Sciencexpress Report

Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Plants

Ben Field and Anne E. Osbourn*

Department of Metabolic Biology, John Innes Centre, Colney Lane, Norwich, NR4 7UH, UK.

*To whom correspondence should be addressed. E-mail: anne.osbourn@bbsrc.ac.uk

Operons are clusters of unrelated genes with related functions that are a feature of prokaryotic genomes. Here we report on an operon-like gene cluster in the plant *Arabidopsis thaliana* that is required for triterpene synthesis (the thalianol pathway). The clustered genes are co-expressed, as in bacterial operons. However, despite the resemblance to a bacterial operon, this gene cluster has been assembled from plant genes by gene duplication, neo-functionalization and genome reorganization rather than by horizontal gene transfer from bacteria. Furthermore, recent assembly of operon-like gene clusters for triterpene synthesis has occurred independently in divergent plant lineages (*Arabidopsis* and oat). Thus selection pressure may act during the formation of certain plant metabolic pathways to drive gene clustering.

Triterpenes protect plants against pests and diseases and are also important drugs and anticancer agents (*1–4*). Like sterols, these compounds are synthesized from the isoprenoid pathway by cyclization of 2,3-oxidosqualene (*1*, *3*). The *Arabidopsis* genome contains thirteen predicted oxidosqualene cyclase (OSC) genes (*3*, *5*). Of these, one encodes cycloartenol synthase (CAS), which is required for sterol biosynthesis, and another encodes lanosterol synthase (LAS), which is conserved across the eudicots and whose function in plants is unknown (Fig. 1A). The eleven remaining OSCs fall into two major clades (I and II) (Fig. 1A). These OSCs produce various triterpenes when expressed in yeast. However their function in *Arabidopsis* is unknown. The OSCs in clade I have close homologs in other eudicots, while those in clade II appear to be restricted to the Brassicaceae family, and show homology to a single OSC from *Brassica napus*.

Oat (*Avena* spp.), a monocot which diverged from the eudicots ~ 180 million years ago, produces defense-related triterpenes known as avenacins. The first committed step in avenacin synthesis is catalyzed by the OSC β-amyrin synthase (encoded by *Sad1*) (*6*). *Sad1* is hypothesized to have arisen from a duplicated monocot cycloartenol synthase-like gene after the separation of wheat and oat ~25 million years ago (*6*, *7*). The second step in avenacin biosynthesis is mediated by SAD2, a member of the newly described monocot-specific CYP51H subfamily of cytochrome P450 enzymes (CYP450s) (*8*). *Sad1* and *Sad2* are embedded in a gene cluster that includes genes required for acylation, glucosylation and other steps in the pathway (*2*, *7*). The avenacin biosynthesis genes are tightly regulated and expressed only in the root epidermis, the site of accumulation of the pathway end-product (*6*, *8*). The avenacin gene cluster lies within a region of the oat genome lacking synteny with rice and other cereals (*7*).

We examined the genomic regions around each of the 13 *Arabidopsis* OSC genes in the *Arabidopsis* genome to establish whether genes for triterpene synthesis might be clustered (*27*). Four OSC genes are flanked by genes predicted to encode other classes of enzymes implicated in secondary metabolism. These four OSCs all belong to clade II, which appears to have undergone accelerated evolution compared to other *Arabidopsis* OSCs (Fig. 1A). We focused on a region containing four contiguous genes predicted to encode an OSC (*At5g48010*), two CYP450s (*At5g48000* and *At5g47990*) and a BAHD family acyltransferase (*At5g47980*) (Fig. 1B). The expression of all four genes is highly correlated (Fig. 1C) and occurs primarily in the root epidermis (fig. S1), suggesting that the genes are functionally related (*9*).

The OSC gene within this region, *At5g48010*, converts 2,3-oxidosqualene to the triterpene thalianol when expressed in yeast (*10*). However thalianol has not been reported in plants. We detected low levels of thalianol in roots but not leaves of wild type *Arabidopsis* (Fig. 2C, 2D), consistent with the expression of *At5g48010*. Thalianol was not detectable in root extracts of a null insertion mutant of *At5g48010* (*thas1-1*) (Fig. 2E) indicating that the *At5g48010* gene product [hereafter named thalianol synthase (THAS)] is required for synthesis of thalianol in *Arabidopsis* roots. Overexpression of THAS in *Arabidopsis* resulted in thalianol accumulation in leaves (Fig. 2F) and in dwarfing (Fig. 3A). Mutations in the gibberellin (GA), brassinosteroid (BR) and primary sterol pathways can result in dwarfing (*11–13*). However the dwarf phenotype of the THAS-overexpressing lines was not rescued

by application of GA or BR and the sterol content of these plants was not significantly altered (P>0.05, Student's t-test, n=3), suggesting that thalianol may be detrimental to plant growth.

GC-MS analysis of wild type root extracts revealed additional peaks that were absent in *THAS* mutants (Fig. 4A; wild type and *thas1-1* panels). Since these peaks were dependent on *THAS*, this suggested that thalianol (**1**) is converted to unknown downstream products in wild type *Arabidopsis* plants (peaks **2a, 2b, 2c, 3a** and **3b**). Therefore, the co-expressed genes adjacent to *At5g48010* (*THAS*) were examined to determine if they function in thalianol modification. We analyzed T-DNA insertion mutants in *At5g48000*, which is immediately adjacent to *THAS*. *At5g48000* is predicted to encode a CYP450 (CYP708A2) belonging to the functionally uncharacterized CYP708 family, a CYP450 family specific to the Brassicaceae (*14*). Root extracts of mutants affected at *At5g48000* (*thah1-1* and *thah1-2*) showed increased thalianol levels compared to the wild type (Fig. 4A and fig. S3, A and B), suggesting that the CYP450 encoded by *At5g48000* is required for conversion of thalianol to a downstream product. Furthermore, we observed that peaks **2a-3b** were absent in root extracts of *thah1-1* (Fig. 4A) and so may correspond to downstream pathway intermediates.

The second CYP450 in this region (*At5g47990*) (Fig. 1B) belongs to the CYP705 family, another functionally uncharacterized Brassicaceae-specific CYP450 family (*15*). The CYP705 and CYP708 families belong to the CYP71 and CYP85 clans respectively demonstrating that *At5g47990* is not a tandem duplicate of *At5g48000*. Null insertion mutants and RNA interference (RNAi) knock down lines for *At5g47990* had enhanced levels of peaks **2a** and **2b** relative to wild type plants (Fig. 4A and fig. S3, C and D). Peaks **2a, 2b** and **2c,** with similar ion fragmentation patterns, were identified as thalian-diol ([3*S*,13*S*,14*R*]-malabarica-8,17,21-trien-3,?-diol) (fig. S4) and are likely to represent different thalian-diol isomers. Peaks **2a-2c** were not present in root extracts of *thas1-1*, confirming that production of thalian-diol depends on THAS (Fig. 4A). These peaks were also absent from root extracts of *thah1-1* (Fig. 4A), implicating *At5g48000* in thalian-diol biosynthesis. Overexpression of THAH with THAS in *Arabidopsis* leaves resulted in conversion of thalianol to thalian-diol (fig. S6A) and overexpression of THAH in *thah1-1* restored the thalianol pathway (fig. S6B). On the basis of these data we concluded that *At5g48000* encodes thalianol hydroxylase (hereafter referred to as THAH). Plants that overaccumulate thalian-diol are dwarfed (fig. S6B), suggesting that thalian-diol, like thalianol, is detrimental to growth when produced in the above-ground parts of the plant.

The increased levels of thalian-diol in *thad1-1* suggested that *At5g47990* was required for conversion of thalian-diol to a further downstream product. We observed that peaks **3a** and **3b**, which are present in wild type plants, are absent in *thas1-1*, *thah1-1* and *thad1-1* (Fig. 4A). These two peaks were identified as isomers of desaturated thalian-diol ([3*S*,13*S*,14*R*]-malabarica-8,15,17,21-tetraen-3,?-diol) (fig. S5). Conversion of thalian-diol to desaturated thalian-diol involves introduction of a double bond at carbon 15 (Fig. 4). CYP450 enzymes can catalyze desaturation reactions of this kind (see ref. 16). On the basis of these data we concluded that *At5g47990* encodes thalian-diol desaturase (hereafter referred to as THAD).

These data show that *THAS*, *THAH*, and *THAD* are contiguous co-expressed genes encoding biosynthetic enzymes required for three consecutive steps in the synthesis and modification of thalianol (Fig. 4B). The fourth gene in the cluster (*At5g47980*) is predicted to encode a BAHD acyltransferase. As for THAS, THAH and THAD, this enzyme also belongs to a Brassicaceae-specific enzyme subgroup. Since *At5g47980* has a very similar expression pattern to *THAS*, *THAH* and *THAD* and is implicated in secondary metabolism, it is likely to be required for modification of desaturated thalian-diol. However, we have not detected acylated desaturated thalian-diol in *Arabidopsis* root extracts. This may be because this compound is further modified, sequestered or present at very low levels.

The avenacin gene cluster in oat (*Avena* spp.) confers broad spectrum resistance to fungal pathogens (*2*). We tested whether the *Arabidopsis* thalianol gene cluster was also defense-related. We challenged the roots of mutant and wild type plants with strains of fungal and bacterial plant pathogens (*Alternaria brassicicola*, *Botrytis cinerea* and *Pseudomonas syringae* pv tomato DC3000) but saw no discernible differences in disease progression (fig. S7). However, examination of data from a recent survey of genome-wide polymorphisms in *Arabidopsis* (*17*) revealed that the thalianol pathway genes represent one of the most conserved regions of the genome. This is the hallmark of a recent selective sweep and implies that this gene cluster confers an important (and as yet unidentified) selective advantage.

Genes for most metabolic pathways are not clustered in plants. However clustering facilitates the inheritance of beneficial combinations of genes (*7*, *18*, *19*); furthermore, disruption of metabolic gene clusters can lead to accumulation of deleterious intermediates (*20*). The observations that ectopic over-accumulation of thalianol (Fig. 2A) or thalian-diol (fig. S6B) lead to severe dwarfing in *Arabidopsis* are consistent with a need for tight coordinate regulation of the pathway. Interestingly, lines that accumulate elevated levels of either of these compounds have

significantly longer roots than the wild type (Fig. 3B), suggesting distinct and organ-specific effects of thalianol and thalian-diol on plant growth. We note that the four co-expressed genes within the thalianol gene cluster have marked histone H3 lysine 27 trimethylation (H3K27me3), while the immediate flanking genes do not (*21*), suggesting that clustering may also facilitate coordinate regulation of the gene cluster at the chromatin level.

Despite the fact that oats and *Arabidopsis* both contain gene clusters for triterpene synthesis - the avenacin and thalianol clusters, respectively - these two gene clusters are unlikely to share a common origin. This is supported by the fact that the oat *Sad1* and *Sad2* genes do not have orthologs in *Arabidopsis* and are monocot specific (*6–8*). Furthermore there is no evidence for horizontal transfer of either gene cluster from microbes or elsewhere. Phylogenetic analysis suggests that an ancestral gene cluster formed in *Arabidopsis* around the progenitor of the lineage-specific OSC clade II (Fig. 5). Sequential rearrangements, duplications and gene loss presumably led to formation of the present-day thalianol cluster. Cluster formation may have been accompanied by the rapid expansion and functional diversification of the lineage-specific OSC clade II, along with the lineage-specific CYP702/708, CYP705 and acyltransferase gene families. Additionally, while THAS makes thalianol, the other OSCs in clade II produce different triterpene products when expressed in yeast (*3*, *5*). Some of these OSCs may be components of other functional triterpene gene clusters (Fig. 1A) as suggested by genome-wide co-expression of CYP450s (*22*).

An obvious assumption may be that gene clusters of the kind that we have observed were inherited from early evolutionary progenitor species. However our data clearly indicate that the thalianol and avenacin gene clusters are the products of separate and recent evolutionary events. This finding suggests that eukaryotic genomes are capable of remarkable plasticity and can assemble operon-like gene clusters *de novo*, which raises intriguing questions about the molecular mechanisms and evolutionary pressures that have acted to cause these gene clustering arrangements to assemble and become fixed. Comparative genomics will now enable us to trace the origins of such gene clusters and so to gain insights into the mechanisms that drive their formation. A further intriguing question is concerned with why genes for some metabolic pathways are clustered while others are not. Our identification of two triterpene gene clusters [for thalianol in *Arabidopsis* (this paper) and for avenacins in oat (*2*, *6–9*)] implies that triterpene pathways may be predisposed to gene clustering. There are two other examples of gene clusters for plant defense compounds (for rice diterpenes and maize benzoxazinoids) (*18*, *23*, *24*). As we learn more about why genes for some metabolic pathways are clustered and others are not we may need to redefine our understanding of plant metabolism.

References and Notes

1. K. Hostettmann, A. Marston, *Saponins* (Cambridge Univ. Press, Cambridge, UK, 1995).
2. K. Papadopoulou, R. E. Melton, M. Leggett, M. J. Daniels, A. E. Osbourn, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12923 (1999).
3. D. R. Phillips, J. M. Rasbery, B. Bartel, S. P. Matsuda, *Curr. Opin. Plant Biol.* **9**, 305 (2006).
4. K. T. Liby, M. M. Yore, M. B. Sporn, *Nat. Rev. Cancer* **7**, 357 (2007).
5. I. Abe, *Nat. Prod. Rep.* **24**, 1131 (2007).
6. K. Haralampidis *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13431 (2001).
7. X. Qi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8233 (2004).
8. X. Qi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18848 (2006).
9. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
10. G. C. Fazio, R. Xu, S. P. Matsuda, *J. Am. Chem. Soc.* **126**, 5678 (2004).
11. M. Koornneeff, J. H. Vanderveen, *Theor. Appl. Genet.* **58**, 257 (1980).
12. T. Yokota, *Trends Plant Sci.* **2**, 137 (1997).
13. S. D. Clouse, *Plant Cell* **14**, 1995 (2002).
14. B. Hamberger, J. Bohlmann, *Biochem. Soc. Trans.* **34**, 1209 (2006).
15. D. R. Nelson, M. A. Schuler, S. M. Paquette, D. Werck-Reichhart, S. Bak, *Plant Physiol.* **135**, 756 (2004).
16. A. E. Rettie, A. W. Rettenmeier, W. N. Howald, T. A. Baillie, *Science* **235**, 890 (1987).
17. J. O. Borevitz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12057 (2007).
18. A. Gierl, M. Frey, *Planta* **213**, 493 (2001).
19. S. Wong, K. H. Wolfe, *Nat. Genet.* **37**, 777 (2005).
20. P. Mylona *et al.*, *Plant Cell*, Epub Jan 18 (2008).
21. X. Zhang *et al.*, *PLoS Biol.* **5**, e129 (2007).
22. J. Ehlting, N. J. Provart, D. Werck-Reichhart, *Biochem. Soc. Trans.* **34**, 1192 (2006).
23. R. Jonczyk *et al.*, *Plant Physiol.* Epub Jan 11 (2008).
24. K. Shimura *et al.*, *J. Biol. Chem.* **282**, 34013 (2007).
25. P. Zimmermann, M. Hirsch-Hoffmann, L. Hennig, W. Gruissem, *Plant Physiol.* **136**, 2621 (2004).
26. K. Toufighi, S. M. Brady, R. Austin, E. Ly, N. J. Provart, *Plant J.* **43**, 153 (2005).
27. Materials and methods are available as supporting material on *Science* Online.

28. We thank D. Baulcombe, L. Dolan, R. Field, S. Kopriva, K. Papadopoulou, P. Shaw, A. Smith, D. Studholme and T. Wang for comments, JIC Metabolite Services for

metabolite analysis, L. Peña-Rodríguez and M. Rejzek for advice on chemistry, R. Melton and members of the Sainsbury Laboratory for assistance with the pathogenicity work, and the Biotechnology and Biological Sciences Research Council (UK) for funding.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1154990/dc1
Materials and Methods
Figs. S1 to S6
Table S1
References

8 January 2008; accepted 7 March 2008
Published online 20 March 2008; 10.1126/science.1154990
Include this information when citing this paper.

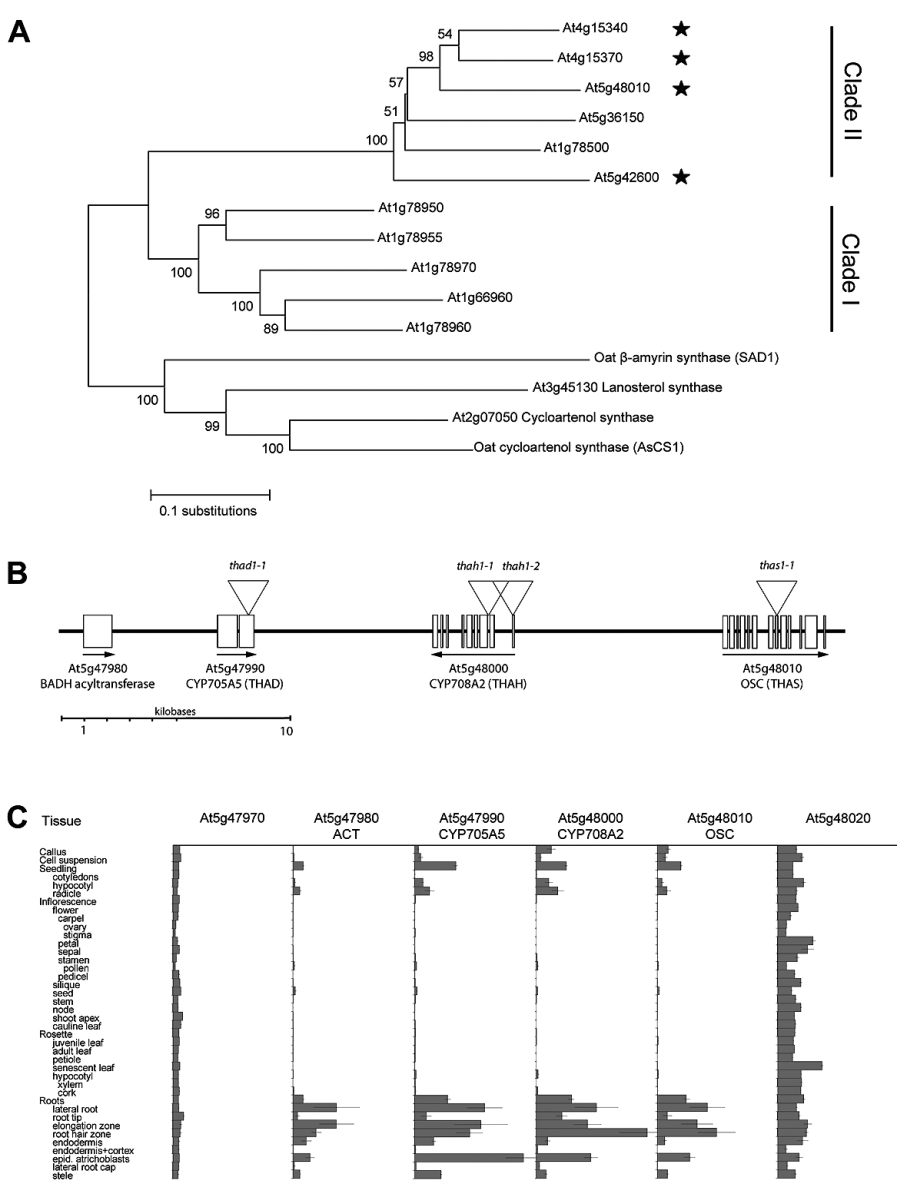
Fig. 1. (**A**) Neighbor-joining tree of *Arabidopsis* and oat OSC enzymes (% bootstrap support indicated). *Arabidopsis* OSCs are indicated with AGI gene codes. Those genes residing in candidate metabolic gene clusters are starred. Oat OSC NCBI accession numbers: SAD1 (CAC84558), AsCS1 (CAC84559). (**B**) Map of the triterpene gene cluster on *Arabidopsis* chromosome 5. T-DNA insertion mutants are indicated. (**C**) Microarray expression profiles of the genes in Fig. 1B and the two immediately flanking genes *At5g47970* and *At5g48020* (neither of which are implicated in secondary metabolism). Absolute expression values are shown with identical scales (± SE, n≥3). Data were retrieved from Genevestigator (*25*). The genes in Fig. 1B were highly co-expressed across 392 microarray experiments (average r=0.86). The flanking genes were not co-expressed with genes in the cluster region or with one another (r <0.15). Co-expression analysis was performed at the Bio-Array Resource (*26*) with data from NASCArrays (http://arabidopsis.info/).

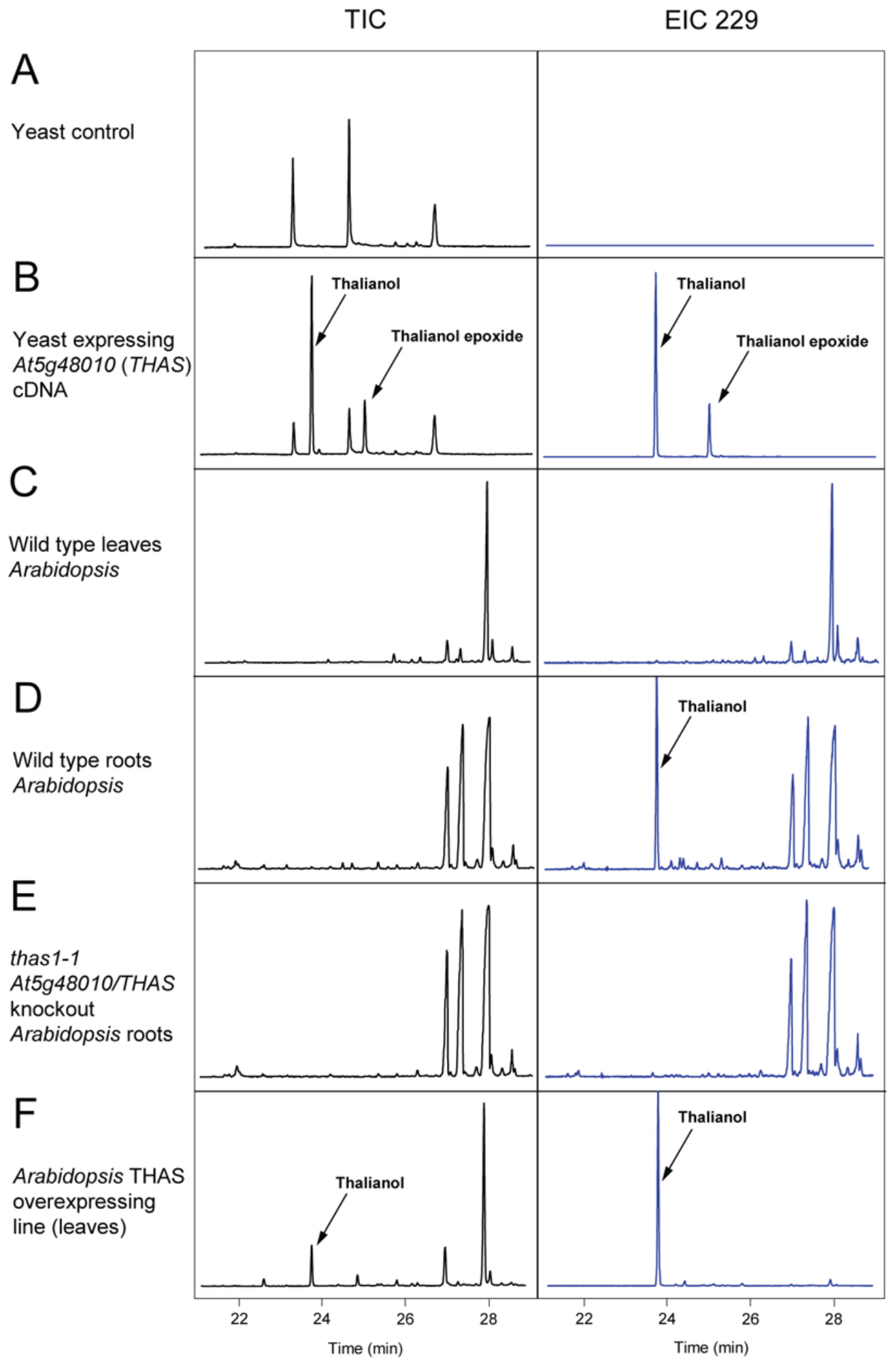
Fig. 2. Extracts from yeast and *Arabidopsis* were analyzed for triterpene content by GC-MS: TIC, total ion chromatograms; EIC 229, extracted ion chromatograms at m/z 229. Data are representative of at least two separate experiments, each with triplicate samples. (**A**) yeast empty vector control; (**B**) yeast expressing the *At5g48010* cDNA; (**C**) leaf and (**D**) root extracts from wild type *Arabidopsis*; (**E**) root extracts from an *At5g48010* (*thas1-1*) knockout line, (**F**) leaf extracts from an *Arabidopsis* line overexpressing *THAS*. The ion fragmentation pattern of thalianol was as reported (fig. S2) (*10*). Thalianol epoxide, a product of bis-oxidosqualene cyclization, was detected in yeast only. The chromatograms are scaled to the highest peak. Unlabelled peaks are sterols.

Fig. 3. (**A**) Plants overexpressing thalianol synthase (THAS) are dwarfed. (**B**) Roots from 7 day old plants that accumulate thalianol (*thah1-1*) or thalian-diol (*thad1-1*) are significantly longer than those of the wild type or *thas1-1* (which lacks the entire pathway). Plants that overexpress THAS, and thus have elevated levels of thalianol, also have significantly longer roots than the control. Error bars are ± SE, n=68-90 for three replicate experiments.

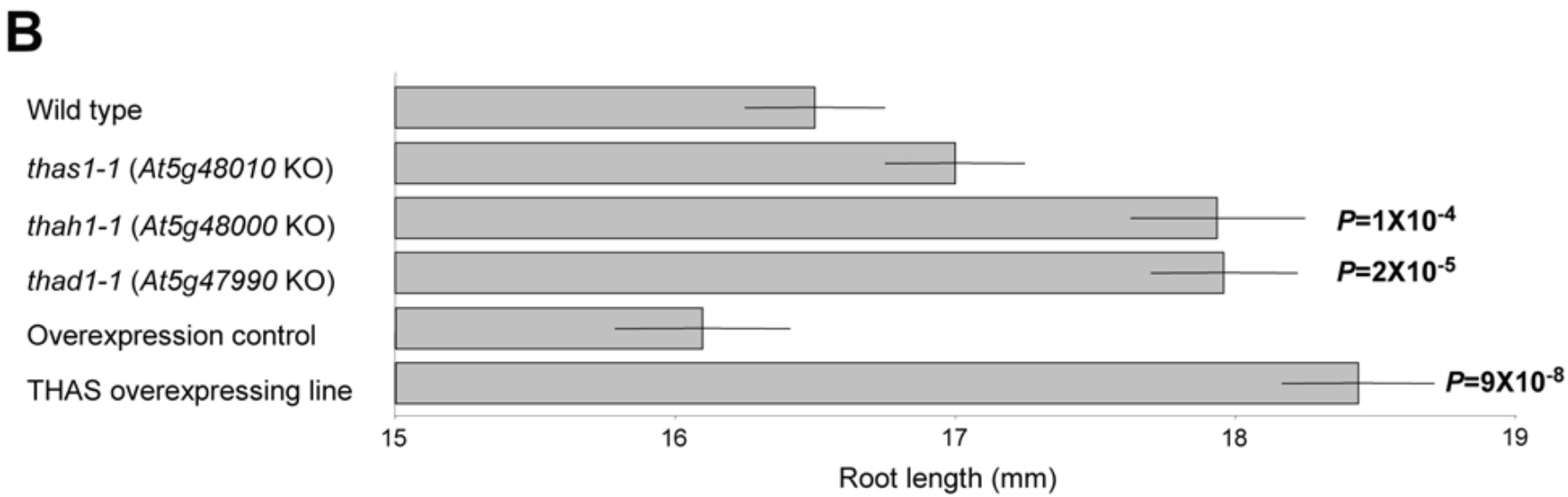
Fig. 4. (**A**) Detection of thalianol (**1**) and other pathway intermediates in root extracts from wild type and T-DNA insertion lines. TIC, total ion chromatograms; EIC 227 and 229, extracted ion chromatograms at m/z 227 and 229, respectively. (**B**) Scheme of the thalianol pathway showing the structures of 2,3-oxidosqualene, thalianol (**1**), thalian-diol (**2a, 2b** and **2c**), and desaturated thalian-diol (**3a** and **3b**) (see figs. S2, S4 and S5 for respective ion fragmentation patterns). The hydroxyl group introduced to thalianol by THAH to give thalian-diol is drawn in red. GC-MS ionization data indicate that this hydroxyl group is located at one of the four available carbon positions in rings B or C. Peaks **2a-2c** are isomers of thalian-diol and are likely to differ in the position of the hydroxyl group. Because of the low levels of these compounds in *Arabidopsis* root extracts we were unable to determine the precise position of the hydroxyl group in these isomers by NMR. The chromatograms are scaled to the highest peak. The peaks between 26 and 28 minutes (TIC/EIC) are plant sterols. The data are representative of at least two separate experiments, each with triplicate samples.

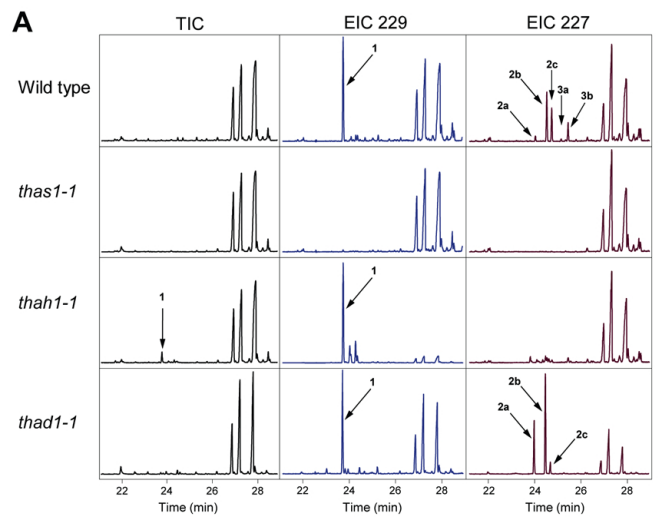
Fig. 5. Model of thalianol cluster evolution. The OSC tree is as in Fig. 1. Colored circles next to each OSC indicate the presence of adjacent genes encoding specific classes of other biosynthetic enzymes (see key). The colored diamonds indicate points at which common ancestors of genes for these other specific classes of enzyme are hypothesized to have become located in the vicinity of an ancestral OSC gene. The reconstruction minimizes the number of rearrangement and gene loss events required to reach the present-day chromosomal arrangement. At5g48010 (the OSC that lies within the functional gene cluster reported in this paper) is indicated in bold. The existence of other triterpene gene clusters is inferred by association of other clade II OSCs with genes for other enzymes implicated in secondary metabolism.

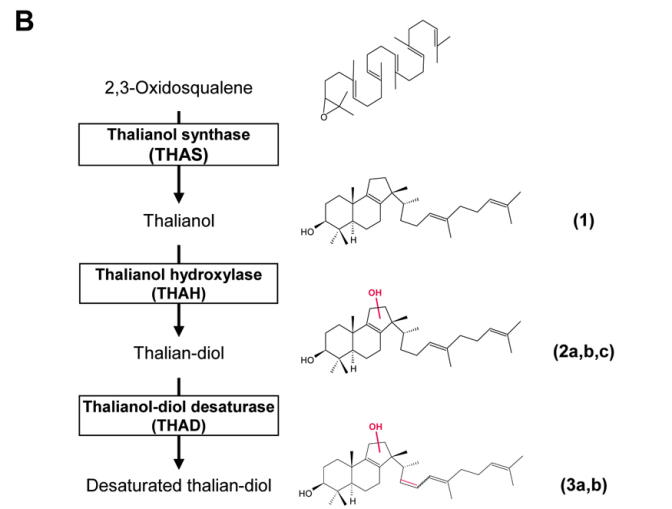

A
At4g15340
At4g15370
At5g48010
At5g36150
At1g78500
At5g42600
At1g78950
At1g78955
At1g78970
At1g66960
At1g78960
Oat β-amyrin synthase (SAD1)
At3g45130 Lanosterol synthase
At2g07050 Cycloartenol synthase
Oat cycloartenol synthase (AsCS1)
54
98
57
51
100
96
100
100
89
100
99
100
Clade II
Clade I
0.1 substitutions
B
thad1-1
thah1-1
thah1-2
thas1-1
At5g47980
BADH acyltransferase
At5g47990
CYP705A5 (THAD)
At5g48000
CYP708A2 (THAH)
At5g48010
OSC (THAS)
kilobases
1
10
C
Tissue
At5g47970
At5g47980
ACT
At5g47990
CYP705A5
At5g48000
CYP708A2
At5g48010
OSC
At5g48020
Callus
Cell suspension
Seedling
cotyledons
hypocotyl
radicle
Inflorescence
flower
carpel
ovary
stigma
petal
sepal
stamen
pollen
pedicel
silique
seed
stem
node
shoot apex
cauline leaf
Rosette
juvenile leaf
adult leaf
petiole
senescent leaf
hypocotyl
xylem
cork
Roots
lateral root
root tip
elongation zone
root hair zone
endodermis
endodermis+cortex
epid. atrichoblasts
lateral root cap
stele

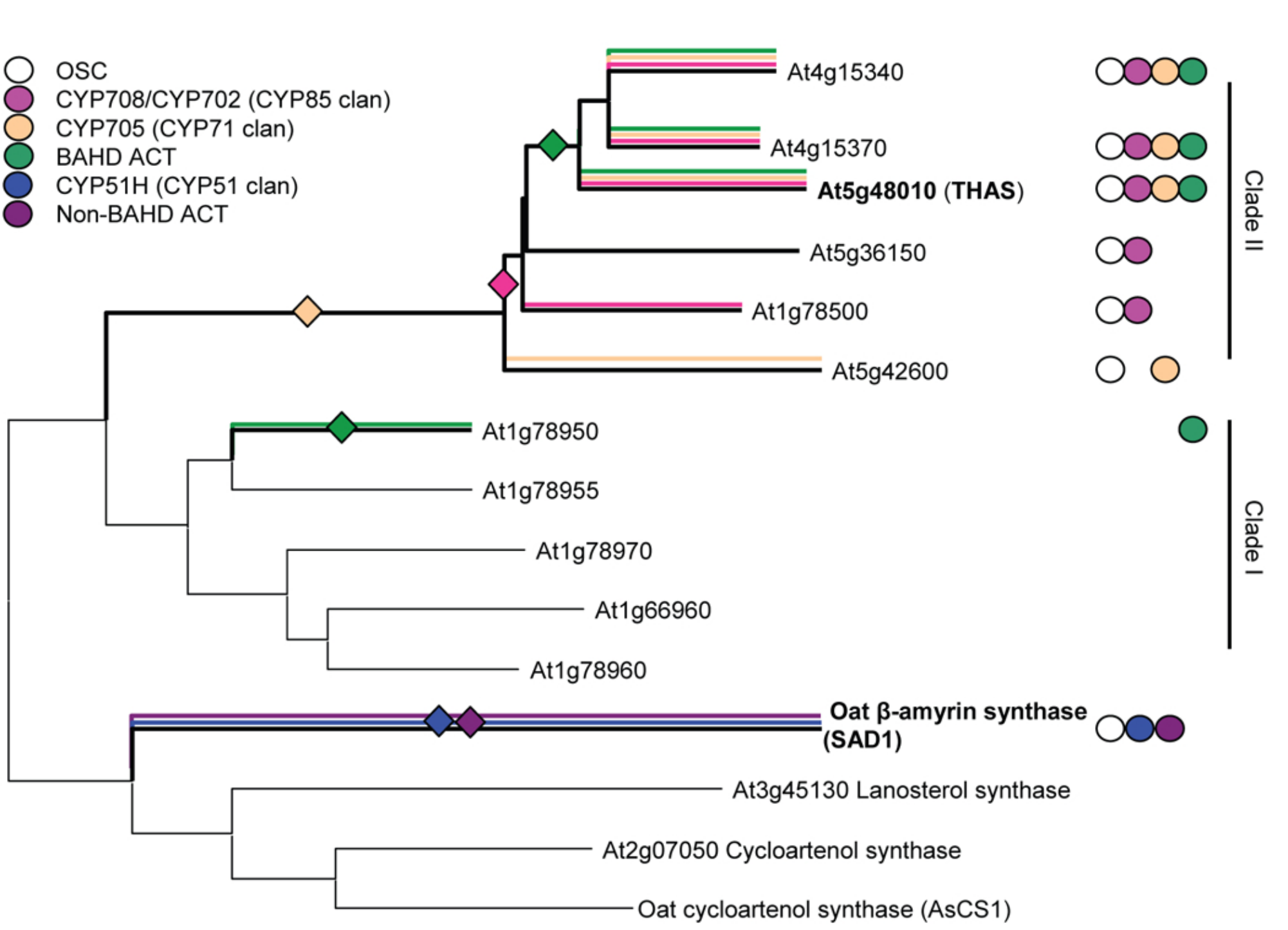

TIC
EIC 229
A
Yeast control
B
Yeast expressing
At5g48010 (THAS)
cDNA
Thalianol
Thalianol epoxide
C
Wild type leaves
Arabidopsis
D
Wild type roots
Arabidopsis
Thalianol
E
thas1-1
At5g48010/THAS
knockout
Arabidopsis roots
F
Arabidopsis THAS
overexpressing
line (leaves)
Thalianol
22
24
26
28
Time (min)


A
Wild type
THAS overexpressing line
Wild type
THAS overexpressing line


B
Wild type
thas1-1 (At5g48010 KO)
thah1-1 (At5g48000 KO)
thad1-1 (At5g47990 KO)
Overexpression control
THAS overexpressing line
P=1X10-4
P=2X10-5
P=9X10-8
15
16
17
18
19
Root length (mm)


A
TIC
EIC 229
EIC 227
Wild type
thas1-1
thah1-1
thad1-1
1
2a
2b
2c
3a
3b
22
24
26
28
Time (min)


B
2,3-Oxidosqualene
Thalianol synthase (THAS)
Thalianol
(1)
Thalianol hydroxylase (THAH)
Thalian-diol
(2a,b,c)
Thalianol-diol desaturase (THAD)
Desaturated thalian-diol
(3a,b)
OH
HO
H


OSC
CYP708/CYP702 (CYP85 clan)
CYP705 (CYP71 clan)
BAHD ACT
CYP51H (CYP51 clan)
Non-BAHD ACT
At4g15340
At4g15370
At5g48010 (THAS)
At5g36150
At1g78500
At5g42600
Clade II
At1g78950
At1g78955
At1g78970
At1g66960
At1g78960
Clade I
Oat β-amyrin synthase (SAD1)
At3g45130 Lanosterol synthase
At2g07050 Cycloartenol synthase
Oat cycloartenol synthase (AsCS1)