Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family

Johannes A. Hofberger1,2, Eric Lyons3, Patrick P. Edger4, J. Chris Pires4, and M. Eric Schranz1,*

1Biosystematics Group, Wageningen University & Research Center, Wageningen, Gelderland, The Netherlands
2Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Noord-Holland, The Netherlands
3School of Plant Sciences; iPlant Collaborative, BIOS Institute, University of Arizona
4Division of Biological Sciences, Bond Life Sciences Center, University of Missouri

*Corresponding author: E-mail: eric.schranz@wur.nl.
Accepted: October 21, 2013

Abstract

Plants share a common history of successive whole-genome duplication (WGD) events retaining genomic patterns of duplicate gene copies (ohnologs) organized in conserved syntenic blocks. Duplication was often proposed to affect the origin of novel traits during evolution. However, genetic evidence linking WGD to pathway diversification is scarce. We show that WGD and tandem duplication (TD) accelerated genetic versatility of plant secondary metabolism, exemplified with the glucosinolate (GS) pathway in the mustard family. GS biosynthesis is a well-studied trait, employing at least 52 biosynthetic and regulatory genes in the model plant Arabidopsis.

In a phylogenomics approach, we identified 67 GS loci in Aethionema arabicum of the tribe Aethionemae, sister group to all mustard family members. All but one of the Arabidopsis GS gene families evolved orthologs in Aethionema and all but one of the orthologous sequence pairs exhibit synteny. The 45% fraction of duplicates among all protein-coding genes in Arabidopsis was increased to 95% and 97% for Arabidopsis and Aethionema GS pathway inventory, respectively. Compared with the 22% average for all protein-coding genes in Arabidopsis, 52% and 56% of Aethionema and Arabidopsis GS loci align to ohnolog copies dating back to the last common WGD event. Although 15% of all Arabidopsis genes are organized in tandem arrays, 45% and 48% of GS loci in Arabidopsis and Aethionema descend from TD, respectively. We describe a sequential combination of TD and WGD events driving gene family extension, thereby expanding the evolutionary playground for functional diversification and thus potential novelty and success.

Key words: comparative genomics, systems biology, whole-genome duplication, functional diversification, Brassicaceae.

Introduction

Gene duplication has played an important evolutionary role in angiosperm adaptation and success, for example, by contributing to regulatory and enzymatic pathways involved in generating more than 200,000 diverse biochemical plant secondary metabolites found in the Angiosperm lineage (Hartmann 2007). Functional diversification refers to processes of gene duplication followed by sub- or neofunctionalization of the enzymes encoded by duplicate copies (Ohno 1970; Roth et al. 2007; Wang et al. 2011c), mediating specificities to extended classes of substrates or catalysis of novel reactions (Stehle et al. 2008). Fast expansion of gene copy number occurs in various ways. In this study, we focus on whole-genome duplication (WGD), tandem duplication (TD), and gene transposition duplication (GTD). For example, approximately 45% of the Arabidopsis nuclear protein-coding genes have been affected by such processes (Bowers et al. 2003; Rizzon et al. 2006; Huang et al. 2012). In this study, we investigated the impact of gene duplication to the diversification of plant secondary metabolites exemplified with glucosinolate (GS) biosynthesis. GS biosynthesis is a well-studied key trait shared by all Brassicales including the mustard family (Brassicaceae) crown group (Schranz et al. 2011) and its sister lineage Aethionemae. Comparative genomics analysis unraveled a history of successive paleopolyploidy events commonly shared by almost all Angiosperms (Bowers et al. 2003). The Arabidopsis lineage underwent at least five polyploidy events in the history of life, two preceding and three following Angiosperms radiation (Bowers et al. 2003; Jiao et al. 2011). The most recent WGD is commonly referred to as At-α and occurred approximately 30–60 Ma in the ancestor of all Brassicaeae, including the sister group Aethionemae (Edger PP, Pires Jc, submitted for publication). As a result, pairwise
synteny regions are scattered throughout the genome (genomic blocks), defined as copies of consecutive ohnologs derived from At-α (Bowers et al. 2003). It is known that polyploidy is succeeded by a genome-wide process of biased fractionation, preferentially targeting one subgenome to retain clusters of dose-sensitive genes organized in functional modules (Thomas et al. 2006). Furthermore, several studies have established a potential link of polyploidy to natural variation due to differential expression of ohnolog copies (Wang et al. 2011c), seed and flower origin and diversification (De Bodt et al. 2005; Irish and Litt 2005; Jiao et al. 2011), morphological complexity (Freeing and Thomas 2006), and survival of plant lineages at the Cretaceous-Tertiary extinction event (Fawcett et al. 2009). In this study, we provide solid evidence for the link of WGD to pathway expansion of a distinct key trait relevant for herbivore defense and hence highly connected to fitness. Interestingly, polyploidy also affects other kinds of duplication, creating networks of factors with mutual influence. Recent studies have shown an interaction between polyploidy and the fractionation rate of tandem duplicate copies in both Arabidopsis and Brassica rapa (having undergone an additional genome triplication). Hence, we analyzed short-sequence duplications to utilize the evolutionary significance of different duplication classes.

TD of short sequences can be caused by unequal crossing-over or template slippage during DNA repair, producing tandem arrays (TARs) of homologous genes in close genomic vicinity (Kane et al. 2010). Depending on the number of allowed gene spacers, TAR genes include about 10–15% of the Arabidopsis thaliana genome (0 and 10 spacers, respectively) (Rizzon et al. 2006). Comparison of TARs in Arabidopsis and rice revealed enrichment of genes encoding membrane proteins and function in biotic and abiotic stress (Rizzon et al. 2006). Notably, the impact of TD to trait evolution has been elucidated in multiple taxa, including disease resistance in Solanaceae (Parniske et al. 1999) and Brassicaceae (Leister 2004). Likewise, TD played a role in the evolution of signal transduction, for example, the expansion and functional diversification of the F-box type transcriptional activator gene family in Fabaceae (Belliery-Rabelo et al. 2013). Moreover, TD is an important factor for increasing versatility of defense response in Brassicaceae. In GS biosynthesis, subfunctionalization of TAR genes is evident for 2-oxoglutarate-dependent dioxygenases (AOP) (Kliebenstein et al. 2001), flavin-monooxygenases (FMOGSOX) (Li et al. 2008) and methylthiokylmalate synthases (MAM) (Kroymann et al. 2003; Heidel et al. 2006; Textor et al. 2007). In this study, we integrate previous findings to dissect the influence of polyploidy with TD and GTD in the last 30–60 Ma of GS pathway expansion since Aethionema and Arabidopsis lineage divergence.

Duplicate gene copies can move to a new genomic location. The observed frequency of gene movements explains the observed erosion of synteny between plant genomes during evolution (Wicker et al. 2010), defining the limits of synteny-based approaches for ortholog detection. Gene movements are often caused by transposition. GTD events occur when a single nontransposon gene relocates to a new position, and segregants contain duplicates (Freeing 2009). Although transposable elements (TEs) account for approximately 10% of the Arabidopsis genome (Huang et al. 2012) and show nonrandom association to syntenic blocks (Hughes et al. 2003), 14% of all protein-coding genes in Arabidopsis transposed at least once during Rosid evolution (Freeing et al. 2008; Woodhouse et al. 2011). Importantly, a novel genomic context of the transposed copy potentially influences rates of gene expression (Wang et al. 2013) and might thereby contribute to the phenotypic consequences of the duplication event (Kliebenstein 2008). Accordingly, TE activity was shown to foster variation of NBS-resistance proteins in grape (Malacarne et al. 2012) as well as natural growth variation and expansion of ERF family transcriptional regulators in Arabidopsis (Nakano et al. 2006; Vlad et al. 2010). In contrast, evolutionary dynamics of GTD events affecting genetic versatility of plant secondary metabolism has not yet been investigated.

GS comprise a class of secondary plant metabolites derived from amino acids and sugars, part of a two-component chemical defense against herbivory in Brassicaceae (Rodman 1998; Windsor et al. 2005; Beekwilder et al. 2008). Myrosinase enzymes are the other component of the defense system and confer GS hydrolysis activity. They are released from the vacuole upon tissue damage, producing a plethora of GS degradation products such as nitriles, isothiocyanates, thiocyanates, and epithioalkanes with various bioactivities (Rask et al. 2000; Bones and Rossiter 2006). GS are of particular interest for human health because they can inhibit carcinogen activation (Hecht 2000; Nakajima et al. 2001) and carcinogenesis by triggering cell cycle arrest and stimulating apoptosis (Wittstock et al. 2003; Hayes et al. 2008). The observed variation in GS biochemistry across Brassicaceae is due to the differences in biochemistry among their amino acid precursors (Fahay et al. 2001; Windsor et al. 2005) and allows GS grouping to four distinct classes. Oxidative deamination of Phe and Tyr initiates biosynthesis of indolic GS (I); Trp is the substrate for indolic GS production (II); Ala, Val, Leu, and Ile are precursors for biosynthesis of aliphatic GS (III) (Mithen et al. 2010). Although aromatic and aliphatic GS have been detected in other eudicot families including Phytolaccaceae, Euphorbiaceae, and Pittosporaceae (Rodman et al. 1996; Fahey et al. 2001), indolic GS are Brassicales specific. Met-derived GS form a fourth class of GS (IV), referring to a subset of aliphatic GS specific to the Brassicales crown group, including the sister group Aethionemae. The utilization of Trp- and Met-derived amino acids for GS production may be tied to pathway expansion caused by ancient WGD events (Schranz et al. 2011).

The genus Aethionema of the tribe Aethionemae is an ideal group for comparative genomics of polyploidy and GS pathway evolution. First, it shares the composite GS chemotype.
observed in the larger and more diverse Brassicaceae crown group (Schranz et al. 2012). Second, phylogenetic analysis highly support the tribe Aethionemae as the earliest diverged clade and extant sister to the crown group Brassicaceae (Couvreur et al. 2010) with an estimated split of the two lineages approximately 30–60 Ma. However, a high degree of interspecies synteny (see Results) is maintained. Third, the most recent WGD event identified in the lineage of Arabidopsis (referred to as At-\( \alpha \)) predated the divergence of Arabidopsis and Aethionema. Furthermore, it was not succeeded by an additional species-specific genome polyploidization, preventing additional fractionation of synteny (Bowers et al. 2003; Haudry et al. 2013) (Edger PP, Pires Jc, submitted for publication). In contrast, B. rapa underwent an additional genome triplication event (Wang et al. 2011b), complicating efforts to analyze the potential impact of At-\( \alpha \) on the evolution of the GS pathway inventory.

**Materials and Methods**

*Aethionema arabicum* Genome Assembly and Set of Annotated Genes

Sequence assembly and annotation of the *Aethionema arabicum* genome was obtained from Haudry et al. (2013).

RNA Isolation and Sequencing

*Aethionema arabicum* RNA was isolated from fresh apical meristematic tissue or very young leaves using an RNeasy Plant Mini Kit (Qiagen, Valencia, CA). Samples were kept on liquid nitrogen before RNA isolation. The optional step of heating the lysis solution to 65 \( ^\circ \)C was used to maximize RNA yield. RNA was eluted into a final volume of 100 \( \mu \)l RNase-free water. Total mass of RNA and quality was estimated using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). Samples were deemed acceptable if RIN (RNA integrity number) scores were greater than 8.0. A minimum of 20 \( \mu \)g of total RNA was required for library building and sequencing. RNA-seq (Wang et al. 2009) paired-end libraries with average fragment lengths of 250 bp were constructed, and each library was sequenced on a single lane of an Illumina GAIIx sequencer flow cell (Illumina, San Diego, CA) to generate a minimum of 3 gigabases of 75 bp, paired-end sequences.

Database of *A. thaliana* Genes Retaining an Ohnologous Copy Dating Back to the At-\( \alpha \) WGD Event

First, we generated a spreadsheet with information on all 33,323 *A. thaliana* nuclear genes annotated in the TAIR database v10, including 1) Arabidopsis gene identifiers (AGIs), 2) locus type, 3) locus name, and 4) short description of encoded function. Second, we integrated optional affiliation to 5) syntenic block and 6) ohnolog copy dating back to At-\( \alpha \) WGD event as described previously (Freeling and Thomas 2006). The corresponding authors did not account for every gene in their analysis and inferred the genomic location of ohnolog blocks dating back to At-\( \alpha \) using the TIGR *Arabidopsis* genome annotation v5 from 2005. Third, we added an additional column (i), to indicate coverage of the gene in Feeling’s study (yes/not considered/not present in TIGR5).

Database for GS Biosynthetic Gene Identification in *A. thaliana* and *Aet. arabicum*

Files containing the coding sequences representing the complete set of GS biosynthetic and regulatory (AtGS) genes in *A. thaliana* (Sonderby et al. 2010) were acquired from the TAIR database v10 (www.arabidopsis.org, last accessed November 9, 2013). We highlighted the AGIs in the spreadsheet covering all nuclear genes in *Arabidopsis* (see Materials and Methods).

Interpolation of Novel Putative GS Biosynthetic Genes and Retained At-\( \alpha \) Ohnolog Pairs in *A. thaliana*

We utilized similarity among ohnologous gene copies. We employed the spreadsheet containing information on genomic location of AtGS genes and \( \alpha \)-blocks with optional retained duplicates therein. We visually screened for all ohnolog copies of AtGS genes not sharing annotation as AtGS genes themselves. Differential expression of ohnolog pairs was tested using the botanical array resource (http://bar.utoronto.ca/welcome.htm, last accessed November 8, 2013). We included all ohnolog copies for our analysis to create an extended AtGS gene set. For the spreadsheet, see supplementary table S1, Supplementary Material online. Previous approaches of ohnologous gene pair identification did not consider every protein-coding gene (Bowers et al. 2003; Thomas et al. 2006). To minimize resulting errors for analysis of AtGS loci, we performed a BlastP screen without a cut-off e value, querying all AtGS genes with nonretained At-\( \alpha \) ohnologs against all other *Arabidopsis* genes with nonretained At-\( \alpha \) ohnologs. Highest-scoring sequence pairs sharing genomic location within converse copies of the same \( \alpha \)-block were tested for synteny and positives defined as additional pairs of retained At-\( \alpha \) ohnolog copies (marked as “our addition/OA” in table 1 and supplementary table S1, Supplementary Material online).

Database of *A. thaliana* Tandem Arrayed Genes

A database of *A. thaliana* coding sequences organized in TARs was generated for the TAIR annotation v10 as described previously (Rizzon et al. 2006), using a low-stringency approach with a number of \( N = 10 \) allowed gene pacers. Information was updated to TAIR10 and included to the spreadsheet covering all *Arabidopsis* nuclear genes (supplementary table S1, Supplementary Material online).

Database of *A. thaliana* GS Genes Affected by GTD

A database of the epoch-independent positional history of all *Arabidopsis* genes was generated as described previously for...
TAIR9 (Woodhouse et al. 2011). We updated all putative GTD copies to TAIR10. Woodhouse et al. scored gene duplicates as transposed based on a function of synteny across taxa in the direction \( A. \text{thaliana} \rightarrow A. \text{lyrata} \rightarrow \text{Carica papaya} \rightarrow \text{Populus trichocarpa} \rightarrow \text{Vitis vinifera} \). For analysis of Brassicaceae genome evolution, methodical restrictions apply due to the low resolution within that clade, covered by only two tribes. Thus, we screened the genomic context of AtGS genes within a narrow window of 3 kb for flanking TE-like sequences, using the GEvo function from the CoGe comparative genomics platform (http://genomevolution.org/CoGe/GEvo.pl, last accessed November 8, 2013) (Lyons and Freeling 2008). Graphical

Table 1
Retained At-\(\alpha\) Ohnolog Duplicate Gene Pairs in \(\text{Arabidopsis}\) and \(\text{Aethionema}\) GS Pathway Inventory

| Protein Name | AGI | \(\alpha\)-Block | Evident SSD | AabdID | Syntelog | % Identity | Col-0 → Aab |
|--------------|-----|-----------------|------------|--------|----------|------------|--------------|
| UGT74C1      | AT2G31790 | A02N051 | TD | Aab37175 | Yes | 79.44 | 6 → 10 |
| [UGT like]   | AT1G05670 | A02N051 | TD | Aab31930 | Yes | 81.11 | 6 → 10 |
| FMO-GSOX-2   | AT1G62540 | A03N117 | TD | Aab10869 | Yes | 76.6 | 11 → 8 |
| [FMO like]   | AT1G12130 | A03N117 | TD | Aab13543 | Yes | 65.09 | 11 → 8 |
| CYP79F2      | AT1G16400 | A05N062 | TD | — | — | — | 8 → 9 |
| CYP79C1      | AT1G79370 | A05N062 | GTD | Aab34143 | Yes | 76.8 | 8 → 9 |
| SOT16        | AT1G74100 | A05N186 | TD | Aab14278 | Yes | 91.07 | 3 → 3 |
| SOT17        | AT1G18590 | A05N186 | TD | Aab19675 | Yes | 86.05 | 3 → 3 |
| SUR1         | AT2G20610 | A10N194 | TD | Aab30136 | Yes | 89.15 | 3 → 2 |
| [SUR like]   | AT4G28420 | A10N194 | TD | Aab31155 | Yes | 57.93 | 3 → 2 |
| CYP79B2      | AT4G39950 | A10N257 | GTD | Aab18705 | Yes | 81.38 | 8 → 9 |
| CYP79B3      | AT2G22330 | A10N257 | GTD | Aab19477 | Yes | 81.4 | 8 → 9 |
| GGP1         | AT4G30530 | A10N314 | TD | Aab24374 | Yes | 87.6 | 5 → 5 |
| [GGP like]   | AT2G23960 | A10N314 | TD | Aab11021 | Yes | 61.38 | 5 → 5 |
| GSTF11       | AT3G03190 | A12N102 | TD | Aab14996 | Yes | 77.1 | 4 → 4 |
| [GSTF12]     | AT5G17220 | A12N102 | TD | Aab14791 | Yes | 77.84 | 4 → 4 |
| Cosubstrate pathways |
| [AAO3]       | AT2G27150 | A02N01A | GTD | Aab2016 | Yes | 77.67 | 2 → 2 |
| AAO4         | AT1G04580 | A02N01A | TD | Aab24896 | Yes | 79.58 | 2 → 2 |
| APK1         | AT2G14750 | A10N02A | TD | Aab32150 | Yes | 83.75 | 2 → 2 |
| APK2         | AT4G39940 | A10N02A | TD | Aab17804 | Yes | 86.05 | 2 → 2 |
| Side-chain elongation |
| BCAT4        | AT3G19710 | A08N074 | TD | Aab21007 | Yes | 75 | 6 → 6 |
| [BCAT7]      | AT1G50090 | A08N074 | TD | Aab22548 | Yes | 76.9 | 6 → 6 |
| IPMI1        | AT3G58990 | A11N226 | TD | Aab13092 | Yes | 83 | 3 → 3 |
| [IPMI like]  | AT2G43090 | A11N226 | TD | Aab19619 | Yes | 85.99 | 3 → 3 |
| BCAT3        | AT3G49680 | A19N002 | TD | Aab33782 | Yes | 76.02 | 6 → 6 |
| [BCAT5]      | AT5G65780 | A19N002 | TD | Aab23605 | Yes | 75.3 | 6 → 6 |
| BAT5         | AT4G12030 | A20N095 | TD | Aab23285 | Yes | 76.21 | 2 → 2 |
| [BAT like]   | AT4G22840 | A20N095 | TD | Aab23321 | Yes | 91.82 | 2 → 2 |
| TF regulation |
| OBP2         | AT1G07640 | A02N142 | TD | Aab18330 | Yes | 80.28 | 2 → 2 |
| [OBP like]   | AT2G28810 | A02N142 | TD | Aab24559 | Yes | 70.03 | 2 → 2 |
| MYB122       | AT1G74080 | A05N185 | TD | Aab14276 | Yes | 57.56 | 6 → 4 |
| MYB51        | AT1G18570 | A05N185 | TD | Aab19683 | Yes | 59.89 | 6 → 4 |
| IQD1         | AT3G09710 | A14N046 | TD | Aab18852 | Yes | 65.59 | 2 → 2 |
| [IQD2]       | AT5G03040 | A14N046 | TD | Aab18368 | Yes | 77.39 | 2 → 2 |
| MYB28        | AT5G61420 | A26N034 | TD | Aab12163 | Yes | 67.39 | 6 → 4 |
| MYB29        | AT5G07690 | A26N034 | TD | Aab33585 | Yes | 65.13 | 6 → 4 |

*Squared brackets indicate ohnolog copies of GS biosynthetic genes without GO!-annotation to GS biosynthetic process.

TD refers to members of TARs and GTD refers to a history of transposition in \(\text{Arabidopsis}\).

Predicted \(\text{Aethionema}\) CDS.

Change of gene family locus count in \(\text{Arabidopsis} \rightarrow \text{Aethionema}\) order.
highlights of TE-like sequences have been customized by choosing “show other features” in the “results visualization” screen. By that means, we confirmed AtGS genes that transposed at least once during lineage evolution as defined by Woodhouse et al. and identified further GTDs missed by that approach due to lack of synteny data (i.e., GTDs predating *Vitis* speciation and recent GTDs of Brassicaceae-specific genes; marked by asterisks in table 2). Information on GTD events was added to an additional column in supplementary table S1, Supplementary Material online.

Analysis of Putative GS Genes Not Affected by TD, GTD, or At-α Ohnolog Retention in *Arabidopsis*

We performed additional analysis of AtGS loci beyond the aforementioned types of duplication by considering more ancient WGD events. Information on *Arabidopsis* genome-wide distribution of ohnolog duplicate pairs dating back to the At-β and At-γ WGD events (Bowers et al. 2003) were added to an additional column in supplementary table S1, Supplementary Material online. GS genes were referenced accordingly. Remnants did not show significant similarities to any other locus in *Arabidopsis* by definition, and evolutionary stability was confirmed using the *Arabidopsis* transpositional history database (http://nature.berkeley.edu/freelinglab, last accessed November 8, 2013) and data on AtGS syntelogs in *B. rapa* (Wang et al. 2011a).

| Protein Name | AGI | α-Block | AabID | Syntelog | % Identity | Lineage Specific? |
|--------------|-----|---------|-------|----------|------------|------------------|
| [AAO3]       | AT2G27150 | A02NOA1 | Aab27016 | Yes | 77.67 | No | 2 → 2 |
| CV79C1       | AT1G79370 | A05NO62 | Aab34143 | Yes | 76.8 | No | 8 → 9 |
| APK2         | AT4G39940 | A10NOA2 | Aab17804 | Yes | 86.05 | No | 2 → 2 |
| CV79B2       | AT4G39950 | A10N257 | Aab17805 | Yes | 81.38 | No | 8 → 9 |
| CV79B3       | AT2G23330 | A10N257 | Aab19477 | Yes | 81.4 | No | 8 → 9 |

Orthologous Gene Identification of *Arabidopsis* GS Biosynthetic Genes in *Aet. Arabicum*

We considered multiple lines of evidence for identification of orthologs between *A. thaliana* GS loci and *Aet. arabicum*. We defined orthologous pairs of *A. thaliana* and *Aet. arabicum* GS loci as reciprocal best hits (RBHs) within a given region of gene colinearity (synteny). First, we screened for regions in the

| Protein Name | AGI | α-Block | AabID | Syntelog | % Identity | Lineage Specific? |
|--------------|-----|---------|-------|----------|------------|------------------|
| CV79F1       | AT1G16410 | A05 | Aab27579 | Yes | 72.79 | *Aethionema* | 8 → 9 |
| CV79C2       | AT1G58260 | A03 | Aab17711 | Yes | 71.85 | *Aethionema* | 8 → 9 |
| CV83A1       | AT4G13770 | A15 | Aab32506 | Yes | 69.67 | No | 2 → 3 |
| CV83B2       | AT5G57220 | A22 | Aab30975 | No | 82.8 | *Aethionema* | 2 → 3 |

| Protein Name | AGI | α-Block | AabID | Syntelog | % Identity | Lineage Specific? |
|--------------|-----|---------|-------|----------|------------|------------------|
| CV79A2       | AT5G05260 | A14 | Aab36760 | Yes | 73.37 | No | 8 → 9 |
| CHY1         | AT5G65940 | A19 | Aab05851 | Yes | 80.75 | No | 2 → 2 |
| FMO-GSOX-1   | AT1G65860 | A25 | Aab30109 | Yes (minimum) | 58.45 | No | 11 → 8 |

aAsterisks mark GTDs inferred by flanking TE-like sequences using GEVs.
bPredicted *Aethionema* CDS.
cChange of gene family locus count in *Arabidopsis* → *Aethionema* order.
Aet. arabicum genome displaying synteny to genomic regions in A. thaliana harboring GS loci, using the “Synfind” function with standard parameters from the CoGe comparative genomics system (www.genomeevolution.org, last accessed November 8, 2013) (Tang and Lyons 2012). Second, we determined RBHs between A. thaliana GS genes and Aet. arabicum genes within the syntenic regions from (i), using BlastP with a minimum query coverage of $N=0.5$ and a cut-off $e$-value of $1E-10$. Third, we queried all putative Aet. arabicum GS loci against the Aet. arabicum genome in a BlastP screen with a cut-off $e$-value of $1E-30$. We screened for subject sequences not sharing the query sequence scaffold and identified syntenic regions in A. thaliana. If a GS biosynthetic gene was present in syntenic A. thaliana region (BlastP with a cut-off $e$-value of $1E-30$), we defined the aligned Aet. arabicum subject sequence as ortholog to the A. thaliana query sequence.

Tandem Arrayed Gene Copy Identification of Putative GS Biosynthetic Genes in Aet. arabicum
We queried all putative Aet. arabicum GS loci against the Aet. arabicum genome in a BlastP screen with a cut-off $e$-value of $1E-30$. For identification of TDs, GS query sequences were grouped with the subset of respective subject sequences located within a window of $N=10$ allowed gene spacers to form Aet. arabicum superfamilies of putative TAR genes. TDs were visualized using the MAFFT package (http://mafft.cbrc.jp/alignment/software/, last accessed November 8, 2013) (Kato et al. 2002). We further confirmed Aet. arabicum GS genes expression by querying the RNA-seq data. Transcriptome data were mined for expression of GS genes using TBlastX with a cut-off $e$-value of $1E-10$ (data not shown).

Identification of Lineage-Specific GS GTDs Comparing Putative GS Biosynthetic Genes in Aet. arabicum and A. thaliana
We queried all putative Aet. arabicum GS loci against the Aet. arabicum genome in a BlastP screen with a cut-off $e$-value of $1E-30$. For identification of GTDs following divergence of these lineages, we screened for subject sequences not sharing the query sequence scaffold and identified syntenic regions in A. thaliana. If GS biosynthetic gene is absent in syntenic A. thaliana region (BlastP with a cut-off $e$-value of $1E-30$), we defined the aligned Aet. arabicum subject sequence as lineage-specific GTD copy.

Phylogenetic and Similarity Analysis
A number of Arabidopsis flavin-monoxygenases involved in GS biosynthesis (FMO GS-OX) are encoded in clusters consisting of retained ohnolog copies as well as both tandem- and gene transposition duplicates. To visualize the evolution of FMO-like sequences in Brassicales, Carica. papaya, and Tarenaya hassleranii (Cheng et al. 2013), FMO orthologs from these species were obtained using the CoGe comparative genomics system. A phylogenetic tree was constructed using the maximum-likelihood method with PhyML 3.1 software (Guindon et al. 2010), employing the Le/Gascuel (LG) model for amino acid substitution. Protein sequence similarity analysis were performed using the Needle program from the EMBOSS software package (http://emboss.sourceforge.net/, last accessed November 8, 2013) (Rice et al. 2000).

Genome Data Visualization and Statistics
Fisher’s exact test for count data was performed using the R package for statistical computing (www.r-project.org, last accessed November 8, 2013). Circular visualization of genome data was performed using the circos package (www.circos.ca, last accessed November 8, 2013) (Krzywinski et al. 2009) and graphically edited with the GIMP-package (www.gimp.org, last accessed November 8, 2013).
two ohnolog copies each. We thereby created an extended set of 64 putative AtGS genes (figs. 1 and 3). Among genes located within α-block boundaries, we found an At-α ohnolog retention rate of 59% (36/61) for the extended AtGS set (fig. 1), which is more than double of the observed 22% average rate for ohnolog retention among all Arabidopsis protein-coding loci harbored within the boundaries of α-blocks (fig. 2B).

Quantification of TD Influence to GS Pathway Evolution in Arabidopsis

In the next step, we quantified the impact of TD to GS pathway versatility in Arabidopsis. Minor changes were made in the list of Arabidopsis TAR genes by Rizzon et al. (2006) due to the gene updates to TAIR10 (supplementary table S1, Supplementary Material online). We mined the 1,497 Arabidopsis TARs comprising 4,034 duplicate gene copies for AtGS.

Table 3

| Protein Name | AGI | α-Block | Evident SSD* | AabIDb | Syntelog | % Identity | Col-0→Aab |
|--------------|-----|---------|--------------|--------|----------|------------|------------|
| **Core-structure formation** | | | | | | | |
| AOP1         | AT4G03070 | A01 | TD/GTD | Aab37231 | Yes | 70.03 | 2→1 |
| AOP3         | AT4G03050 | A01 | TD/GTD | — | — | — | 2→1 |
| UGT74-like   | Aab specific | A02 | TD | Aab37178 | Yes | 82.05 | 6→10 |
| UGT74-like   | Aab specific | A02 | TD | Aab37179 | Yes | 77.63 | 6→10 |
| GSTF10       | AT2G30870 | A02 | TD | Aab28612 | Yes | 91.59 | 4→4 |
| FMO-GSOX-3   | AT1G62560 | A03 | TD | Aab10867 | Yes | 71.9 | 11→8 |
| FMO-GSOX-4   | AT1G62570 | A03 | TD | Aab10866 | Yes | 55.2 | 11→8 |
| FMO-GSOX-5   | AT1G12140 | A03 | TD | Aab13546 | Yes | 71.9 | 11→8 |
| CYP79C2      | AT1G58260 | A03 | TD | Aab17711 | Yes | 71.85 | 8→9 |
| **Cosubstrate pathways** | | | | | | | |
| CYP79F1      | AT1G16410 | A05 | TD | Aab27579 | Yes | 72.79 | 8→9 |
| SOT18        | AT1G74090 | A05 | TD | Aab14277 | Yes | 83.9 | 3→3 |
| GSTU20       | AT1G78370 | A05 | TD | Aab7000 | Yes | 67.29 | 5→6 |
| CYP83B1      | AT4G31500 | A10 | TD/GTD | Aab12019 | No | 92.73 | 2→3 |
| GSL-OH       | AT2G25450 | A10 | TD/GTD | — | — | — | 1→0 |
| CYP79A2      | AT5G05260 | A14 | TD/GTD | Aab36760 | Yes | 73.37 | 8→9 |
| CYP83A1      | AT4G13770 | A15 | TD/GTD | Aab32506 | Yes | 69.67 | 2→3 |
| FMO-GSOX-1   | AT1G58680 | A25 | TD/GTD | Aab30109 | Yes | 58.45 | 11→8 |
| CHY1         | AT5G65940 | A19 | TD/GTD | Aab05851 | Yes | 80.75 | 2→2 |
| GSH1         | AT4G23100 | A20 | NA | Aab22781 | Yes | 91.81 | 2→2 |
| BZO1         | AT1G65880 | A25 | TD | Aab31601 | Yes | 70.04 | 2→4 |
| IM3D         | AT1G31180 | A06 | TD | Aab31602 | Yes | 69.4 | 2→4 |
| IIM2         | AT2G43100 | A11 | TD | Aab19630 | Yes | 78.71 | 2→1 |
| IMD1         | AT5G14200 | A12 | TD/GTD | Aab14760 | Yes | 89.5 | 2→1 |
| IIL1         | AT4G13430 | A15 | NA | Aab18132 | Yes | 93.9 | 1→1 |
| **Side-chain elongation** | | | | | | | |
| MYB76        | AT5G07700 | A26 | TD | — | — | — | 6→4 |

**Table 3**

Genes with Nonretained At-α Ohnolog Duplicate Gene Copy in Arabidopsis and Aethionema GS Pathway Inventory

Notes.—NA, not applicable.
*TD (Tandem Duplicate) refers to members of tandem arrays and GTD (Gene Transposition Duplication) refers to a history of transposition in Arabidopsis.
aPredicted Aethionema CDS
bChange of gene family locus count in Arabidopsis→Aethionema order

Quantification of TD Influence to GS Pathway Evolution in Arabidopsis

In the next step, we quantified the impact of TD to GS pathway versatility in Arabidopsis. Minor changes were made in the list of Arabidopsis TAR genes by Rizzon et al. (2006) due to the gene updates to TAIR10 (supplementary table S1, Supplementary Material online). We mined the 1,497 Arabidopsis TARs comprising 4,034 duplicate gene copies for AtGS.
Forty-five percent (29/64) of AtGS genes are members of TARs, compared with a genome-wide average of 15% (figs. 2B and D). Contribution of GTDs to GS pathway evolution in Arabidopsis

We quantified the influence of GTD to GS pathway evolution in Arabidopsis. Initially, a list of 4,575 genes with putative origin due to a GTD was proposed for TAIR9 (Woodhouse et al. 2011). Our update to TAIR10 retained 4,539 loci clearly referenced to transposition events (supplementary table S1, Supplementary Material online), illustrating a 14% average for GTD genes among protein-coding loci in Arabidopsis (fig. 2B). Among those, we confirmed all 13 references to AtGS loci (table 2), using the GEvo function from the CoGe platform (see Materials and Methods). We thereby discovered
Fig. 2.—Duplicate distribution among (A, B) Arabidopsis protein-coding genes compared with (C, D) AtGS and (E, F) Aethionema GS loci. Shown are retained ohnologs (green), tandem duplicates (blue), and gene transposition duplicates (orange). GS metabolic versatility resulted from a combination of increased ohnolog retention and TD rates.
Fig. 3.—Ideogram of Arabidopsis thaliana chromosomes with GS biosynthetic genes. Circos plot visualizing the evolutionary contribution of different duplication types to GS pathway inventory in Arabidopsis and Aethionema. (A) Inner chromosome scale (Mb). (B) Arabidopsis thaliana GS biosynthetic genes. Gray text indicates genomic location outside ohnolog blocks. Black text indicates genomic location within ohnolog blocks but nonretained ohnolog copy. Green text indicates retained pairs of ohnolog copies with missing GO annotation to GS biosynthetic process shown in edged brackets. Orange text indicates single copy genes without clear paralogs in both species. (C) Blue circles indicate genes organized in TARs (i). Red circles indicate genes with transpositional history (ii). Purple circles indicate loci sharing (i) and (ii). (D) Number of rectangles indicates number of homologs present in the Aethionema arabicum draft genome (0–4). Color of rectangles indicates presence (black) or absence (red) of synteny between A. thaliana and Aet. arabicum in the genomic context of the target gene. (E) Arabidopsis thaliana chromosomes with labels showing GS biosynthetic genes. Bands for genes retained in ohnolog pairs are connected with colors of corresponding ohnolog blocks, as defined by Bowers et al. (2003). (F) Genomic location of ohnolog block copies harboring GS biosynthetic genes in A. thaliana, connected by gray bands. All ranges are in scale.
that four additional AtGS genes (marked by asterisks in table 2) lost all syntenic anchor genes in genomic proximity (±1,000 kb) but are surrounded by TE-like sequences. Thus they may have transposed following the At-α WGD event (marked by asterisks in table 2). These additional genes may be Brassicaceae specific but lost secondarily in A. lyrata. This might explain their absence in the Arabidopsis gene transpositional history database (that mainly scores pre-α GTD events due to the lack of further Brassicaceae synteny data necessary for scoring of post-α GTDs). Hence, the total fraction of GTD copies among AtGS genes sums up to 27% (18/67) (fig. 2D).

Analysis of GS Genes Not Affected by TD, Ohnolog Retention, or Transposition

In addition, we performed a more in-depth analysis of the three AtGS genes lacking TD, retained At-α ohnolog copy or evidence for transposition during evolution of the Arabidopsis lineage (table 7). Among those, MYB34 is the only locus retaining an ohnologous copy dating back to the At-B WGD event, leaving two putative nonduplicate genes in AtGS pathway inventory: GSH1 and IIL1, functioning in GS cosubstrate pathways and side-chain elongation, respectively (table 3). To confirm the observed evolutionary stability of these genes, we identified syntelogs in V. vinifera (IIL1) and B. rapa (GSH1), respectively. Syntelogs in Vitis proof that IIL1 did not transpose since the birth of the Rosids. Therefore, this gene represents a very ancient unigenie. In case of duplication before Vitis lineage evolution, all copies were lost subsequently before radiation of the Rosid clade. In contrast, GSH1 may be Brassicaceae-specific unigene that likewise lost all duplicates with above-threshold similarity.

Table 5

| α-Block | Protein Name | GO! | Similarity in Col-0 (%) | AabID | Similarity in Aab (%) | Expression change |
|---------|--------------|-----|-------------------------|-------|-----------------------|------------------|
| A05N062 | CYP79C1      | Yes | AT1G79370                | 60.90 | NA                    | 0.7              |
|         | CYP79F2      | Yes | AT1G164000               | 60.90 | NA                    | 0.7              |
| A05N185 | MYB122       | Yes | AT1G74080                | 68.80 | Aab14276              | 0.6              |
|         | MYB51        | Yes | AT1G18570                | 68.80 | Aab19683              | 0.6              |
| A05N186 | SOT17        | Yes | AT1G18590                | 83.50 | Aab19675              | 0.8              |
|         | SOT16        | Yes | AT1G74100                | 83.50 | Aab19675              | 0.8              |
| A10N0A1 | APK1         | Yes | AT2G14750                | 67.50 | Aab32150              | 0.7              |
|         | APK2         | Yes | AT4G39940                | 67.50 | Aab32150              | 0.7              |
| A10N257 | CYP79B3      | Yes | AT2G22330                | 92.10 | Aab19477              | 0.9              |
|         | CYP79B2      | Yes | AT4G39950                | 92.10 | Aab19477              | 0.9              |
| A26N034 | MYB29        | Yes | AT5G07690                | 72.30 | Aab33585              | 0.1              |
|         | MYB28        | Yes | AT5G61420                | 72.30 | Aab12163              | 0.1              |

NOTE.—MeJA, methyljasmonic acid; NA, not applicable.

GS Biosynthetic Gene Identification from Draft Aet. arabicum Genome

On the basis of the Aet. arabicum genome v1.0 and 37,839 annotated genes (Haudry et al. 2013), we identified homologs of A. thaliana loci coding for GS biosynthetic and regulatory genes. Combining reciprocal best BlastP hits with LAST screens for large scale gene colinearity/synteny (100 kb–1.2 Mb) (employed by the Synfind algorithm, see Materials and Methods), we found putative Aet. arabicum orthologs covering 57 of the 64 proposed AtGS genes with an observed nucleotide sequence identity of 45–94% (tables 1, 3, and 4). Among those, seven loci gave rise to an additional 10 further paralogs due to TD and GTD in Aethionema (see later), thereby extending the copy number of six multigene families to a total of 67 putative AabGS genes. The mRNA sequencing data for Aet. arabicum supported the evidence that all 67 putative AabGS genes were expressed (data not shown).

GS Gene Families with Expanded Copy Number in Aethionema

Among the 10 novel paralogs, eight were identified as descendants from TD events. Intriguingly, the Arabidopsis methylthiomalate synthase array MAM1/MAM-L underwent
a further duplication in Aethionema, retaining four MAM-like loci (supplementary fig. S1, Supplementary Material online). Likewise, we observed a further TD of the Arabidopsis benzoate-CoA ligase array BZO1/BZO-like in Aethionema, adding two paralogs to the set of putative AabGS genes (table 8). Notably, both clusters encode functions in GS side-chain elongation (MAM) or cosubstrate pathways (BZO), indicating the connection of TD to metabolic versatility in both Arabidopsis and Aethionema. Furthermore, TD extended the gene inventory of GS core-structure modification in two cases. First, we detected one additional duplicate of the Arabidopsis tau-type glutathione-s-transferase array GSTU19-23 in Aethionema (table 8). Second, we identified extension of the UGT-like superfamily that is present with five members in Arabidopsis and organized in two TARs of distant genomic location (fig. 3B). Intriguingly, both regions represent the sister copies of an-ß or a-02 (fig. 3F). Furthermore, both UGT-like TARs comprise neighboring pairs of the At-α

Table 6

Intraspecies Protein Similarities for At-α Ohnolog Pairs Not Sharing GS Annotation, Shown with Differential Expression in Arabidopsis Following MeJA Treatment

| α-Block | Protein Namea | GO!b | AGI | Similarity in Col-0 (%) | AabIDc | Similarity in Aab (%) | Expression change | Time → | 0.5 | 1 | 3 |
|---------|---------------|------|-----|------------------------|--------|----------------------|------------------|-------|-----|---|---|
| A02N0A2 | AAA04         | Yes  | AT1G04580 | 84.10 | Aab24896 | 79.90 | -1.7 | -0.2 | -0.4 |
|         | [AA03]        | No   | AT2G27150 | 82.90 | Aab27016 | 81.10 | 0.8  | -0.2 | 0.1 |
| A02N051 | UGT74C1       | Yes  | AT2G31790 | 22.30 | Aab37175 | 19.50 | 0.4  | 0.4  | 0.3 |
|         | [UGT-like]    | No   | AT1G05670 | 21.90 | Aab31930 | 22.40 | 0.3  | 0.1  | 0  |
| A02N142 | OBP2          | Yes  | AT1G07640 | 68.50 | Aab18330 | 34.30 | -0.1 | 0.1  | -0.2 |
|         | [OBP-like]    | No   | AT2G28810 | 67.80 | Aab24559 | 66.80 | -0.1 | -0.4 | -0.2 |
| A03N117 | FMO-GSOX-2    | Yes  | AT1G62540 | 75.90 | Aab10869 | 70.60 | -0.3 | 0.3  | 0.4 |
|         | [FMO-like]    | No   | AT1G12130 | 75.10 | Aab13543 | 74.60 | -0.8 | 0.6  | 2.1 |
| A08N074 | BCAT4         | Yes  | AT3G19710 | 72.20 | Aab21007 | 76.80 | 0.1  | 0.2  | 0.5 |
|         | [BCAT7]       | No   | AT1G05090 | 72.10 | Aab22550 | 72.00 | 0.4  | 0   | 0.7 |
| A10N194 | SUR1          | Yes  | AT2G20610 | 84.00 | Aab31155 | 69.90 | 0.6  | 1    | 0.8 |
|         | [SUR-like]    | No   | AT4G24820 | 83.80 | Aab30136 | 82.60 | 0.5  | 0.6  | -2.5 |
| A10N314 | GGPI          | Yes  | AT4G30530 | 84.90 | Aab24374 | 83.40 | 0.8  | 1    | 1.4 |
|         | [GGP-like]    | No   | AT2G23960 | 83.90 | Aab11021 | 82.00 | -0.2 | -0.7 | -0.7 |
| A11N226 | IPMI1         | Yes  | AT3G58990 | 80.60 | Aab13092 | 73.20 | 0.2  | 0.6  | 0.9 |
|         | [IPM-like]    | No   | AT2G43090 | 80.60 | Aab19619 | 80.00 | 0    | 0.1  | 0.2 |
| A12N102 | GSTF11        | Yes  | AT3G03190 | 83.60 | Aab14996 | 41.80 | 1.4  | 1.7  | 1.4 |
|         | [GST12]       | No   | AT5G17220 | 83.60 | Aab14791 | 82.20 | -0.5 | 0.2  | 1.1 |
| A14N046 | IQD1          | Yes  | AT3G09570 | 67.10 | Aab18852 | 52.60 | -0.2 | -0.3 | -0.1 |
|         | [IQD2]        | No   | AT5G03040 | 67.10 | Aab18368 | 52.50 | -0.2 | -0.2 | -0.4 |
| A19N002 | BCAT3         | Yes  | AT3G49860 | 8.20  | Aab32782 | 79.90 | 0    | 0.1  | 0  |
|         | [BCATS]       | No   | AT5G65780 | 8.20  | Aab32605 | 80.00 | -0.1 | -0.2 | 0  |
| A20N095 | BAT5          | Yes  | AT4G12030 | 78.90 | Aab32285 | 63.30 | 0.4  | -0.1 | 0.4 |
|         | [BAT-like]    | No   | AT4G22840 | 78.90 | Aab32321 | 63.30 | 0    | 0    | 0.6 |

Ø 67.52 | Ø 62.10

Table 7

Putative Single-Copy Genes in Aethionema and Arabidopsis GS Pathway Inventory

| AGI     | Name   | α-Block | Retained At-ß/γ Ohnolog | Most Ancient Syntelog | Closest Paralog | BlastP e Value | Name   | α-Block | Retained At-ß/γ Ohnolog |
|---------|--------|---------|-------------------------|-----------------------|-----------------|----------------|--------|---------|-------------------------|
| AT4G13430 | IIL1 | A15 | No | Vitis vinifera | AT4G26970 | 1.00E–16 | AC02 | A22N121 | No |
| AT4G23100 | GSH1 | A20 | No | Brassica rapa | AT1G193220 | 0.19 | ARF1 | A05 | No |
| AT5G60890b | MYB34 | — | B20N001 | V. vinifera | AT1G74080 | 4.00E–62 | MYB122 | A05N185 | B20N004 |

aGene transpositional duplicate copy.

bAbsent in Aethionema.
Table 8
Tandem Duplicate Genes in Arabidopsis and Aethionema GS Pathway Inventory

| Protein Name | AGI   | Block | AabID | Syntelog | % Identity | Lineage Specific | Col-0 → Aab |
|--------------|-------|-------|-------|----------|------------|-----------------|--------------|
| UGT74C1      | AT2G31790 | A02N051 | Aab37175 | Yes | 79.44 | No | 6 → 10 |
| [UGT-like]   | AT1G05670 | A02N051 | Aab37178 | Yes | 82.05 | Aethionema | 6 → 10 |
| UGT-like_oa | AT1G05675 | A02 | Aab37179 | Yes | 77.63 | Aethionema | 6 → 10 |
| UGT74E2_oa  | AT1G05680 | A02N053 | Aab37180 | Yes | 78.95 | Aethionema | 6 → 10 |
| FMO-GSOX-2  | AT1G62540 | A03N117 | Aab10869 | Yes | 76.6 | No | 11 → 8 |
| FMO-GSOX-3  | AT1G62560 | A03 | Aab10867 | Yes | 71.9 | No | 11 → 8 |
| FMO-GSOX-4  | AT1G62570 | A03 | Aab10866 | Yes | 55.2 | No | 11 → 8 |
| FMO-like_oa | AT1G62580 | A03 | — | — | — | (no GS gene) | 10 → 7 |
| FMO-like_oa | AT1G62600 | A03 | — | — | — | (no GS gene) | 10 → 7 |
| FMO-like_oa | AT1G62620 | A03 | — | — | — | (no GS gene) | 10 → 7 |
| [FMO-like]   | AT1G12130 | A03N117 | Aab13543 | Yes | 65.09 | No | 11 → 8 |
| FMO-like_oa | AT1G12140 | A03 | Aab13546 | Yes | 71.9 | No | 11 → 8 |
| CYP79F2      | AT1G62540 | A03N117 | Aab13549 | Yes | 66.66 | (no GS gene) | 10 → 7 |
| CYP79F1      | AT1G62540 | A03 | Aab14278 | Yes | 91.07 | No | 3 → 3 |
| GSTF10       | AT2G30870 | A01 | Aab28612 | Yes | 85.99 | No | 3 → 3 |
| GSTF9_oa     | AT2G30860 | A02 | Aab19619 | Yes | 87.6 | No | 5 → 5 |
| CYP79C2      | AT1G62540 | A03 | Aab23605 | Yes | 75.3 | No | 6 → 6 |
| IPMI-like_oa | AT2G30900 | A01 | Aab19630 | Yes | 85.99 | No | 3 → 3 |
| IPMI2        | AT2G31000 | A01 | Aab23605 | Yes | 75.3 | No | 3 → 3 |
| [BCAT7]      | AT1G60090 | A03N104 | Aab22548 | Yes | 69 | No | 6 → 6 |
| [BCAT7]      | AT1G60090 | A03 | Aab22550 | Yes | 78.12 | (no GS gene) | 5 → 5 |
| [IPMI-like]  | AT2G30900 | A01 | Aab19619 | Yes | 87.6 | No | 5 → 5 |
| [IPMI-like]  | AT2G30900 | A01 | Aab23605 | Yes | 75.3 | No | 5 → 5 |
| GSTF9_oa     | AT2G30860 | A02 | Aab19619 | Yes | 85.99 | No | 3 → 3 |
| CYP79C2      | AT1G62540 | A03 | Aab19630 | Yes | 87.6 | No | 3 → 3 |
| [BCAT7]      | AT1G60090 | A03 | Aab23605 | Yes | 75.3 | No | 6 → 6 |
| CYP79C2      | AT1G62540 | A03 | Aab19630 | Yes | 87.6 | No | 3 → 3 |
| [BCAT7]      | AT1G60090 | A03 | Aab23605 | Yes | 75.3 | No | 6 → 6 |

Tandem duplicates of syntenic anchor genes retaining an At-α ohnolog

AOP1       | AT4G03070 | A01 | Aab37231 | Yes | 70.03 | Arabidopsis | 2 → 1 |
AOP3       | AT4G03050 | A01 | — | — | — | Arabidopsis | 2 → 1 |
GSTF10     | AT2G30870 | A02 | Aab28612 | Yes | 91.59 | No | 4 → 4 |
GSTF9_oa   | AT2G30860 | A02 | Aab28613 | Yes | 89.76 | (no GS gene) | 4 → 4 |
CYP79C2    | AT1G62540 | A03 | Aab17711 | Yes | 71.85 | No | 8 → 9 |
CYP-like_oa| AT1G62620 | A03 | Aab17712 | Yes | 60.71 | (no GS gene) | 8 → 9 |
UGT74B1    | AT1G24100 | A05 | Aab07827 | Yes | 80.65 | Aethionema | 6 → 10 |
MYB29      | AT5G07690 | A02N034 | Aab33585 | Yes | 65.13 | Arabidopsis | 6 → 4 |
MYB76      | AT5G07700 | A02 | — | — | — | Arabidopsis | 6 → 4 |

(continued)
sequences (table 8). Therefore, the diversity of draft genome (tables 1, 3, and 4), leading to the contraction of six multigene families and loss of one single-copy gene.

Four of those loci, namely AOP3, CYP79F2, IMD3, and MYB76, are likewise absent in the B. rapa genome and may therefore be specific to A. thaliana and more closely related species (Wang et al. 2011a). AOP1/AOP3 and CYP79F1/2 represent two neighboring TARs with evident subfunctionalization in Arabidopsis (Kliebenstein et al. 2001; Prasad et al. 2012). Although AOP3 functions in GS side-chain elongation in Arabidopsis (Kliebenstein et al. 2001), CYP79F2 encodes an enzyme involved in core structure formation of long-chain aliphatic GS. Furthermore, overexpression of the MYB76 transcription factor correlates with increased levels of both long-chained and short-chained aliphatic GS in Arabidopsis (Gigolashvili et al. 2008). However, experiments with Arabidopsis myb76 T-DNA insertion lines to date failed to show any significant change in GS chemotype, making a strict requirement of MYB76 for GS biosynthesis unlikely (Gigolashvili et al. 2008). Moreover, IMD3 encodes a predicted enzyme with proposed functional redundancy to (as well as strong coexpression with) IMD1, a protein that was shown to be involved in GS accumulation in Arabidopsis (Hirai et al. 2007; Wentzell et al. 2007; Gigolashvili et al. 2009; Sawada et al. 2009; He et al. 2010). Therefore, absence of IMD1 (IMD3) in B. rapa (Aet. arabicum) supports the hypothesized capability of mutual phenotype rescue among IMD1/3 double knock-outs in Brassicaceae, eventually preventing significant alterations of GS chemotype due to fractionation of IMD-like genes in Aethionema.

The other three of the seven AtGS loci that lack a clear ortholog in Aethionema are not found in the B. rapa genome: MYB34, CYP81F2, and GSL-OH (tables 3 and 4). Therefore, they represent Aethionema lineage-specific gene losses. MYB34 was shown to control indolic GS biosynthesis in Arabidopsis (Celenza et al. 2005). Interestingly, overexpression of MYB34 in Arabidopsis partially rescued the altered GS chemotype caused by MYBS1 knockout (Gigolashvili et al. 2007).

| Protein Name | AGI    | $\pi$-Block | AabID$^a$ | Syntelog | % Identity$^a$ | Lineage Specific? | Col-0 $\rightarrow$ Aab$^b$ |
|--------------|--------|-------------|-----------|-----------|---------------|------------------|--------------------------|
| CYP71B10 _oa| AT5G57260 | A22         | Aab25774  | Yes       | 73.21         | (no GS gene)    | 2 $\rightarrow$ 1     |
| BZO1         | AT1G65880 | A25         | Aab31601  | Yes       | 70.04         | No               | 2 $\rightarrow$ 4     |
| BZO-like _oa | AT1G65890 | A25         | Aab31602  | Yes       | 69.4          | Aethionema       | 2 $\rightarrow$ 4     |
|              |         |             | Aab31603  | Yes       | 68.85         | (no GS gene)    | 2 $\rightarrow$ 4     |
|              |         |             | Aab31604  | Yes       | 67.83         | (no GS gene)    | 2 $\rightarrow$ 4     |

Tandem duplicates of genes outside the boundaries of $\pi$-blocks

MAM1

| AT5G23010  | —       | Aab12229    | Yes       | 72.31 | No   | 2 $\rightarrow$ 4 |
| A1B23020   |         | Aab12225    | Yes       | 70.67 | No   | 2 $\rightarrow$ 4 |
| AT5G23020  | —       | Aab12226    | Yes       | 68.36 | Aethionema | 2 $\rightarrow$ 4 |

AtGS genes:

45% TD (29/64) 46% TD (31/67)

$^a$Squared brackets indicate ohnolog copies of GS biosynthetic genes without GO! annotation to GS biosynthetic process. The "_oa" suffix indicates tandem duplicate copies without GO! annotation to GS biosynthetic process. These genes have not been considered for the tandem duplicate count of GS loci in both organisms.

$^b$Underlined items refer to offspring of one pre-$\pi$ TD event.

$^c$Predicted Aethionema CDS.

$^d$In case of Aethionema-specific TAR expansion, the corresponding Arabidopsis sequence for identity comparison was determined based on both genomic location and homology criteria.

$^e$Change of gene family locus count in Arabidopsis $\rightarrow$ Aethionema order.

Aethionema GS Loci without Orthologs in Aethionema

In seven cases, RBH- and synteny-based evidence was not sufficient to clearly assign orthologs to AtGS loci in the Aet. arabicum draft genome (tables 1, 3, and 4), leading to the contraction of six multigene families and loss of one single-copy gene.

Arabidopsis GS Loci with Orthologs in Aethionema

In seven cases, RBH- and synteny-based evidence was not sufficient to clearly assign orthologs to AtGS loci in the Aet. arabicum draft genome (tables 1, 3, and 4), leading to the contraction of six multigene families and loss of one single-copy gene.

Table 8 Continued

| Protein Name | AGI    | $\pi$-Block | AabID$^a$ | Syntelog | % Identity$^a$ | Lineage Specific? | Col-0 $\rightarrow$ Aab$^b$ |
|--------------|--------|-------------|-----------|-----------|---------------|------------------|--------------------------|
Because of functional redundancy of MYB51/34, the loss of MYB34 likely does not cause significant changes in GS chemotype in *Aethionema* lineages. In contrast, *CYP81F2* and *GSL-OH* encode functions associated with secondary modification of the GS core structures (Sonderby et al. 2010). *CYP81F2* has been shown to control the indole GS modifier1 quantitative trait loci in *Arabidopsis*, catalyzing the conversion of indole-3-yl-methyl GS to 4-hydroxy-indole-3-yl-methyl GS (Pfalz et al. 2009). Notably, these metabolites play a significant role in MAMP-triggered immunity in *Arabidopsis* (Bednarek et al. 2009). Among various known cytochrome p450s active in indolic GS biosynthesis, *CYP81F2* is the only locus impairing callose deposition after detection of the non-self-infection in *Arabidopsis* (Clay et al. 2009). On the basis of these findings, we concluded a *CYP81F2*-specific onset of sub/neofunctionalization from GS biosynthesis toward plant innate immunity after divergence of the *Arabidopsis* and *Aethionema* lineages, thereby mitigating fatal consequences of the absent ortholog for GS chemotype in *Aethionema*.

Moreover, we determined the absence of the 2-oxoacid-dependent oxygenase activity GSL-OH in *Aethionema* (table 3). GSL-OH is necessary for biosynthesis of 2-hydroxybut-3-enyl GS in *Arabidopsis* (Wentzell et al. 2007; Hansen et al. 2008) and present in the *B. rapa* genome (Wang et al. 2011a). Noteworthy, GSL-OH was the only multigene family member within the extended AtGS set whose loss in *Aethionema* was not accompanied by copy number expansion of one or more paralogs (table 3). Accordingly, we concluded an *Aethionema*-specific loss of this locus after divergence from the *Arabidopsis* lineage, creating measurable differences in GS chemotype among *Arabidopsis* and *Aethionema*. Consistently, we could not detect traces of 2-hydroxybut-3-enyl GSs in *Aet. arabicum* root-, leaf-, and seed extract using uHPLC (data not shown).

---

**Fig. 4.**—Phylogenetic relationships among FMO proteins. Col-0, Aab, and papaya refer to *Arabidopsis thaliana* (circles), *Aethionema arabicum* (triangles), and *Carica papaya* (colorless), respectively. Boxes indicate annotation to GS metabolic activity in *Arabidopsis*. *Tarenaya hasslerania* (diamonds) represents the closest-related outgroup of Brassicaceae, including the sister clade Aethionemae. Stars: At-α WGD event. Blue: proteins encoded by members of TARS. Red: protein encoded by GTD locus. The At-α WGD leads to duplication of a FMO locus, resulting in two clades comprising all FMO-like sequences of both *Aethionema* and *Arabidopsis*. Thus, FMO versatility has been promoted by a combination of increased degree of ohnolog retention and TD events.
GS Gene Families with Lower Copy Number in Aethionema

Considering the Aethionemae-specific loss of GSL-OH, six putative GS-annotated multigene families display a lower copy number in Aethionema (AOPx, CYP79x, CYP81x, GSLx, IMDx, and MYBx) (tables 1, 3, 4, and 8). In sum, their total gene count increased from 20 genes in Aethionema to 27 observed in Arabidopsis, thereby mediating a 35% increase (tables 1, 4, and 8).

Although IMD3 and GSL-OH possess GTD copies in Arabidopsis, these loci are absent in Aethionema. In contrast, AOP12, IMD13, MYB29/76, and CYP81F2 with its CYP-like neighbor AT1G58265 comprise TARs in Arabidopsis but are likewise absent in Aethionema (table 8) and B. rapa (Wang et al. 2011a).

Therefore, the underlying TD events may be Arabidopsis specific. Thus, TD facilitated GS pathway expansion in the Arabidopsis lineage after split from the tribe Aethionemae.

Furthermore, we found evidence for Arabidopsis-specific TD of three neighboring FMO-like loci (FMO GS-OX2_{2,3}) (figs. 3 and 4), leading to lower copy number in Aethionema. These genes were lost in B. rapa (Wang et al. 2011a), illustrating a degree of plasticity across Brassicaceae. FMO-like loci comprise a multigene family with five members annotated to GS biosynthesis in Arabidopsis mapping to three distant genomic locations (fig. 3B). Among those, two regions are embedded in ohnolog copies of α-block A02 (fig. 3E) and contain the retained α-pair of AT1G2540 (FMO GS-OX2) and AT1G12130 (FMO-like) (fig. 3B). The latter is not annotated to GS biosynthesis in Arabidopsis. However, AT1G12130 is member of a FMO-like 4-gene TAR with its 3′ neighbor FMO GS-OX5 involved in aliphatic GS biosynthesis (Li et al. 2008). The third genomic region in Arabidopsis harboring a FMO-like sequence with encoded function in GS metabolism is defined by AT1G65860 (FMO GS-OX1), representing a transposed duplicate gene copy (fig. 4). Interestingly, FMO GS-OX1_{1,4} share broad substrate specificity and catalyze the conversion from methythioalkyl GS to the related methylsulfinylalkyl GS independent of chain length. In contrast, FMO GS-OX5 shows substrate specificity for 8-methylthiooctyl GS (Hansen et al. 2007; Li et al. 2008). This example is similar to the case of UGT-like loci (see earlier) and again illustrates the combination of ohnolog retention with tandem- and GTD leading to increased GS pathway versatility in Brassicaceae (fig. 4).

Deduction of Total Duplicate Frequencies in Aethionema and Arabidopsis GS Pathway Inventory and Comparison to Arabidopsis Genome-Wide Average

We found the fraction of retained ohnolog duplicate gene pairs among Arabidopsis (56%) and Aethionema (52%) GS biosynthetic and regulatory genes significantly increased compared with the genome-wide average in Arabidopsis (22%) (fig. 2, table 9). Moreover, 46% (31/67) of AabGS genes are organized in TARs (fig. 2F), compared with a 22% average for all protein-coding genes in Arabidopsis, thereby significantly surpassing the TAR coverage rate of 45% (29/64) observed for AtGS loci (fig. 2B and D, table 9).

For duplication by gene transposition, we detected 27% (17/67) of affected AabGS loci (fig. 2). In summary, we found no significant enrichment of GTD events among GS pathway inventory in both species (table 9).

Table 9

| Protein-coding genes | Arabidopsis Genome | Arabidopsis GS Genes \(^b\) | Aethionema GS Genes |
|----------------------|--------------------|--------------------------|-------------------|
| 27,206               | 64                 | 67                       |
| Retained At-α ohnologs | 6,038/22\%     | 36/56\%                 | 35/52\%           |
| P value              | 3.87E–09          | 3.24E–07                 | 1.26E–07          |
| Tandem duplicates    | 4,022/15\%        | 29/45\%                 | 32/48\%           |
| P value              | 5.71E–09          | 9.14E–10                 |                   |
| GTDs                 | 3,879/14\%        | 17/27\%                 | 18/27\%           |
| P value              | 0.01066           | 0.007462                 |                   |
| Sum duplicates       | 12,132/45\%       | 61/95\%                 | 65/97\%           |
| P value              | 2.20E–16          | 2.20E–16                 |                   |

\(^a\) Fisher’s exact test on count data.
\(^b\) Extended set (fig. 1).

Discussion

The Aethionemae/Brassicaceae crown-group/sister-group lineages split about 30–60 Ma shortly after the last common WGD event and independently evolved ever since (Couvreur et al. 2010; Mithen et al. 2010; Schranz et al. 2011). Notably, the radiation process evident for the Brassicaceae lineage created about 3,700 species (Couvreur et al. 2010). In contrast, the species-poor Aethionemae lineage (Schranz et al. 2012) is well established as most ancient Brassicaceae extant sister and may therefore possess a more “ancient” genome organization when compared with Arabidopsis. This facilitates the recognition and quantification of common factors underlying rapid innovation of complex traits shared by both species. We exploit novel genomics resources for evolutionary analysis of the complete GS pathway inventory in both A. thaliana and Aet. arabicum to utilize the impact of different kinds of duplication classes to diversification of plant secondary metabolites.

In a comparative genomics approach, we employ the phylogenetic relation of A. arabicum and A. thaliana to identify key factors driving GS pathway divergence. In this context, we establish GSs genetics/genomics as a scaffold to incorporate further phenotypic data for better understanding the impact of duplication to rapid evolution of novel key traits. In
Arabidopsis, several GS genes retained duplicate gene copies dating back to the last WGD event but lacking annotation to GS metabolic processes (fig. 1). Illustrating high degrees of protein similarities among these ohnolog copy pairs and/or similar responses in gene regulation following GS pathway induction (tables 5 and 6), we identified 12 novel putative Arabidopsis genes associated to GS biosynthesis (fıgs. 1 and 3). Given the fact that these loci remained unknown despite their putative relevance for an experimentally very well-studied trait-like GS biosynthesis, we highlight the importance of considering ohnolog copies when analyzing a plethora of other highly diverged multigene pathways (i.e., terpenoid biosynthesis). We thereby provided an easy-to-follow framework on how to use existing data on WGD in Arabidopsis to better understand the networks of functional redundancy, especially involving genes that are targeted for knock-out experiments in functional studies.

Evolutionary analysis of homologous GS loci in Arabidopsis and Aethionema found a majority (all but two) comprising duplicate groups organized in multigene families (fıgs. 2 and 3). This underlined the dominant role of duplication for creation and expansion of biochemical diversity in plant (secondary) metabolism.

Clear orthologs of seven Arabidopsis GS genes are absent in the Aethionema draft genome (due to three Aethionema-specific GS gene losses and four Arabidopsis-specific TDs). Evolution of 10 additional Aethionema paralogs (two due to gene transposition and eight due to TD events, fig. 3) lead to an almost 100% conserved GS pathway inventory across the crown group/sister group system. This sheds light upon the relevance of genome plasticity for key trait maintenance despite of scattered gene losses. To test this hypothesis, we indicate the requirement of further research on additional multigene pathways in a deeper phylogenetic resolution. Identification of Aethionema GS gene homologs allowed confirming the increased frequency of duplicates in lineages that diverged more than 30–60 Ma. The absence of lineage-specific polyploidy events in either species facilitated the comparative analysis of genes duplicated due to the common ancient WGD events (particularly At-α) as well as lineage-specific gene tandem and transposition duplications. Partitioning the duplicate genes set in GS pathway inventory revealed significant enrichments of retained At-α ohnologs and tandem duplicates (but not GTD events) in both species compared to the average observed for protein-coding genes in Arabidopsis (table 2). We therefore conclude that WGD and TD facilitated the early and continued evolution of GS biosynthesis in the mustard family. To our knowledge, this is the first study providing distinct indications on a genetics level for the connection of WGD to the emergence of key traits in planta.

Various duplicates of different GS gene families code for proteins encoding functions in consecutive steps of GS biosynthesis (Kliebenstein et al. 2001; Hansen et al. 2007). Among GS biosynthetic and regulatory genes, pairs of retained At-α ohnolog duplicates in distant genomic location further expand to TARs (fig. 3). In Arabidopsis, the S-oxygenase activity FMO is provided by a pair of retained ohnologs on distant arms on chromosome 1 (fıgs. 3 and 4). Both copies evolved further tandem duplicates with different substrate specificities (Li et al. 2008). Different groups of substrates are products of SOT-type sulfotransferases provided by another retained ohnolog pair on At1 with additional TD copies sharing annotation to GS production (Bowers et al. 2003; Piotrowski et al. 2004) (fig. 3). The reaction delivering substrates for GS SOT-type sulfoconjugates is catalyzed by UGT-type proteins, likewise encoded by a pair of retained ohnologs that evolved multiple tandem and gene transposition duplicates in both Aethionema and Arabidopsis (fig. 3). It is thus inferred that sub-functionalization of both TD and retained At-α ohnolog pairs caused functional diversification of GS biosynthetic and regulatory elements. Showing mutual influence of ohnolog retention and TD rate across a crown group–sister group system, we describe a complex network of gene duplication fostering the expansion of a composite trait, thereby contributing to the means of mutation and selection to create evolutionary innovation in a limited timeframe. Evidence of the model of evolution by gene duplication can be found in comparative GS pathway analysis. Thus, GS may provide a framework for investigating the expansion of complex traits.

Supplementary Material

Supplementary figures S1 and S2 and table S1 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org).

Acknowledgments

I am grateful for the input and support of Erik van den Bergh, Nicole van Dam, Mike Freeling, Tom Mitchell-Olds, Benjamin Schweißinger, Cyril Zipfel, and Martin Parniske. Thanks to Chiara Vercesi for her assistance with graphical editing of the figures. Likewise, I want to acknowledge the contributions of two anonymous reviewers. This work was supported by a Netherlands Organization for Scientific Research (NWO) Ecogenomics grant to M.E.S.

Literature Cited

Bednarek P, et al. 2009. A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. Science 323:101–106.

Beekwilder J, et al. 2008. The impact of the absence of aliphatic glucosinolates on insect herbivory in Arabidopsis. PLoS One 3:e2068.

Bellény-Rabelo D, Oliveira AE, Venancio TM. 2013. Impact of whole-genome and tandem duplications in the expansion and functional diversification of the F-box family in legumes (Fabaceae). PLoS One 8:e55127.

Bones AM, Rossiter JT. 2006. The enzymic and chemically induced decomposition of glucosinolates. Phytochemistry 67:1053–1067.
Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

Celenza JL, et al. 2005. The Arabidopsis ATR1 Myb transcription factor controls indolic glucosinolate homeostasis. Plant Physiol. 137:253–262.

Cheng S, et al. 2013. The Tarenaya hassleriana genome provides insight into reproductive trait and genome evolution of crucifers. Plant Cell 25:2813–2830.

Cheong JJ, Choi YD. 2003. Methyl jasmonate as a vital substance in plants. Trends Genet. 19:409–413.

Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM. 2009. Glucosinolate metabolites required for an Arabidopsis innate immune response. Science 323:95–101.

Couvreur TL, et al. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol Biol Evol. 27:55–71.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 20:591–597.

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

Fawcett JA, Maere S, Van de Peer Y. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:435–453.

Freeling M. 2009. Gene-balanced duplications, like tetra-ploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

Freeling M, et al. 2008. Many or most genes in Arabidopsis transposed after the order of the Brassicales. Genome Res. 18:1924–1937.

Gigolashvili T, Engqvist M, Yatusevich R, Muller C, Flugge UI. 2008. HAG2/MYB29 modulates the ERF gene family in Arabidopsis thaliana. Mol Genet Genomics. 277:53–64.

Gigolashvili T, et al. 2007. The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in Arabidopsis thaliana. Plant J. 50:886–901.

Gigolashvili T, et al. 2009. The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana. Plant Cell 21:1813–1829.

Guinond S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hansen BG, Klebenstein DJ, Halkier BA. 2007. Identification of a flavin-monoxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in Arabidopsis thaliana. New Phytol. 177:627–642.

Huang CR, Burns KH, Boeke JD. 2012. Active transposition in genomes. Annu Rev Genet. 46:651–675.

Hughes AL, Friedman R, Ekollu V, Rose JR. 2003. Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. Mol Phylogenet Evol. 29:410–416.

Hirai MY, et al. 2007. Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci U S A. 104:6478–6483.

Huang CR, Burns KH, Boeke JD. 2012. Active transposition in genomes. Annu Rev Genet. 46:651–675.

Irish VF, Litt A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. Curr Opin Genet Dev. 15:454–460.

Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Kane J, Freeing M, Lyons E. 2010. The evolution of a high copy gene array in Arabidopsis. J Mol Evol. 70:531–544.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kliebenstein DJ. 2008. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. PLoS One 3:e1838.

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

Krommann J, Donnerhacker S, Schnabelrauch D, Mitchell-Olds T. 2003. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci U S A. 100(2 Suppl):14587–14592.

Krywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet. 20:116–122.

Li J, Hansen BG, Ober JA, Klebenstein DJ, Halkier BA. 2008. Subclade of flavin-monoxygenases involved in aliphatic glucosinolate biosynthesis. Plant Physiol. 148:1721–1733.

Lyons E, Freeing M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53:661–673.

Malacarne G, et al. 2012. Deconstruction of the (kale)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. PLoS One 7:e29762.

Mithen R, Bennett R, Marquez J. 2010. Glucosinolate biochemical diversity and innovation in the Brassicales. Phytochemistry 71:2047–2086.

Nakajima M, Yoshida R, Shimada N, Yamazaki H, Yokoi T. 2001. Inhibition and inactivation of human cytochrome P450 enzymes by phenethyl isothiocyanate. Drug Metab Dispos. 29:1110–1113.

Nakano T, Suzuki K, Fujimura T, Shinshi H. 2006. Genome-wide analysis of the ERF gene family in Arabidopsis and rice. Plant Physiol. 140:411–422.

Naur P, et al. 2003. CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes metabolizing oximes in the biosynthesis of glucosinolates in Arabidopsis. Plant Physiol. 133:63–72.

Ohno S. 1970. Evolution by gene duplication. New York: Springer Publishing Group, p. 160.

Pennings M, et al. 1999. Homologues of the Cf-9 disease resistance gene (Hcr9s) are present at multiple loci on the short arm of tomato chromosome 1. Mol Plant Microbe Interact. 12:93–102.

Pfalz M, Vogel H, Krommann J. 2009. The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole
glucosinolate structures and aphid resistance in Arabidopsis. Plant Cell 21:985–999.
Piotrowski M, et al. 2004. Desulfoglucosinolate sulfotransferases from Arabidopsis thaliana catalyze the final step in the biosynthesis of the glucosinolate core structure. J Biol Chem. 279:50717–50725.
Prasad KV, et al. 2012. A gain-of-function polymorphism controlling complex traits and fitness in nature. Science 337:1081–1084.
Rask L, et al. 2000. Myrosinase: gene family evolution and herbivore defense in Brassicaceae. Plant Mol Biol. 42:93–113.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.
Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. PLoS Comput Biol. 2:e115.
Rodman J. 1998. Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. Am J Bot. 85:997.
Rodman JE, Karol KG, Price RA, Sytsma KJ. 1996. Molecules, morphology, and Dahlgren’s expanded order capparales. Syst Bot. 21:289–307.
Roth C, et al. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol. 308:58–73.
Sawada Y, et al. 2009. Omics-based approaches to methionine side chain elongation in Arabidopsis: characterization of the genes encoding methythioalkylmalate isomerase and methythioalkylmalate dehydrogenase. Plant Cell Physiol. 50:1181–1190.
Schránz ME, Edger PP, Fries JC, van Dam NM, Wheat CW. 2011. Comparative genomics in the Brassicales: ancient genome duplications, glucosinolate diversification and pterinae herbivore radiation. In: Edwards D, Batley J, Parkin I, Kole C, editors. Genetics, genomics and breeding in crop plants. Boca Raton (FL): CRC press. p. 206–218.
Schránz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr Opin Plant Biol. 15:147–153.
Sonderby IE, Geu-Flores F, Halkier BA. 2010. Biosynthesis of glucosinolates—gene discovery and beyond. Trends Plant Sci. 15:283–290.
Steife F, Brandt W, Schmidt J, Milikowski C, Strack D. 2008. Activities of Arabidopsis sinapoylglucose: malate sinapoyltransferase shed light on functional diversification of serine carboxypeptidase-like acyltransferases. Phytochemistry 69:1826–1831.
Tang H, Lyons E. 2012. Unleashing the genome of Brassica rapa. Front Plant Sci. 3:172.
Textor S, de Kraker JW, Hause B, Gershenzon J, Tokuhisa JG. 2007. MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. Plant Physiol. 144:60–71.
Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.
Vlad D, Rappaport F, Simon M, Loudet O. 2010. Gene transposition causing natural variation for growth in Arabidopsis thaliana. PLoS Genet. 6:e1000945.
Wang H, et al. 2011a. Glucosinolate biosynthetic genes in Brassica rapa. Gene 487:135–142.
Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in Arabidopsis. PLoS Genet. 9:e1003255.
Wang X, et al. 2011b. The genome of the mesopolyploid crop species Brassica rapa. Nat Genet. 43:1035–1039.
Wang Y, et al. 2011c. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS One 6:e28150.
Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57–63.
 Wentzell AM, et al. 2007. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 3:1687–1701.
 Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. 20:1229–1237.
 Windsor AJ, et al. 2005. Geographic and evolutionary diversification of glucosinolates among near relatives of Arabidopsis thaliana (Brassicaceae). Phytochemistry 66:1321–1333.
 Wittstock U, Kliebenstein DJ, Lambrix V, Reichelt M, Gershenzon J. 2003. Glucosinolate hydrolysis and its impact on generalist and specialist insect herbivores. In: Romeo JT, editor. Integrative phytoc hemistry: from ethnobotany to molecular ecology. Amsterdam (The Netherlands): Elsevier. p. 101–126.
 Woodhouse MR, Tang H, Freeling M. 2011. Different gene families in Arabidopsis thaliana transposed in different epochs and at different frequencies throughout the rosid. Plant Cell 23:4241–4253.

Associate editor: Yves Van De Peer