Phytochem Rev (2009) 8:255–268
DOI 10.1007/s11101-008-9097-1

Methylthioalkylmalate synthases: genetics, ecology and evolution

Markus Benderoth · Marina Pfalz · Juergen Kroymann

Received: 30 January 2008 / Accepted: 10 April 2008 / Published online: 31 May 2008

Abstract Glucosinolates display an enormous amount of structural variation, both within and between species. This diversity is thought to have evolved in response to challenges imposed on plants by their biotic environment. During the past decade, glucosinolates and myrosinase-catalyzed glucosinolate hydrolysis have become excellent examples for understanding functional diversification in plant secondary metabolism and plant defence. Methylthioalkylmalate (MAM) synthase genes and enzymes are central to the diversification of aliphatic glucosinolate structures in *Arabidopsis thaliana* and related plants. This review summarizes efforts to elucidate how *MAM*-mediated diversity in aliphatic glucosinolate structures is generated and maintained. It also attempts to put variability in methionine carbon chain elongation during glucosinolate biosynthesis into an ecological and evolutionary context.

Keywords Complex traits · Evolutionary dynamics · Glucosinolate metabolism · Natural variation · Plant–insect interactions

The authors M. Benderoth and M. Pfalz have contributed equally to the manuscript.

M. Benderoth · M. Pfalz · J. Kroymann (✉)
AG Genetics and Evolution, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8, 07745 Jena, Germany
e-mail: kroymann@ice.mpg.de

Abbreviations

IPMS	Isopropylmalate synthase
MAM	Methylthioalkylmalate synthase
NIL	Near-isogenic line
QTL	Quantitative trait locus/loci
RIL	Recombinant inbred line

Introduction

Glucosinolates and myrosinases are almost exclusively found in plants from the order Capparales. This order consists of more than a dozen families, including the Brassicaceae and the Capparaceae (Rodman 1991a, b; Rodman et al. 1996). More than 130 different glucosinolates have been characterized; however glucosinolate composition varies remarkably between and within species (Daxenbichler et al. 1991; Fahey et al. 2001; Kliebenstein et al. 2001a; Windsor et al. 2005; Heidel et al. 2006). In *Arabidopsis thaliana*, nearly 40 different glucosinolates have been identified (Kliebenstein et al. 2001a; Reichelt et al. 2002). These glucosinolates are generated from methionine, tryptophan or phenylalanine, with methionine-derived (aliphatic) glucosinolates being the predominant glucosinolate class represented in *A. thaliana*.

Comparative analyses of glucosinolate profiles among Arabidopsis accessions reveal extensive qualitative and quantitative variation (Kliebenstein et al. 2001a; Pfalz et al. 2007). Quantitative trait locus

(QTL) mapping studies in various Arabidopsis recombinant inbred line (RIL) populations (Magrath et al. 1994; Mithen et al. 1995; Campos de Quiros et al. 2000; Kliebenstein et al. 2001b, c; Keurentjes et al. 2006; Wentzell et al. 2007) show consistently that two major genetic loci, methylthioalkylmalate (*MAM*) and *AOP* (Fig. 1), account for most of the variability in aliphatic glucosinolate structures. Enzymes encoded at *AOP* are responsible for the modification of the side chain structure (Kliebenstein et al. 2001b), while the *MAM* locus (also referred to as *ELONG*, *GS-ELONG* or *GSL-ELONG*) controls variability in aliphatic glucosinolate carbon chain length.

MAM encodes a small family of MAM synthases (Kroymann et al. 2001, 2003; Benderoth et al. 2006) in *A. thaliana* and related Brassicaceae. MAM synthases catalyze the condensation of ω-methylthio-2-oxoalkanoic acids derived from methionine with acetyl-CoA to form methylthioalkylmalic acids. Subsequent isomerization and oxidative decarboxylation reactions lead to the generation of ω-methylthio-2-oxoalkanoic acids with extended carbon chains (Fig. 2). Together, condensation, isomerization and oxidative decarboxylation are referred to as the methionine carbon chain elongation cycle, with an extension of one methylene group per cycle. MAM substrate specificity (Table 1) determines whether the reaction products of a given cycle enter the biosynthetic pathway generating the glucosinolate core structure or whether they undergo additional cycles of carbon chain elongation. *A. thaliana* MAM1 and MAM2 are involved in the formation of aliphatic glucosinolates with short carbon chains (Kroymann et al. 2001; Textor et al. 2004; Benderoth et al. 2006), while MAM3 catalyzes condensation reactions in the biosynthesis of aliphatic glucosinolates with long carbon chains (Field et al. 2004; Textor et al. 2007; Knoke et al. 2008). MAM synthases determine variability during the earliest stage of aliphatic glucosinolate biosynthesis and play a central role in glucosinolate diversity. Hence, evolutionary forces acting on *MAM* have a fundamental impact on aliphatic glucosinolate composition.

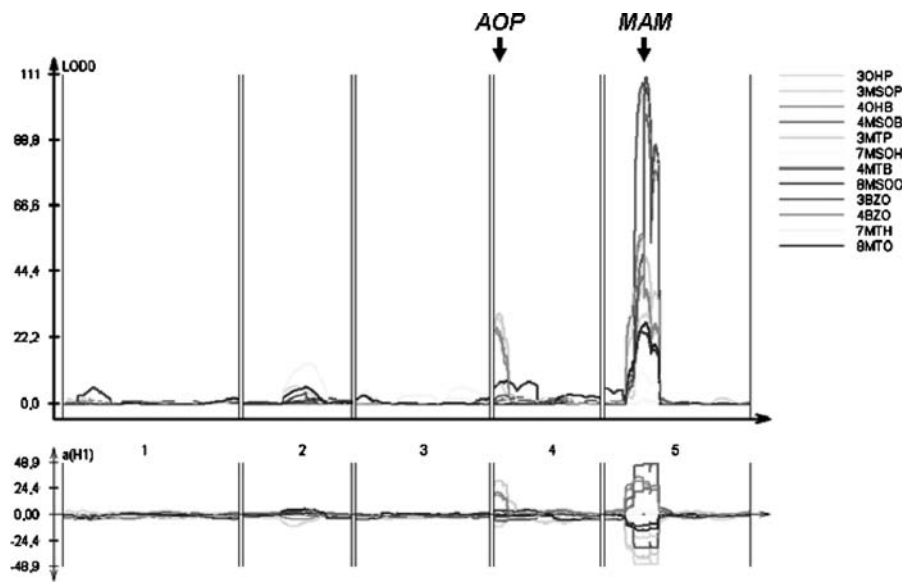

Fig. 1 QTL for aliphatic glucosinolates in seeds from Col-0 × L*er* RILs. Top: The *MAM* locus on chromosome five influences the quantity of all aliphatic glucosinolates. 3-hydroxypropyl (3OHP), 3-methylsulfinylpropyl (3MSOP), 3-methylthiopropyl (3MTP) and 3-benzoyloxypropyl (3BZO) originate from homomethionine (3C), 4-hydroxybutyl (4OHB), 4-methylsulfinylbutyl (4MSOB), 4-methylthiobutyl (4MTB) and 4-benzoyloxybutyl (4BZO) from dihomomethionine (4C), 7-methylsulfinylheptyl (7MSOH) and 7-methylthioheptyl (7MTH) from pentahomomethionine (7C) and 8-methylsulfinylocytyl (8MSOO) and 8-methylthiooctyl (8MTO) from hexahomomethionine (8C). Bottom: Additive genetic effects for homo- and pentahomomethionine-derived glucosinolates are opposite to those for dihomo- and hexahomomethionine-derived glucosinolates. QTL mapping was conducted with Windows QTL cartographer 2.5 (Wang et al. 2001–2004) and is based on log-transformed HPLC data from 297 Col-0 × L*er* RILs (John D'Auria and Juergen Kroymann, unpublished data). For all traits, the significance threshold was ~2.5 LOD units, based on 1,000 permutations of the data

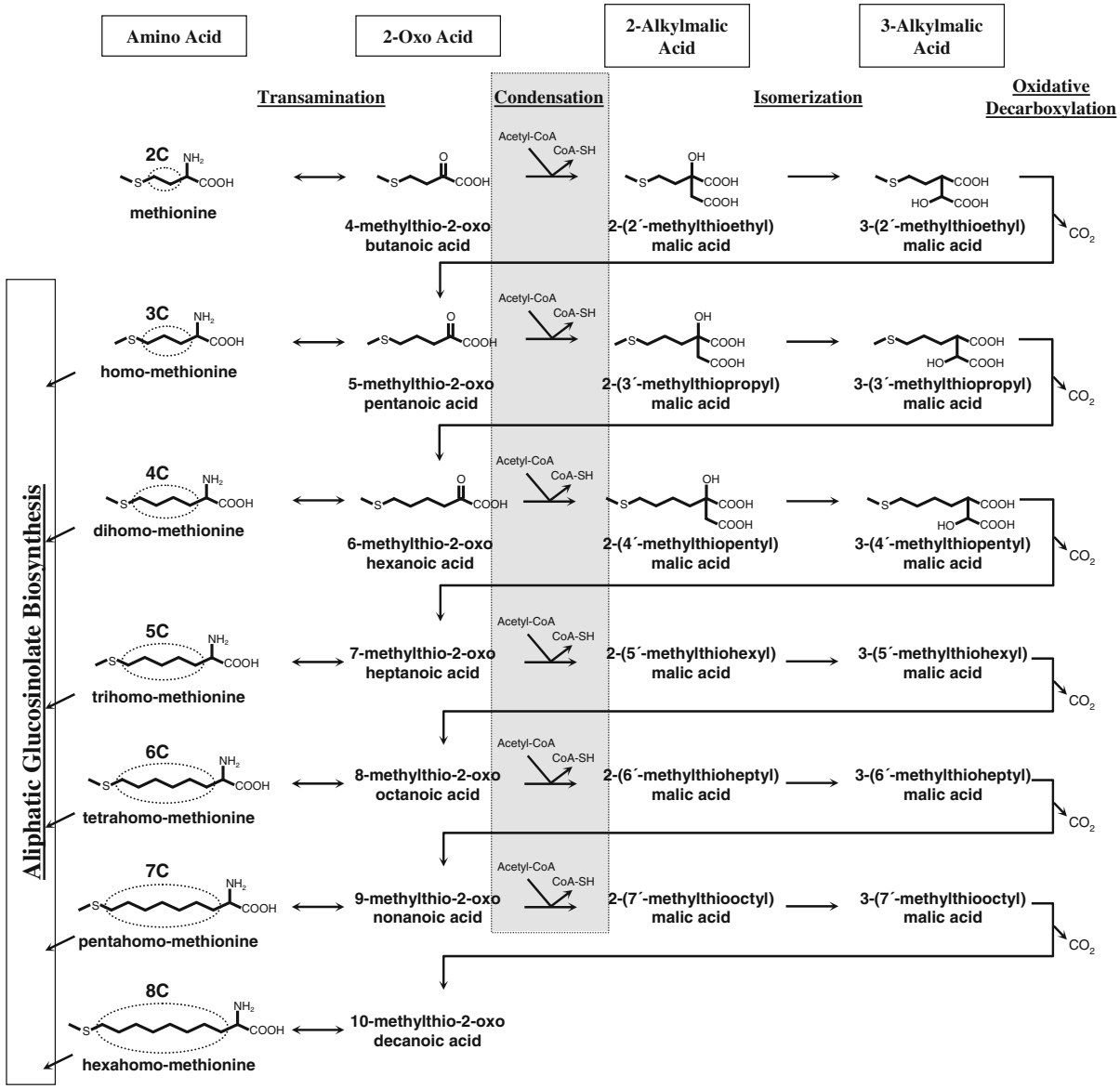

Fig. 2 The methionine carbon chain elongation cycle. MAM synthases encoded at the *MAM* locus catalyze the committed step in methionine carbon chain elongation. Two subsequent reactions, an isomerization and an oxidative decarboxylation complete a cycle. MAM substrate specificity (Table 1) determines whether ω-methylthio-2-oxoalkanoic acids remain in the carbon chain elongation cycle or enter the biosynthetic pathway that generates the glucosinolate core structure. Here, all reactions occurring in *A. thaliana* are depicted; in other Brassicaceae fewer or more cycles are possible

Cloning of the *MAM* QTL

The widely used *A. thaliana* accession Landsberg *erecta* (L*er*) accumulates homomethionine-derived (3C) glucosinolates. In contrast, aliphatic glucosinolates in Columbia (Col-0), recognized for having provided the blueprint of the Arabidopsis genome (Arabidopsis Genome Initiative 2000), originate mainly from dihomomethionine (4C). This biochemical difference between Col-0 and L*er* enabled initial mapping of the responsible genetic locus to a region of approximately 140 kb on chromosome 5 (Magrath et al. 1994; Campos de Quiros et al. 2000). Within this region, two tandemly arranged genes were identified as

Table 1 MAM substrate specificities in glucosinolate biosynthesis

Assays were conducted with heterologously expressed *MAM* genes and generic and/or artificial substrates

MAM synthase	Substrates	Elongation cycle(s)	References
A. thaliana MAM1	2C, 3C, (4C)	1, 2, (3)	Kroymann et al. (2001), Textor et al. (2004), Benderoth et al. (2006)
A. thaliana MAM2	2C	1	Benderoth et al. (2006)
A. thaliana MAM3	2C–7C	1–6	Textor et al. (2007), Knoke et al. (2008)
A. lyrata MAMa	2C	1	Benderoth et al. (2006)
B. stricta MAMa			

candidates, based on their sequence similarity with isopropylmalate synthase (IPMS) genes. IPMS catalyzes the condensation of 2-oxoisovalerate with acetyl-CoA to form isopropylmalate in leucine biosynthesis, a reaction similar to the MAM-catalyzed condensation of ω-methylthio-2-oxoalkanoic acids with acetyl-CoA in glucosinolate biosynthesis. The two candidate genes were termed *MAM1* and *MAM-L* (meanwhile often referred to as *MAM3*), respectively (Kroymann et al. 2001).

In the Col-0 accession, *MAM1* and *MAM3* are separated by ca. 11.5 kb of intervening sequence. High-resolution mapping was employed to separate functional effects of these candidate genes. Col-0 was crossed with CL5, an RIL from the Col-0 × L*er* population (Lister and Dean 1993). This particular line was chosen because it shared approximately 70% of its genome with the Col-0 accession but has the L*er* allele at the *MAM* locus, thus enabling fine-mapping with near-isogenic lines (NILs). In Col-0 × CL5 F_2 progeny one line was identified that had recombined between *MAM1* and *MAM3*. This recombinant line had the Col-0 *MAM1* genotype and was heterozygous at *MAM3*. Its glucosinolate phenotype resembled the parental Col-0 profile closely, with 4C glucosinolates predominating. Likewise, progeny from this line produced mainly 4C glucosinolates when they had the Col-0 *MAM1* and the L*er MAM3* genotype. Hence, the biochemical difference in short-chain aliphatic glucosinolate composition was attributable to *MAM1* and not *MAM3* (Kroymann et al. 2001). Further evidence for the role of MAM1 in methionine carbon chain elongation was obtained with *MAM1* mutants, deficient in dihomomethionine-derived glucosinolates (Haughn et al. 1991), and with biochemical assays of heterologously expressed *MAM1*, which showed that the encoded protein has the capacity of condensing ω-methylthio-2-oxoalkanoic acids with acetyl-CoA (Kroymann et al. 2001; Textor et al. 2004; Benderoth et al. 2006).

Origin of *MAM* genes

Several groups of enzymes catalyze condensation reactions between 2-oxo acids and acetyl-CoA. These enzymes belong to enzyme class EC 2.3.3.-. Examples are citrate synthase (EC 2.3.3.1), which condenses oxaloacetate with acetyl-CoA in the TCA cycle, malate synthase (EC 2.3.3.9), which is responsible for the condensation of glyoxylate with acetyl-CoA in the glyoxylate cycle, and 2-IPMS (EC 2.3.3.13) involved in leucine biosynthesis.

Among the genes that encode enzymes of class EC 2.3.3.-, *MAM* genes share the highest degree of sequence similarity with 2-*IPMS* genes. Phylogenetic analyses reveal a close relationship between *MAM* and *IPMS* genes, with *MAM*s having evolved from *IPMS* (Fig. 3). Similar gene structures of *MAM* and *IPMS* corroborate this close relationship; most intron positions are identical in *MAM* and *IPMS* (Kroymann et al. 2001). However, the first *IPMS* exon corresponds to the first two exons in *MAM* genes, and the last four *IPMS* exons are fused to a single exon in *MAM*s. Exon fusion was accompanied with a net loss of ca. 250 bp in the 3′ portion of *MAM* genes, and may have functional significance since this region encodes a conserved allosteric leucine binding site in IPMS (de Kraker et al. 2007). As a consequence, Arabidopsis *IPMS* genes consist of 12 exons and *MAM* genes of 10 exons, with *IPMS* exons 2–8 and *MAM* exons 3–9 corresponding to one another (Fig. 4). Furthermore, all *MAM* genes encode an N-terminal plastid targeting signal (Kroymann et al. 2001), suggesting that methionine carbon chain elongation occurs—like leucine biosynthesis (Hagelstein and Schultz 1993)—in the plastids, a

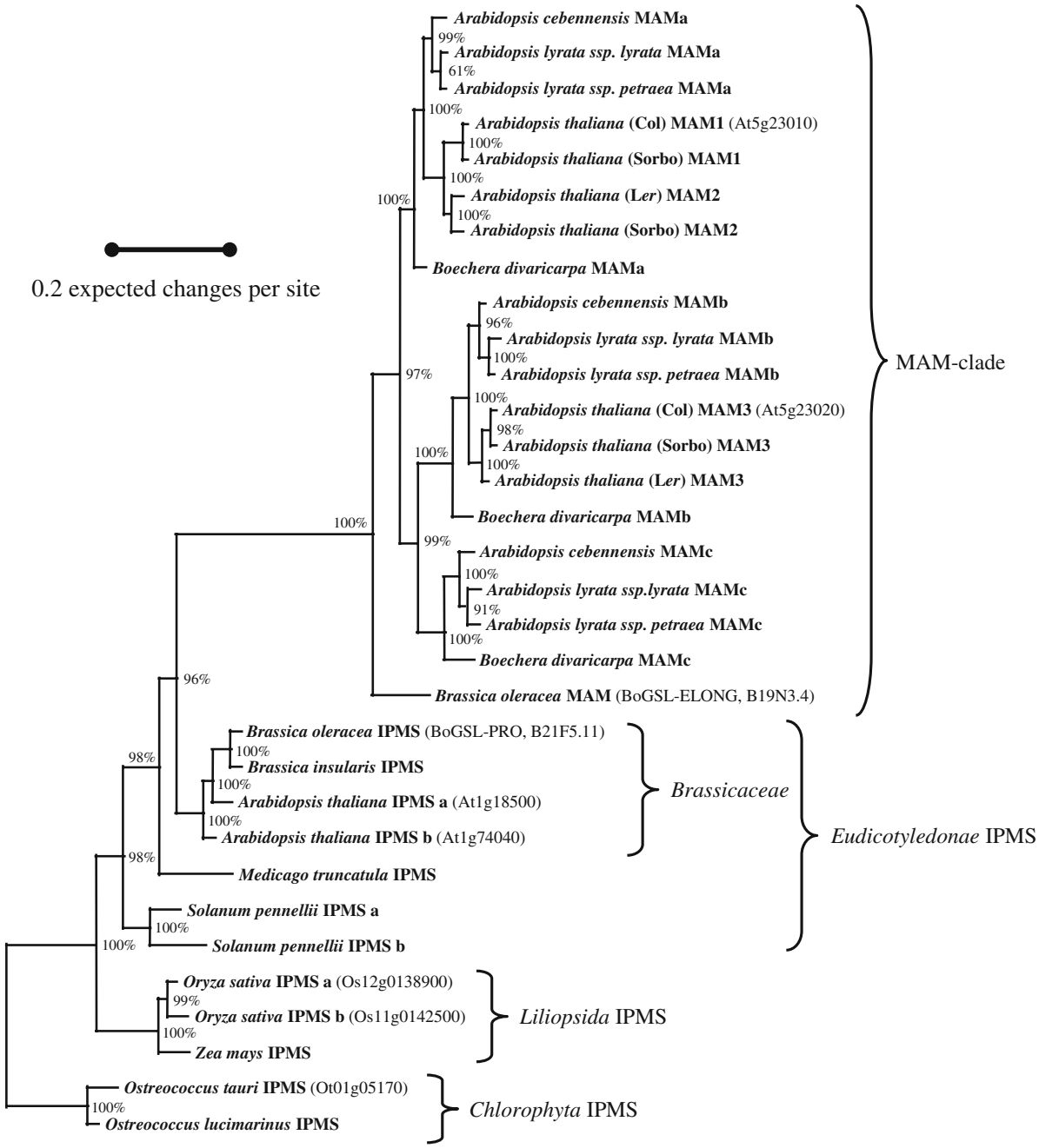

Fig. 3 Bayesian phylogenetic analysis of MAM and IPMS sequences. Amino acid sequences were aligned with ClustalW. Target peptides and alignment gaps were excluded. The tree was constructed with MrBayes v 3.1.2 (Huelsenbeck and Ronquist 2001) using a mixed amino acid substitution model approach. Data structure was best explained by the wag-model (Whelan and Goldman 2001) with a posterior probability of 1.000. The Markov-Chain-Monte-Carlo simulation was performed in three parallel runs with four chains each for 1,000,000 iterations with a sample frequency of 200 and a burnin fraction of 0.05. Convergence diagnostic over all three runs (PSRF) was 1.000. Shown are posterior probabilities for all internal nodes. Note that *Medicago truncatula IPMS* is incorrectly annotated in the databases; here, a corrected annotation was assembled from clone mth2-19b5 for alignment and tree construction

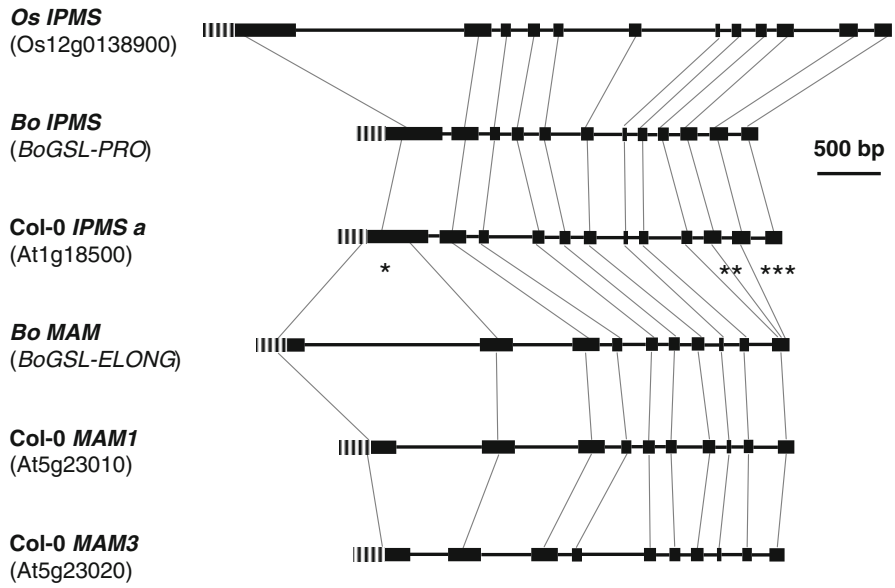

Fig. 4 Exon-intron structure of *IPMS* and *MAM* genes. *IPMS* genes from higher plants consist of 12, *MAM* genes of ten exons. Differences between Arabidopsis *IPMS* and *MAM* gene structures involve the split of *IPMS* exon I (*), the fusion and partial loss of *IPMS* exons IX, X and XI (**), and the loss of *IPMS* exon XII (***). Plastid targeting sequences are represented by dashed lines. Os, *Oryza sativa*; Bo, *B. oleracea*; Col, *A. thaliana* accession Col-0

prediction corroborated by the demonstration of MAM activity in chloroplast-enriched extracts of *Eruca sativa* leaves (Falk et al. 2004) and by recent immunolocalization experiments with MAM3 antibodies (Textor et al. 2007).

Database searches reveal that *IPMS* genes are present in eubacteria and archaebacteria, in green algae, higher plants, and fungi, but are absent in animals. Among the prokaryotes, *IPMS* genes from cyanobacteria appear to be most closely related to those from higher plants and green algae (Fig. 5), suggesting that plant *IPMS* genes originate from the cyanobacterial endosymbionts that gave rise to plastids, with a transfer of the cyanobacterial/plastid copy to the plant nucleus (Martin and Herrmann 1998).

Variability in the genetic composition of the *MAM* locus in *A. thaliana*

To elucidate the genomic basis of carbon chain length control, *MAM* genes and flanking DNA were sequenced from 25 Arabidopsis accessions randomly selected from the species' natural distribution (Kroymann et al. 2003). This comparative analysis revealed extensive variation in the molecular composition of the *MAM* locus among Arabidopsis accessions (Fig. 6). In addition to *MAM1* and *MAM3*, which are present in Col-0, other accessions harbored a third gene, referred to as *MAM2* (Kroymann et al. 2003). Sequence comparison enabled the reconstruction of the archetypical configuration of the *MAM* locus in *A. thaliana*, consisting of tandemly arranged *MAM2*, *MAM1* and *MAM3* genes, all transcribed in the same direction. However, secondary gene deletions have occurred frequently in the history of the *A. thaliana MAM* cluster. In some accessions (including Col-0) the *MAM2* gene was deleted. In other accessions, the *MAM1* gene was deleted, either partially, involving promoter and 5′-portion of the coding sequence (e.g. in L*er*), or completely. Finally, the Lm-2 accession contains a chimerical gene with 5′ *MAM2* fused to 3′ *MAM1* sequence, possibly caused by a deletion of the intervening region. Nonetheless, in contemporary accessions at least one of these genes has been retained. Likewise, all tested natural accessions contain a functional *MAM3* gene.

Sequence exchange between *MAM1* and *MAM2* genes is another factor contributing to variability in the gene composition of the *MAM* locus. In some accessions, sequence transfer has occurred from *MAM2* to *MAM1*, in others sequence portions have been shifted from *MAM1* to *MAM2*. In the most extreme cases, represented by two accessions from Tajikistan (Hodja and Condara), the gene at the position originally occupied by *MAM1* has been almost completely converted to a *MAM2*-like gene. However, another accession from Tajikistan, Sorbo, was not affected by gene deletion or conversion events. This accession has functional *MAM2* and

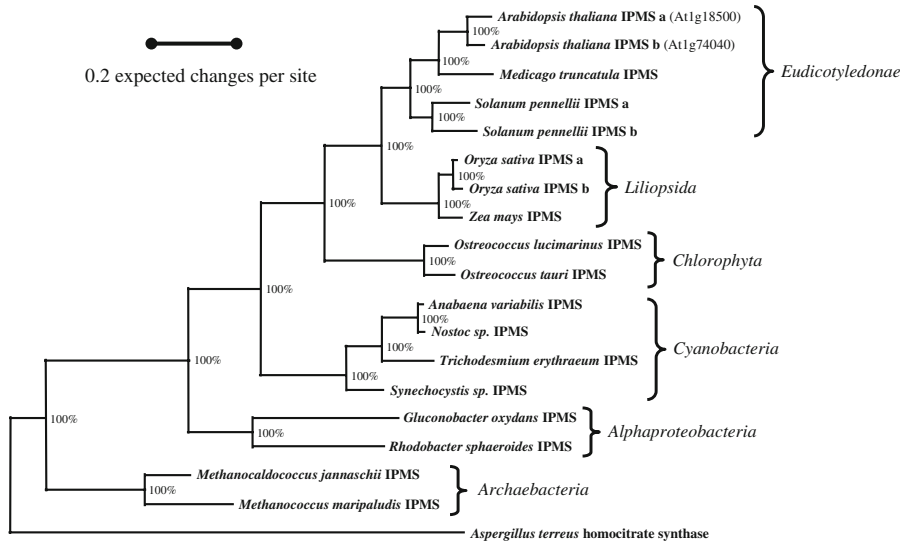

Fig. 5 Evolutionary origin of plant IPMS sequences inferred with Bayesian phylogenetic analysis. Calculations were based on ClustalW-aligned amino acid sequences. Target peptides and alignment gaps were excluded. The analysis was conducted as explained in the legend to Fig. 3. Again, data structure was best explained by the wag-model (Whelan and Goldman 2001) with a posterior probability of 0.999. Convergence diagnostic over all three runs (PSRF) was 1.000. Posterior probabilities for all internal nodes are shown

MAM1 genes and accumulates 4C glucosinolates, indicating that MAM1 overrides MAM2 function. Likewise, all accessions with at least one functional *MAM1*-like gene accumulate short chain aliphatic glucosinolates generated from dihomomethionine, while accessions without a functional *MAM1* produce short-chain aliphatic glucosinolates almost exclusively from homomethionine. Taken together, these data show that the *MAM* cluster in *A. thaliana* is subject to dynamic evolutionary change.

Quantitative effects of the *MAM1/MAM2* polymorphism on glucosinolate profiles

The consequences of the *MAM1/MAM2* polymorphism on Arabidopsis glucosinolate profiles were investigated in a set of NILs, which were developed from Col-0 × CL5 progeny with recombination breakpoints in the vicinity of the *MAM* QTL (Kroymann et al. 2001, 2003). All individual aliphatic glucosinolates were found to be influenced by the *MAM1/MAM2* polymorphism (Fig. 7). Col-0 *MAM1* genotypes produced more aliphatic glucosinolates derived from dihomo- (4C), trihomo- (5C), tetrahomo- (6C) and pentahomomethionine (7C), while L*er MAM2* genotypes accumulated larger quantities of homo- (3C) and hexahomomethionine (8C) derivatives. Moreover, the *MAM1/MAM2* polymorphism also caused differences in the concentration of total aliphatic glucosinolates, with L*er MAM2* genotypes accumulating 60% more aliphatic glucosinolates in leaves and 20% more in seeds than Col-0 *MAM1* genotypes (Kroymann et al. 2003; Fig. 8).

Even though MAM1 and MAM2 lack the capacity to catalyze the condensation reactions in advanced methionine carbon chain elongation cycles (Benderoth et al. 2006), the QTL peaked in all cases at an interval containing *MAM1* or *MAM2*, but not *MAM3*. Hence, the *MAM1/MAM2* polymorphism influenced not only aliphatic glucosinolates with short carbon chains but also affected glucosinolates with long carbon chains. The biochemical basis for this effect of *MAM1* versus *MAM2* on long-chain aliphatic glucosinolate accumulation is not yet understood.

A complex influence of the *MAM* locus on glucosinolate profiles was also seen in studies with another Arabidopsis RIL population, Bay-0 × Sha (Kliebenstein et al. 2006; Wentzell et al. 2007). eQTL mapping and network analysis (further explained in Kliebenstein 2008) suggest that this influence is, to a certain degree, caused by changes in transcript levels.

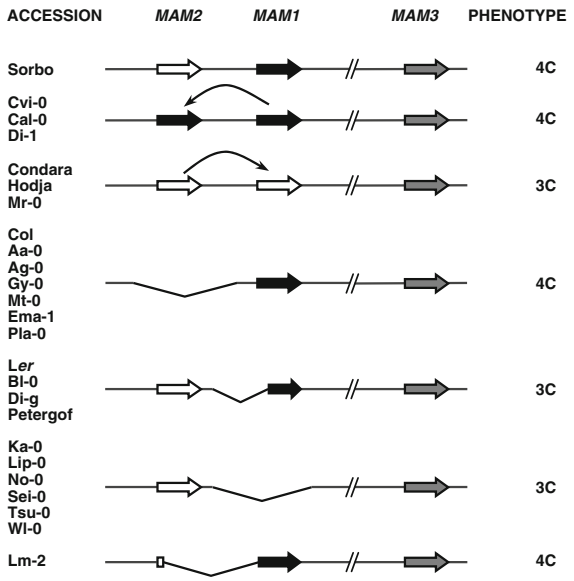

Fig. 6 Schematic representation of the *MAM* cluster in Arabidopsis accessions. The organization of the *MAM* cluster is highly variable. Partial or complete deletions of *MAM2* or *MAM1* have occurred frequently. In Lm-2, 5′-*MAM2* sequence is fused to 3′-*MAM1* sequence. Furthermore, sequence information has been transferred between loci. In Cvi-0, Cal-0 and Di-1, the gene at the *MAM2* position has been partially converted into a *MAM1*-like sequence; in Condara, Hodja and Mr-0, *MAM1* has been converted into a *MAM2*-like gene. All tested accessions have a *MAM3* gene. The predominant aliphatic glucosinolate class is indicated in the right column (3C: homomethionine-derived, 4C: dihomomethionine-derived)

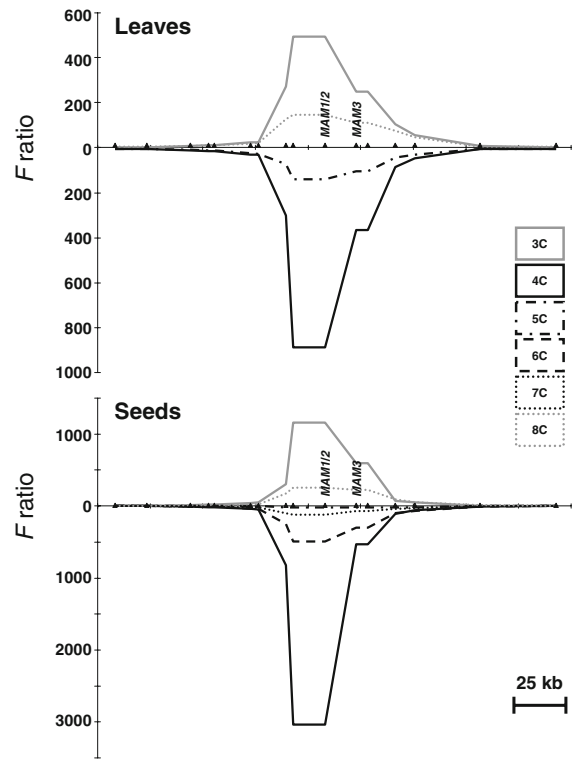

Fig. 7 Single-marker QTL analysis of aliphatic glucosinolates with different carbon chain lengths in a 205-kb region of the Arabidopsis genome. Fifty-eight NILs with recombination breakpoints in the *MAM* region were assayed for trait values in leaves (top) and seeds (bottom). QTL for homo- (3C), dihomo- (4C), trihomo- (5C), tetrahomo- (6C), pentahomo- (7C) and hexahomomethionine (8C)-derived glucosinolates are centred at a 15-kb non-recombinant region that contains Col-0 *MAM1* or L*er MAM2*, but not *MAM3*. Col-0 *MAM1* genotypes produce more 4C–7C derivatives than L*er MAM2* genotypes. By contrast, L*er MAM2* genotypes accumulate higher quantities of 3C and 8C derivatives. Small black triangles show marker positions based on the Col-0 sequence; the position of *MAM1*/*MAM2* and *MAM3* is indicated. *F* ratios quantify statistical significance of a QTL at a given marker position. Note that leaves contain lower concentrations of glucosinolates than seeds; therefore, low-abundant glucosinolates are difficult to detect by HPLC in leaves

Ecological consequences of the *MAM1*/*MAM2* polymorphism

Two lepidopteran insects, *Spodoptera exigua* and *Plutella xylostella*, were tested for differences in larval performance contingent on the genotype at the *MAM* locus in Col-0 × CL5 NILs. *S. exigua* is referred to as a generalist due to possession of a broad host range; it has the ability to feed on a variety of plants from different families. In contrast, *P. xylostella* has a narrow host range utilizing almost exclusively plants from the crucifer family. *S. exigua* was found to respond to the *MAM1*/*MAM2* polymorphism. Its performance was lower on the *MAM2* genotype, and larvae caused approximately 17% less damage than on the *MAM1* genotype (Kroymann et al. 2003). Again, single marker analysis showed that the resistance QTL peaked in the interval containing *MAM1*/*MAM2* but not *MAM3* (Fig. 8). For *P. xylostella*, no significant difference was found with ANOVA. Nonetheless, larvae performed on average better on *MAM2* genotypes (Kroymann et al. 2003). Furthermore, re-analysis of this data set with regression showed that *Plutella* larval herbivory was positively correlated with leaf aliphatic glucosinolate content (Kliebenstein et al. 2005). Hence, the *MAM1*/*MAM2* polymorphism had contrasting effects on *S. exigua* versus *P. xylostella*. *MAM2* genotypes were

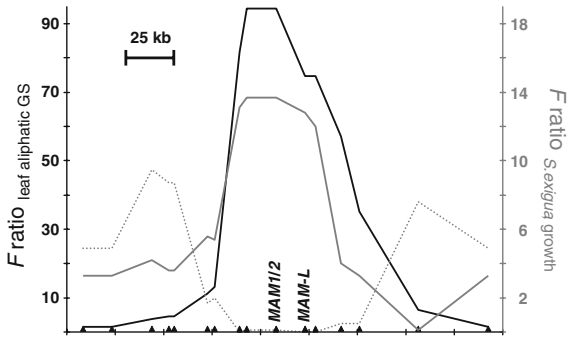

Fig. 8 Single-marker QTL analysis of leaf total aliphatic glucosinolates (black line), resistance against *S. exigua* (solid grey line) and growth rate (dotted grey line). QTL for leaf aliphatic glucosinolates and for resistance against *S. exigua* centre at a 15-kb non-recombinant interval containing L*er MAM2* or Col-0 *MAM1*. NILs with a L*er MAM2* accumulate more aliphatic glucosinolates and are more resistant against *Spodoptera* than the *MAM1* genotype. Two QTL for biomass accumulation were detected upstream and downstream of the *MAM* cluster, but there is no indication for a growth rate QTL at the *MAM* genes (Kroymann et al. 2003; Kroymann and Mitchell-Olds 2005)

better defended against the generalist insect herbivore, whereas *MAM1* genotypes appeared to suffer less damage from the specialist. However, although the *MAM1*/*MAM2* polymorphism controls these differences in insect performance, it is not yet clear which biochemical parameter ultimately accounts for the observed effects, due to the complex influence of the *MAM* locus on aliphatic glucosinolate profiles. Differences in resistance could be attributable to leaf total aliphatic glucosinolate concentration, levels of individual glucosinolates, differences in glucosinolate composition, or any combination of these factors.

Since *MAM2* genotypes produced approximately 60% more leaf and 20% more seed glucosinolates than *MAM1* genotypes, this raised the question whether increased glucosinolate production in the *MAM2* genotype was paid for with a reduction in plant growth rate. Such allocation costs can occur when defenses are energetically expensive, so that genotypes with strong defenses have fewer resources to invest in growth and reproduction (Purrington 2000; Tian et al. 2003). Therefore, dry weight of plant rosettes was measured at the pre-reproductive state. Quantitative analyses indicated the presence of growth rate QTL upstream and downstream of the *MAM* locus, but there was no trace of a significant genotype effect on biomass accumulation at the *MAM* locus itself (Fig. 8, Kroymann et al. 2003; Kroymann and Mitchell-Olds 2005). Thus, allocation costs appear to be of minor importance for the *MAM1*/*MAM2* polymorphism in *A. thaliana*.

Generation of glucosinolate diversity

Sequencing of the MAM locus from the close *A. thaliana* relatives *A. petraea* ssp. *lyrata* (=*A. lyrata*), *Arabidopsis petraea* ssp. *petraea* (= *A. petraea*) and *Arabidopsis cebennensis*, and the more distant relative *Boechera divaricarpa* revealed that these species all possessed three MAM genes in the MAM cluster. These genes were termed MAMa, MAMb and MAMc (Fig. 9, Benderoth et al. 2006). Phylogenetic analyses showed that *A. thaliana* MAM1 and MAM2 originated from the duplication of a MAMa gene after *A. thaliana* diverged from its congeners (Fig. 3). *MAMb* genes were found to be orthologous to *A. thaliana* MAM3. A MAMc does not exist in the Arabidopsis genome, but is present in other Arabidopsis species and in *Boechera*. Hence, this gene was lost in *A. thaliana* after divergence from its congeners.

Analyses of nucleotide substitution patterns revealed that *MAM1* had accumulated an excess of non-synonymous nucleotide substitutions (i.e., substitutions that alter codon meaning) after the *MAMa* duplication, indicating positive (Darwinian) selection. The biochemical characteristics of MAM1, MAM2 and MAMa were investigated to infer which properties of *MAM1* were targeted by positive selection. *MAMa* from *A. lyrata* and from *Boechera stricta* (a close relative of *B. divaricarpa*), *MAM2* from the Arabidopsis accession L*er*, and *MAM1* from the accession Sorbo were heterologously expressed in *Escherichia coli*. Enzyme assays were carried out to investigate substrate specificity, ATP-, metal ion- and pH dependence (Benderoth et al. 2006). Major differences were found only for the enzymes' substrate specificities (Table 1). *A. petraea* and *B. stricta MAMa*, and L*er MAM2* all utilized 4-methylthio-2-oxobutanoic acid (2C) for condensation with acetyl-CoA, but none accepted ω-methylthio-2-oxoalkanoic acids with more than two methylene groups as a substrate. By contrast, Sorbo MAM1 accepted 4-methylthio-2-oxobutanoic acid (2C), 5-methylthio-2-oxopentanoic acid (3C) and, with low activity, also 6-methylthio-2-oxohexanoic

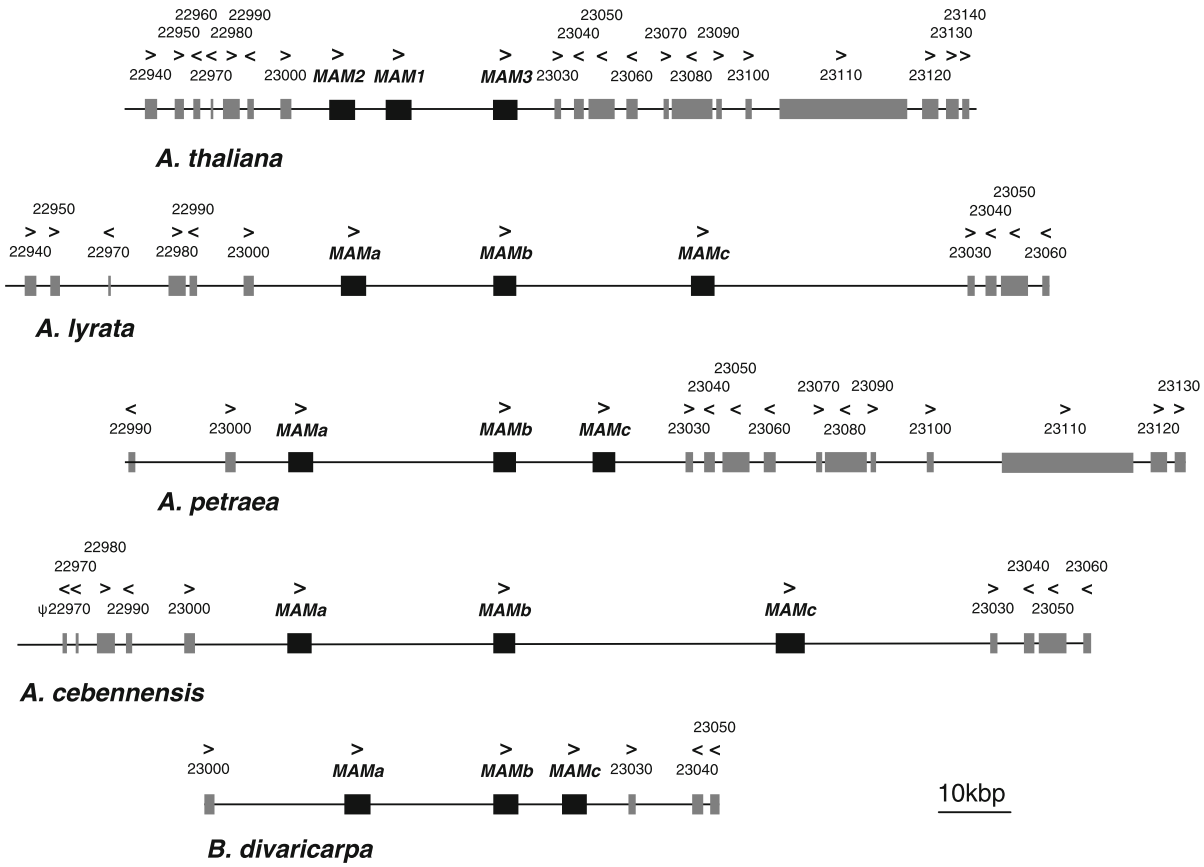

Fig. 9 Organization of the *MAM* region in *A. thaliana* relatives. The *MAM* cluster contains three genes (*MAMa*, *MAMb* and *MAMc*) in *A. lyrata*, *A. petraea*, *A. cebennensis* and *B. divaricarpa*. The regions flanking *MAM* are largely co-linear. For comparison, the archetypical configuration for *A. thaliana* is included. Numbers refer to AGI annotations (Arabidopsis Genome Initiative 2000), and the direction of transcription is indicated

acid (4C) as a substrate. Thus, MAMa and MAM2 function only in the first cycle of carbon chain extension, whereas MAM1 has acquired additional capacity to carry out condensation reactions in subsequent chain elongation cycles, equivalent to a biochemical neofunctionalization. MAM1 and MAM2 substrate specificities are matched by *A. thaliana* glucosinolate phenotypes. Accessions with a functional *MAM1* gene accumulate 4C glucosinolates, whereas accessions without a functional *MAM1* (but with a functional *MAM2*) generate mainly 3C aliphatic glucosinolates (Kroymann et al. 2003). Thus, gene duplication, biochemical neofunctionalization and positive selection account for the generation of diversity in the carbon chain lengths of aliphatic glucosinolates in *A. thaliana*.

The same processes are duplicated along the basal branches of the *MAM* gene tree (Fig. 3). *MAMa*, *MAMb* and *MAMc* have originated by gene duplication events, and are functionally diversified. *MAMa* controls short-chain aliphatic glucosinolates (Benderoth et al. 2006) and *MAMb* is, like its *A. thaliana* ortholog *MAM3* (Textor et al. 2007; Knoke et al. 2008), presumably involved in the biosynthesis of long-chain aliphatic glucosinolates. However, the function of *MAMc* is not yet known. Analyses of nucleotide substitution patterns indicate positive selection along the deep branches of the *MAM* tree, connecting *MAMa*, *MAMb* and *MAMc* (Benderoth et al. 2006). Hence, the same events that explain diversity in short-chain glucosinolates, i.e., gene duplication and biochemical neofunctionalization, driven by positive selection, appear to account for the diversification in aliphatic glucosinolate carbon chain lengths in general.

Dihomomethionine (4C) differs by only one methylene group from homomethionine (3C). At first

glance, this seems to be a minor difference. But the consequences of this small chemical difference can be profound when taken within the context of the glucosinolate-myrosinase system. After methionine carbon chain elongation and glucosinolate core structure generation, enzymes encoded at other genetic loci act to modify the carbon chain (Kliebenstein et al. 2001a, b). These loci are, like *MAM*, also polymorphic in *A. thaliana*, i.e., they harbour alleles whose gene products have diverse biochemical activities and cause different modifications of the aliphatic carbon chain. Likewise, modifying proteins like ESP (Lambrix et al. 2001) and ESM1 (Zhang et al. 2006) can alter glucosinolate breakdown identity during myrosinase-catalyzed glucosinolate hydrolysis. Taken together, different combinations of alleles at glucosinolate biosynthesis and hydrolysis loci generate quite different blends of glucosinolate hydrolysis products (Kliebenstein et al. 2005), and these may, in turn, cause different responses in attacking insect herbivores and other enemies.

Maintenance of glucosinolate diversity

While gene duplication, neofunctionalization and positive selection contribute to the generation of metabolic diversity, these processes do not satisfyingly explain how and why this diversity is maintained. Positive selection on *MAM1* indicates that this gene has provided a fitness advantage to its carriers after duplication of an ancestral *MAMa* gene. MAM1 overrides MAM2 function, and MAM1 has retained the capacity to function in the first methionine carbon chain elongation cycle (Kroymann et al. 2003; Benderoth et al. 2006). Thus, the MAM2 is not required to sustain MAM1 activity. Why then is *MAM2* still present in *A. thaliana*?

It could be possible that *MAM2* is in the process of becoming a pseudo-gene, but that the period after MAM1 neofunctionalization was too short for a complete degeneration of *MAM2*. For two reasons, this is not likely. First, estimates based on the analysis of nucleotide substitution rates in the Brassicaceae (Yang et al. 1999; Koch et al. 2001) suggest that the *MAMa* duplication occurred more than 10^5 generations ago. Second, a degenerating gene is expected to accumulate nucleotide substitutions since selection no longer acts to eliminate deleterious mutations. Mutations in open reading frames are called synonymous when the amino acid sequence remains unchanged and non-synonymous when the codon usage is altered. Because of the nature of the genetic code, a coding sequence has more non-synonymous than synonymous positions. Therefore, in a degenerating gene non-synonymous changes are more likely to occur than synonymous substitutions. But after correction for the number of non-synonymous and synonymous positions, the ratio between non-synonymous and synonymous changes in a degenerating gene is expected to be close to 1. However, when *MAM2* was compared to *A. thaliana MAM1* or *A. lyrata MAMa*, *MAM2* exhibited an excess of synonymous relative to non-synonymous changes, indicating purifying selection (Benderoth et al. 2006). In conclusion, MAM2 function appears to be preserved in *A. thaliana*.

What else could explain why *MAM2* was retained in *A. thaliana*? Insect herbivory assays had shown that NILs with the L*er MAM2* performed ca. 17% better against *S. exigua* larvae than lines with the Col-0 *MAM1* (Kroymann et al. 2003). Thus, *MAM2* can provide a selective advantage over *MAM1* under certain conditions. In addition, secondary gene deletion events and exchange of sequence information between paralogous *MAM1* and *MAM2* loci have occurred frequently in the history of *A. thaliana* (Fig. 6). Deletion of *MAM1* or conversion of *MAM1* into a *MAM2*-like sequence both result in a switch from a 4C glucosinolate profile to a profile dominated by 3C glucosinolates. If such a switch occurred in a local population of plants that were otherwise genetically uniform, this could confer a—temporal—selective advantage to the novel genotypes, provided that the local herbivore community was accustomed to a particular blend of glucosinolates. Since *A. thaliana* propagates mostly by selfing and local populations were founded frequently in the history of this species, such a scenario is not unlikely, and might account for the large proportion of derived genotypes at the *MAM* locus.

Furthermore, the composition of local herbivore communities varies temporally, with different classes of insects—specialists, generalists, and non-feeders (i.e., herbivorous insects not utilizing glucosinolate-containing plants as a host)—occurring in different frequencies over time. This may lead to fluctuating selection on glucosinolate profiles, with periods during which phenotypes with a particular glucosinolate

composition or with high glucosinolate levels increase in frequency, alternating with periods during which other types of glucosinolate profiles or low levels are advantageous. As a consequence, one would expect some form of equilibrium between different glucosinolate phenotypes. Indeed, among 51 Arabidopsis accessions whose glucosinolate profiles have been analyzed (Kliebenstein et al. 2001a; Kroymann et al. 2003; Pfalz et al. 2007), 29 produced aliphatic glucosinolates predominantly from homomethionine (3C) and 22 from dihomomethionine (4C). These data are not significantly different from the hypothesis that both phenotypes have equal frequencies in *A. thaliana* ($N = 51$; $df = 1$, $\chi^2 = 1.27$, n.s.). Also, two independent statistical tests of molecular population genetics found evidence for balancing selection acting on the *MAM2* gene (Kroymann et al. 2003). Balancing selection refers to evolutionary scenarios that maintain more genetic variation in a population than expected under neutrality (Nordborg and Innan 2002). First, a positive Tajima's *D* indicated significantly more intermediate frequency nucleotide polymorphisms segregating at *MAM2* than expected (Tajima 1989). Second, a McDonald and Kreitman test (McDonald and Kreitman 1991) showed that too many amino acids segregated in *A. thaliana* MAM2, when compared to *MAMa* from *A. lyrata*. However, this *MAM2* polymorphism has no impact on glucosinolate identity (Kroymann et al. 2003; Benderoth et al. 2006), suggesting that non-neutrality at *MAM2* is caused by selection on glucosinolate quantity and not quality, a hypothesis that remains to be tested.

While gene duplication, biochemical neofunctionalization and positive selection account for the generation of metabolic diversity at the *MAM* locus, secondary gene deletions, gene conversion and balancing selection appear to maintain biochemical diversity. Of course, deletion of *MAM1* and conversion of *MAM1* into a *MAM2*-like sequence are both one-way streets from an archetypical *MAM2–MAM1–MAM3* configuration. Likewise, deletion of *MAM2* or conversion of *MAM2* into a *MAM1*-like gene both prevents future switches of chain-length phenotypes. Therefore, the Arabidopsis *MAM* locus appears to be in a process during which paralogous genes are being sorted among lineages, ultimately leading to plants that harbor, in combination with *MAM3*, either a *MAM1* or a *MAM2* gene.

Similarities and differences in chain-length variation between *Arabidopsis* and *Brassica*

In *Brassica oleracea* and other members of the genus, aliphatic glucosinolates can be generated from homo-, dihomo- and trihomomethionine (Velasco and Becker 2000). As in *A. thaliana*, there is natural variation for homo- versus dihomomethionine-derived glucosinolates among different accessions. However, in contrast to *A. thaliana*, accumulation of homomethionine- and accumulation of dihomomethionine-derived glucosinolates do not mutually exclude each other. Hence, *B. oleracea* accessions can produce aliphatic glucosinolates from homomethionine (3C), dihomomethionine (4C) or from homomethionine *and* dihomomethionine (3C + 4C). This biochemical polymorphism is caused by variation at two different genetic loci, *BoGSL-ELONG* and *BoGSL-PRO* (Magrath et al. 1994; Li et al. 2001). *BoGSL-ELONG* harbors a typical *MAM* gene with ten exons (Fig. 4), and, in addition, a *MAM* pseudogene (Li et al. 2001; Gao et al. 2005). The *B. oleracea MAM* gene is closely related to *MAM* genes from other Brassicaceae (Fig. 3). Genetic data indicate that *BoGSL-ELONG* is responsible for the generation of dihomomethionine-derived glucosinolates (Li and Quiros 2002). However, *B. oleracea MAM* does not belong to any of the *MAMa*, *b* or *c* subclades found in *A. thaliana* and close relatives (Fig. 3). The second *Brassica* locus, *BoGSL-PRO*, contains a gene with a typical 12-exon *IPMS* structure (Fig. 4), and phylogenetic analyses support a close phylogenetic relationship with other *IPMS* genes (Fig. 3). This gene is supposedly involved in the generation of homomethionine-derived glucosinolates (Gao et al. 2006). This suggests that (i) the gene duplication events leading to *MAMa*, *b* and *c* occurred after *Arabidopsis*, *Boechera* and *Brassica* diverged from a common ancestor, (ii) the ability to utilize 3C precursors for carbon chain elongation evolved independently in *Brassica* and *Arabidopsis* and (iii) the *Brassica* MAM function responsible for generating 3C glucosinolates evolved de novo from a *MAM* progenitor gene. To confirm these hypotheses it will be necessary to sample further *MAM* (and also *IPMS*) genes from close and distant *Arabidopsis* and *Brassica* relatives and to analyse the biochemical properties of the encoded proteins.

Conclusions

The composition of genes at the *MAM* locus varies between and within cruciferous species, causing substantial diversity in glucosinolate profiles. Different types of selection act on *MAM* gene family members, and different factors account for generating glucosinolate variability and for maintaining this diversity. Comparative analyses suggest that particular MAM substrate specificities have evolved repeatedly in different genera of glucosinolate-producing plants. The model plant *A. thaliana* has been invaluable for making progress in dissecting the genetic, functional and ecological basis of glucosinolate diversity. These discoveries have already facilitated the identification and cloning of genes from the glucosinolate-myrosinase system in cruciferous species for which an ab initio approach proves much more difficult, such as crop plants or wild Arabidopsis relatives. Further functional and evolutionary studies, involving additional species from the Brassicaceae but also from the Capparaceae, the second large family in the Capparales order that is capable of using methionine homologs as glucosinolate precursors, will help to better understand the complexity and evolutionary dynamics of variation in plant secondary metabolism.

Acknowledgements We thank John D'Auria for critical reading of the manuscript and gratefully acknowledge financial support from the Max Planck Society and the DFG (grants KR2237/2-1, KR2237/2-2 and KR2237/2-3 to JK).

References

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J (2006) Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci USA 103:9118–9123

Campos de Quiros H, Magrath HCR, McCallum D, Kroymann J, Schnabelrauch D, Mitchell-Olds T, Mithen R (2000) Quantitative trait loci and the biochemical genetics of glucosinolate biosynthesis in *Arabidopsis thaliana*. Theor Appl Genet 101:429–437

Daxenbichler ME, Spencer GF, Carlson DG, Rose GB, Brinker AM, Powell RG (1991) Glucosinolate composition of seeds from 297 species of wild plants. Phytochemistry 30:2623–2638

de Kraker JW, Luck K, Textor S, Tokuhisa JG, Gershenzon J (2007) Two *Arabidopsis thaliana* genes (*IPMS1* and *IPMS2*) encode isopropylmalate synthase, the branchpoint step in the biosynthesis of leucine. Plant Physiol 143: 970–986

Fahey JW, Zalcmann AT, Talalay P (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. Phytochemistry 56:5–51

Falk KL, Vogel C, Textor S, Bartram S, Hick A, Pickett JA, Gershenzon J (2004) Glucosinolate biosynthesis: demonstration and characterization of the condensing enzyme of the chain elongation cycle in *Eruca sativa*. Phytochemistry 65:1073–1084

Field B, Cardon G, Traka M, Botterman J, Vancanneyt G, Mithen R (2004) Glucosinolate and amino acid biosynthesis in Arabidopsis. Plant Physiol 135:828–839

Gao M, Li G, McCombie RW, Quiros CF (2005) Comparative analysis of a transposon-rich *Brassica oleracea* BAC clone with its corresponding sequence in *A. thaliana*. Theor Appl Genet 111:949–955

Gao M, Li G, Potter D, McCombie RW, Quiros CF (2006) Comparative analysis of methylthioalkylmalate synthase (MAM) gene family and flanking DNA sequences in *Brassica oleracea* and *Arabidopsis thaliana*. Plant Cell Rep 25:592–598

Hagelstein P, Schultz G (1993) Leucine synthesis in spinach chloroplasts: partial characterization of 2-isopropylmalate synthase. Biol Chem Hoppe-Seyler 374:1105–1108

Haughn GW, Davin L, Giblin M, Underhill EW (1991) Biochemical genetics of plant secondary metabolites in *Arabidopsis thaliana*: the glucosinolates. Plant Physiol 97:217–226

Heidel A, Clauss M, Kroymann J, Savolainen O, Mitchell-Olds T (2006) Natural variation in *MAM* within and between populations of *Arabidopsis lyrata* determines glucosinolate phenotype. Genetics 173:1629–1636

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754–755

Keurentjes JJB, Fu J, de Vos CHR, Lommen A, Hall RD, Bino RJ, van der Plas LHW, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nat Genet 38:842–849

Kliebenstein DJ (2008) A quantitative genetics and ecological model system: understanding the aliphatic glucosinolate biosynthesis network via QTLs. Phytochem Rev. doi: 10.1007/s11101-008-9102-8

Kliebenstein DJ, Kroymann J, Brown P, Figuth A, Pedersen D, Gershenzon J, Mitchell-Olds T (2001a) Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol 126:811–825

Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T (2001b) Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. Plant Cell 13:681–693

Kliebenstein DJ, Gershenzon J, Mitchell-Olds T (2001c) Comparative quantitative trait loci mapping of aliphatic,

indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. Genetics 159: 359–370

Kliebenstein DJ, Kroymann J, Mitchell-Olds T (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. Curr Opin Plant Biol 8:264–271

Kliebenstein DJ, West MAL, van Leeuwen H, Loudet O, Doerge RW, St. Clair DA (2006) Identification of QTLs controlling gene expression networks defined a priori. BMC Bioinformatics 7:308

Knoke B, Textor S, Gershenzon J, Schuster S (2008) Modelling aliphatic glucosinolate diversity in *Arabidopsis thaliana*. Phytochem Rev (this issue)

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

Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T (2003) Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus. Proc Natl Acad Sci USA 100:14587–14592

Kroymann J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95–98

Kroymann J, Textor S, Tokuhisa JG, Falk KL, Bartram S, Gershenzon J, Mitchell-Olds T (2001) A gene controlling variation in *Arabidopsis thaliana* glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiol 127:1077–1088

Lambrix V, Reichelt M, Mitchell-Olds T, Kliebenstein DJ, Gershenzon J (2001) The *Arabidopsis* epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. Plant Cell 13:2793–2807

Li G, Quiros CF (2002) Genetic analysis, expression and molecular characterization of *BoGSL-ELONG*, a major gene involved in the aliphatic glucosinolate pathway of Brassica species. Genetics 162:1937–1943

Li G, Riaz S, Goyal S, Abel S, Quiros CF (2001) Inheritance of three major genes involved in the synthesis of aliphatic glucosinolates in *Brassica oleracea*. J Amer Soc Hort Sci 126:427–431

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers. Plant J 4:745–750

Magrath R, Bano F, Morgner M, Parkin I, Sharpe A, Lister C, Dean C, Lydiate D, Mithen RF (1994) Genetics of aliphatic glucosinolates. I. Side chain elongation in *Brassica napus* and *Arabidopsis thaliana*. Heredity 72:290–299

Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol 118:9–17

McDonald JH, Kreitman M (1991) Adaptive evolution at the *Adh* locus in *Drosophila*. Nature 251:652–654

Mithen R, Clarke J, Lister C, Dean C (1995) Genetics of aliphatic glucosinolates. III. Side chain structure of aliphatic glucosinolates in *Arabidopsis thaliana*. Heredity 74:210–215

Nordborg M, Innan H (2002) Molecular population genetics. Curr Opin Plant Biol 5:69–73

Pfalz M, Vogel H, Mitchell-Olds T, Kroymann J (2007) Mapping of QTL for resistance against the crucifer specialist insect herbivore *Pieris brassicae* in a new Arabidopsis inbred line population, Da(1)-12 × Ei-2. PLoS ONE 2:e578

Purrington CB (2000) Costs of resistance. Curr Opin Plant Biol 3:305–308

Reichelt M, Brown PD, Schneider B, Oldham NJ, Stauber E, Tokuhisa J, Kliebenstein DJ, Mitchell-Olds T, Gershenzon J (2002) Benzoic acid glucosinolate esters and other glucosinolates from *Arabidopsis thaliana*. Phytochemistry 59:663–671

Rodman JE (1991a) A taxonomic analysis of glucosinolate-producing plants, part 1: phenetics. Syst Bot 16:598–618

Rodman JE (1991b) A taxonomic analysis of glucosinolate-producing plants, part 2: cladistics. Syst Bot 16:619–629

Rodman JE, Karol KG, Price RA, Sytsma KJ (1996) Molecules, morphology, and Dahlgren's expanded order Capparales. Syst Bot 21:289–307

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Textor S, Bartram S, Kroymann J, Falk KL, Hick A, Pickett JA, Gershenzon J (2004) Biosynthesis of methionine-derived glucosinolates in *Arabidopsis thaliana*: recombinant expression and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain elongation cycle. Planta 218:1026–1035

Textor S, de Kraker JW, Hause B, Gershenzon J, Tokuhisa G (2007) MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. Plant Physiol 144:60–71

Tian D, Traw MB, Chen J, Kreitman M, Bergelson (2003) Pleiotropic fitness cost of *Rpm1* resistance in *Arabidopsis thaliana*. Nature 423:74–77

Velasco L, Becker HC (2000) Variability for seed glucosinolates in a germplasm collection of the genus *Brassica*. Genet Res Crop Evol 47:231–238

Wang S, Basten CJ, Zeng ZB (2001–2004) Windows QTL cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. www.statgen.ncsu.edu.qtlcart/WQTLCart.htm

Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ (2007) Linking metabolic QTLs with network and *cis*-eQTLs controlling biosynthetic pathways. PLoS Genet 3:e162

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Windsor AJ, Reichelt M, Figuth A, Svatoš A, Kroymann J, Kliebenstein DJ, Gershenzon J, Mitchell-Olds T (2005) Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae). Phytochemistry 66:1321–1333

Yang YW, Lai KN, Tai PY, Li WH (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

Zhang Z, Ober JA, Kliebenstein DJ (2006) The gene controlling the quantitative trait locus *epithiospecifier modifier1* alters glucosinolate hydrolysis and insect resistance in *Arabidopsis*. Plant Cell 18:1524–1536