D. (1999). A new D-type cyclin of *Arabidopsis thaliana* expressed during lateral root primordia formation. Planta **208,** 453-462.

Devos, D., and Valencia, A. (2001). Intrinsic errors in genome annotation. Trends Genet. **17,** 429-431.

Duronio, R.J., Brook, A., Dyson, N., and O'Farrell, P.H. (1996). E2F-induced S phase requires *cyclin E*. Genes Dev. **10,** 2505-2513.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package) version 3.5c*. Seattle: Department of Genetics, University of Washington.

Friedman, R., and Hughes, A.L. (2001). Gene duplication and the structure of eukaryotic genomes. Genome Res. **11,** 373-381.

Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A., Aytekin-Kurban, G., Bekiranov, S., Fajardo, J.E., Eswar, N., Sanchez, R., Sali, A., and Gaasterland, T. (2001). Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. Nature Genet. **27,** 337-340.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids. Symp. Ser. **41,** 95-98.

Hata, S. (1991). cDNA cloning of a novel $cdc2^{+}$/CDC28-related protein kinase from rice. FEBS Lett. **279,** 149-152.

Hemerly, A., de Almeida Engler, J., Bergounioux, C., Van Montagu, M., Engler, G., Inzé, D., and Ferreira, P. (1995). Dominant negative mutants of the Cdc2 kinase uncouple cell division from iterative plant development. EMBO J. **14,** 3925-3936.

Henikoff, S., and Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. Proteins **17,** 49-61.

Herrmann, C.H., and Mancini, M.A. (2001). The Cdk9 and cyclin T subunits of TAK/P-TEFb localize to splicing factor-rich nuclear speckle regions. J. Cell Sci. **114,** 1491-1503.

Hirano, K., Hirano, M., Zeng, Y., Nishimura, J., Hara, K., Muta, K., Nawata, H., and Kanaide, H. (2001). Cloning and functional expression of a degradation-resistant novel isoform of $p27^{Kip1}$. Biochem. J. **353,** 51-57.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292,** 195-202.

Joubès, J., Chevalier, C., Dudits, D., Heberle-Bors, E., Inzé, D., Umeda, M., and Renaudin, J.-P. (2000). CDK-related protein kinases in plants. Plant Mol. Biol. **43,** 607-620.

Joubès, J., Lemaire-Chamley, M., Delmas, D., Walter, J., Hernould, M., Mouras, A., Raymond, P., and Chevalier, C. (2001). A new C-type cyclin-dependent kinase from tomato expressed in dividing tissues does not interact with mitotic and G1 cyclins. Plant Physiol. **126,** 1403-1415.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics **14,** 846-856.

Lessard, P., Bouly, J.-P., Jouannic, S., Kreis, M., and Thomas, M. (1999). Identification of cdc2cAt: a new cyclin-dependent kinase expressed in *Arabidopsis thaliana* flowers. Biochim. Biophys. Acta **1445,** 351-358.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26,** 1107-1115.

Magyar, Z., Mészáros, T., Miskolczi, P., Deák, M., Fehér, A., Brown, S., Kondorosi, E., Athanasiadis, A., Pongor, S., Bilgin, M., Bakó, L., Koncz, C., and Dudits, D. (1997). Cell cycle phase specificity of putative cyclin-dependent kinase variants in synchronized alfalfa cells. Plant Cell **9,** 223-235.

Magyar, Z., Atanassova, A., De Veylder, L., Rombauts, S., and Inzé, D. (2000). Characterization of two distinct DP-related genes from *Arabidopsis thaliana*. FEBS Lett. **486,** 79-87.

Mumberg, D., Wick, M., Bürger, C., Haas, K., Funk, M., and Müller, R. (1997). Cyclin E_T, a new splice variant of human cyclin E with a unique expression pattern during cell cycle progression and differentiation. Nucleic Acids Res. **25,** 2098-2105.

Pavy, N., Rombauts, S., Déhais, P., Mathé, C., Ramana, D.V.V., Leroy, P., and Rouzé, P. (1999). Evaluation of gene prediction software using a genomic data set: application of *Arabidopsis thaliana* sequences. Bioinformatics **15,** 887-899.

Pedersen, A.G., and Nielsen, H. (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, eds (Menlo Park: American Association for Artificial Intelligence Press), pp. 226-233.

Porceddu, A., Stals, H., Reichheld, J.-P., Segers, G., De Veylder, L., De Pinho Barrôco, R., Casteels, P., Van Montagu, M., Inzé, D., and Mironov, V. (2001). A plant-specific cyclin-dependent kinase is involved in the control of G_2/M progression in plants. J. Biol. Chem. **276,** in press.

Porter, D.C., and Keyomarsi, K. (2000). Novel splice variants of cyclin E with altered substrate specificity. Nucleic Acids Res. **28,** e101.

Raes, J., and Van de Peer, Y. (1999). ForCon: a software tool for the conversion of sequence alignments. EMBnet.news **6** (http://www.ebi.ac.uk/embnet.news/vol6_1).

Ramírez-Parra, E., Xie, Q., Boniotti, M.B., and Gutierrez, C. (1999). The cloning of plant E2F, a retinoblastoma-binding protein, reveals unique and conserved features with animal G_1/S regulators. Nucleic Acids Res. **27,** 3527-3533.

Renaudin, J.-P., Doonan, J.H., Freeman, D., Hashimoto, J., Hirt, H., Inzé, D., Jacobs, T., Kouchi, H., Rouzé, P., Sauter, M., Savouré, A., Sorrell, D.A., Sundaresan, V., and Murray, J.A.H. (1996). Plant cyclins: a unified nomenclature for plant A-, B- and D-type cyclins based on sequence organization. Plant Mol. Biol. **32,** 1003-1018.

Rogers, K.T., Higgins, P.D.R., Milla, M.M., Phillips, R.S., and Horowitz, J.M. (1996). DP-2, a heterodimeric partner of E2F: identification and characterization of DP-2 proteins expressed *in vivo*. Proc. Natl. Acad. Sci. USA **93,** 7594-7599.

Rouzé, P., Pavy, N., and Rombauts, S. (1999). Genome annotation: which tools do we have for it? Curr. Opin. Plant Biol. **2,** 90-95.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16,** 944-945.

Schiex, T., Moisan, A., and Rouzé, P. (2001). EuGène: an eukaryotic gene finder that combines several sources of evidence. In Computational Biology: Selected Papers, (Lecture Notes in Computer Science, Vol. 2066), O. Gascuel, and M.-F. Sagot, eds (Berlin: Springer-Verlag), pp. 111-125 [ISBN 3-540-42242-0].

Sekine, M., Ito, M., Uemukai, K., Maeda, Y., Nakagami, H., and Shinmyo, A. (1999). Isolation and characterization of the E2F-like gene in plants. FEBS Lett. **460,** 117-122.

Soni, R., Carmichael, J.P., Shah, Z.H., and Murray, J.A.H. (1995). A family of cyclin D homologs from plants differentially controlled by growth regulators and containing the conserved retinoblastoma protein interaction motif. Plant Cell **7,** 85-103.

Sorrell, D.A., Menges, M., Healy, J.M.S., Deveaux, Y., Amano, X., Su, Y., Nakagami, H., Shinmyo, A., Doonan, J.H., Sekine, M., and Murray, J.A.H. (2001). Cell cycle regulation of cyclin-dependent kinases in tobacco cultivar Bright Yellow-2 cells. Plant Physiol. **126,** 1214-1223.

Sun, Y., Flannigan, B.A., Madison, J.T., and Setter, T.L. (1997). Alternative splicing of cyclin transcripts in maize endosperm. Gene **195,** 167-175.

Sun, Y., Dilkes, B.P., Zhang, C., Dante, R.A., Carneiro, N.P., Lowe, K.S., Jung, R., Gordon-Kamm, W.J., and Larkins, B.A. (1999). Characterization of maize (*Zea mays* L.) Wee1 and its activity in developing endosperm. Proc. Natl. Acad. Sci. USA **96,** 4180-4185.

Stals, H., and Inzé, D. (2001). When plant cells decide to divide. Trends Plant Sci. **6,** 359-364.

Swaminathan, K., Yang, Y., Grotz, N., Campisi, L., and Jack, T. (2000). An enhancer trap line associated with a D-class cyclin gene in *Arabidopsis*. Plant Physiol. **124,** 1658-1667.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22,** 4673-4680.

Tolstrup, N., Rouzé, P., and Brunak, S. (1997). A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res. **25,** 3159-3163.

Umeda, M., Bhalerao, R.P., Schell, J., Uchimiya, H., and Koncz, C. (1998). A distinct cyclin-dependent kinase-activating kinase of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **95,** 5021-5026.

Van de Peer, Y., and De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. **10,** 569-570.

Van de Peer, Y., Frickey, T., Taylor, J.S., and Meyer, A. (2001). Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene**,** submitted.

Vision, R.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in *Arabidopsis*. Science **290,** 2114-2117.

Wegener, S., Hampe, W., Herrmann, D., and Schaller, H.C. (2000). Alternative splicing in the regulatory region of the human phosphatases CDC25A and CDC25C. Eur. J. Cell Biol. **79,** 810-815.

Yamaguchi, M., Umeda, M., and Uchimiya, H. (1998). A rice homolog of Cdk7/MO15 phosphorylates both cyclin-dependent protein kinases and the carboxy-terminal domain of RNA polymerase II. Plant J. **16,** 613-619.

Zhang, K., Letham, D.S., and John, P.C.L. (1996). Cytokinin controls the cell cycle at mitosis by stimulating the tyrosine dephosphorylation and activation of $p34^{cdc2}$-like H1 histone kinase. Planta **200,** 2-12.

Zheng, N., Fraenkel, E., Pabo, C.O., and Pavletich, N.P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev. **13,** 666-674.

[Chapter 7]

Investigating ancient duplication events in the *Arabidopsis* genome

Jeroen Raes†, Klaas Vandepoele†, Cedric Simillion, Yvan Saeys and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium;
* *Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*
† *The two authors contributed equally to this work.*

Published in: Journal of Structural and Functional Genomics **3**, 117-129 (2003)

Abstract

The complete genomic analysis of *Arabidopsis thaliana* has shown that a major fraction of the genome consists of paralogous genes that probably originated through one or more ancient large-scale gene or genome duplication events. However, the number and timing of these duplications still remains unclear, and several different hypotheses have been put forward recently. Here, we reanalyzed duplicated blocks found in the *Arabidopsis* genome described previously and determined their date of divergence based on silent substitution estimations between the paralogous genes and, where possible, by phylogenetic reconstruction. We show that previously used methods based on averaging protein distances of heterogeneous classes of duplicated genes lead to unreliable conclusions and that a large fraction of blocks duplicated much more recently than assumed previously. We found clear evidence for one large-scale gene or even complete genome duplication event somewhere between 70 to 90 million years ago. Traces pointing to a much older (probably more than 200 million years) large-scale gene duplication event could be detected. However, for now it is impossible to conclude wether these old duplicates are the result of one or more large-scale gene duplication events.

Introduction

For over 30 years, geneticists, evolutionists and, more recently, developmental biologists have been debating on the number of genome duplications in the evolution of animal lineages and its impact on major evolutionary transitions and morphological novelties. Thanks to the recent progress made in gene mapping studies and large-scale genomic sequencing, the debate has been livelier than ever before. Indeed, huge amounts of sequence data have become available, amongst which the complete genome sequences of invertebrates, such as *Drosophila melanogaster*, *Caenorhabditis elegans*, and vertebrates, such as pufferfish and human, while others are being finalized. With these data at our disposition, we expect to address the ancient questions and hypotheses regarding genome duplications, as formulated by pioneers like J.B.S. Haldane (who already in 1933 contemplated the benefits and evolutionary impact of polyploidy events) and S. Ohno. However, a great deal of controversy still exists on the prevalence of genome duplications in certain lineages. For example, the classic hypothesis of Ohno (1970) that at least one genome duplication occurred in the evolution of the vertebrates has not been evidenced yet. Several theories, which differ in the proposed number of duplications as well as in their timing, have been proposed, but without confirmation (Skrabanek and Wolfe, 1998; Hughes, 1999; Wolfe, 2001). More recently, a putatively ancient fish-specific genome duplication before the teleost radiation has been the subject of lively debate (Robinson-Rechavi et al., 2001; Taylor et al., 2001a, 2001b; Van de Peer et al., this issue). Given the already controversial nature of the occurrence and date of these genome duplications in vertebrates, their precise role in the evolution of new body plans (Holland, 1992) or in speciation (Lynch and Conery, 2001; Taylor et al., 2001c) remains even more speculative.

For plants, controversy about ancient genome duplications has long been nearly nonexisting. Polyploidy seems to have occurred frequently in plants. Up to 80% of angiosperms are estimated to be polyploid, with variation from tetraploidy (maize) and hexaploidy (wheat) to 80-ploidy (*Sedum suaveolens*) (for a review, see Leitch et al., 1997). Because of the complexity of many plant genomes and lack of sequence data, research on plant genome evolution was basically restricted to experimental techniques (Wendel, 2000) and, until very recently, few computational analyses had been performed to investigate the prevalence and timing of older large-scale duplications and their impact on plant evolution.

In 1996, the plant community decided to determine the complete genome sequence of *Arabidopsis thaliana*. This model plant was chosen because it has a small genome with a high gene density and seemed to be an ‘innocent’ diploid. However, during and even before this huge enterprise, some indications were found that large-scale duplications had occurred (Kowalski et al., 1994; Paterson et al., 1996; Terryn *et al*, 1999; Lin et al., 1999; Mayer et al., 1999). After bacterial artificial chromosome sequences representing approximately 80% of the genome had been analyzed, almost 60% of the genome was found to contain duplicated genes and regions (Blanc et al., 2000). This phenomenon could only be explained by a complete genome duplication event, an opinion shared by the *Arabidopsis* Genome Initiative (2000). Previously, comparative studies of bacterial artificial chromosomes between *Arabidopsis* and soybean (Grant et al., 2000) and between *Arabidopsis* and tomato (Ku et al., 2000) had led to similar notions.

In the latter study, two complete genome duplications were proposed: one 112 and another 180 x 10^6 years ago (MYA). Vision et al. (2000) rejected the single-genome duplication hypothesis by dating duplicated blocks through a molecular clock analysis. Several different age classes among the duplicated blocks were found, ranging from 50 to 220 MYA and at least four rounds of large-scale duplications were postulated. One of these classes, dated approximately 100 MYA, grouped nearly 50% of all the duplicated blocks, suggesting a complete genome duplication at that time (Vision et al., 2000). However, the dating methods used for these gene duplications were based on averaging evolutionary rates of different proteins, which was later criticized because of their high sensitivity to rate differences (Sankoff, 2001; Wolfe, 2001). Because the same methodology was also used by Ku et al. (2000), their results should also be considered with caution. On the other hand, Vision et al. (2000) discovered overlapping blocks, a phenomenon that can be explained only by multiple duplication events. Neither Blanc et al. (2000) nor the *Arabidopsis* Genome Initiative (2000) detected these overlapping blocks.

Using a different method of dating based on the substitution rate of silent substitutions, Lynch and Conery (2000) discovered that most *Arabidopsis* genes had duplicated approximately 65 MYA, which brings us back to a single polyploidy event. However, no duplicated blocks of genes, but only paralogous gene pairs were taken into account.

Apparently, the evolutionary history of the first fully sequenced plant seems a lot more complex than originally expected. There is no clear answer on whether one single or multiple polyploidy events took place nor when they occurred. The results of the different analyses seem to be highly dependent of the methods used. For this reason, we reinvestigated the ancient large-scale gene duplications described by Vision et al. (2000) by applying two alternative dating methodologies on several of the more anciently duplicated blocks found in their study. Furthermore, we compared the results obtained to pinpoint the strengths and weaknesses of the methodology used in the two studies.

Materials and Methods

Strategy

The original goal was to reinvestigate whether one or several ancient large-scale gene duplication(s) had occurred in the evolution of *Arabidopsis thaliana*. Furthermore, because Vision et al. (2000) dated one of the large-scale duplication events as approximately 200 x 10^6 years old, we were curious to see whether this event pre- or postdated the monocot-dicot split, which is estimated to have occurred at about that time: 170-235 MYA (Yang et al., 1999) and 143-161 MYA (Wikström et al., 2001). We focused on the blocks that according to Vision et al. (2000) originated during this ancient round of duplication and consisted of six regions in the genome (class F). We mapped these regions to a more up-to-date data set (see below) and subjected them to two dating methodologies: dating based on synonymous substitution rates and molecular phylogeny. The former was done with three different approaches to estimate synonymous substitution rates, namely those of Li (1993), of Nei and Gojobori (1986) and of Yang and Nielsen (2000).

Molecular phylogeny-based dating was performed through the construction of evolutionary trees by the Neighbor-joining method (Saitou and Nei, 1987). By using these different approaches, the possibility of drawing wrong conclusions caused by weaknesses of one particular method is minimized.
However, during the course of this study, it became clear that the most ancient blocks described by Vision et al. (2000) contained genes that had duplicated much more recently. Because the dating methodology of Vision et al. (2000) had been criticized before (Wolfe, 2001; Sankoff, 2001), we subsequently focused on two sets of 10 blocks of two younger age classes, D and E, estimated to be 140 and 170 x 10^6 years old, respectively. These data sets were chosen in such a way that they represented a wide distribution in block size (number of anchor points) as well as amino acid substitution rate (dA) within each age class.

Data set of duplicated genes

From the complete set of segmentally duplicated blocks defined by Vision et al. (2000) that consisted of 103 regions with seven or more duplicated genes, we analyzed selected blocks covering the three oldest classes. This selection consisted of all six blocks from class F (200 x 10^6 years old), 10 from class E (170 x 10^6 years old) and 10 from class D (140 x 10^6 years old). Because the original data set (i.e., the chromosomal DNA sequences) represented a preliminary version of the *Arabidopsis* genome sequence (incomplete and not always correctly assembled), the positions of these duplicated blocks were transferred to a data set that had been built recently. This new data set consisted of a genome-wide non-redundant collection of *Arabidopsis* protein-encoding genes, which were predicted with GeneMark.hmm (Lukashin and Borodvsky, 1998; genome version of January 18th, 2000 (v180101), downloaded from the Institute for Protein Sequences center [Martiensried, Germany; ftp://ftpmips.gsf.de/cress/]). In addition to the protein sequence, the position and orientation of the genes within the *Arabidopsis* genome were determined.
Within this protein set, all pairs of homologous gene products between two chromosomes were determined and the result stored in a matrix of (m, n) elements (m and n being the total number of genes on a certain chromosome). Two proteins were considered as homologous if they had an E-value < 1^{e-50} within a BLASTP (Altschul et al., 1997) sequence similarity search (Friedman and Hughes, 2001).
The synchronization of our data set with the blocks detected by Vision et al. (2000) was done using their supplementary data (website: http://www.igd.cornell.edu/~tvision/arab/science_supplement.html). Initially, for a set of anchor points (i.e. pairs of duplicated genes), defining a duplicated block (Vision et al., 2000), the corresponding protein couples were detected in our data set and then these protein couples were localized in the matrix. To check whether these proteins were indeed part of a segmentally duplicated block, an automatic and manual detection was performed. The automatic detection was done with a new tool (Vandepoele et al., 2002), primarily based on discovering clusters of diagonally organized elements (representing duplicated blocks) within the matrix of homologous gene products. Similar to the strategy of Vision et al. (2000), tandem repeats were remapped before defining a duplicated block.

An overview of blocks analyzed in this study, together with the number of anchor points (the pairs of duplicated genes that make up the duplicated block), is presented in Table 1.

Dating based on Ks

Blocks of duplicated genes were dated using the NTALIGN program in the NTDIFFS software package (Conery and Lynch, 2001). This package first aligns the DNA sequence of two mRNAs based on their corresponding protein alignment and then calculates Ks by the method of Li (1993). We calculated Ks also with two alternative dating methodologies (Nei and Gojobori, 1986; Yang and Nei, 2000) based on the same alignments. These two methods are implemented in the PAML (Yang, 1997) phylogenetic analysis package. The time since duplication was calculated as $T=Ks/2\lambda$, with λ being the mean rate of synonymous substitution; in *Arabidopsis* the estimation is λ = 6.1 synonymous subsitutions per 10^9 years (Lynch and Conery, 2000). The mean Ks value (average of the estimates obtained by the three methods) for each block was derived for each duplicated pair. These values were then used to calculate the mean Ks for each block, excluding outliers using the Grubbs test (Grubbs, 1969; Stefansky, 1972) with a 99% confidence interval.

Phylogenetic analysis

The public databases (PIR, GenBank/EMBL/DBJ, Swiss-PROT) were scanned for homologues of the anchor points using BLASTP (Altschul et al., 1997). When homologs were found in other species next to the *Arabidopsis* paralogs, the gene family was selected for phylogenetic analysis. Protein sequences were subsequently aligned with ClustalW (Thompson et al., 1994). Duplicates or sequences that were too short were removed from the data set. After manual optimization of the alignment and reformatting using BioEdit (Hall, 1999) and ForCon (Raes and Van de Peer, 1999), the more conserved positions of the alignment were subjected to phylogenetic analysis. Trees were constructed based on Poisson or Kimura distances using the Neighbor-joining algorithm as implemented in the TREECON package (Van de Peer and De Wachter, 1997).
Supplementary data such as sequences, accession numbers, alignments and trees can be obtained from the authors upon request.

Results

Dating based on Ks

In contrast to mutations that result in amino acid changes (nonsynonymous substitutions), silent or synonymous substitutions do not affect the biochemical properties of the protein. As such they are generally believed not to be subjected to natural selection and, consequently, to evolve in a (nearly) neutral, clock-like way (Li, 1997). Absolute dating based on synonymous substitution rates (Ks) should be more accurate than dating based on the estimation of genetic distances between duplicated protein sequences.

Table 1. Re-analysis of the duplicated blocks as described by Vision *et al.* (2000)

Vision et al. (2000)						This study						
Block number	Chr1[a]	Chr2[a]	Anchors	dA	Age class	Age in MY	Anchors[b]	Ks[c]	Ks[d]	Ks[e]	Mean age[f]	SD
15	1	3	7	0.8975	F	200	7	1.8641	2.5378	2.1679	213	92
25	1	5	7	0.8012	F	200	6	1.6757	1.7008	2.5515	160	27
37	1	5	11	0.8146	F	200	17	0.8386	0.8138	0.9698	72	19
39	1	3	8	0.8375	F	200	7	1.6053	1.9744	1.8768	170	62
57	2	3	7	0.8521	F	200	7	2.9251	3.2702	2.4395	269	64
59	2	5	15	0.8473	F	200	18	1.8078	2.3744	2.0642	191	70
34	1	5	23	0.7165	E	170	27	0.8723	0.8308	0.8900	71	18
71	3	5	31	0.6814	E	170	70	0.7933	0.8262	0.8312	67	19
100	4	5	20	0.6899	E	170	15	1.8656	1.9727	2.1682	170	45
78	3	5	26	0.701	E	170	35	0.7382	0.7551	0.8475	64	11
47	2	5	8	0.7397	E	170	8	1.8475	3.0169	2.1072	218	87
16	1	3	8	0.6562	E	170	7	0.8390	0.8536	1.0224	74	19
55	2	5	14	0.685	E	170	9	1.7585	2.0966	1.8341	162	32
9	1	3	24	0.6947	E	170	20	0.9098	0.9966	1.1350	83	20
87	3	4	11	0.7231	E	170	8	1.6049	1.8936	2.1889	164	67
48	2	3	11	0.7045	E	170	8	1.7175	1.9716	2.0465	162	56
6	1	5	30	0.6106	D	140	30	0.7754	0.8138	0.9228	69	17
30	1	3	92	0.5262	D	140	106	0.8047	0.8325	0.9668	71	20
95	4	5	88	0.5592	D	140	61	0.7337	0.7884	0.8707	65	10
17	1	1	153	0.5684	D	140	167	0.8110	0.8175	0.8983	69	18
92	4	5	97	0.6064	D	140	107	0.8741	0.8849	1.0507	77	25
33	1	4	18	0.5381	D	140	11	1.6283	1.6707	1.5669	133	26
5	1	4	13	0.5631	D	140	6	1.5232	1.5657	1.5324	126	16
73	3	5	26	0.5855	D	140	25	0.7965	0.8187	0.9105	69	15
93	4	5	42	0.6263	D	140	28	0.7719	0.8174	0.9010	68	16
26	1	4	35	0.5273	D	140	42	0.8719	0.8946	1.0867	78	23

[a] Chromosome numbers on which the two duplicated blocks are found.
[b] Number of anchor points in blocks detected in this study.
[c] Ks values calculated according to Li (1993).
[d] Ks values calculated according to Nei and Gojobori (1986).
[e] Ks values calculated according to Yang and Nielsen (2000).
[f] Mean age of the block was derived from the mean Ks, excluding outliers (see Materials and Methods).

However, because of rapid saturation of synonymous sites, dates of older (Ks>1) divergences/ duplications will become unreliable (Li, 1997).

We calculated Ks values with three different methods for all pairs of duplicated genes in 26 old blocks (classes D, E, and F, estimated to have originated between 140 and 200 MYA; Vision et al., 2000). From these values we calculated the duplication date of each block. The results of this analysis are given in Table 1.

Interestingly, several block duplications were dated to be much younger than what was found by Vision et al. (2000). For example, a duplication between chromosome 1 and 5, denoted as block 37 and based on 11 gene pairs (17 in our study; Table 1), was found to have occurred 72 MYA, and not 200 MYA. The distribution of the Ks values of the duplicated pairs in this block, calculated with the three different methods, confirmed our hypothesis that this is a younger block. With only a few exceptions, almost all duplicated pairs seemed to have Ks values between 0.5 and 1 synonymous substitutions per synonymous site, and this for the three methods used (Fig. 1). For three pairs of genes within the duplicated block, the situation is less clear (Fig. 1). No results were obtained with the method of Li (1993), probably because the duplicated gene sequences are too divergent to calculate a Ks value using this method, whereas the two other methods gave extremely high or no Ks values. One possible explanation is a higher synonymous mutation rate specific for these genes, because fluctuations in Ks have been reported before (Li, 1997; Zeng et al., 1997). Another possible explanation could be that these genes originated earlier than the other genes in that block and that the situation observed is due to differential deletions of alternate members of duplicated tandem pairs (Friedman and Hughes, 2001). For this reason, these gene pairs were not included in the calculation of the duplication date of the whole block (see Materials and Methods).

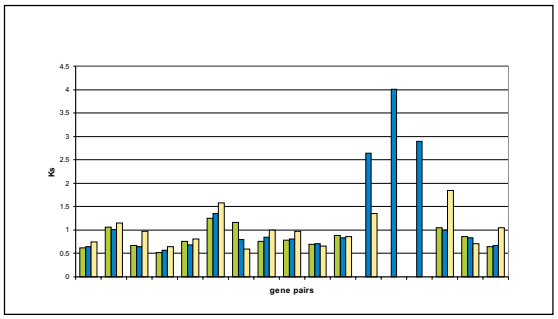

Figure 1. Distribution of Ks values for duplicated genes as found in block 37, calculated with the methods of Li *et al.* (purple bars), Nei and Gojobori (red bars) and Yang and Nielsen (yellow bars).

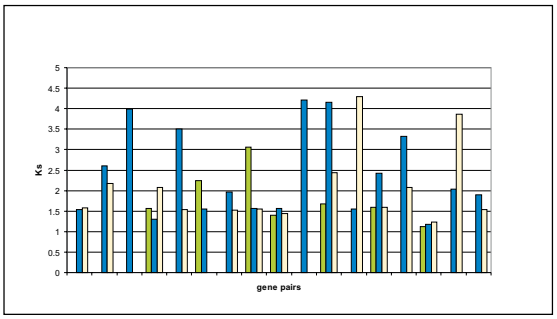

Figure 2. Distribution of Ks values for duplicated genes found in block 59, calculated with the methods of Li *et al.* (purple bars), Nei and Gojobori (red bars) and Yang and Nielsen (yellow bars).

However, most blocks of age class F had significantly higher Ks values and consequently older divergence dates, which indeed points to a more ancient large-scale duplication event. This observation was strengthened by the fact that, with a few exceptions, duplicated blocks of this age class had less anchor points (Table 1) and Ks values seemed to fluctuate more between members of the same block (see, for example, the distribution of block 59, estimated to have duplicated approximately 190 MYA; Fig. 2).

The latter is probably due to saturation of synonymous substitutions, by which larger errors in Ks estimation are introduced, causing values of Ks >1 to be unreliable.
In our evaluation of class E blocks (170 MYA; Vision et al., 2000), the situation is even more peculiar. From the 10 blocks we selected, a large part again seemed to be much younger than what was derived based on dA values. Five out of 10 blocks seemingly originated only approximately 70 MYA, less than half the age calculated by Vision et al. (2000). Here also, the distribution of Ks values clearly showed that a large majority of duplicated pairs in these blocks belonged to the same, much younger age class, with only a few exceptions (data not shown). However, the other half of the 10 selected blocks seem to be older.
In the class D sample, dated 140 x 10^6 years old by Vision et al. (2000), eight out of 10 blocks seemed to have duplicated approximately 70 MYA. The distribution of Ks values within one block again gave similar results as above: most pairs had Ks values between 0.5 and 1, with a minor fraction of exceptions (data not shown).
Although only a subset of the complete set of duplicated blocks of age classes D and E were analyzed, many blocks appeared to be much younger than proposed by Vision et al. (2000). Preliminary results of a more rigorous analysis seem to confirm our findings (unpublished results).

Dating by phylogenetic analysis

Absolute dating methods based on substitution numbers per site are very useful in high-throughput analyses, such as those by Lynch and Conery (2000) and Vision et al. (2000), but they have some serious drawbacks. Inferred divergence dates based on amino acid substitutions are not as quickly underestimated due to saturation, although saturation at the amino acid level has been demonstrated (Van de Peer et al., 2002). However, when using this technique, there is a serious risk of overestimating the age of more rapidly evolving blocks, or underestimating the age of blocks containing more slowly evolving proteins. The use of synonymous mutation rates is probably favourable because these positions evolve at nearly neutral rates and, so, give a more reliable estimate in the case of fast or slowly evolving genes. Unfortunately, these analyses are compromised for older duplications because of the rapid saturation of these sites.
To validate the results, an alternative technique was applied, namely relative dating using phylogenetic methods. If a duplication occurred before the monocot-dicot split, this could be proven by a tree topology (Fig. 3a), in which the two dicot members of a gene family each group with a monocot sequence. If, however, the two *Arabidopsis* duplicates originated more recently, i.e. after the dicot-monocot split, the two dicot branches should be sister sequences, outgrouped by their monocot ortholog (Fig. 3b).
Even if certain sequences are still missing from the databases (because of gene loss or nondetection), conclusions can be drawn. For example, the tree topology presented in Figure 3c could only be explained by a duplication that occurred before the monocot-dicot split.

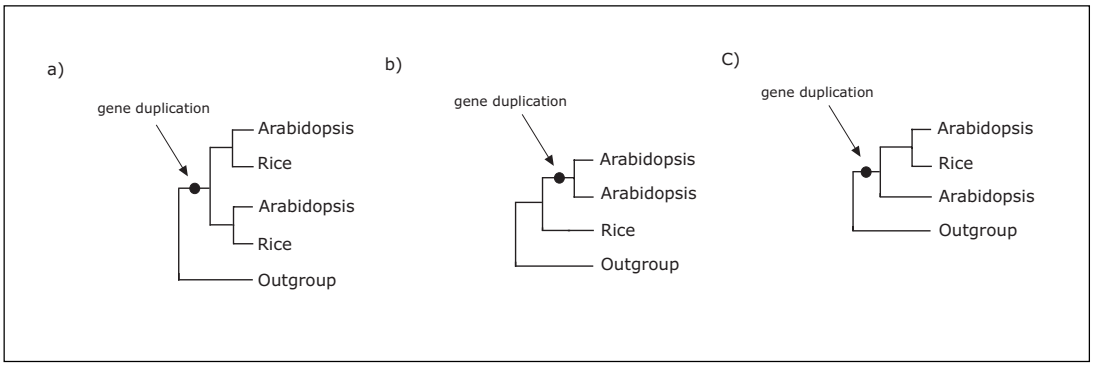

Figure 3. a) Expected tree topology for genes formed by a gene/genome duplication event prior to the split of monocots and dicots. b) Expected tree topology for genes formed by a gene/genome duplication event that occurred after the split of monocots and dicots and specific to *Arabidopsis*. c) Even if only one of the paralogs is known, due to gene loss or absence in the databases, the gene duplication can be inferred.

For all the anchor points of the oldest blocks (F), we searched the protein databases for homologs in other plant species to construct evolutionary trees. Unfortunately, it was impossible to construct trees for many of the duplicated genes, the main reason being the absence of homologs from plant species other than *Arabidopsis* in the databases. Furthermore, the sequences often contained too few conserved positions to get statistically significant results (i.e. high bootstrap values).
An overview of constructed trees and conclusions is presented in Table 2. Gene families for which no homologues from other species than *Arabidopsis thaliana* could be found in the databases are not shown.

Table 2. Gene families selected for phylogenetic analysis for each paralogous block, belonging to age class F (Vision et al., 2000; 200 MYA)

Block[a]	Family[b]	Sites[c]	Conclusion	Reason
15	Unknown	279	None	No statistical support
25	-	-	None	No trees possible due to absence of sequences from other species
37	Calmodulin	105	None	No statistical support
	Calmodulin-like	112	Probably younger than the split between eurosids I and eurosids II	Genetic distance
	Glutamine synthase	314	Younger than split with asteridae and older than *Arabidopsis-Brassica* divergence (see Fig. 3)	Topology with statistical support
39	Unknown	287	None	Too few monocot sequences for this family
57	DOF Zinc-finger	85	None	Highly inequal rates of evolution between duplicates
	GATA transcription factor	148	Older than monocot-dicot split (see Fig. 4)	Topology with statistical support
	Apetala 2	81	None	No statistical support
	Expansin	180	None	No statistical support
59	Protein phosphatase 2C	174	None	Too few monocot sequences available
	Putative Rab5 interacting protein	100	Probably younger than monocot-dicot split	Genetic distance
	Cyclophilin	141	None	No statistical support
	Phosphoprotein phosphatase 1	305	None	No statistical support
	Apetala 2 (see also B57)	81	None	No statistical support

[a] Block number as defined by Vision et al. (2000).
[b] Name of the family analyzed, as far as could be deduced from the description line of the entries.
[c] Length of sequence alignment used for tree construction.

Although we could not draw conclusions on many of the genes/blocks, we would like to consider some of the constructed trees. A first interesting result was obtained from the analysis of the gluthatione synthase gene family: it has two members on chromosomes 1 and 5 that are part of block 37, which is a duplicated block of class F (200 MYA; Vision et al., 2000) but, according to our estimation, it had duplicated approximately 72 MYA.

The tree topology (Fig. 4) for this family clearly showed that the duplication that yielded the two duplicates occurred before the divergence of *Arabidopsis* and *Brassica*, but after the split between Asteridae and Rosidae. In consequence, the duplication between these two genes must have happened between 15-20 (Yang et al., 1999; Koch et al., 2001) and 135 MYA (the latter value being the mean of two estimations, 112-156 MYA [Yang et al., 1999]) and 114-125 MYA [Wikström et al., 2001]), which is in accordance with our findings for this block.

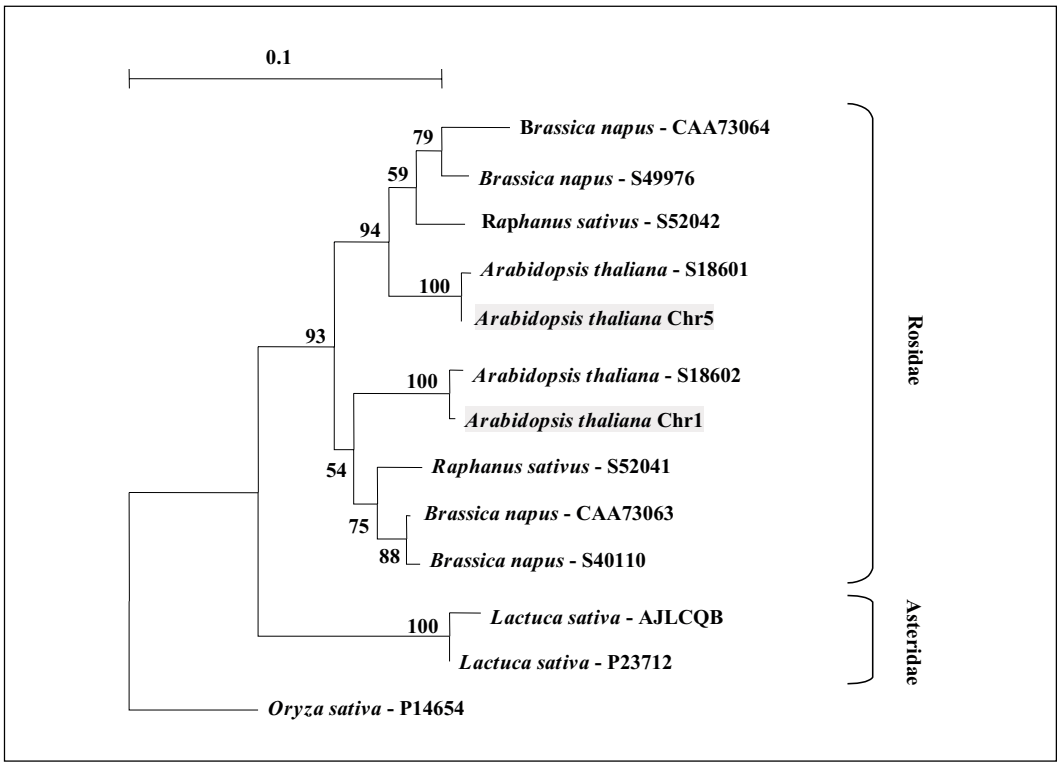

Figure 4. Neighbor-joining tree of the gluthamine synthase family, inferred from Poisson-corrected evolutionary distances. Shaded sequences belong to the analyzed duplicated blocks. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

A second tree of interest is that of the GATA transcription factor family with a pair of duplicates on chromosomes 2 and 3 that belong to block 57, also of age class F. It was very hard to date this block with our dating methods, because the sequences were apparently saturated for synonymous substitutions. However, all Ks values calculated for pairs in this block were above 2.2 synonymous substitutions per synonymous site (see Table 1), suggesting that this block is genuinely old.

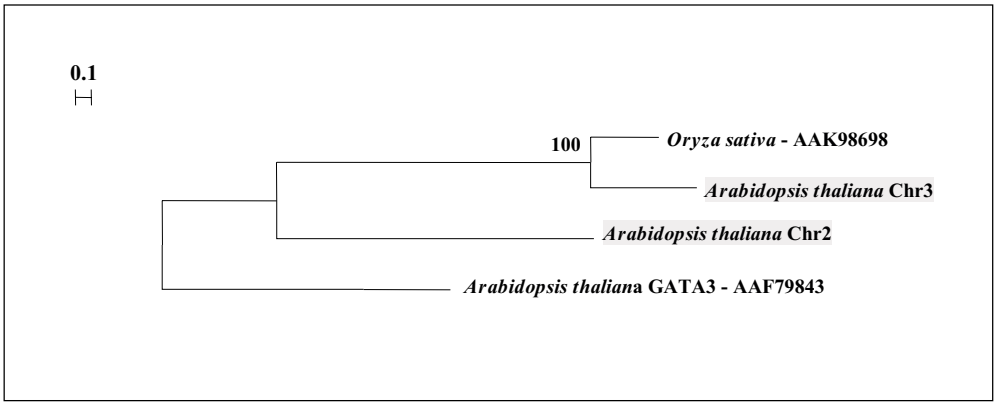

Figure 5. Neighbor-joining tree of the GATA family of transcription factors, inferred from Poisson corrected evolutionary distances. Shaded sequences belong to the analyzed duplicated blocks. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

When we investigated the topology of the GATA family (Fig. 5), we observed a topology similar to that described in Figure 3c: although there is only one monocot sequence, this topology could only be explained if the duplication that gave rise to the two *Arabidopsis* genes occurred before the monocot-dicot split. This would mean that this block occurred at least 190 MYA (Yang et al., 1999; Wilkström et al., 2001).

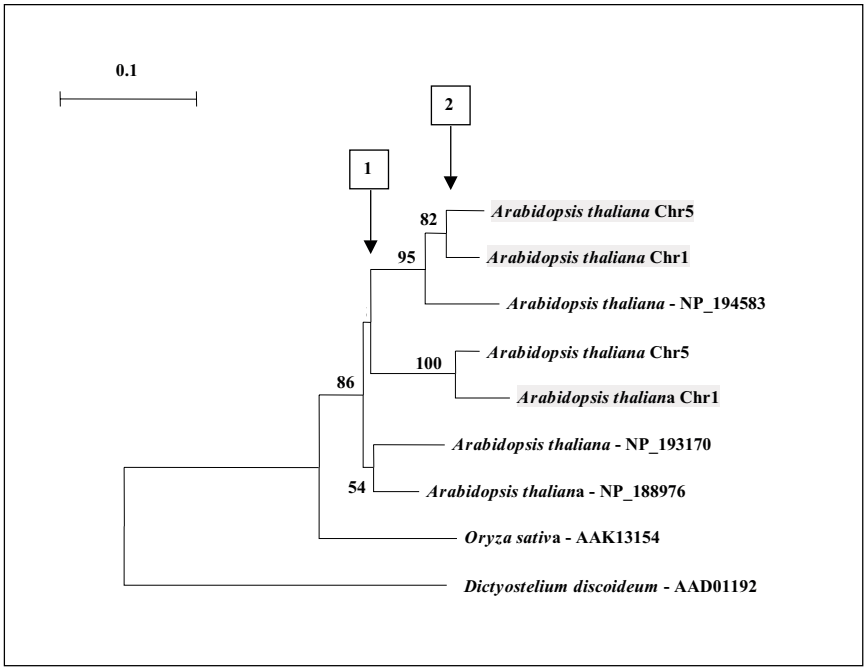

Figure 6. Neighbor-joining tree of the casein kinase family, using Poisson correction for evolutionary distance calculation. Shaded sequences belong to the analyzed duplicated blocks. Arrows indicate (1) a tandem duplication and (2) the block duplication. Bootstrap values (above 50%) are shown in percentages at the internodes. Scale = evolutionary distance in substitutions per amino acid.

In some cases, evolutionary distances can be informative of duplication dates. As illustration, an example from the age class D (140 MYA; Vision et al., 2000) is given. Figure 6 shows the topology of the casein kinase gene family that has two members on both chromosomes 1 and 5, all four of them belonging to the same duplicated block 6.
Using Ks-based dating, we determined that this block duplicated approximately 70 MYA, with approximately 80% of the Ks values in this block being smaller than 1. As can be seen from the tree topology, the two members of block 6 first originated (probably) through tandem duplication (arrow 1) and then through a larger-scale duplication including the other members of that block (arrow 2). Both these events happened after the monocot-dicot split, as can be derived from the fact that the group containing these four proteins is outgrouped by a rice sequence. The evolutionary distance from each of the duplicates to the block duplication point is approximately 0.025 amino acid substitutions per site, whereas the evolutionary distance between the genes originating by tandem duplication is approximately 0.158 amino acid substitutions per site. The average evolutionary distance between the sequences of rice and *Arabidopsis* is approximately 0.206 amino acid substitutions per site, meaning that, if a divergence date for monocots and dicots of 190 MYA (Yang et al., 1999; Wilkström et al., 2001) and a molecular clock-like evolution of this protein were assumed, the block duplication would have happened somewhere 46 MYA (with $\lambda = K/2T = 0.206$ substitutions per site/ 380 MY = 5.42×10^{-4} substitutions per site/MY). This value is much closer to our estimation based on Ks than that of 140 MYA obtained by Vision et al. (2000).

Discussion

Currently, three different methods to date gene duplication events are generally used: absolute dating based on synonymous substitution rates, absolute dating based on nonsynonymous substitution rates or protein-based distances, and relative dating through the construction of phylogenetic trees. Here, we provide some evidence that protein distances are not very reliable for large-scale dating of heterogeneous classes of proteins. For example, classes containing blocks of the same age based on mean protein distance (classes D, E, and F; Vision et al., 2000) seem to be very heterogeneous in age when dating is based on synonymous substitution rates. Protein-based distances are known to vary considerably among proteins (e.g. Easteal and Collet, 1994); therefore, duplicated blocks that contain a larger fraction of fast-evolving genes will have a relatively high mean protein distance between the paralogous regions and appear older than they actually are. In our opinion, the use of synonymous and, consequently, neutral substitutions for evolutionary distance calculations is more reliable. However, there is one important caveat: dating based on silent substitutions can only be applied when $Ks < 1$. A $Ks > 1$ points to saturation of synonymous sites and can no longer be used to draw any reliable conclusions regarding the origin of duplicated genes or blocks. In this case, a solution could be relative dating with phylogenetic means. Although the dating is rather crude, it offers a way of determining duplication dates relative to known divergences. The main problem here, however, is the availability of plant sequence data. Only a few duplicated pairs had enough orthologs in the public databases to allow any conclusions to be drawn.

Furthermore, if orthologs would be found, the sequences may not be very suitable for phylogenetic analysis. Consequently, it seems that phylogenetic inference cannot yet be as widely applied to plant as to animal genomes (e.g., Wang and Gu, 2000; Friedman and Hughes, 2001; Van de Peer et al., 2001). However, as soon as more sequence data from key species such as mosses, ferns and monocots become available, this approach may become more useful.

From the three oldest age classes defined by Vision et al. (2000), only one (F) seems to contain many old duplicated blocks, whereas several blocks of the two other age classes have seemingly been duplicated approximately 70-90 MYA. In our opinion, the hypothesis of Vision et al. (2000) that at least four large-scale duplications have occurred is far from being proven. In contrast with the multimodal distribution of large-scale gene duplication, our results show that a major fraction of blocks has duplicated approximately at the same time and has probably originated by a complete genome duplication. On the other hand, a fraction of block duplications seems much older than the others. Unfortunately, because synonymous sites were saturated and trees were not reliable enough, these duplications could not be dated more accurately. Although these old duplicated blocks are scattered throughout the genome (Table 1), it is hard to prove that they are the result of a single duplication event.

The question of whether large-scale gene duplications have occurred before the divergence of monocots and dicots still remains to be answered. Some of these events are probably anterior to the monocotyl-dicotyl split, as suggested by the GATA transcription factor topology (Fig. 5). Large-scale gene duplication events prior to the monocot-dicot split may have led to the origin of flowering or even of seed plants: Duplications of (sets of) developmentally important genes could have given the opportunity to develop new reproductive organs and strategies and consequently cause reproductive isolation, which may have resulted in speciation. The ongoing accumulation of sequence data delivered by several plant expressed sequence tags and genome sequencing projects will provide the means to answer the questions regarding the prevalence and timing of gen(om)e duplications in the evolution of plants and will hopefully help elucidating the role of these events in the diversification and evolution of plant species.

Acknowledgments

The authors would like to thank Eric Bonnet, Sven Degroeve and John S. Taylor for helpful discussions and Martine De Cock for help with the manuscript. K.V. and C.S. are indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie for a predoctoral fellowship. Y.V.d.P. is a Research Fellow of the Fund for Scientific Research (Flanders).

Note added in proof

Since acceptance of this paper, novel tools to identify heavily degenerated block duplications allowed us to find evidence for the recent genome duplication described in this study. The occurence of two additional, but probably no more, ancient genome duplications in *Arabidopsis* was also demonstrated (Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M. and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana.* Proc Natl Acad Sci USA **99**, 13627-13632).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389-3402.

The *Arabidopsis* Genome Initiative(AGI) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12,** 1093-1101.

Conery, J.S., and Lynch, M. (2001). Nucleotide substitutions and the evolution of duplicate genes. Pac Symp Biocomput**,** 167-178.

Easteal, S., and Collet, C. (1994). Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. Mol Biol Evol **11,** 643-647.

Friedman, R., and Hughes, A.L. (2001). Pattern and timing of gene duplication in animal genomes. Genome Res **11,** 1842-1847.

Grant, D., Cregan, P., and Shoemaker, R.C. (2000). Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc Natl Acad Sci U S A **97,** 4168-4173.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. Technometrics **11,** 1-21.

Haldane, J.B.S. (1933). The part played by recurrent mutation in evolution. The american naturalist **67,** 5-19.

Hall, T.A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser**,** 95-98.

Holland, P. (1992). Homeobox genes in vertebrate evolution. Bioessays **14,** 267-273.

Hughes, A.L. (1999). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J Mol Evol **48,** 565-576.

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

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

Ku, H.M., Vision, T., Liu, J., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A **97,** 9121-9126.

Leitch, I.J., and Bennet, M.D. (1997). Polyploidy in angiosperms. Trends Plant Sci **2,** 470-476.

Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36,** 96-99.

Li, W.-H. (1997). Molecular evolution. (Sunderland, Mass.: Sinauer Associates).

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M., Feldblyum, T.V., Buell, C.R., Ketchum, K.A., Lee, J., Ronning, C.M., Koo, H.L., Moffat, K.S., Cronin, L.A., Shen, M., Pai, G., Van Aken, S., Umayam, L., Tallon, L.J., Gill, J.E., Venter, J.C., and et al. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana* [see comments]. Nature **402,** 761-768.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26,** 1107-1115.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso,, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S.-A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y.-J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Klein Lankhorst, R., Rose, M., Hauf, J., Kötter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysschaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L.,

Foggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M.-A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blöcker, H., Scharfe, M., Grimm, M., Löhnert, T.-H., Dose, S., de Haan, M., Maarse, A., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fartmann, B., Granderath, K., Dauner, D., Herzl, A., Neumann, S., Argiriou, A., Vitale, D., Liguori, R., Piravandi, E., Massenet, O., Quigley, F., Clabauld, G., Mündlein, A., Felber, R., Schnabl, S., Hiller, R., Schmidt, W., Lecharny, A., Aubourg, S., Chefdor, F., Cooke, R., Berger, C., Montfort, A., Casacuberta, E., Gibbons, T., Weber, N., Vandenbol, M., Bargues, M., Terol, J., Torres, A., Perez-Perez, A., Purnelle, B., Bent, E., Johnson, S., Tacon, D., Jesse, T., Heijnen, L., Schwarz, S., Scholler, P., Heber, S., Francs, P., Bielke, C., Frishman, D., Haase, D., Lemcke, K., Mewes, H.W., Stocker, S., Zaccaria, P., Bevan, M., Wilson, R.K., de la Bastide, M., Habermann, K., Parnell, L., Dedhia, N., Gnoj, L., Schutz, K., Huang, E., Spiegel, L., Sehkon, M., Murray, J., Sheet, P., Cordes, M., Abu-Threideh, J., Stoneking, T., Kalicki, J., Graves, T., Harmon, G., Edwards, J., Latreille, P., Courtney, L., Cloud, J., Abbott, A., Scott, K., Johnson, D., Minx, P., Bentley, D., Fulton, B., Miller, N., Greco, T., Kemp, K., Kramer, J., Fulton, L., Mardis, E., Dante, M., Pepin, K., Hillier, L., Nelson, J., Spieth, J., Ryan, E., Andrews, S., Geisel, C., Layman, D., Du, H., Ali, J., Berghoff, A., Jones, K., Drone, K., Cotton, M., Joshu, C., Antonoiu, B., Zidanic, M., Strong, C., Sun, H., Lamar, B., Yordan, C., Ma, P., Zhong, J., Preston, R., Vil, D., Shekher, M., Matero, A., Shah, R., Swaby, I'K., O'Shaughnessy, A., Rodriguez, M., Hoffman, J., Till, S., Granat, S., Shohdy, N., Hasegawa, A., Hameed, A., Lodhi, M., Johnson, A., Chen, E., Marra, M., Martienssen, R., and McCombie, W.R. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402,** 769-777.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3,** 418-426.

Ohno, S. (1970). Evolution by Gene Duplication. (Berlin;Heidelberg;New York: Springer-Verlag).

Paterson, A.H., Lan, T.H., Reischmann, K.P., Chang, C., Lin, Y.R., Liu, S.C., Burow, M.D., Kowalski, S.P., Katsar, C.S., DelMonte, T.A., Feldmann, K.A., Schertz, K.F., and Wendel, J.F. (1996). Toward a unified genetic map of higher plants, transcending the monocot- dicot divergence. Nat Genet **14,** 380-382.

Raes, J., and Van de Peer, Y. (1999). ForCon: A software tool for the conversion of sequence alignments. EMBNet.news **6,** (http://www.ebi.ac.uk/embnet.news/vol6_1/).

Robinson-Rechavi, M., Marchand, O., Escriva, H., and Laudet, V. (2001). An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. Curr Biol **11,** R458-459.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4,** 406-425.

Sankoff, D. (2001). Gene and genome duplication. Curr Opin Genet Dev **11,** 681-684.

Skrabanek, L., and Wolfe, K.H. (1998). Eukaryote genome duplication - where's the evidence ? curr opin genet dev **8,** 694-700.

Stefansky, W. (1972). Rejecting outliers in factorial designs. Technometrics **14,** 469-479.

Taylor, J.S., Van De Peer, Y., Braasch, I., and Meyer, A. (2001a). Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos Trans R Soc Lond B Biol Sci **356,** 1661-1679.

Taylor, J.S., Van de Peer, Y., and Meyer, A. (2001b). Genome duplication, divergent resolution and speciation. Trends Genet **17,** 299-301.

Taylor, J.S., Van de Peer, Y., and Meyer, A. (2001c). Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Curr Biol **11,** R1005-R1007.

Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van Den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Van Montagu, M., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445,** 237-245.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22,** 4673-4680.

Van de Peer, Y., and De Wachter, R. (1997). Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. CABIOS **13,** 227-230.

Van de Peer, Y., Taylor, J.S., Braasch, I., and Meyer, A. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol **53,** 436-446.

Van de Peer, Y., Frickey, T., Taylor, J.S., and Meyer, A. (2002). Dealing with saturation at the amino acid level: A case

study involving anciently duplicated zebrafish genes. Gene **295**, 205-211.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res. **12,** 1792-1801.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in *Arabidopsis*. Science **290,** 2114-2117.

Wang, Y., and Gu, X. (2000). Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol **51,** 88-96.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol **42,** 225-249.

Wikstrom, N., Savolainen, V., and Chase, M.W. (2001). Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci **268,** 2211-2220.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet **2,** 333-341.

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

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13,** 555-556.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17,** 32-43.

Zeng, L.W., Comeron, J.M., Chen, B., and Kreitman, M. (1998). The molecular clock revisited: the rate of synonymous vs. replacement change in Drosophila. Genetica **103,** 369-382.

[Chapter 8]

Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico

Jeroen Raes and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
**Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*

Published in: Applied Bioinformatics, in press.

Abstract

The (large-scale) duplication of genes increases the amount of genetic material on which evolution can work, and has been considered of major importance for the development of biological novelties or to explain important evolutionary transitions that have occurred during biological evolution. Recently, much research has been devoted to the study of the evolutionary and functional divergence of duplicated genes. Since the majority of genes are part of gene families, there is considerable interest in predicting differences in function between duplicates and assessing the functional redundancy of genes within gene families. In this review, we discuss the strengths and limitations of older and novel approaches to investigate the evolution of duplicated genes *in silico*.

Introduction

In his now classic book 'Evolution by Gene Duplication', published in 1970, Ohno claimed that if evolution had been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged, while big leaps in evolution would have been impossible without the creation - through duplication - of many new gene loci with previously nonexistent functions. During the last few decennia it became clear that, from an evolutionary point of view, most genes are indeed not unique but are part of larger families of related genes. These gene families have originated by duplication of an ancestral gene, after which these duplicated genes in turn have duplicated. It is now generally believed that extensive gene duplication has been responsible for increased genomic and phenotypic complexity (e.g. Aburomia et al., 2003; Meyer and Van de Peer, 2003)

Although there is some evidence that gene duplication is a continuous and very frequently occurring process (Lynch and Conery, 2000), more and more genomic data seem to suggest that many duplicates have been formed during some major large-scale gene duplication events. Entire genome duplication events have been postulated for (members of the) the three major eukaryotic kingdoms. Based on a genome-wide analysis of the yeast *Saccharomyces cerevisiae*, Wolfe and Shields (1997) postulated a duplication of the entire yeast genome about 100 MYA, although this event was dated much older (200-300 MYA) by others (Friedman and Hughes, 2001). About 13% of the yeast genome still consists of duplicated genes, resulting from this polyploidy event (Seoighe and Wolfe, 1999).

For animals, the first indications about large-scale duplications early in the vertebrate lineage were found by the analysis of *Hox* genes (Holland, 1994). *Hox* genes encode DNA-binding proteins that specify cell fate along the anterior-posterior axis of bilaterian animal embryos and occur in one or more clusters of up to 13 genes per cluster (Gehring, 1998). It is thought that the ancestral *Hox* gene cluster arose from a single gene by a number of tandem duplications. The observation that protostome invertebrates, as well as the deuterostome cephalochordate *Amphioxus,* possess a single *Hox* cluster while Sarcopterygia, a monophyletic group including lobe-finned fish such as the coelacanth and lungfishes, amphibians, reptiles, birds, and mammals have four clusters (Holland and Garcia-Fernandez, 1996; Holland, 1997) supports the hypothesis of **2 R**ounds (events) of entire genome duplications early in vertebrate evolution. Additional evidence comes from the detection and dating of duplicated blocks in the human genome (McLysaght et al., 2002), large-scale phylogenetic analysis of gene families (Gu et al., 2002) and analysis of gene clusters such as the major histocompatibility complex MHC region (Spring, 1997; Abi-Rached et al., 2002). However, in general, phylogenetic evidence for the 2R hypothesis is hard to find and the 2R hypothesis is still vividly debated (Spring, 2002; Furlong and Holland, 2002; Larhammar et al., 2002; Friedman and Hughes, 2003).

A few years ago, 'extra' *Hox* gene clusters have been discovered in fish. Amores and co-authors (1998) described the existence of seven *Hox* clusters in zebrafish (*Danio rerio*), and additional *Hox* clusters have also been described for medaka (*Oryzias latipes;* Naruse et al., 2000), the African cichlid fish *Oreochromis niloticus* (Málaga-Trillo and Meyer 2001), and the pufferfish *Fugu rubripes* (Aparicio et al., 1997).

Box: Glossary of terms

Alloploidy: fusion of the genomes of two different species.
Codon bias: unequal codon usage frequencies for amino acids between different genes or organisms.
Differential gene loss: reciprocal deletion of genes in duplicated segments.
Gene conversion: also called nonreciprocal recombination, a process leading to nondivergence of homologous loci in a species.
Homologs: genes that share a common ancestor.
In silico: method to study biological questions using computational means rather than laboratory (*in vitro*) and animal experiments (*in vivo*).
Nonsynonymous substitutions: nucleotide substitution that leads to amino acid replacement.
Orthologs: homologous genes that originated through speciation (see figure 1).
Paralogs: homologous genes that originated through duplication (see figure 1).
Polyploidy: doubling of the copy number of each chromosome in a species.
Positive selection: selection fixing advantageous mutations.
Pseudogene: nonfunctional gene due to accumulation of deleterious mutations.

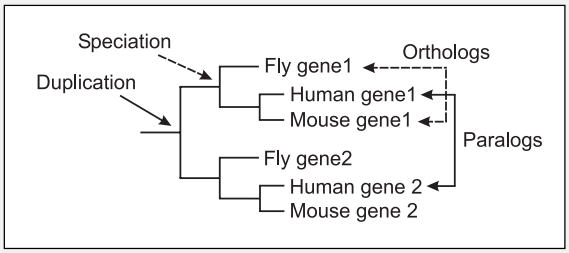

Figure 1. Hypothetical tree depicting a duplication event, followed by speciation. Paralogs arise through duplication (full arrows), while orthologs arise through speciation (dashed arrows).

Purifying selection: selection against deleterious mutations (also called negative selection).
Subfunctionalization: process in which duplicated genes divide functions originally exerted by the ancestral gene.
Synonymous substitutions: nucleotide substitution that does not lead to amino acid replacement.
Transitions: substitution of a purine (A,G) by another purine or a pyrimidine (C,T) by another pyrimidine.
Transversions: substitution of a purine (A,G) by a pyrimidine (C,T) or vice versa.

All these data strongly point to an additional *Hox* cluster duplication in ray-finned fishes that occurred before the divergence of zebrafish, medaka and pufferfish, at least 100 Myr ago (Nelson, 1994). Furthermore, mapping data suggest that duplications are not limited to *Hox* clusters, and that large chromosome segments or entire chromosomes are duplicated (Amores et al., 1998; Force et al., 1999; Woods et al., 2000; Postlethwait et al., 2000). In the meantime, many other multigene families have been described that have more genes in fish than in other vertebrates (Wittbrodt et al., 1998; Postlethwait et al., 2000; Taylor et al., 2001a, 2001b). Moreover, tree topologies clearly support a fish-specific genome duplication that has occurred early in the evolution of ray-finned fish (Taylor et al., 2003).

In plants, early analyses based on the - at that time - unfinished genome sequence of *Arabidopsis thaliana* already showed that large-scale gene duplication, probably a complete genome duplication occurred in the evolution of this model plant (e.g. Terryn et al., 1999; Blanc et al., 2000; Paterson et al., 2000), an opinion later shared by the *Arabidopsis* Genome Initiative (AGI, 2000). Vision et al. (2000) investigated this genome duplication by looking at large regions (“blocks”) of genes that showed statistically significant colinearity with other regions in the genome. They rejected a single-genome duplication hypothesis because of the discovery of overlapping blocks, a phenomenon that can only be attributed to multiple duplication events. By dating these duplicated blocks, these authors postulated up to four different large-scale gene duplication events, ranging from 50 to 220 MYA. One of these classes, dated approximately 100 MYA, grouped nearly 50% of all the duplicated blocks, suggesting a complete genome duplication at that time (Vision *et al.*, 2000). However, the dating methods used in this study were later criticized (Wolfe, 2001; Raes et al., 2003).

A recent reanalysis of the *Arabidopsis thaliana* genome by Simillion et al. (2002) considered heavily degenerated block duplications. These ancient duplicated blocks can no longer be recognized by directly comparing both segments due to differential gene loss, but can still be detected through indirect comparison with other segments. When these so-called hidden duplications are taken into account to describe the duplication landscape in *Arabidopsis*, many homologous genomic regions can be found in five up to eight copies suggesting three polyploidization events in the evolutionary past of *Arabidopsis thaliana*. Furthermore, about 28% of the genes in *Arabidopsis* are retained duplicates, resulting from these ancient large-scale gene duplication events, the youngest one estimated to have occurred about 75 million years ago (Simillion et al., 2002).

The evolution of novel gene functions

Large-scale gene or entire genome duplication events such as the ones described above have been considered very important for biological evolution because they provide a way to double the genetic material on which evolution can work (Ohno, 1970; Holland, 1994; Sidow, 1996; Prince and Pickett, 2002; Holland, 2003). Indeed, since duplicated genes are redundant, one of the copies is, at least in theory, freed from functional constraint, and can therefore evolve a new function. The classical model, put forward by Ohno (1970) predicts that mutations in the second copy are selectively neutral and will either turn the gene into a non-functional pseudogene, or alternatively, turn the duplicate gene into a gene with a new function, due to a series of non-deleterious random mutations. This model of gene evolution has been widely adopted as an explanation for the evolution of novel genes and gene functions but has been criticized, mainly because little evidence has been found for genes that have obtained novel functions this way. Several alternative models for gene evolution after duplication events have been proposed (Hughes, 1994; 1999; Walsh, 1995; Nowak et al., 1997; Gibson and Spring, 1998; Wagner, 1998; Force et al., 1999). For example, Hughes (1994) and Force et al. (1999) argue that when a gene with multiple functions is duplicated, the duplicates are redundant only for as long as each retains the ability to perform all ancestral roles. When one of the duplicates experiences a mutation that prevents it from carrying out one of its ancestral roles, the other duplicate is no longer redundant. According to Force et al.'s (1999) 'duplication-degeneration-complementation (DDC)' model, degenerative mutations preserve rather than destroy duplicated genes but also change their functions - or at least restrict them - to become more specialized. Gibson and Spring (1998) have argued that alteration of a single domain in a multidomain protein might lead to nonfunctional complexes that exhibit a so-called 'dominant-negative phenotype'. Their model is based on the observation that, for several genes, point mutations lead to a much more severe phenotype than when the (duplicated) gene is simply knocked out. In this case, one would expect selection against deleterious point mutations resulting in the retention of the gene. As a matter of fact, the gene is not only retained, it is also kept redundant. Although these models explain gene retention rather than gene evolution, keeping the genes around increases the chance for functional divergence later on, by e.g. positive selection (e.g. Zhang et al., 1998; Duda et al., 1999; Hughes et al., 2000) or subfunctionalization (Li, 1980; Piatigorsky and Wistow, 1991; Hughes, 1994; Force et al., 1999; Stoltzfus, 1999; Wagner, 2002).

Likewise, processes like gene conversion might also retard the functional divergence of duplicated genes, while at the same time prevent pseudogenisation of a redundant copy (Li, 1997).
In this review, we discuss some older and novel *in silico* approaches to study the evolution of duplicated genes, mainly focussing on the coding part of the gene, in order to find traces that might imply functional divergence after duplication. Figure 2 summarizes these different approaches, starting from two paralogs, but extending the set of sequences according to the method used.

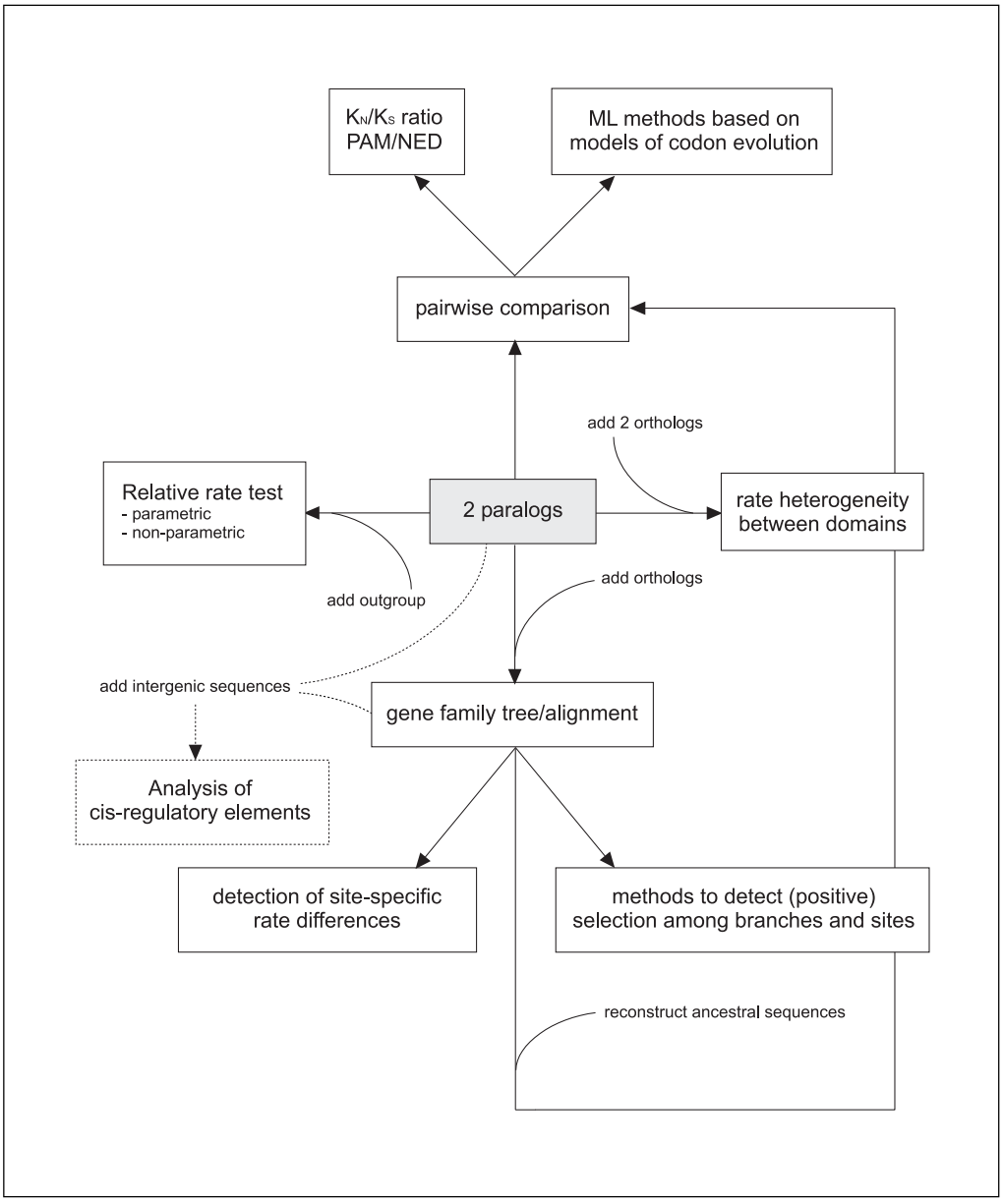

Figure 2. Overview of the different *in silico* approaches to study possible functional divergence at the coding level between two duplicated genes. Simple approaches are often based on the comparison of only two paralogs, while more sophisticated analyses are usually based on a larger collection of sequences. See text for details.

Detecting functional divergence

Relative-rate tests

One of the simplest ways to study the evolution of duplicated genes is to investigate whether one of the duplicates has evolved at a faster rate after duplication, compared to a reference or outgroup sequence, using a so-called relative-rate test (Margoliash, 1963; Sarich and Wilson, 1973). An increase in rate of evolution could be explained by relaxed functional constraints eventually turning one of the duplicates into a pseudogene, due to accumulation of deleterious mutations. On the other hand, an increase in rate could also point to positive selection by which the gene evolves a new function. In general, relative-rate tests can be divided into two main categories: parametric and non-parametric. Parametric rate tests use a model of evolution to account for multiple substitutions, in order to compute branch lengths more accurately.

To this end, many alternatives and improvements have been proposed over the years, using distance (e.g. Wu and Li, 1985; Takezaki et al., 1995; Robinson et al., 1998) and likelihood (e.g. Felsenstein, 1988; Muse and Weir, 1992) approaches. Non-parametric tests have the advantage that they will not be influenced by the choice of a, possibly wrong, substitution model (Nei and Kumar 2000). The non-parametric rate test of Tajima (1993) compares two sequences with an outgroup sequence and counts the number of unique substitutions in both lineages. When both genes evolve under the molecular clock hypothesis (Zuckerkandl and Pauling, 1965), both genes are expected to have accumulated a similar number of 'unique' substitutions. On the other hand, when one of the duplicates has accumulated a significantly larger number of substitutions, the molecular clock does not apply and one of the paralogs is inferred to have experienced an increased evolutionary rate.

In several studies, rate differences between duplicates have been investigated. Hughes and Hughes (1993) did not detect any significant rate differences when investigating 17 recently duplicated genes in the tetraploid frog *Xenopus laevis*. Cronn et al. (1999) compared 16 paralogous loci in allotetraploid cotton and did not detect any significant rate difference after duplication, except for one locus, where pseudogenisation of one of the duplicates after the alloploidy event was suspected. In a study of 19 gene families in fish and mammals, Robinson-Rechavi and Laudet (2001) detected four families with a significant rate difference between duplicates. Kondrashov et al. (2002) analyzed 101 paralogous pairs in prokaryotes and eukaryotes and found about five with a significant rate difference. Zhang et al. (2002a) recently compared rates of 105 duplicated gene pairs on chromosomes 2 and 4 of *Arabidopsis thaliana*. Only three of these showed a significant rate difference after duplication at the protein level. In conclusion, according to most of the studies only a very small fraction of the duplicates show an increase in evolutionary rate after duplication, possibly pointing to relaxed functional constraints or positive selection. One of the few studies contradicting this finding was performed by Van de Peer and coauthors (2001), who examined 26 anciently duplicated genes in zebrafish, and observed an accelerated rate in about half of the duplicates using a nonparametric rate test.

However, only two of 14 duplicated fish genes from the study of Robinson-Rechavi and Laudet showed an accelerated rate. The drawback of both studies is, as with others (see above), the small number of duplicates investigated. Furthermore, the selection of genes might have been biased. For example, the majority of genes investigated by Van de Peer et al. (2001) are transcription factors. Whether this bias is responsible for the high fraction of duplicates that evolve at unequal rates remains to be investigated.

Detecting positive selection

A second way to study the evolution of genes after duplication *in silico* is to compare the rate of nonsynonymous substitutions, i.e. substitutions leading to amino acid replacements (K_N), with the rate of synonymous substitutions, i.e. substitutions that do not lead to amino acid replacement (K_S). The ratio of these two values, called w, provides a measure for the selection pressure on the protein product of a gene. A value of $\omega<1$ indicates purifying or negative selection that keeps the amino acid sequence from changing since most amino acid changes are disadvantageous, while $\omega=1$ indicates neutral evolution (Kimura, 1983). When $\omega>1$, this implies that natural selection favours amino acid replacements and as a result nonsynonymous substitutions are fixed at a higher rate than synonymous substitutions. A value for ω significantly greater than 1 can thus be an indication for the evolution of the gene towards a new function.
Traditionally, ω is measured over all sites of a gene. To estimate the number of nonsynonymous and synonymous rates, different approaches exist. In general, these can be divided into two classes: approximate (counting) methods, which estimate K_S and K_N for pairs of sequences, and Maximum Likelihood methods, which are usually based upon an explicit codon-substitution model, using a multiple sequence alignment and a phylogenetic tree. Approximate methods are based on counting the number of observed nonsynonymous and synonymous substitutions per nonsynonymous and synonymous site, after which a correction for multiple substitutions is applied. The simplest methods, such as the one of Nei and Gojobori (1986), assume equal nucleotide frequencies and no bias in the direction of change, while others take into account different rates of transitions and transversions (Li et al., 1985; Li, 1993; Pamilo and Bianchi, 1993; Ina, 1995; Comeron, 1995). A recently developed method also compensates for codon bias and unequal nucleotide frequencies (Yang and Nielsen, 2000).
The first Maximum Likelihood methods using explicit codon substitution models that allowed estimating K_N and K_S were developed in 1994 (Goldman and Yang, 1994; Muse and Gaut, 1994). These methods take into account biases in codon usage, base frequency, and transition/transversion ratio. Furthermore, the likelihood framework has the advantage of providing a statistical test to determine whether K_N is significantly higher than K_S. Using a Likelihood Ratio Test (LRT), one can compare the likelihood values under two hypotheses, in this case H_0 where ω is fixed to 1, and H_1 where ω is estimated as a free parameter. The rejection of the null model in the LRT, combined with an estimation of $\omega>1$, indicates positive or adaptive selection (Yang and Bielawski, 2000).

Although different methods have been developed to detect positive selection based on ω, it must be noted that the ratio of nonsynonymous over synonymous mutations can only be used to detect positive selection for recently duplicated genes. Once the gene has adapted to its specific function, purifying selection is expected to predominate, allowing the number of synonymous substitutions per site to catch up and eventually exceed the number of nonsynonymous substitutions per site (Hughes, 1999; Nei and Kumar, 2000; see further).

Using the methods described above, several examples of positive selection have been described in duplicated genes such as the primate ribonuclease (Zhang et al., 1998; Zhang et al., 2002b), mammalian immunoglobulin (Tanaka and Nei, 1989), pregnancy-associated glycoprotein (Hughes et al., 2000), and gastropod conotoxin genes (Duda and Palumbi, 1999). A more extensive overview of paralogous as well as orthologous genes for which positive selection has been detected can be found in Yang and Bielawski (2000).

On the other hand, several large-scale analyses showed that functional divergence through positive selection was not as ubiquitous as previously thought. Hughes and Hughes (1993) detected no positive selection in their analysis of 17 duplicated genes of *Xenopus laevis*, using the method of Nei and Gojobori (1986). Lynch and Conery (2000) observed 328 duplicated pairs with $\omega>1$ in a Maximum Likelihood analysis (Goldman and Yang, 1994) of 9870 pairs in several different eukaryotes. Zhang and co-workers (2002a), using the same technique, did not detect any genes under positive selection among 242 duplicated gene pairs on chromosome 2 and 4 in *Arabidopsis thaliana*. Kondrashov and co-workers (2002) found that the large majority of duplicates is under purifying selection, using the method of Pamilo and Bianchi (1993) and Li (1993) in an analysis of 4233 recently duplicated gene pairs in 26 bacterial, 6 archaeal and 7 eukaryotic genomes. Studies looking for positive selection without restricting to paralogs had also only limited success. Endo et al. (1996) applied the Nei and Gojobori (1986) test on 3595 groups of homologous genes and found only 17 groups of genes to have been under positive selection (with $\omega>1$ for a majority of all pairwise comparisons within a group). Sharp (1997), comparing 363 pairs of genes in mouse and rat, found only one gene, i.e. interleukin-3, with $\omega>1$.

The question remains whether positive selection is more rare than expected, or whether the developed methodologies are often incapable to reliably detect it. At least in one case, the shortcomings of the $\omega>1$ test to detect positive selection were clearly demonstrated. In a two-time point study on HIV drug resistance, Crandall and co-workers (1999) analyzed differences in ω for the protease gene in eight patients using the Nei and Gojobori (1986) method. They showed that in only two cases positive selection could be detected, while parallel adaptive substitutions leading to drug resistance were observed in five out of eight patients.

Problems in detecting positive selection

Sequence bias

A first problem in detecting positive selection is that the estimation of K_N and K_S is influenced by sequence composition (e.g. GC content) and codon biases (Smith et al., 1994). Several analyses discussed above used a simple method that does not compensate for biases in sequence content. More complex methods try to account for these biases and allow for, in general, more accurate estimations of ω (Bielawski et al., 2000).

The episodic nature of selection

Another problem is that positive selection is of an episodic nature, which means that, after a period of positive selection, purifying selection usually blurs the substitution pattern indicative of positive selection (Hughes, 1999; Nei and Kumar, 2000). As a result, positive selection cannot be detected anymore 30-50 million years after gene duplication using the ratio of K_N over K_S (Hughes, 1999; Hughes et al., 2000). To address this problem, three approaches have been used. A first approximate method evaluates whether nonsynonymous mutations occur in such a way as to change protein charge or polarity to a greater extent than is expected under random substitution. This method involves the computation of the proportion of radical nonsynonymous differences (p_{NR}) per radical nonsynonymous site versus the proportion of conservative nonsynonymous differences per conservative nonsynonymous site (p_{NC}). When $p_{NR} > p_{NC}$, nonsynonymous differences occur in such a way as to change the property of interest to a greater extent than expected at random (Hughes et al., 1990). Since this method looks at nonsynonymous sites only and the resulting amino acid changes, the occurrence of positive selection should be evident for a much longer period. It should be noted though that this method might be less sensitive to detect positive selection than looking at the K_N/K_S ratio (Vacquier et al., 1996; Hughes, 1999). Furthermore, a recent study showed that this measure is heavily influenced by the transition-transversion ratio and amino acid composition of the investigated sequences (Dagan et al., 2002). Therefore, inferences on positive selection based on this method should be treated with caution.
The second strategy is based on the reconstruction of ancestral sequences at the internodes of the phylogenetic tree. Given a substitution model and a tree topology, ancestral sequences can be inferred through a variety of parsimony (Eck and Dayhoff, 1966; Fitch 1971; Maddison and Maddison, 1992; Swofford, 2002), distance (Zhang and Nei, 1997), maximum likelihood (Yang et al., 1995; Schluter, 1995; Koshi and Goldstein, 1996; Pagel, 1999; Pupko et al., 2000; 2002) and hierarchical Bayesian (Huelsenbeck and Bollback, 2001) approaches. By comparing these ancestral sequences, ω can be measured along a specific branch (between two ancestral nodes, or an ancestral node and an endnode) on the tree, corresponding with a more specific period in evolution. Although not explicitly looking at duplicated genes, Liberles et al. (2001) detected about 4% of 8690 chordate and embryophyte gene families investigated to have at least one branch in which $\omega>1$ using this approach.

A third strategy relies on the above-mentioned Maximum Likelihood approach using codon models, which allow for ω to vary among branches of the tree. Using a Likelihood Ratio Test (LRT), one can compare the likelihood values under two hypotheses, in this case H_0 where ω is fixed, and H_1 where ω is estimated as a free parameter for (a) specific branch(es). If ω is estimated to be >1 for the chosen branch(es) and the LRT gives a significant result, this is indicative for positive selection in that branch (Yang, 1998). This technique was successfully applied to duplicated ribonuclease genes, thereby confirming earlier results (Bielawski and Yang, 2003).

Positive selection acts locally

Another major reason that might explain the low prevalence of detectable positive selection lies in the fact that, in general, ω is measured as an average over all sites of a gene. This implies that, if only a fraction of sites is under positive selection, their detection is complicated. Not all amino acids of a protein are functionally important and therefore these can evolve in a more neutral way, while others do have important structural and functional roles and are under strong purifying selection. One can imagine that after duplication, e.g. only the domains involved in substrate binding specificity are under positive selection, while all the other sites retain their original evolutionary rates, obscuring the former sites when looking at the K_N/K_S ratio for the gene as a whole. For example, Hughes and Nei (1988) detected ω values >1 in the antigen recognition region of the Major Histocompatibility Complex, while other regions of the genes had values for ω less than 1. Endo and co-workers (1996) also recognized the possibility of region-restricted positive selection, and also used a second, sliding window method to look for evidence of positive selection, to avoid averaging over the entire gene, an approach also followed by Duda and Palumbi (1999). Fares and co-workers (2002) further improved this kind of approach by estimating the appropriate window size and by detecting saturation at synonymous sites.
Positive selection can also be limited to a few dispersed amino acids. For this reason, methods were developed that allow detecting positive selection at single amino acid sites. A first method is based on inferring ancestral sequences for a given tree topology by testing neutrality (ω=1) for each codon site using the numbers of synonymous and nonsynonymous changes detected throughout the tree. Using this method, positive selection on specific sites of the Human Leukocyte Antigen (HLA) gene was detected, yielding two new putative antigen recognition sites (Suzuki and Gojobori, 1999). This method is now also implemented in a publicly available software package for UNIX called ADAPTSITE (Suzuki et al., 2001). Another application of a similar technique can be found in Bush et al. (1999) who examined positive selection in individual codons for the H3 hemagglutinin gene of the human influenza virus A.
In addition, Maximum Likelihood models were developed that allow for heterogeneous selection pressure among sites. They also allow hypothesis testing as described above, using classes of sites that have different values of w. Models implementing discrete as well as continuous (gamma, beta) ω distributions are provided. For example, one can compare (using a LRT) a model in which sites have a continuous distribution of ω values between 0 and 1 with a model having one extra class of sites exists in which ω is freely estimated.

If the LRT is significant and sites in the extra class have an ω>1, positive selection on a subset of sites is assumed. This method allowed the detection of positive selection in several genes, where earlier methods had failed (Nielsen and Yang, 1998; Yang et al., 2000). Using a Bayesian approach, the posterior probability for each site to belong to a class of ω values can be calculated, and by consequence the sites under positive selection can be identified (Nielsen and Yang, 1998; Yang et al., 2000).

Only recently, methods have been developed to combine detection of lineage- and site-specific positive detection (Yang and Nielsen, 2002). As in the lineage-specific methods, a branch can be selected, for which positive selection should be tested (the so-called 'foreground' branch). All other branches are referred to as 'background' branches. Two models were developed. The first model (referred to as the "A" model) is based on four classes of sites, namely two classes containing sites with ω_0=0 (class 0) or ω_1=1 (class1), representing sites that are not under positive selection, and two classes allowing (background) sites of the ω_0 and ω_1 class to change to a third (estimated) ω_2 >1 in the foreground branch, respectively (sites going from purifying to positive selection (ω_0->ω_2) in class 2 and sites going from neutral evolution to positive selection (ω_1->ω_2) in class 3). The second ("B") model allows also for sites under positive selection in the background lineages, as ω_0 and ω_1 are estimated freely over the entire phylogenetic tree. These models have been applied successfully to detect positive selection after gene duplication in the phytochrome, Troponin C and chalcone synthase gene families, for which the previous models did not detect positive selection (Yang and Nielsen, 2002; Bielawski and Yang, 2003; Yang et al., 2002). A new model is currently under development, which is less restrictive and allows a class of sites with two independent estimations of ω for the two branches following the duplication event, in order to model site-specific divergence in selective pressure following duplication. This model further refines the possibilities of the previous ones, and has been successfully applied to a number of gene families (Joseph P. Bielawski, pers. comm.). These recent models, together with the Bayesian identification of sites under positive selection are very promising and will hopefully allow very detailed study of functional divergence after duplication.

All Maximum Likelihood approaches using codon models described above are implemented in the PAML package (Yang, 1997), which is publicly available for UNIX, Windows and Apple Macintosh operating systems.

One of the most recent developments is the use of "stand-alone" Bayesian approaches to detect positively selected mutations at specific sites and lineages. Nielsen and Huelsenbeck (2002) developed a method based on mapping mutations on the phylogenetic tree (Nielsen, 2002), that gave similar results to the Yang and Nielsen (2002) Maximum Likelihood approach. However, this approach allows the further exploration of the evolutionary history of the investigated genes. As an example, they showed, rather unexpectedly, that in the Influenza hemagglutinin protein, positively selected amino acid changes tended to be mostly conservative, instead of the expected radical substitutions.

Other methods to detect functional divergence

Several methods have been developed to detect functional divergence after duplication on the premise of rate shifts of specific positions or regions of the protein. It is postulated that when new functions are acquired by amino acid substitutions, the selective constraints upon these positions will also change, which in turn will lead to a difference in substitution rate at these sites (the so-called type-I functional divergence; Gu, 1999). One of the first methods to detect rate changes was developed by Gu (1999, 2001), and uses a coefficient of statistical divergence (q_I) to measure the functional divergence between two paralogous clusters of a tree. q_I is defined as the decrease in rate correlation between the two clusters, and was initially estimated using a simple algorithm based on a Poisson model of molecular evolution. Gu also developed a probabilistic model with two possible states for each site: S_0 when the site is 'functional-divergence-unrelated', meaning that the evolutionary rate of that site is the same between two clusters, and S_1 ('functional-divergence-related') when there is no rate correlation between clusters and altered functional constraints are hypothesized. In this model, q_I can be interpreted as the probability P(S) of a site being in the 'functional divergence state'. Using a Maximum Likelihood approach, q_I and the other parameters of the model (the gamma shape parameter a and branch lengths) are estimated after which a Likelihood Ratio Test can be used to discern between the null hypothesis that there is no rate difference between the same sites of two clusters (H_0: q_I=0) and the alternative hypothesis H_1: q_I>0. The method also allows to analyze three or more clusters at the same time and incorporates a Bayesian approach to predict sites which are likely responsible for the functional divergence. It was successfully applied to several vertebrate gene families (for an overview, see Gaucher et al., 2002). In addition, methods to detect type II functional divergence are proposed. In type II divergence, there is no detectable rate difference between clusters, but sites have functionally diverged shortly after duplication at certain sites, resulting in radical amino acid property differences at these positions between clusters, although the functional constraint (which is reflected by the evolutionary rate) became similar again, as soon as these changes had occurred (Gu, 2001). The algorithms were recently embedded in a software package called DIVERGE, featuring a graphical user interface for Windows and Linux operating systems. This program also allows to map these sites on a 3D-structure, if available, in order to facilitate the understanding of the functional importance of discovered critical sites (Gu and Vander Velden, 2002).
Gaucher et al. (2001) used statistical quantiles to detect functionally important sites in elongation factors by comparing the bacterial EF-Tu proteins with their eukaryotic (and functionally diverged) EF-1a counterparts. Sites that had a rate difference between the two groups of more than 2 standard deviations in the distribution of rate differences per site were considered to be candidate sites responsible for the difference in function. Subsequently, they mapped these positions on the known tertiary structure of these proteins. By correlating this position with the known functional divergence of the proteins, they were able to propose putative functions (e.g. tRNA and cytoskeleton interaction) for these sites. Liberles (2001) proposed two alternative measures of adaptive evolution. A first method consists of calculating the ratio between the number of Point Accepted Mutations (PAM) and the Neutral Evolutionary Distance (NED; Peltier et al., 2000).

The latter distance is based on the proportion of conserved twofold degenerate codons. These codons are chosen because the differences between each of these codons are represented solely by transitions at the third codon position (Peltier et al., 2000), making the NED more clocklike than K_S, where transitions and transversions, which occur with different probabilities, are considered. Nevertheless, in general, it is expected that PAM/NED ratios are similar to K_N/K_S ratios, as also observed by Liberles (2001). A second method, the Sequence Space Assessment (SSA) statistic, measures the fraction of amino acid sites that have undergone substitution along a certain branch, compared to the total number of sites that are variable at one or more branches in the tree (normalized for the number of taxa).

Dermitzakis and Clark (2001) modified a method designed by Tang and Lewontin (1999) that measures within-protein rate heterogeneity in duplicated genes. This method, called Paralog Heterogeneity Test, was developed in particular to detect subfunctionalization (see introduction) at the protein domain level. In other words, it detects whether in one paralog, one region of the protein has evolved more rapidly than that same region in the other paralog. The method works by comparing each paralog to a respective ortholog using a sliding window approach where a Q-value is measured for each variable site in the alignment at a certain window position. This Q-value is a measure for the density of sequence variability in that window. By comparing the Q-values of both paralogs, regions that differ in variability can be determined. The software tools also contain a script to perform randomization tests in order to calculate the significance of the obtained results. The authors applied their method to several mouse and human gene families and detected several cases in which two regions of a protein evolved at a different rate in two paralogs, which may point to subfunctionalization. A similar method, using user-defined regions instead of a sliding window was also described by Marín et al. (2001).

Functional divergence at the regulatory level

Although this review focuses on the analysis of the protein coding part of a gene, novel gene functions do not only arise by modification of the coding region, but also by changing its expression. As the expression of genes is, at least partly, dependent on the presence of transcription factor binding sites in regulatory regions, mutations in these elements can alter the expression domain of genes. For example, subfunctionalisation has been proposed to act mainly at the regulatory level, where the reciprocal loss of different regulatory elements can lead to functional divergence through expression in e.g. different organs or stages of development (Force et al., 1999). The *in silico* investigation of promoter regions of duplicated genes should allow to unravel the evolution of regulation after duplication. The most straightforward approach would be to align promoters using standard alignment tools, and look for patterns of loss and gain of regulatory motifs. Unfortunately, these alignment methods are rather rigid and when, for example, the motif position or order is changed, or sequences are too divergent, methods based on sequence alignment have serious difficulties of aligning homologous regulatory regions.

New techniques such as the detection of overrepresented motifs by word counting or probabilistic methods and especially methods such as phylogenetic footprinting, which take into account the phylogenetic relationships of genes, do consider this dynamic nature of promoters and allow to investigate whether loss or gain of certain regulatory motifs might have led to the functional divergence of duplicated genes. Nevertheless, although recent approaches seem promising, in general, unambiguous identification of regulatory elements is far from straightforward. The delineation of promoters is even harder, due to its complex nature, and *in silico* promoter prediction is still in its infancy (Rombauts et al., 2003).

Conclusions

The function of a gene is usually determined by a rather complex combination of the three-dimensional structure of the protein it encodes, and its spatio-temporal expression determined by its cis-regulatory elements. In addition, other processes such as post-translational and –transcriptional modifications, transport and cellular context also play an important role in the definition of a gene's function. Duplicated genes provide an excellent tool to study gene function and how genes diverge in function. After duplication, one gene copy is redundant and, freed from functional constraint, can evolve a new function. Numerous models have been put forward to explain the retention and functional divergence of genes and the study of these processes, bringing together fundamental evolutionary research and more applied functional genomics, has now become a rapidly growing field of research. Although the *in silico* determination of functional difference between two duplicated genes is inevitably compromised by the complex nature of what defines a gene's function, as discussed here, much progress has been made in the last few years and many novel approaches have become available to study the functional diversification of genes. By formulating testable working hypotheses, these *in silico* methods can speed up and focus research in many different domains.

Acknowledgements

The authors would like to thank J.P. Bielawski for helpful discussions and for sharing unpublished results.

References

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. (2002). Evidence of en bloc duplication in vertebrate genomes. Nat Genet **31,** 100-105.

Aburomia, R., Khaner, O., and Sidow, A. (2003). Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. J Struct Funct Genom **3**, 45-52.

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M., and Postlethwait, J.H. (1998). Zebrafish hox clusters and vertebrate genome evolution. Science **282,** 1711-1714.

Aparicio, S., Hawker, K., Cottage, A., Mikawa, Y., Zuo, L., Venkatesh, B., Chen, E., Krumlauf, R., and Brenner, S. (1997). Organization of the Fugu rubripes Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat Genet **16,** 79-83.

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

Bielawski, J.P., and Yang, Z. (2003). Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genom **3**, 201-212.

Bielawski, J.P., Dunn, K.A., and Yang, Z. (2000). Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics **156,** 1299-1308.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12,** 1093-1101.

Bush, R.M., Fitch, W.M., Bender, C.A., and Cox, N.J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol **16,** 1457-1465.

Comeron, J.M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol **41,** 1152-1159.

Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C., and Salzman, N.P. (1999). Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol **16,** 372-382.

Cronn, R.C., Small, R.L., and Wendel, J.F. (1999). Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci U S A **96,** 14406-14411.

Dagan, T., Talmor, Y., and Graur, D. (2002). Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol Biol Evol **19,** 1022-1025.

Dermitzakis, E.T., and Clark, A.G. (2001). Differential selection after duplication in mammalian developmental genes. Mol Biol Evol **18,** 557-562.

Duda, T.F., Jr., and Palumbi, S.R. (1999). Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. Proc Natl Acad Sci U S A **96,** 6820-6823.

Eck, R.V., and Dayhoff, M.O. (1966). Atlas of Protein Sequence and Structure. (Silver Spring, MD: National Biomedical Research Foundation).

Endo, T., Ikeo, K., and Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Mol Biol Evol **13,** 685-690.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet **22,** 521-565.

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool **20,** 406-416.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151,** 1531-1545.

Friedman, R., and Hughes, A.L. (2001). Pattern and timing of gene duplication in animal genomes. Genome Res **11,** 1842-1847.

Friedman, R., and Hughes, A.L. (2003). The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol **20,** 154-161.

Furlong, R.F., and Holland, P.W. (2002). Were vertebrates octoploid? Philos Trans R Soc Lond B Biol Sci **357,** 531-544.

Gaucher, E.A., Miyamoto, M.M., and Benner, S.A. (2001). Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc Natl Acad Sci U S A **98,** 548-552.

Gaucher, E.A., Gu, X., Miyamoto, M.M., and Benner, S.A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci **27,** 315-321.

Gehring, W.J. (1998). Master control genes in development and evolution: the homeobox story. (New Haven: Yale University Press).

Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet **14,** 46-49; discussion 49-50.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11,** 725-736.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol **16,** 1664-1674.

Gu, X. (2001). A site-specific measure for rate difference after gene duplication or speciation. Mol Biol Evol **18,** 2327-2330.

Gu, X., and Vander Velden, K. (2002). DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics **18,** 500-501.

Gu, X., Wang, Y., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet **31,** 205-209.

Holland, P.W. (1997). Vertebrate evolution: something fishy about Hox genes. Curr Biol **7,** R570-572.

Holland, P.W. (2003). More genes in vertebrates? J Struct Funct Genom **3**, 75-84.

Holland, P.W., and Garcia-Fernandez, J. (1996). Hox genes and chordate evolution. Dev Biol **173,** 382-395.

Holland, P.W., Garcia-Fernandez, J., Williams, N.A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. Dev Suppl, 125-133.

Huelsenbeck, J.P., and Bollback, J.P. (2001). Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol **50,** 351-366.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci **256,** 119-124.

Hughes, A.L. (1999). Adaptive evolution of genes and genomes. (New York: Oxford University Press).

Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335,** 167-170.

Hughes, M.K., and Hughes, A.L. (1993). Evolution of duplicate genes in a tetraploid animal, xenopus laevis. Mol Biol Evol **10,** 1360-1369.

Hughes, A.L., Ota, T., and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol **7,** 515-524.

Hughes, A.L., Green, J.A., Garbayo, J.M., and Roberts, R.M. (2000). Adaptive diversification within a large family of recently duplicated, placentally expressed genes. Proc Natl Acad Sci U S A **97,** 3319-3323.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol **40,** 190-226.

Kimura, H. (1983). The neutral theory of molecular evolution. (Cambridge: Cambridge University Press).

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002). Selection in the evolution of gene duplications. Genome Biol **3,** research0008.1-0008.9.

Koshi, J.M., and Goldstein, R.A. (1996). Probabilistic reconstruction of ancestral protein sequences. J Mol Evol **42,** 313-320.

Larhammar, D., Lundin, L.G., and Hallbook, F. (2002). The Human Hox-bearing Chromosome Regions Did Arise by Block or Chromosome (or Even Genome) Duplications. Genome Res **12,** 1910-1920.

Li, W.H. (1980). Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95,** 237-258.

Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36,** 96-99.

Li, W.H. (1997). Molecular Evolution. (Sunderland, Massachusets: Sinauer).

Li, W.H., Wu, C.I., and Luo, C.C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2,** 150-174.

Liberles, D.A. (2001). Evaluation of methods for determination of a reconstructed history of gene sequence evolution. Mol Biol Evol **18,** 2040-2047.

Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., and Benner, S.A. (2001). The adaptive evolution database (TAED). Genome Biol **2,** research0028.1-0028.6.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Maddison, W.P., and Maddison, D.R. (1992). MacClade: analysis of phylogeny and character evolution (Sunderland, MA: Sinauer).

Malaga-Trillo, E., and Meyer, A. (2001). Genome Duplications and Accelerated Evolution of Hox Genes and Cluster Architecture in Teleost Fishes. Am Zool **41,** 676-686.

Margoliash, E. (1963). Primary structure and evolution of cytochrome *c*. Proc Natl Acad Sci U S A **50,** 672-679.

Marin, I., Fares, M.A., Gonzalez-Candelas, F., Barrio, E., and Moya, A. (2001). Detecting changes in the functional constraints of paralogous genes. J Mol Evol **52,** 17-28.

McLysaght, A., Hokamp, K., and Wolfe, K.H. (2002). Extensive genomic duplication during early chordate evolution. Nat Genet **31,** 200-204.

Meyer, A., and Van de Peer, Y. (2003). Natural selection merely modified while redundancy created' - Susumu Ohno's idea of the evolutionary importance of gene and genome duplications. J Struct Funct Genom **3,** vii-ix.

Muse, S.V., and Weir, B.S. (1992). Testing for equality of evolutionary rates. Genetics **132,** 269-276.

Muse, S.V., and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11,** 715-724.

Naruse, K., Fukamachi, S., Mitani, H., Kondo, M., Matsuoka, T., Kondo, S., Hanamura, N., Morita, Y., Hasegawa, K., Nishigaki, R., Shimada, A., Wada, H., Kusakabe, T., Suzuki, N., Kinoshita, M., Kanamori, A., Terado, T., Kimura, H., Nonaka, M., and Shima, A. (2000). A detailed linkage map of medaka, Oryzias latipes: comparative genomics and genome evolution. Genetics **154,** 1773-1784.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3,** 418-426.

Nei, M., and Kumar, S. (2000). Molecular evolution and phylogenetics. (New York: Oxford University Press).

Nelson, J.S. (1994). Fishes of the world, 3rd edition. (New York: Wiley).

Nielsen, R. (2002). Mapping mutations on phylogenies. Syst Biol **51,** 729-739.

Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148,** 929-936.

Nielsen, R., and Huelsenbeck, J.P. (2002). Detecting positively selected amino acid sites using posterior predictive P-values. Pac Symp Biocomput**,** 576-588.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature **388,** 167-171.

Ohno, S. (1970). Evolution by Gene Duplication. (Berlin;Heidelberg;New York: Springer-Verlag).

Pagel, M. (1999). Inferring the historical patterns of biological evolution. Nature **401,** 877-884.

Pamilo, P., and Bianchi, N.O. (1993). Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol **10,** 271-281.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.X., Katsar, C.S., Lan, T.H., Lin, Y.R., Ming, R., and Wright, R.J. (2000). Comparative genomics of plant chromosomes. Plant Cell **12,** 1523-1540.

Peltier, M.R., Raley, L.C., Liberles, D.A., Benner, S.A., and Hansen, P.J. (2000). Evolutionary history of the uterine serpins. J Exp Zool **288,** 165-174.

Piatigorsky, J., and Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. Science **252,** 1078-1079.

Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W.S. (2000). Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res **10,** 1890-1902.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet **3,** 827-837.

Pupko, T., Pe'er, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol **17,** 890-896.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., and Friedman, N. (2002). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. Bioinformatics **18,** 1116-1123.

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., and Van de Peer, Y. (2003). Investigating ancient duplication events in the Arabidopsis genome. J Struct Funct Genom **3,** 117-129.

Robinson, M., Gouy, M., Gautier, C., and Mouchiroud, D. (1998). Sensitivity of the relative-rate test to taxonomic sampling. Mol Biol Evol **15,** 1091-1098.

Robinson-Rechavi, M., and Laudet, V. (2001). Evolutionary rates of duplicate genes in fish and mammals. Mol Biol Evol **18,** 681-683.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., and Van de Peer, Y. (2003). Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. Plant Phys **(in press)**.

Sarich, V.M., and Wilson, A.C. (1973). Generation time and genomic evolution in primates. Science **179,** 1144-1147.

Schluter, D. (1995). Uncertainty in ancient phylogenies. Nature **377,** 108-110.

Seoighe, C., and Wolfe, K.H. (1999). Yeast genome evolution in the post-genome era. Curr Opin Microbiol **2,** 548-554.

Sharp, P.M. (1997). In search of molecular darwinism. Nature **385,** 111-112.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev **6,** 715-722.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van De Peer, Y. (2002). The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci U S A **99,** 13627-13632.

Smith, J.M. (1994). Estimating selection by comparing synonymous and substitutional changes. J Mol Evol **39,** 123-128.

Spring, J. (1997). Vertebrate evolution by interspecific hybridisation—are we polyploid? FEBS Lett **400,** 2-8.

Spring, J. (2002). Genome duplication strikes back. Nat Genet **31,** 128-129.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J Mol Evol **49,** 169-181.

Suzuki, Y., and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. Mol Biol Evol **16,** 1315-1328.

Suzuki, Y., Gojobori, T., and Nei, M. (2001). ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics **17,** 660-661.

Swofford, D. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). (Sunderland, Massachusetts: Sinauer Associates).

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135,** 599-607.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic tests of the molecular clock and linearized trees. Mol Biol Evol **12,** 823-833.

Tanaka, T., and Nei, M. (1989). Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol **6,** 447-459.

Tang, H., and Lewontin, R.C. (1999). Locating regions of differential variability in DNA and protein sequences. Genetics **153,** 485-495.

Taylor, J.S., Van de Peer, Y., and Meyer, A. (2001b). Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Curr Biol **11,** R1005-1008.

Taylor, J.S., Van de Peer, Y., Braasch, I., and Meyer, A. (2001a). Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos Trans R Soc Lond B Biol Sci **356,** 1661-1679.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. Genome Res **13,** 382-390.

Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van Den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Van Montagu, M., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in Arabidopsis thaliana by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett **445,** 237-245.

Vacquier, V.D., Swanson, W.J., and Lee, Y.H. (1997). Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? J Mol Evol **44,** S15-22.

Van de Peer, Y., Taylor, J.S., Braasch, I., and Meyer, A. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol **53,** 436-446.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in Arabidopsis. Science **290,** 2114-2117.

Wagner, A. (1998). The fate of duplicated genes: loss or new function? Bioessays **20,** 785-788.

Wagner, A. (2002). Assymetric functional divergence of duplicate genes in yeast. Mol Biol Evol **19,** 1760-1768.

Walsh, J.B. (1995). How often do duplicated genes evolve new functions? Genetics **139,** 421-428.

Wittbrodt, J., Meyer, A., and Schartl, M. (1998). More genes in fish? Bioessays **20,** 511-512.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet **2,** 333-341.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387,** 708-713.

Woods, I.G., Kelly, P.D., Chu, F., Ngo-Hazelett, P., Yan, Y.L., Huang, H., Postlethwait, J.H., and Talbot, W.S. (2000). A comparative map of the zebrafish genome. Genome Res **10,** 1903-1914.

Wu, C.I., and Li, W.H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA **82,** 1741-1745.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13,** 555-556.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15,** 568-573.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17,** 32-43.

Yang, Z., and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15,** 496-503.

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19,** 908-917.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141,** 1641-1650.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155,** 431-449.

Yang, J., Huang, J., Gu, H., Zhong, Y., and Yang, Z. (2002). Duplication and adaptive evolution of the chalcone synthase genes of Dendranthema (Asteraceae). Mol Biol Evol **19,** 1752-1759.

Zhang, J., and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol **44,** S139-146.

Zhang, J., Rosenberg, H.F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A **95,** 3708-3713.

Zhang, L., Vision, T.J., and Gaut, B.S. (2002a). Patterns of Nucleotide Substitution Among Simultaneously Duplicated Gene Pairs in Arabidopsis thaliana. Mol Biol Evol **19,** 1464-1473.

Zhang, J., Zhang, Y.P., and Rosenberg, H.F. (2002b). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet **30,** 411-415.

Zuckerkandl, E., and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins, V. Bryson and H.J. Vogel, eds (New York: Academic Press).

[Chapter 9]

Discussion

The last decade has seen the advent of an enormous amount of sequence data, both from large-scale EST and complete genome sequencing projects. The same period is marked by the rise of bioinformatics, a research field aimed at all aspects of the art of extracting knowledge from this huge pile of raw data. Thanks to the development of gene prediction programs and similarity-based search tools, an unseen wealth of genes and gene families is discovered. New advances in evolutionary analysis tools allow the in-depth analysis of the history of duplications and divergences lying at the origin of the expansion of these gene families. Furthermore, the availability of complete genome sequences provides a global and exhaustive view at the complete set of candidate genes for a particular function. In-depth sequence analysis of these genes can result in the *in silico* prediction of signal peptides, conserved domains, regulatory elements as well as secondary and even tertiary protein structure (Chapters 1, 3, 4 and 6).

The rapid increase of data and high demand by the research community forced the development of highly automated systems for gene discovery and functional annotation. This high-throughput approach came at a price: since the first publication of many genome sequences, numerous reports of erroneous annotation have been published, both at the functional and structural level (see Chapter 1).

Although many reasons exist for the incorrect assignment of function, the main danger lies in the fact that, due to the automated use of sequence similarity for functional annotation, existing erroneous annotations quickly propagate through the databases, becoming themselves a source of further error proliferation (Brenner, 1999; Aubourg and Rouzé, 2001). At the structural level, one of the main problems is the intrinsic compositional difference between genomes, resulting in the fact that the most optimal gene finding strategy or tool for one species can produce significantly inferior results in another. This was especially true for the *Arabidopsis* genome, being the first plant genome to be sequenced. Software developed for animal genomes did not perform as well in this model plant, due to differences in base composition, codon usage, splice site consensuses, et cetera (Pavy et al., 1999).

Several of the results presented in this PhD thesis relate directly to this problem: due to the numerous errors found in the first automatic annotation of the *Arabidopsis* genome, manual reannotation of genes was a slow, tedious, but necessary step in the analysis of genes and their families. It was in this respect that initiatives such as the GeneFarm project were set up (Chapter 2). The value of manual, expert annotation was quickly acknowledged by the *Arabidopsis* research community, as can be seen from the websites dedicated to this subject that are hosted by TAIR (Rhee et al., 2003) and MIPS (Schoof et al., 2002), as well as the numerous family-wise reannotation papers that have been published in recent years.

However, during the course of this PhD, an important number of EST and full-length cDNA sequences was generated for *Arabidopsis*, leading to a significant improvement in annotation quality (e.g. Haas et al., 2002). This extrinsic gene prediction approach showed to be - although more expensive and labour-intensive than pure *in silico* annotation pipelines - a qualitative high-troughput alternative to manual annotation. In the future, the extrinsic annotation approach will even gain importance in plant research, as large amounts of ESTs have been and are being sequenced for a great number of species from different taxonomic groups. In addition, already a second plant genome sequence is available: that of the model monocot plant rice (Goff et al., 2002; Yu et al., 2002). Thanks to extrinsic approaches using EST data and the transfer of annotation from *Arabidopsis* and rice to other genomes, together with a steady improvement of gene prediction tools, future annotation pipelines will probably show a much higher accuracy. For transfer of annotation to work, however, a reliable annotation must be achieved for the two current reference genomes. For this reason, the many general or family specific re-annotation projects (based on expert intervention or large-scale cDNA sequencing) lay the foundations of the future annotation of *Populus* and *Medicago*.

The family-wise annotation of genes provides - besides the obvious gene structure - the opportunity to gain insight in the function of these genes and processes of evolution and functional divergence. The exhaustive annotation of a gene family within a genome allows one to get a clear view at the toolbox of genes at the disposition of the organism for a specific function. Sequence analysis of a gene family can provide an insight into differences/similarities in function between members and give indications at possible redundancy between closely related duplicates. Phylogenetic studies allow the classification of genes in related groups and present hypotheses on the evolution of the family and the pathways it is involved in. In addition, it allows the derivation of correct orthology-paralogy relationships between family members in different organisms. The transfer of function between orthologous genes has been shown to be a reliable way of *in silico* functional annotation (Eisen, 1998). In this respect, the family analyses presented in this thesis (Chapters 3, 4, 5 and 6) constitute the foundations of current and future research in the respective families and pathways, by pointing out target genes for specific functions and giving insights into putative redundant genes within species.

The analysis of gene families also allows to investigate the role of duplication in evolution. Both large and small-scale duplication events lie at the basis of the current diversity of genes found in *Arabidopsis*. It was shown that at least one, and probably three genome duplication have marked the evolutionary history of this model plant (Chapter 7). In addition, many examples of recent tandem duplication were found in studies presented here, together with more ancient duplications, of which the origin is blurred through gene loss, translocations and genome rearrangements (Chapters 3, 4 and 6). However, it appears to be very difficult to draw general conclusions of the (relative) impact of these different events. In the families studied here, no clear effect was seen of the complete genome duplications on the gene families in question nor on the pathway they acted in, or at least not the ‘large’ effect, that one would expect to see, given the magnitude of the event.

The reason for this could lie in the fact that the knowledge about function, interaction and expression of the genes investigated remains probably too limited to actually conclude something. In addition, it could also be that to see the importance of these events, one should look at a level even beyond that of single pathways, as the genome-wide duplication of genes acts on the organism as a whole. Finally, it should be noted that up till now, although many theories on the impact of genome duplication on speciation and evolution of novelty exist, no convincing evidence has been found favouring one or the other. Consequently, one would have to take into account the sobering possibility that the impact of these genome duplications is not as profound as has been previously thought. However, given the paucity of functional data available, I personally prefer to give the current hypotheses the benefit of the doubt (a practice also known as “hope-driven science”).

On the other hand, when looking at a smaller scale, the effect of gene duplication on single genes has been described in literature for more and more (isolated) cases now. Subfunctionalisation at the regulatory level is increasingly observed under the form of differences in expression patterns of duplicated genes, while at the same time smaller functional differences (e.g. substrate or binding specificity) are more and more observed between duplicates (Prince and Pickett, 2002). On the other hand, cases of apparent complete functional redundancy are also observed, albeit of course easier to detect one difference than to prove there is none. In recent years, great advances have been made to investigate these events *in silico*. The thorough computational analysis of gene families will allow to formulate hypotheses on the diversification of genes, and consequently on the reasons behind the expansion or apparent redundancy between genes, i.e. whether duplicated genes are subject to differences in evolutionary rate, to regulatory or coding-level subfunctionalisation, positive or purifying selection. By complementing these hypotheses with wet-lab experiments to investigate differences in expression (RT-PCR, Northern, micro-array,…), interaction (Y1H, Y2H, Y3H, ChIP…) or substrate specificity (metabolic assays), we will be one step closer in discovering the fate of duplicated genes and their role in evolution.

References

Aubourg, S., and Rouzé, P. (2001). Genome annotation. Plant Physiol Biochem **39,** 181-193.

Brenner, S.E. (1999). Errors in genome annotation. Trends Genet **15,** 132-133.

Eisen, J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8,** 163-167.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa L.* ssp. *japonica*). Science **296,** 92-100.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol **3,**research0029.1-0029.12.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. (1999). Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. Bioinformatics **15,** 887-899.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet **3,** 827-837.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res **31,** 224-228.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., and Mayer, K.F. (2002). MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30,** 91-93.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (*Oryza sativa L.* ssp. *indica*). Science **296,** 79-92.

[Addenda]

The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice.

Klaas Vandepoele[†], Yvan Saeys[†], Cedric Simillion, Jeroen Raes and Yves Van de Peer[*]

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

[*] *Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*
[†] *The two authors contributed equally to this work.*

Published in: Genome Research **12**, 1792-1801 (2002)

Abstract

It is expected that one of the merits of comparative genomics lies in the transfer of structural and functional information from one genome to another. This is based on the observation that, although the number of chromosomal rearrangements that occur in genomes is extensive, different species still exhibit a certain degree of conservation regarding gene content and gene order. It is in this respect that we have developed a new software tool for the Automatic Detection of Homologous Regions (ADHoRe). ADHoRe was primarily developed to find large regions of microcolinearity, taking into account different types of microrearrangements such as tandem duplications, gene loss and translocations, and inversions. Such rearrangements often complicate the detection of colinearity, in particular when comparing more anciently diverged species. Application of ADHoRe to the complete genome of Arabidopsis and a large collection of concatenated rice BACs yields more than 20 regions showing statistically significant microcolinearity between both plant species. These regions comprise from 4 up to 11 conserved homologous gene pairs. We predict the number of homologous regions and the extent of microcolinearity to increase significantly once better annotations of the rice genome become available.

Transcriptome analysis during cell division in plants.

Peter Breyne[†], Rozemarijn Dreesen[†], Klaas Vandepoele, Lieven De Veylder, Frank Van Breusegem, Lindy Callewaert, Stephane Rombauts, Jeroen Raes, Bernard Cannoot, Gilbert Engler, Dirk Inzé[*] and Marc Zabeau

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

[*] *Author for correspondence (e-mail diinz@psb.ugent.be; fax: +32 9 331 3809)*
[†] *The two authors contributed equally to this work.*

Published in: Proceedings of the National Academy of Sciences USA **99**:14825-14830 (2002)

Abstract

Using synchronized tobacco Bright Yellow-2 cells and cDNA-amplified fragment length polymorphism-based genomewide expression analysis, we built a comprehensive collection of plant cell cycle-modulated genes. Approximately 1,340 periodically expressed genes were identified, including known cell cycle control genes as well as numerous unique candidate regulatory genes. A number of plant-specific genes were found to be cell cycle modulated. Other transcript tags were derived from unknown plant genes showing homology to cell cycle-regulatory genes of other organisms. Many of the genes encode novel or uncharacterized proteins, indicating that several processes underlying cell division are still largely unknown.

Molecular characterization of *Arabidopsis* PHO80-Like-Proteins, a novel class of CDKA;1 binding cyclins

Juan Antonio Torres Acosta , Janice de Almeida Engler, Jeroen Raes, Zoltan Magyar, Ruth De Groodt, Dirk Inzé *, Lieven De Veylder

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

* *Author for correspondence*

Manuscript in preparation. Intended for: Plant Physiology

Abstract

Cyclin regulatory proteins interact with Cyclin-Dependent Kinases (CDKs) to control the progression through the cell cycle. In *Arabidopsis thaliana,* 34 cyclin genes, grouped into different classes (A-, B-, D-, H- and T-type cyclins), have been described. Here we report the isolation and characterization of a novel class of seven *Arabidopsis* cyclin genes, designated *PLPs*. All PLP cyclins share a highly conserved 100 amino acids central region ("Cyclin box") displaying a significant homology to the PHO80 cyclin from *Saccharomyces cerevisiae* and the related G_1 cyclins from *Trypanosome cruzi* and *T. brucei*. In agreement, *PLP4;2* was able to complement a PHO80 mutant yeast strain. PLP cyclins interact with CDKA;1 *in vivo* and *in vitro* as shown by yeast two-hybrid analysis and co-immunoprecipitation experiments. In addition, PLP proteins were demonstrate to co-localize with CDKA;1 in the nucleus of interphase cells, strongly suggesting the formation of a CDKA;1/PLP complex in planta. As *PLP* expression is restricted to proliferating tissues but also can be found in differentiating and mature tissues, we postulate that in analogy with other systems PLP cyclins are involved in the linkage between cell division, cell differentiation, and the nutritional status of the cell.

Old theories and New Functions: One hundred years studying the evolutionary consequences of gene and genome duplication

John S. Taylor[1,*] and Jeroen Raes[2]

[1] Department of Biology, University of Victoria, Victoria, BC, V8W 3N5, Canada
[2] Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[*] *Author for correspondence*

Manuscript in preparation. Intended for: Current Genomics

Abstract

Almost a century ago, when comparative genomics was the study of chromosome numbers, the evolutionary implications of genome duplication appear to have been nearly as hot a topic as today. Kuwada (1911) proposed that the production of innumerable races of *Zea mays* has a certain relation to the duplication of chromosomes and Tischler, (1915) observed a correlation between chromosome variation and external morphology in a diversity of plant species. By the 1930s the concept of genes had become established and smaller scale duplication events could be visualised in *Drosophila* polytene chromosomes. Stadler (1929) discovered that barley species in the genera *Avena* and *Triticum* with 21 chromosomes were less prone to harmful mutations than species with seven or 14 chromosomes. He concluded the the frequency of induced mutation in polyploids was low because of gene reduplication. Haldane (1932), Muller (1934), Bridges (1935) and Serebrovsky (1938) all considered the possibility that gene duplicates might be altered (evolve new functions) without disadvantage to the organism. Later, discussion about the possible connection between gene duplication and macroevolution emerged from debates between proponents and opponents of neo-darwinism. Metz (1947) argued that without duplication events we would have to assume that the 'primordial ameoba' was endowed with all the germinal components now present in its descendants, from protozoa to man. Insights into genome duplication in the late sixties drew upon data from isozyme electrophoresis, amino acid sequencing and DNA-RNA hybridization research. Ohno (1970) echoed the sentiment of Metz when he proposed that the creation of new gene loci with previously non-existent functions was a pre-requisite for the creation of metazoans, vertebrates and finally mammals from unicellular organisms. Genome sequencing projects have now shown that unicellular organisms do have fewer genes than vertebrates. However, the connection between speciation, organismal complexity, and gene content remains a contentious issue. Here we review the long history of gene and genome duplication research and the contribution of whole genome sequencing to the debate over the evolutionary importance of gene and genome duplication.

ForCon, an automatic tool for alignment format conversion

Jeroen Raes and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

* *Author for correspondence (e-mail yvdp@psb.ugent.be; fax: +32 9 331 3809)*

Published in: EMBnet.news 6(1)

Abstract

ForCon is a software tool for the conversion of nucleic acid and amino acid sequence alignments that runs on IBM-compatible computers under a Microsoft Windows environment. The program converts alignment formats used by all popular software packages for sequence alignment and phylogenetic tree inference. ForCon is available for free on request from the authors or can be downloaded via internet at URL http://www.psb.ugent.be/~jerae/ForCon/index.html. It is also included in the software package TREECON for Windows (see http://www.psb.ugent.be/bioinformatics/psb/treeconw/treeconw.zip).

List of publications

Raes, J. and Van de Peer, Y. ForCon, a tool to automatically convert sequence alignment formats. (1999) EMBnet.news 6(1)

Vandepoele, K., **Raes, J.**, De Veylder, L., Rouze, P., Rombauts, S. and Inze, D. (2002) Genome-wide analysis of core cell cycle genes in Arabidopsis. The Plant Cell 14, 903-916

Vandepoele, K., Saeys, Y., Simillion, C., **Raes, J.** and Van de Peer, Y. (2002) The Automatic Detection of Homologous Regions (ADHoRe) and its Application to Microcolinearity between Arabidopsis and Rice. Genome Research 12, 1792-1801

Breyne P., Dreesen R., Vandepoele K., De Veylder L., Van Breusegem F., Callewaert L., Rombauts S., **Raes J.**, Cannoot, B., Engler G., Inzé D. and Zabeau M. (2002) Functional analysis of the transcriptome during cell division in plants. Proceedings of the National Academy of Sciences USA 99, 14825-14830

Raes, J., Vandepoele, K., Simillon, C., Saeys, Y. and Van de Peer, Y. (2003) Investigating ancient duplication events in the Arabidopsis genome. Journal of Structural and Functional Genomics 3, 117-129

De Bodt, S., **Raes, J.**, Florquin, K., Rombauts, S., Rouze, P., Theissen, G. and Van de Peer, Y. (2003) Genome-wide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. Journal of Molecular Evolution 56, 573-586

Raes, J. and Van de Peer, Y. (2003) Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates *in silico*. Applied Bioinformatics, in press.

Raes, J., Rohde, A., Christensen, J.H., Van de Peer, Y. and Boerjan, W. Genome-wide characterization of the lignification toolbox in *Arabidopsis*. Plant Physiology, in press.

De Bodt, S., **Raes, J.**, Van de Peer, Y. and Theissen, G. MADS in the post-genomic era. Trends in Plant Sciences, in press.

Acosta Torres, J.A., de Almeira Engler, J., **Raes, J.**, Beemster, G.T.S., De Groodt, R., Inzé, D. and De Veylder, L. Molecular characterization of *Arabidopsis* PHO80 Like Proteins, a novel class of plant cyclins that interact with the CDKA;1 protein. In preparation.

Taylor, J.S. and **Raes, J.** Old theories and New Functions: One hundred years studying the evolutionary consequences of gene and genome duplication. In preparation.