the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84,** 478–490 doi: 10.1111/tpj.13012

Identification and genome organization of saponin pathway genes from a wild crucifer, and their use for transient production of saponins in *Nicotiana benthamiana*

Bekzod Khakimov[1,2], Vera Kuzina[1,†], Pernille Ø. Erthmann[1], Ery Odette Fukushima[3], Jörg M. Augustin[1], Carl Erik Olsen[1], Jelle Scholtalbers[4,5], Hanne Volpin[5,6], Sven Bode Andersen[1], Thure P. Hauser[1], Toshiya Muranaka[3] and Søren Bak[1,*]

[1]*Department of Plant and Environmental Sciences, Copenhagen Plant Science Center, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg CDK-1871, Denmark,*

[2]*Department of Food Science, University of Copenhagen, Rolighedsvej 26, Frederiksberg DK-1958,Denmark,*

[3]*Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita Osaka 565-0871, Japan,*

[4]*EMBL, Heidelberg,Germany,*

[5]*Keygene, Agro Business Park 90, Wageningen 6708 PW,The Netherlands, and*

[6]*Danziger Innovations, Mishmar HaShiva 5029700, Israel*

Received 10 July 2015; revised 12 August 2015; accepted 18 August 2015; published online 01 September 2015.

*For correspondence (e-mail bak@plen.ku.dk).

†Present address: Chr. Hansen, Bøge Allé 10, DK-2970,Hørsholm, Denmark.

SUMMARY

The ability to evolve novel metabolites has been instrumental for the defence of plants against antagonists. A few species in the *Barbarea* genus are the only crucifers known to produce saponins, some of which make plants resistant to specialist herbivores, like *Plutella xylostella*, the diamondback moth. Genetic mapping in *Barbarea vulgaris* revealed that genes for saponin biosynthesis are not clustered but are located in different linkage groups. Using co-location with quantitative trait loci (QTLs) for resistance, transcriptome and genome sequences, we identified two 2,3-oxidosqualene cyclases that form the major triterpenoid backbones. LUP2 mainly produces lupeol, and is preferentially expressed in insect-susceptible *B. vulgaris* plants, whereas LUP5 produces β-amyrin and α-amyrin, and is preferentially expressed in resistant plants; β-amyrin is the backbone for the resistance-conferring saponins in *Barbarea*. Two loci for cytochromes P450, predicted to add functional groups to the saponin backbone, were identified: CYP72As co-localized with insect resistance, whereas CYP716As did not. When *B. vulgaris* sapogenin biosynthesis genes were transiently expressed by CPMV-HT technology in *Nicotiana benthamiana*, high levels of hydroxylated and carboxylated triterpenoid structures accumulated, including oleanolic acid, which is a precursor of the major resistance-conferring saponins. When the *B. vulgaris* gene for sapogenin 3-*O*-glucosylation was co-expressed, the insect deterrent 3-*O*-oleanolic acid monoglucoside accumulated, as well as triterpene structures with up to six hexoses, demonstrating that *N. benthamiana* further decorates the monoglucosides. We argue that saponin biosynthesis in the *Barbarea* genus evolved by a neofunctionalized glucosyl transferase, whereas the difference between resistant and susceptible *B. vulgaris* chemotypes evolved by different expression of oxidosqualene cyclases (OSCs).

Keywords: evolution, metabolic engineering, P450, UDP-glycosyltransferases, oxidosqualene cyclases, synthetic biology, *Barbarea vulgaris*, *Nicotiana benthamiana*.

INTRODUCTION

Through time, plants have evolved an astounding diversity of chemical defences against herbivores and other antagonists based on bioactive compounds of low molecular weight. One such class of compounds are the triterpenoid saponins, which contribute to plant defence against a broad spectrum of insects, pathogens and other herbivores (Osbourn, 1996). Saponins are structurally highly diverse, with more than 200 different structures described (Augustin *et al.*, 2012). The common structural motif of saponins is a hydrophilic glycosidic part coupled to a

hydrophobic triterpenoid part, which gives them detergent-like properties. They are structurally related to sterols and steroid hormones, and can therefore interfere with sterols in cell membranes, which may lead to deleterious pore formation that causes cell death (Augustin *et al.*, 2011).

Triterpene saponins are derived from the omnipresent sterol precursor 2,3-oxidosqualene. The first step in the biosynthesis of saponins is the formation of cyclical triterpene skeletons, catalysed by oxidosqualene cyclases (OSCs; Figure 1a). OSCs are phylogenetically derived from cycloartenol synthases in primary sterol biosynthesis (Figure 2; Phillips *et al.*, 2006; Augustin *et al.*, 2011; Thimmappa *et al.*, 2014), and OSCs thus constitute a branching point between primary sterol metabolism and the synthesis of triterpenoid secondary metabolites, including saponins. Most plants contain multiple OSCs that catalyse the formation of different triterpene backbones. The triterpene skeletons are subsequently oxygenated by various cytochromes P450 (P450s; Bak *et al.*, 2011; Hamberger and Bak, 2013; Thimmappa *et al.*, 2014), and are glycosylated by UDP-glycosyltransferases (UGTs; Augustin *et al.*, 2012).

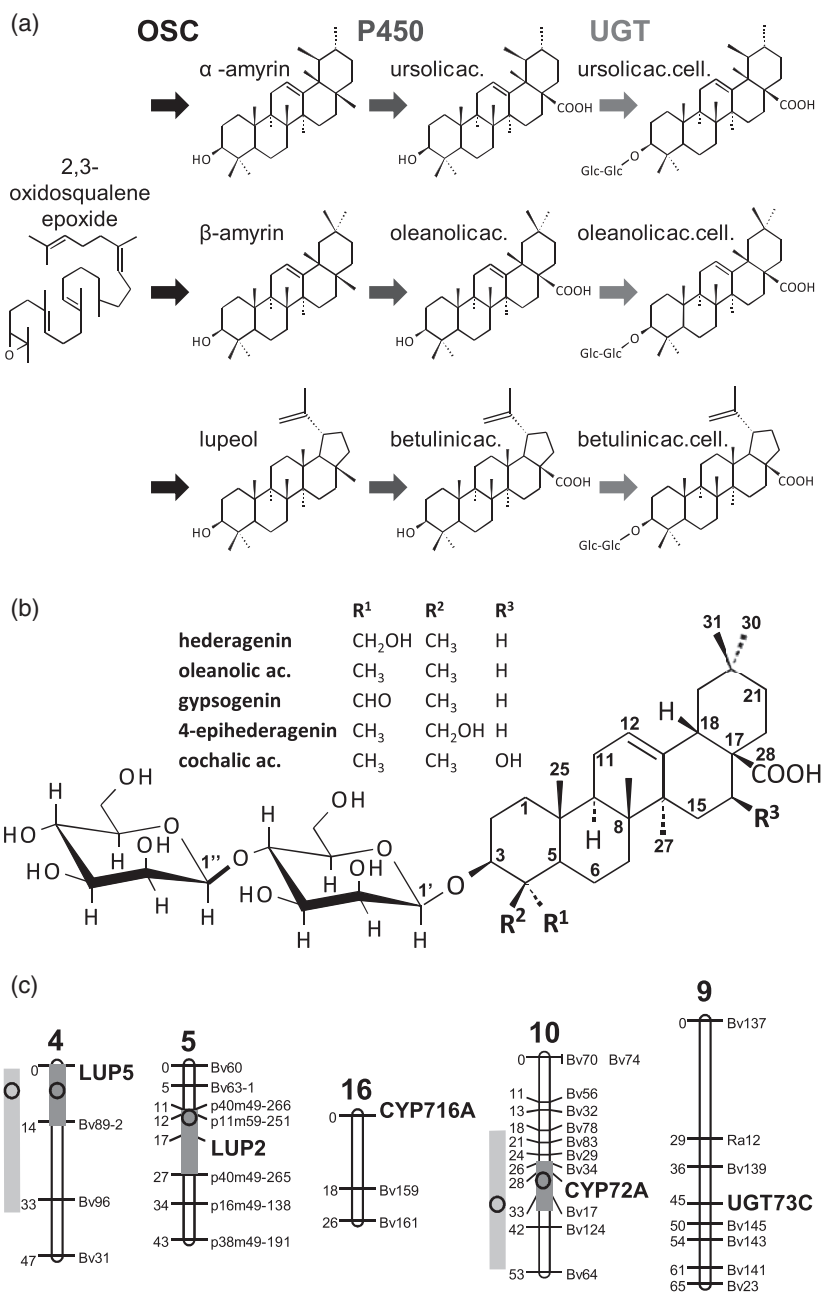

Figure 1. (a) Generic biosynthesis pathway of saponins in *Barbarea vulgaris*, showing the steps catalysed by an oxidosqualene cyclase (OSC), a cytochrome P450 (P450) and a UDP-glycosyltransferase (UGT). (b) The five saponins with known structures from *B. vulgaris*. (c) Localization of biosynthetic genes in four linkage groups (vertical open bars), with genetic distances (cM), and genes, AFLP and SSR markers shown to the left and right of the linkage groups, respectively. Quantitative trait loci (QTLs) for the four known G-type saponins are indicated by light-grey bars, and QTLs for insect resistance are indicated by dark-grey bars. The complete linkage map is shown in Figure S2.

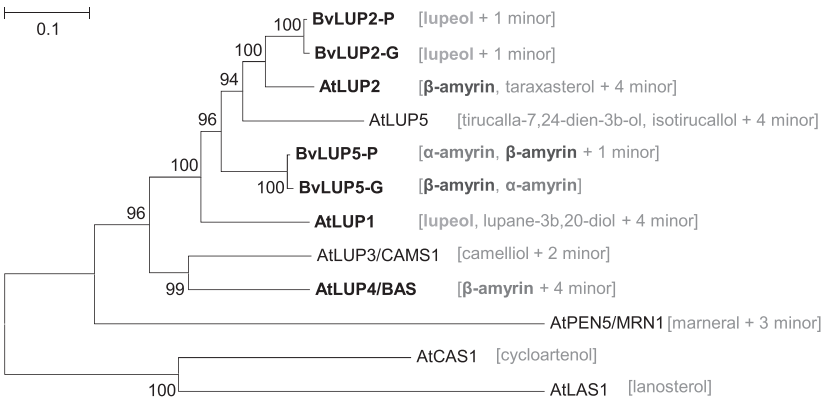

Figure 2. Maximum-likelihood gene tree of the four *Barbarea vulgaris* oxidosqualene cyclases (OSCs) described in this paper, in addition to related *Arabidopsis thaliana* OSCs. OSCs in bold produce lupeol and/or β-amyrin and/or α-amyrin. AtCAS1 and ATLAS1 represent OSCs involved in the biosynthesis of sterols. Main and minor products reported to be produced by *A. thaliana* OSCs are listed in brackets (adapted from Augustin *et al.*, 2011). Numbers on nodes indicate bootstrap values; the alignment underlying the tree is shown in Figure S1.

The large and expanding structural complexity of saponins is thought to derive from the continuous recruitment of new genes from the OSC, P450 and UGT multigene families by gene duplications followed by selection, leading to neofunctionalization (Augustin *et al.*, 2012; Xue *et al.*, 2012; Hamberger and Bak, 2013). Additional gene families, such as methyl and acyl transferases, are also known to be involved in saponin biosynthesis and diversification (Thimmappa *et al.*, 2014). Most plants that produce saponins accumulate a blend of different chemical structures, which may reflect a continuous recruitment and neofunctionalization of enzymatic variants from a limited number of gene families, which may also result from the substrate and/or product promiscuity of the individual enzymes.

Saponins are found in several different and unrelated plant families, and the identification of genes involved in their biosynthesis has revealed that the pathways must have evolved convergently several times. This, and the sheer gene family sizes of OSCs and in particular P450s and UGTs, complicates the use of sequence similarity and identity for gene discovery across plant species, and hampers the identification of their biological roles.

The dependence on several different genes and gene families in biosynthetic pathways for specialized metabolites has raised the question of how these genes are organized in the genome. In some cases, genes are clustered in operon-like structures, as found initially for genes in benzoxaxinoid biosynthesis (Frey *et al.*, 1997), and later for a number of other biosynthetic genes in, for example, *Arabidopsis thaliana* and *Avena strigosa* (Qi *et al.*, 2004; Field and Osbourn, 2008; Chu *et al.*, 2011; Field *et al.*, 2011; Mugford *et al.*, 2013). This has been suggested to secure co-inheritance and coordinated expression, and to prevent pathway disruption, e.g. by hybridization, as this could lead to a build-up of toxic intermediates (Osbourn, 2010; Takos and Rook, 2012).

Within the large cabbage family (Brassicaceae), only a few species in the *Barbarea* genus are known to produce saponins (Nielsen *et al.*, 2010a; Badenes-Perez *et al.*, 2014), and among them the short-lived perennial herb *Barbarea vulgaris* (winter cress). Until now, the structure of five saponins, all derived from β-amyrin, have been identified in *B. vulgaris*: 3-*O*-cellobiosyl-oleanolic acid (oleanolic acid cellobioside), 3-*O*-cellobiosyl-hederagenin (hederagenin cellobioside), 3-*O*-cellobiosyl-gypsogenin (gypsogenin cellobioside), 3-*O*-cellobiosyl-4-epihederagenin (4-epihederagenin cellobioside) and 3-*O*-cellobiosyl-cochalic acid (chochalic acid cellobioside) (Shinoda *et al.*, 2002; Agerbirk *et al.*, 2003a; Nielsen *et al.*, 2010a; Figure 1a). One of the saponins, hederagenin cellobioside, renders the plants resistant to a flea beetle (*Phyllotreta nemorum*), the devastating agricultural pest *Plutella xylostella* (diamondback moth) and probably other species as well (Renwick, 2002; Shinoda *et al.*, 2002; Kuzina *et al.*, 2009, 2011; Nielsen *et al.*, 2010a; Badenes-Perez *et al.*, 2014). Resistant plants also produce oleanolic acid cellobioside, which has a lesser effect than hederagenin cellobioside (Nielsen *et al.*, 2010a). Not all *B. vulgaris* plants are resistant to these insects, and we now know that these plants also produce saponins but of different and currently unknown structures (Kuzina *et al.*, 2011). Resistant and susceptible plants also differ in other traits, including the hairiness of rosette leaves, glucosinolate profiles and frequency of resistance to an oomycete pathogen, and they have therefore been designated as different plant types: the insect-resistant and glabrous plant type is termed G-type, and the susceptible and pubescent plant type is termed P-type (Agerbirk *et al.*, 2003b; Christensen *et al.*, 2014). Plants of G- and P-type are genetically divergent, reproductively partly incompatible and occupy different but geographically overlapping ranges in Eurasia (Christensen *et al.*, 2014).

In the present study, we identified the genomic locations, expression profiles and roles of the OSCs, P450s and UGTs involved in saponin biosynthesis in the insect-resistant G-type and insect-susceptible P-type of *B. vulgaris*. Transient co-expression of the genes from *B. vulgaris* in *Nicotiana benthamiana*, using hyper-translatable Cow Pea Mosaic Virus protein expression system (CPMV-HT) technology (Sainsbury and Lomonossoff, 2008), allowed us to partially reconstitute saponin biosynthesis, resulting in

the accumulation of high levels of oxygenated and glycosylated triterpene structures, including the known insecticidal 3-*O*-oleanolic acid monoglucoside (Augustin *et al.*, 2012).

RESULTS

Using genetic linkage, transcriptome and genome data to find candidate genes

We used quantitative trait loci (QTL) analysis of hybrids between G- and P-type plants to focus our search for candidate genes. Previously, we identified two QTLs for flea beetle resistance that co-localized with QTLs for hederagenin cellobioside and other saponins of the insect-resistant *B. vulgaris* G-type (Kuzina *et al.*, 2011). In the present study, we added more microsatellite markers to the linkage map, based on synteny with the *A. thaliana* genome, and used these for an improved QTL analysis of the *B. vulgaris* genome (Table S1 and S2). Seventeen linkage groups were identified (Figure S2), representing the eight *B. vulgaris* chromosomes (Ørgaard and Linde-Laursen, 2007); linkage groups 4 and 5 may be one, as they merged when the log odds score (LOD) threshold was lowered from 6 to 4.

Three QTLs for flea beetle resistance were identified (Figures 1a and S2), of which two co-localized with QTLs for four of the known β-amyrin derived G-type saponins. A QTL for structurally unknown P-type saponins co-localized with the QTL for G-type saponins on linkage group 10.

Barbarea vulgaris transcriptome and genome data were mined for OSCs, P450s and UGTs, which revealed 17 OSC, 269 P450 and 117 UGT candidate sequences. Many of the candidates were not full length, and were extended using rapid amplification of cDNA ends (RACE) for further analysis. Two OSCs co-localized with the QTLs for insect resistance on linkage groups 4 and 5 (Figures 1b and S2). Based on their phylogenetic clustering with OSCs from *A. thaliana* and syntenic position with *A. thaliana* LUP5, they were designated LUP2 and LUP5, respectively. LUP2 clusters with AtLUP2 with a bootstrap value of 100%, whereas LUP5 clusters with AtLUP5 with a bootstrap value of 96% (Figure 2). The corresponding *AtLUP2* and *AtLUP5* from *A. thaliana* are separated by more than 5 Mb, in agreement with the placement of *LUP2* and *LUP5* in different linkage groups (using an LOD score of 6). Unfortunately, the quality of the available Illumina assembly could not resolve the distance between the two QTLs because of a large cluster of repetitive sequences between the QTLs.

A tandem array of at least five P450s from the CYP72A subfamily co-localized with the QTL for resistance and saponins on linkage group 10 (Figures 1 and S2). Members of the CYP72A subfamily are known to be involved in saponin biosynthesis in *Glycyrrhiza* species (liquorice; Seki *et al.*, 2011). The position of this QTL and the *CYP72* repeat is syntenic to a tandem repeat of eight *CYP72A*s and a *CYP72A* pseudogene on *A. thaliana* chromosome 3. As a result of the high degree of sequence identity between the *CYP72A* repeats in *B. vulgaris*, combined with the short reads obtained from the Illumina sequencing, it was not possible to assemble and annotate these genes and possible pseudogenes, even using PCR strategies.

A *CYP716A* was placed in linkage group 16 outside any of the QTLs for saponin accumulation or flea beetle resistance. CYP716s in, for example, *Medicago truncatula*, *Vitis vinifera* and *Maesa lanceolata* have been shown to catalyse three successive oxidations at the C28 position of β-amyrin, α-amyrin and lupeol to oleanolic acid, ursolic acid and betulinic acid, respectively (Carelli *et al.*, 2011; Fukushima *et al.*, 2011; Moses *et al.*, 2015), and accordingly they are candidates for saponin biosynthesis enzymes in *B. vulgaris*. The CYP716As from the G- and P-types are 95% identical at the amino acid level, and were designated CYP716A80 and CYP716A81, respectively, according to the P450 nomenclature.

Product profiles of OSCs and P450s

To elucidate the biochemical function of the selected OSCs, they were heterologously expressed in *N. benthamiana*, using the CPMV-HT technology, and microsomes containing the OSCs were assayed with 2,3-oxidosqualene as substrate to determine their product profiles. CPMV-HT facilitates the accumulation of high levels of recombinant proteins as a result of the hypertranslation of genes inserted into a plant viral expression vector (Sainsbury and Lomonossoff, 2008). Microsomes containing LUP2 from either the G- and P-types mainly produced lupeol, as shown by GC-MS (Figure 3a). In contrast, LUP5 from the G-type produced mainly β-amyrin, in addition to some lupeol and α-amyrin, whereas LUP5 from the P-type mainly produced a mixture of α-amyrin, β-amyrin and very small quantities of lupeol (Figure 3a). Low concentrations of α-amyrin, β-amyrin and/or lupeol were also detected when microsomes were assayed without adding the substrate 2,3-oxidosqualene (Figure 3a); this background is produced by the *B. vulgaris* OSCs in *N. benthamiana* leaves (Figures 3b and 4), and co-purifies with the microsomes. The microsome data fitted very well with the profiles of lupeol, α-amyrin and β-amyrin when ethyl acetate-extracted metabolites from *N. benthamiana* leaves expressing the OSCs were analysed (Figure 3b). In summary, LUP2s predominantly make lupeol, whereas LUP5s predominantly make β-amyrin and α-amyrin; furthermore, the product ratios of α-amyrin and β-amyrin produced by LUP5 from the G- and P-types differ (Figures 3b and 4).

The two CYP716As (CYP716A80 and CYP716A81 from the G- and P-types, respectively) were expressed in *Saccharomyces cerevisiae*, using a construct containing two copies of the P450s to increase expression levels. The

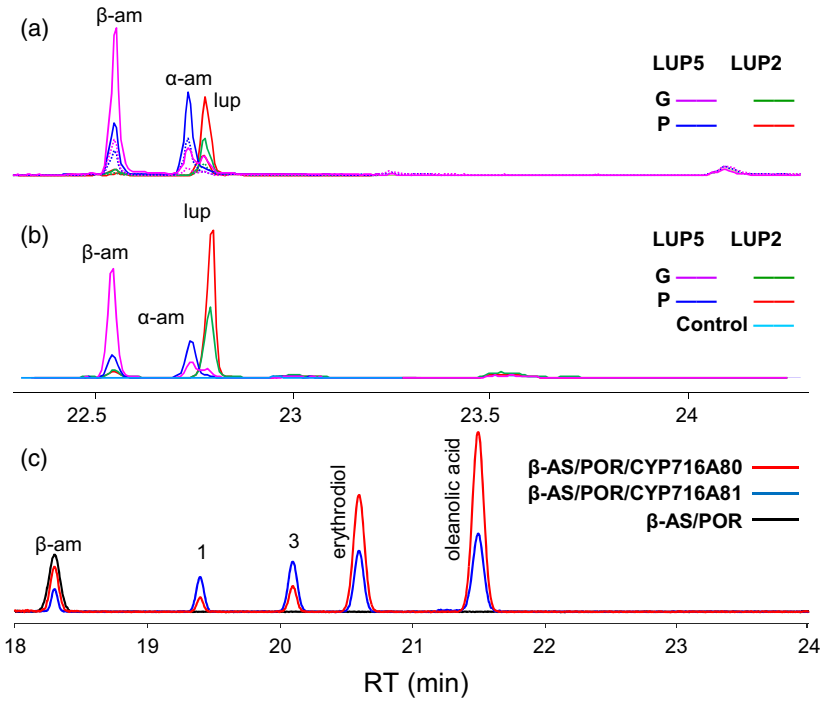

Figure 3. Product profiles of *Barbarea vulgaris* oxidosqualene cyclase (OSC) and cytochrome P450s (P450s). Panels show GC-MS profiles (superimposed extracted ion chromatograms (EICs) of *m/z* 189 and 218) for (a) LUP5s and LUP2s from the G- and P-types, with (solid lines) or without (dotted lines) 2,3-oxidosqualene added to microsomes; microsomes were prepared from *Nicotiana benthamiana* leaves transiently expressing the OSCs.
(b) As for (a), but based on ethyl acetate-extracted metabolites from *N. benthamiana* leaves transiently expressing the OSCs. Peaks with known structures are indicated: α-am, α-amyrin; β-am, β-amyrin; lup, lupeol. Control represents *N. benthamiana* leaves that do not express *B. vulgaris* OSCs.
(c) Total ion current chromatograms (TICs) of GC-MS data from *Saccharomyces cerevisiae* producing saponin backbones by expressing CYP716A80 (G) and CYP716A81 (P) in combination with *Lotus japonicus* β-amyrin synthase (β-AS) and *L. japonicus* cytochrome P450 oxidoreductase (POR). Peaks 1 and 3 represent unknowns 1 and 3 (Table S8), respectively. The *y*-axis of the panels are scaled to the highest peak intensity.

constructs were co-expressed with *Lotus japonicus* β-amyrin synthase, to provide β-amyrin as substrate (Seki *et al.*, 2008), and *L. japonicus* P450 oxidoreductase (POR), to supplement the endogenous *S. cerevisiae* POR. These constructs produced oleanolic acid and erythrodiol, corresponding to the oxidation of β-amyrin at C28 to either the carboxylic acid or the corresponding alcohol (Figure 3b). In addition to the C28 oxidation products, both CYP716As produced similar relative levels of two unknown β-amyrin-derived oxygenated compounds (unknown 1 and unknown 3; Figure 3c). Unknown-1 is predicted to be a β-amyrin-derived alcohol with two hydroxyl groups, whereas unknown 3 presumably is similar to unknown 1 but with an additional double bond (Table S8). CYP716A80 from the G-type produced significantly more erythrodiol and oleanolic acid relative to unknown 1 and unknown 3, compared with CYP716A81 from the P-type. Thus, both CYP716As produce the same four products, but the product ratios of the CYP716As differ between G- and P-types, even though these two P450s are 95% identical at the amino acid level.

Reconstruction of sapogenin and saponin biosynthesis in *N. benthamiana*

By the combinatorial expression of selected *B. vulgaris* OSCs, P450s and UGTs in *N. benthamiana*, using CPMV-HT technology, we attempted to reconstruct part of sapogenin and saponin biosynthesis in *B. vulgaris*. As expected, the product profiles and ratios of the OSCs were in agreement with the microsome analysis above (Figures 3b and 4). Subsequently, the four OSCs were co-expressed with G- and P-type CYP716As (CYP716A80 and CYP716A81, respectively) to produce sapogenins. Finally, UGT73C11 was co-expressed to catalyse sapogenin 3-*O*-glucosylation; this UGT specifically glucosylates sapogenins at C3 and not C28, and does not glucosylate phytosterols or flavonols (Augustin *et al.*, 2012).

Combinations of the LUPs and CYP716As produced oleanolic, ursolic and betulinic acids, which are derived from the C28 carboxylation of β-amyrin, α-amyrin and lupeol, respectively, but at highly varying levels (Figure 4). This is in agreement with the product profiles observed by the expression of CYP716As in *S. cerevisiae* (Figure 3c). In addition, five unknown metabolites, including unknown 1 and unknown 3 observed from the expression of CYP716A80/81 in *S. cerevisiae* (Figure 4), were detected at different ratios. The increase in the number of unknowns, compared with the expression of CYP716As in *S. cerevisiae*, probably reflects that β-amyrin was the only available substrate for oxidation in yeast, whereas in the co-expression with the four *B. vulgaris* OSCs, β-amyrin as well as α-amyrin and lupeol were also present at varying concentrations.

Sapogenin profiles from eight different combinations of LUPs and CYP716As were analysed using GC-MS (Figure 4). When LUP2 from either the G- or the P-type was co-expressed with CYP716A80 (from the G-type), betulinic acid was the major sapogenin detected (>60%). This fits with our results that LUP2 predominantly makes lupeol (>90%) and CYP716A80 predominantly oxidizes the triterpenoid backbones to the corresponding C28 carboxylic

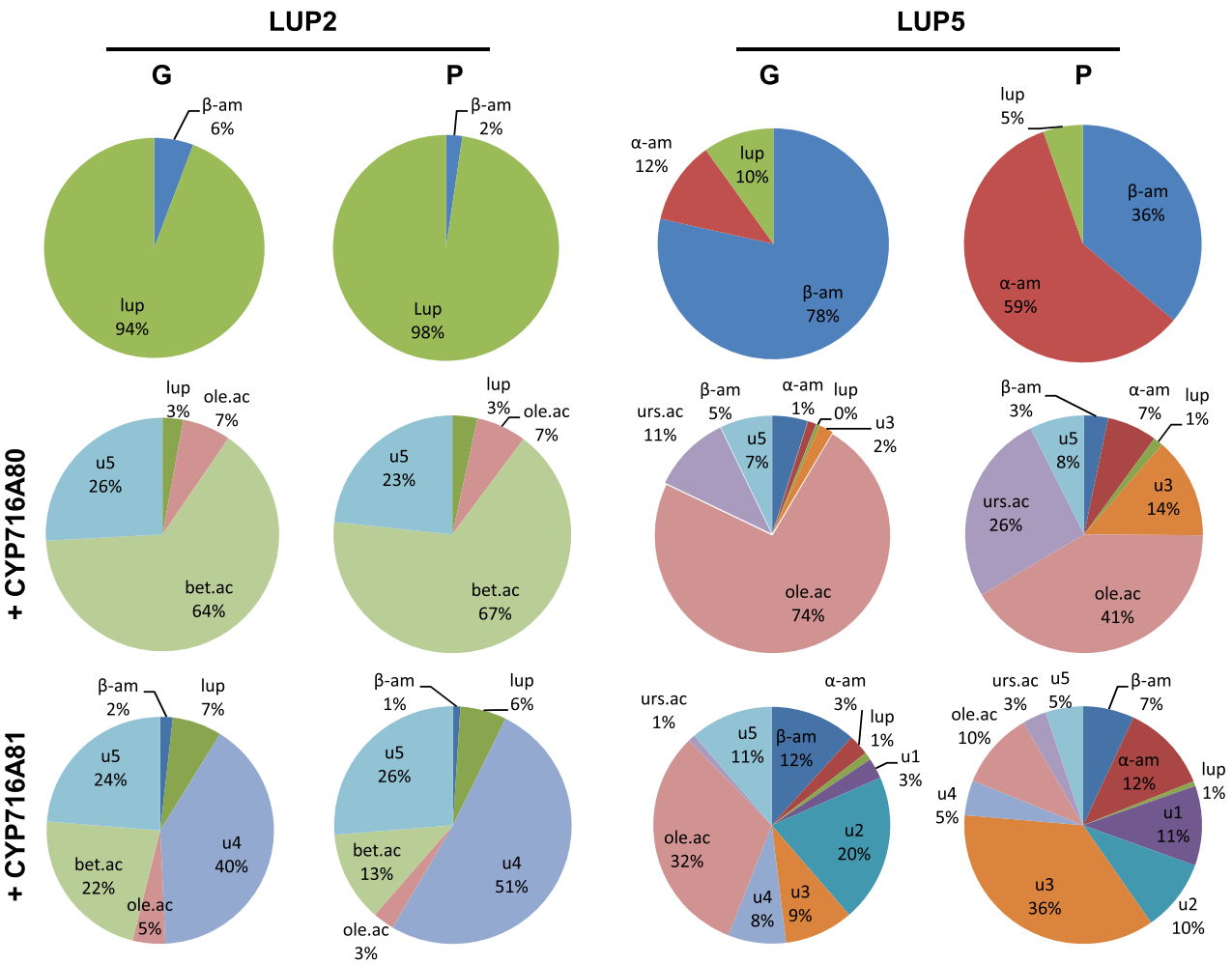

Figure 4. Relative concentrations of sapogenins produced by the transient expression of LUP2 and LUP5 from the G- and P-types alone (first row), and in combination with CYP716A80 (G) and CYP716A81 (P) (second and third rows), as detected by GC-MS. A total of six known triterpenes [α-amyrin (α-am), β-amyrin (β-am), betulinic acid (bet.ac), lupeol (lup), oleanolic acid (ole.ac), ursolic acid (urs.ac)], in addition to five tentatively identified triterpenes [unknowns 1–5 (u1–u5)] were detected and shown for a representative set of samples. Information on unknowns 1–5 are presented in Table S8 and Figure S3.

acid (see above; Figure 4). When LUP2s were co-expressed with CYP716A81 (P-type), the product profile was more complex, with unknown 4 being the most abundant metabolite (>40%) and with unknown 5 being the second most abundant (>20%). The more complex product profiles reflect the lower product specificity of CYP716A81 compared with CYP716A80, as was also observed when they were co-expressed with *L. japonicus* β-amyrin synthase (Figure 3c). In analogy with unknown 1 and 3, we speculate that unknown 4 is a triterpenoid acid derived from lupeol that may have a similar structure to betulinic acid, but with a reduced double bond, and thus has a molecular mass of 458 Da. Likewise, unknown 5 is a lupeol-derived triterpenoid acid, and may have a similar structure as betulinic acid, but with an extra functional group that might be either a hydroxyl or a carboxylic acid (for further information, see Table S8).

LUP5 from the G-type predominantly makes β-amyrin, and accordingly oleanolic acid is the major metabolite accumulating when it is co-expressed with CYP716A80 (G-type; >70%). When LUP5 from the P-type is co-expressed with CYP716A80, both ursolic acid (>25%; derived from α-amyrin) and oleanolic acid (>40%; derived from β-amyrin) are the main products, in agreement with the more complex product profile of P-type LUP5. LUP5s in combination with CYP716A81 (P-type) lead to a relative decrease of the C28 carboxylic acids and an increase in unknowns 1–5, as also seen for LUP2 in combination with CYP716A81. Based on their fragmentation patterns we speculate that they represent further oxidation of the corresponding carboxylic acids and reduction of double bond(s) (for further information and discussion on these unknowns, see Table S8). Unexpectedly, C28 alcohols of lupeol, α-amyrin and β-amyrin were not detected when the OSCs and P450s were

co-expressed, as observed in *S. cerevisiae*. This suggests that, to some extent, the product profile is influenced by the heterologous host.

To convert the sapogenins into saponins, the OSCs and CYPs were combined with UGT73C11. UGTC11 comes from the G-type and is 98% identical to UGT73C10 from the P-type, and they have the same biochemical properties (Augustin *et al.*, 2012). We have analysed most combinations, but for simplicity only two enzyme combinations are discussed here, as they represent gene combinations present *in planta*: (i) LUP5G co-expressed with CYP716A80 (G-type) and UGT73C11 (Figure 5a); and (ii) LUP2P co-expressed with CYP716A81 (P-type) and UGT73C11 (Figure 5b). These combinations were expected to generate 3-*O*-sapogenin monoglucosides, and indeed oleanolic acid 3-*O*-glucoside (peak 20, Figure 5a; Table S3) accumulated when LUP5G was used as the OSC. This was in accordance with the predominant production of oleanolic acid when LUP5G was co-expressed with CYP71A80 (Figure 4). Similarly, 3-*O*-monoglucosides of unknown 4 (peaks 15 and 16, Figure 5b; Table S4) accumulated when LUP2P was used as the OSC, which is in agreement with the predominant production of the lupeol-derived unknown 4 (51%) when LUP2P was co-expressed with CYP71A81 (Figure 4). Likewise, a 3-*O*-monoglucoside of betulinic acid (peak 15, Figure S5c) accumulated, which is in agreement with a predominant production of betulinic acid when LUP2P was co-expressed with CYP716A80 (Figure S5; Table S7).

Surprisingly, a total of 29 putative saponins, ranging from one to six hexose moieties (22 derived from LUP5G and 16 from LUP2P), were detected from the enzyme combinations described in Figure 5, and were tentatively identified based on elution time and MS/MS patterns (Figure 5; Tables S3 and S4). The higher number of putative saponins produced by the LUP5G combinations is in agreement with the more complex triterpenoid product profiles of LUP5G compared with LUP2, and of CYP716A81 compared with CYP716A80 (Figures 3 and 4). Several other combinations of G- and P-type genes of LUP2, LUP5 and CYP716A, as well as with the more unspecific *B. vulgaris* UGT73C13, were also tested. UGT73C13 is phylogenetically very close to UGT73C10 and UGT73C11, but is less specific, and glucosylates at both the C3 and C28 position, whereas UGT73C11 only glucosylates at C3 (Augustin *et al.*, 2012). These combinations showed essentially the same overall product complexity (Figure S5; Tables S5–S7), but with the main predictable exception that less of the 3-*O*-glucosides accumulate when the unspecific UGT73C13 was used.

In order to tentatively determine the position of sugars on the saponin backbones, extracts of *N. benthamiana* plant leaves transfected with (LUP5G +CYP716A80 + UGT73C11; Figure 5a) and (LUP2P + CYP716A80 + UGT73C11; Figure S5C) constructs were treated with sodium hydroxide and re-analysed by LC-MS (Figure 6a and b, respectively; Tables S3 and S7). Mild saponification only cleaved off sugars connected through ester linkages, such as at C28, but not sugars connected through glycosidic bonds, such as at position C3. Several peaks disappeared or diminished upon saponification (indicated by asterisks in Figures 5 and 6; see also

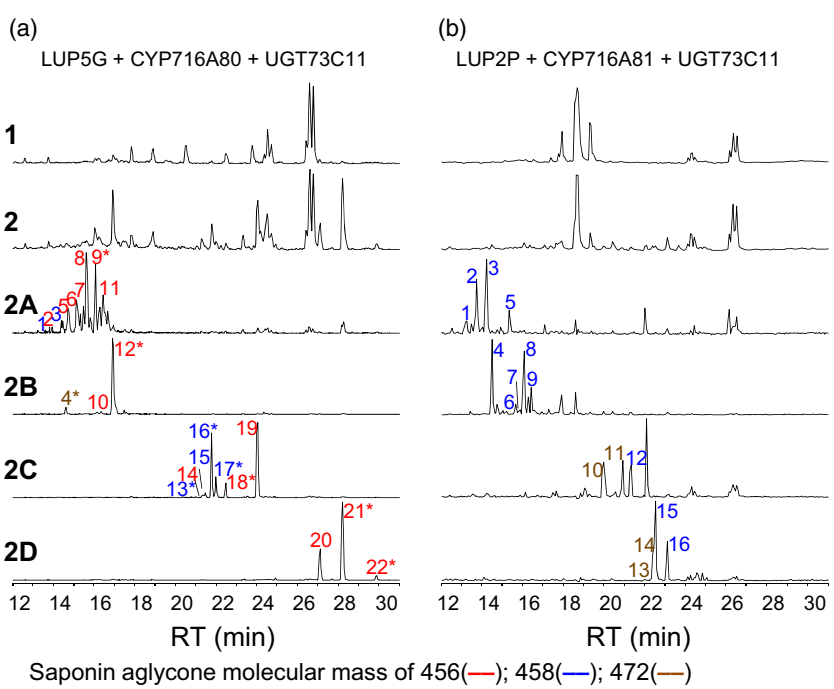

Figure 5. Saponins produced by co-expression of selected oxidosqualene cyclases (OSCs), P450s and a UDP-glycosyltransferase (UGT) from *Barbarea vulgaris* in *Nicotiana benthamiana* leaves. (1) Total ion current chromatograms (TICs) of LC-MS profiles for (a) *LUP5G* and (b) *LUP2P*, in combination with *CYP716A80* and *CYP716A81*, respectively. (2) TICs of *LUPs* and P450s in combination with (a) *UGT73C11* or (b) *UGT73C11*. (2A–2D) extracted ion chromatograms (EICs) of (2), representing saponins with: 2A, five and six hexose moieties [(a) *m/z* 1266–1268, 1308–1310, 1418–1420; (b) *m/z* 1268–1270, 1336–1338, 1428–1430]; 2B, four hexose moieties [(a) *m/z* 1108–1110, 1145–1147, 1190–1192; (b) *m/z* 1107–1109, 1120–1122, 1058–1060]; 2C, two hexose moieties [(a) *m/z* 823–825, 901–903, 1560–1566; (b) *m/z* 633–635, 796–799, 864–867, 902–904, 1564–1566]; and 2D, one hexose moiety [(a) *m/z* 617–619, 1236–1238, 1408–1411; (b) *m/z* 618–620, 666–668, 822–824, 888–890, 1440–1442]. The *y*-axes (ion count) of each chromatogram are log10 scaled to the highest peak intensity.
*Peaks correspond to saponins that significantly decreased or disappeared after sodium hydroxide-based saponification, indicating the presence of sugar moieties at the C28 position of the triterpene aglycones. See Figure S4 for MS/MS and Tables S3 and S4 for detailed information on these peaks.

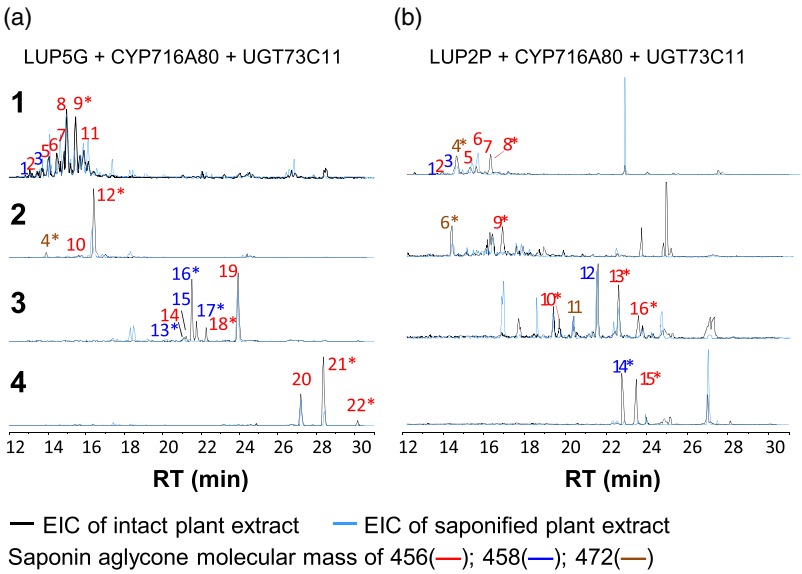

Figure 6. Effect of sodium hydroxide-based mild saponification of methanol extracts from *Nicotiana benthamiana* leaves transiently producing saponins. Extracted ion chromatograms (EICs) of LC-MS profiles for (a) LUP5G plus CYP716A80 plus UGT73C11 and (b) LUP2P plus CYP716A80 plus UGT73C11, using selected *m/z* ions representing saponins with: (1) five and six; (2) four; (3) two; and (4) one sugar moieties, as in Figure 5, before (black) and after (blue) saponification. The *y*-axes (ion count) are log10 scaled to the highest peak intensity.
*Peaks correspond to saponins that significantly decreased or disappeared after sodium hydroxide-based saponification, indicating the presence of sugar moieties at the C28 position. See Figure S6 for MS/MS and Tables S3 and S7 for detailed information on these peaks.

Tables S3 and S7, respectively), whereas others remained unchanged. This clearly indicates the presence of saponins glucosylated both at positions C28 and C3. For example, saponins derived from the aglycone, unknown 5, with five and four hexose moieties (peaks 4 and 6, respectively; Figure 6b; Table S7), diminished as a result of mild saponification.

LUP2s and LUP5s are differentially expressed in G- and P–types

The expression of OSCs, P450s and UGTs from *B. vulgaris* in *S. cerevisiae* and *N. benthamiana* revealed that the enzymes of G- and P-type plants have overlapping product profiles, yet the saponin composition of G- and P-type plants is very different. This suggests that the difference in saponin profile may be caused by differential expression of these genes. qPCR on *B. vulgaris* leaf, petiole and root RNA revealed that LUP5, producing mainly β-amyrin, is predominantly expressed in the G-type leaves, whereas LUP2, producing mainly lupeol, is predominantly expressed in P-type leaves (Figure 7). CYP716A80 from the G-type is expressed at higher levels than the corresponding CYP716A81 in the P-type, and UGT73C11 from the G-type is expressed at a similar level to its ortholog UGT73C10 in the P-type. As expected from the saponin profiles in the G- and P-type plants, the expression profiles of the genes tested follow the abundance of saponins (Augustin *et al.*, 2012), with the highest expression levels in the leaf, lower expression levels in the petiole and lowest expression levels in the roots.

DISCUSSION

A few species in the *Barbarea* genus are the only ones within the large and economically important Brassicaceae family know to produce triterpenoid saponins (Nielsen *et al.*, 2010a; Badenes-Perez *et al.*, 2014). Some of these saponins make *B. vulgaris* G-type plants resistant to important herbivores (Agerbirk *et al.*, 2003a; Kuzina *et al.*, 2009; Augustin *et al.*, 2012), whereas P-type plants, which produce structurally different saponins, are susceptible (Kuzina *et al.*, 2009). To uncover saponin biosynthesis in *B. vulgaris*, and elucidate how it is evolving and has diverged between G- and P-type plants, we have taken a multidisciplinary approach spanning genetics, genomics, transcriptomics, phylogenetics, synthetic biology and analytical chemistry.

Saponin biosynthesis in *Barbarea* evolved by the neofunctionalization of a UGT

Triterpenoid saponin biosynthesis typically relies on the recruitment of OSCs, P450s and UGTs as key enzymes (Seki *et al.*, 2015). We have identified and characterized two different OCSs (LUP2 and LUP5) that differ in product profile and are present in both the G- and P-types (Figures 3, 4 and 7), and two highly similar P450s (CYP716A80 and CYP716A81 in G- and P-type plants, respectively) that catalyse the same reactions but differ in product ratios (Figures 3c, 4 and 7). In a previous study (Augustin *et al.*, 2012), we identified two UDP-glucosyltransferases (UGT73C10 and UGT73C11) that are orthologous and have neofunctionalized to specifically glucosylate sapogenins at the C3 position, the unique position where all saponins are glycosylated. The genes coding for these OSCs, P450s and UGTs are closely related to the homologous genes in *A. thaliana*, yet *B. vulgaris* produces saponins and *A. thaliana* does not.

The four *B. vulgaris* OSCs described here (LUP2G/P and LUP5G/P) are phylogenetically closely related to

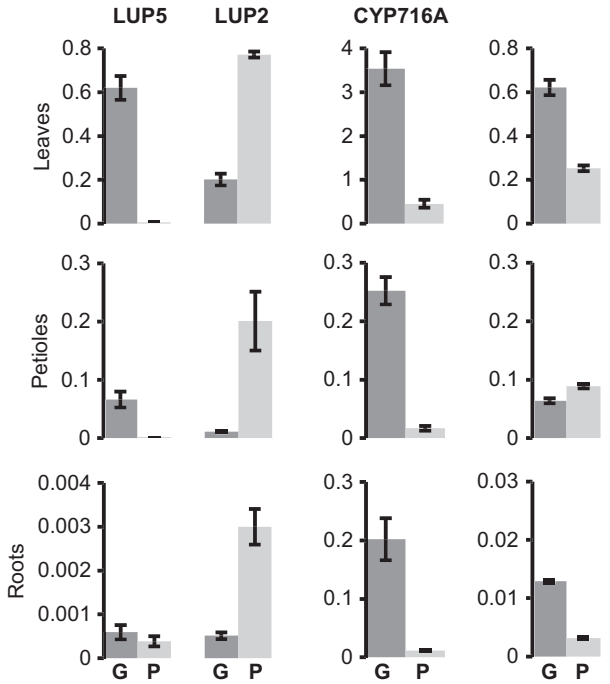

Figure 7. Relative expression of *LUP5* and *LUP2* from G- and P-type *Barbarea vulgaris*, *CYP716A80* (G), *CYP716A81* (P), *UGT73C10* (P) and *UGT73C11* (G) in leaves, petioles and roots of G- and P-type plants. Values for the *UGT73C* are adapted from Augustin *et al.* (2012); the primer combination for *UGT73C10* cannot discriminate between *UGT73C10* and the presumed pseudogene *UGT73C9*. Means ± SDs are shown relative to *ACT2*.

A. thaliana LUP1, LUP2 and LUP5, and accordingly are named after them (Figure 2). As seen for *B. vulgaris* LUP2 and LUP5, *A. thaliana* LUP1, LUP2 and LUP5 are also known to produce α-amyrin and β-amyrin, or lupeol (Augustin *et al.*, 2011; Figure 2). The *A. thaliana* and *B. vulgaris* OSCs are more than 80% identical at the amino acid level, and do not reveal any obvious sites that explain the difference in product specificity. P450s belonging to the CYP716A subfamily are well known to catalyse C28 carboxylations in *A. thaliana* and other species (Carelli *et al.*, 2011; Fukushima *et al.*, 2011; Moses *et al.*, 2015). Whereas CYP716A1 and CYP716A2 are mainly expressed in the roots in *A. thaliana*, CYP716A80 and CYP716A81 are expressed in the leaves of *B. vulgaris* (Figure 7). Furthermore, *B. vulgaris* CYP716As have evolved to hydroxylate at yet unknown positions, as well as to catalyse C28 carboxylations, leading to an increase in sapogenin complexity.

As *A. thaliana* and other crucifers also produce β-amyrin and contain CYP716As, this suggests that the ability to produce saponins in the *Barbarea* genus evolved by novel specificity of a UGT to glucosylate a triterpene at the 3-OH position (Augustin *et al.*, 2012). Whereas *A. thaliana* UGT73C5, the closest homologue to *B. vulgaris* UGT73C10/11, can also glucosylate sapogenins to some extent, it has a much lower regiospecificity and efficiency (Augustin *et al.*, 2012); such a specificity as found in *B. vulgaris* has not yet been reported for other crucifers, to our knowledge. Our hypothesis is further supported by analyses for selection signatures in the DNA sequences of the biosynthetic genes, using branch and site models in CODEML of the PAML package (Yang, 2007). Whereas *UGT73C10* and *UGT73C11* showed signs of positive selection (Augustin *et al.*, 2012), the *B. vulgaris* OSC and P450 sequences did not ($\omega < 1$).

Genes for saponin biosynthesis in *B. vulgaris* are unlinked

Recent research within genome organization has suggested that pathways for specialized metabolites may be organized in operon-like gene clusters, as shown for benzoxaxinoids, cyanogenic glucosides and avenacins (triterpenoid saponins) in oats (*Avena spp.*) (Nutzmann and Osbourn 2014). The mapping of the genes for the OSCs, P450s and UGTs characterized here showed that they are placed in different linkage groups. This has also been found for other well-studied biosynthetic pathways of specialized metabolites, such as anthocyanins and glucosinolates (Kliebenstein and Osbourn, 2012). Both anthocyanin and glucosinolate biosynthetic pathways are known to produce a range of different compounds with small structural differences, and the unlinked organization of the biosynthesis genes may serve to increase structural diversification by recombination and hybridization. We speculate that this is also the case for triterpenoid saponin biosynthesis in *B. vulgaris*. We initially thought that the saponin profiles were rather simple, and that the P-type did not contain saponins, but with increasing analytical capacity, we now know that there are more than 20 different saponin structures in both the G- and the P-types (our preliminary data). The OSC, P450 and UGT genes described here are each located in tandem repeats, where some of the repeats are most likely pseudogenes, and could reflect recent duplication events (Paquette *et al.*, 2000; Hanada *et al.*, 2008; Kondrashov, 2012). These duplications may also serve to increase the saponin structural diversity by creating gene versions with altered substrate and product specificities.

Insect resistance and susceptibility within *B. vulgaris* is determined by differing expression of OSCs

Hederagenin and oleanolic acid cellobioside are the most insecticidal and insect-deterring saponins in *B. vulgaris*, as shown by feeding experiments (Nielsen *et al.*, 2010a,b; Kuzina *et al.*, 2011; Augustin *et al.*, 2012). These two insecticidal saponins are only produced by G-type plants, whereas P-type plants produce other saponins with unknown effects and structures (Kuzina *et al.*, 2011). Hederagenin and oleanolic acid cellobioside are derived from β-amyrin, the main product of LUP5. Surprisingly, both the G- and P-type plants contain functional LUP5s, as shown

by heterologous expression in both *S. cereviceae* and *N. benthamiana*. These LUP5s are 98% identical at the amino acid level, but LUP5G produces a higher percentage of β-amyrin than found in LUP5P (78 and 36%, respectively; Figure 4). In contrast, LUP2 from both the G- and P-types predominantly produces lupeol and only small quantities of β-amyrin (Figure 4). We found a strong difference in expression of these OSCs between resistant and susceptible plants. G-type plants predominantly express *LUP5G*, which leads to the production of β-amyrin, the precursor of the insecticidal hederagenin and oleanolic acid cellobiosides. In contrast, P-type plants hardly express this gene (Figure 7). The expression of *LUP2*, which leads to the production of lupeol, is high in P- but not in G-type plants (Figure 7). The differential expression of *LUP5* and *LUP2* is in agreement with our own transcriptome results and with those from Wei *et al.* (2013), who found *LUP5* in the G-type to be upregulated upon exposure to *P. xylostella* (diamondback moth), whereas *LUP2* was not.

As the insect-resistant G-type and insect-susceptible P-type of *B. vulgaris* do not differ in substrate, nor in the product profiles of UGTs (UGTC10 and UGTC11) or P450s (CYP716A80 and CYP716A81), we propose that the difference in insect resistance between the two *B. vulgaris* chemotypes is caused by the differential expression of *LUP5* and *LUP2*, leading to an accumulation of different triterpenoid backbones in the two types. Agerbirk *et al.* (2003b) suggested that insect resistance was lost in the P-type of *B. vulgaris*; our present results thus suggest that this loss resulted from a change in gene expression and the associated terpenoid backbone structures.

The production of monoglucosylated saponins in *N. benthamiana* leads to hyperglucosylation

By combining *B. vulgaris* OSCs, P450s and UGTs from both the G- and P-types, we have reconstituted part of the biosynthesis of *B. vulgaris* sapogenins and monoglucosylated saponins in *N. benthamiana*. Monoglycosylated saponins have previously been produced by combinational biosynthesis in *S. cerevisiae* (Moses *et al.*, 2014) and by using recombinant UGT73C10 *in vitro* (Augustin *et al.*, 2012). In contrast to *S. cerevisiae*, the monoglucosides produced in *N. benthamiana* were further glycosylated with up to five additional hexoses. Mild saponification, which removes glycoses linked as esters at the C28 position, revealed that both the C3 and C28 positions have been glycosylated, and that these hexoses are subjected to further glycosylation (Figures 5 and 6). Saponins have detergent-like properties and interact with membranes, and this is partly determined by their glycosylation patterns. We speculate that the 3-O-monoglucosylated saponins are cytotoxic to *N. benthamiana*, as *N. benthamiana* readily accumulates the sapogenins without glycosylating at the C28 position. This is supported by the lack of monoglucosylated saponins in *B. vulgaris* tissues, and by the apparent lack of reports of monoglucosylated saponins in other plant species. The toxicity of 3-O-monoglucosylated saponins has also been implicated in abnormal phenotypic changes of *Medicago truncatula* roots: the so-called makibishi phenotype (Pollier *et al.*, 2013). The glycosylation of low-molecular-weight compounds typically serves to detoxify them by changing their solubility and bioactivity (Bowles *et al.*, 2005; Kristensen *et al.*, 2005), and, accordingly, the accumulation of hyperglucosylated saponins in *N. benthamiana* could indicate a detoxification reaction.

Conclusions

We have identified key enzymes in the biosynthesis of saponins in *B. vulgaris* and proposed crucial enzymatic changes in the evolution of saponin biosynthesis in the *Barbarea* genus, the only crucifer genus known to produce saponins. The emergence of a UGT with a positional specificity seems to have enabled the ability to synthesize saponins. The difference in insect resistance between the two *B. vulgaris* chemotypes, G- and P-type, seems to be caused by the differential expression of *LUP2* and *LUP5* with different product profiles. Establishing a high-quality genome sequence, and transformation protocols for directed genome editing and overexpression, is needed to determine *in planta* functions and further elucidate the impact and evolution of specific saponins in plant–insect relationships in *B. vulgaris*. A better resolution of the QTL region with the CYP72A gene candidates on linkage group 10 should facilitate the identification of the P450 that catalyses the C23 hydroxylation to hederagenin, and thus enable *in planta* production of the main insecticidal monoglycosylated saponin. Our metabolic engineering results pave the way for future metabolic engineering of a diversity of triterpenes and saponins for use in plant production and pharmacy, and for breeding insect-resistant crop plants.

EXPERIMENTAL PROCEDURES

Transcriptome and genome sequencing and gene mining

The *B. vulgaris* ssp. *arcuata* P-type transcriptome was sequenced as previously described for the G-type (Kuzina *et al.*, 2011). The two transcriptomes were assembled with CAP3 and MOSAIK, and the CAP3-based assembly was used as a simulated reference genome. Two short-insert DNA libraries (with an insert size of 150–350 bp) from the genomes of the P- and G-types were sequenced using Illumina HiSeq 2000 with a pair-end channel module with 2×100-bp read lengths, and assembled by Eurofins MWG Operon. A combination of BLAST analysis, search for specific domains using HMMER and the Pfam database, and the CLC Genomic Workbench was used to identify OSC, P450 and UGT candidates.

Map construction and QTL analysis

Barbarea vulgaris F_2 population linkage groups were constructed using JOINMAP, as previously described (Kuzina *et al.*, 2011). This was based on 68 simple sequence repeats, 39 amplified fragment

length polymorphism and markers in six genes, of which 39 SSRs were from Kuzina *et al.* (2011), and using an LOD threshold of six in JOINMAP. The QTL analysis for resistance was performed as previously described (Kuzina *et al.*, 2011), using a generalized linear model of binomially distributed data (McCullagh *et al.*, 1989). The traditional logit link function [$\log(p/1-p)$] was substituted with a power function [$(p\beta - 1)/\beta$], with $\beta = 0.1$, to gain more uniformly distributed residuals and a better model fit. The effects of dominance within the two QTLs and interactions between them were non-significant, and were therefore removed from the final model.

Expression in *Saccharomyces cerevisiae*

For expression analysis in *S. cerevisiae* INVSc1 (Invitrogen, now ThermoFisher Scientific, http://www.thermofisher.com), CYP716A80 and CYP716A81 were cloned into pENTR/D-TOPO (Invitrogen) and transferred to pYES-DEST2 (Invitrogen) as well as pELC destination vectors (Seki *et al.*, 2008) using Gateway technology. For expression assays, pYES3-ADH-OSC1 (Seki *et al.*, 2008) was co-transformed with the pELC constructs using Frozen-EX Yeast Transformation IITM (Zymo Research, http://www.zymoresearch.com). Yeast cells were cultured in synthetic complete medium containing 2% glucose at 30°C for 1 day by shaking at 175 rpm. The cells were then collected and resuspended in synthetic complete medium without tryptophan and leucine medium (10 ml), containing 13 μg ml^{-1} hemin and 2% galactose instead of glucose, and cultured at 30°C for 2 days by shaking at 175 rpm.

Expression in *N. benthamiana*

The selected genes were amplified by PCR from cDNA synthesized from *B. vulgaris* RNA and Gateway cloned into pDONR 207 (Invitrogen), and then into the pEAQ-HT-DEST expression vector (Sainsbury *et al.*, 2012). The primers used for amplification are listed in Table S9. The expression constructs were transformed into *Agrobacterium tumefaciens* strain AGL1. One day prior to infiltration, about five colonies of *Ag. tumefaciens* were used as inoculum for 10 ml of LB medium supplemented with appropriate antibiotics. The cultures were grown for 22 h at 28°C to an OD_{600} of ~2.5. The cultures were centrifuged and the collected cells resuspended in infiltration buffer (10 mM $MgCl_2$, 10 mM MES, pH 5.6, and 100 μM acetosyringone), and OD_{600} was adjusted to 1.2. The suspension was incubated for 2 h at room temperature (20–25°C), and slightly shaken. Co-infiltration of genes was performed with cultures mixed in equal density. *Nicotiana benthamiana* plants were grown in soil (Pindstrup substrate no. 2) in a glasshouse with a 16-h light/8-h dark photoperiod at 28°C. Transient transformation of leaves was performed on 3–4-weeks old plants.

Preparation of microsomes

Microsomes were isolated from a single leaf (~0.3 g fresh weight) 5 days post-infiltration, as previously described (Blomstedt *et al.*, 2012). The microsomal pellet was resuspended in 90 μl of resuspension buffer (50 mM Tricine, pH 7.9, 20 mM NaCl, 2 mM DTT), and 20 μl was used for enzyme activity assays in the presence of an additional 50 μl of resuspension buffer, with 1 μl of 100 mM 2,3-oxidosqualene as substrate. Microsome assay solutions were incubated at 30°C for 2 h, with shaking at 600 rpm, and were extracted twice with ethyl acetate at a ratio of 1 : 1 (vol : vol) prior to GC-MS analysis.

Metabolite analysis by GC-MS and LC-MS

Yeast cultures (with a starting volume of 5 ml) were extracted using 3 ml of ethyl acetate (Wako, http://www.wako-chem.co.jp), by 1 min of vortexing, 30 min of sonication (70% intensity; Sharp, http://www.sharp-world.com), followed by centrifuging at 200 ***g*** for 5 min. The organic phase was transferred to a silica-gel column (6 cc; Waters, http://www.waters.com) and eluted with 10 ml of ethyl acetate and 10 ml of chloroform-methanol (1 : 1 dilution). The eluates were placed into an evaporator for 60 min, and re-suspended in 300 μl of chloroform-methanol (1 : 1); 100 μl was transferred to a vial and evaporated in vacuum for 30 min. Finally, the pellet was trimethylsilylated with 50 μl of MSTFA (Sigma-Aldrich, http://www.sigmaaldrich.com) for 30 min at 80°C. GC-MS analysis was performed as previously described (Fukushima *et al.*, 2011). Single *N. benthamiana* leaves (~0.3 g fresh weight) 6 days post-infiltration were ground to a fine powder under liquid N_2 in a mortar and pestle. One half of the powder was placed immediately into a tube with 1.5 ml of 85% methanol for LC-MS analysis, and the other half was placed into a tube with 1.5 ml of ethyl acetate for GC-MS analysis. LC-MS analysis was performed as previously described (Augustin *et al.*, 2012), and GC-MS analysis was also performed as previously described (Khakimov *et al.*, 2013) but with slight modifications of the GC oven temperature gradient (40°C at 0 min, then 12°C min^{-1} to 310°C and hold for 8 min) and solvent delay time (20 min). Raw GC-MS data were processed by the PARAFAC2 method (Khakimov *et al.*, 2012) in order to deconvolute and quantify GC-MS peaks, using MATLAB 8.2.0.701 (2013b; http://www.mathworks.com). Metabolites were identified by comparing retention times and mass spectra with those of authentic standards. Standards of β-amyrin, α-amyrin, lupeol, hederagenin, oleanolic acid and betulinic acid were purchased from Sigma-Aldrich (Fluka); standards of ursolic acid, erythrodiol, uvaol, betulin, ursolic aldehyde, betulinic aldehyde and oleanolic aldehyde were available within the group. Relative concentrations of PARAFAC2 based on deconvoluted peaks were normalized with the norm-2 approach (division of each peak area by the sum of areas of all peaks detected in the same sample).

Gene expression analysis

To estimate gene expression levels, qPCR using the DyNAmo™ Flash SYBR® Green qPCR Kit (Finnzymes, now ThermoFisher Scientific, http://www.thermofisher.com) and a Rotor Gene Q machine (QIAGEN, http://www.qiagen.com) was performed as previously described (Augustin *et al.*, 2012). cDNA synthesized from 2-month-old G- and P-type *B. vulgaris* leaves (Augustin *et al.*, 2012) were used as a template. Primer sequences are listed in Table S9. The expression levels were compared with Actin (Augustin *et al.*, 2012). qPCR products were verified by sequencing (LUP2, Primer15 + Primer17; LUP5, Primer13 + Primer14; CYP716A, Primer15 + Primer16). As a result of the very high degree of sequence identity between UGT73C9 and UGT73C10 they cannot be distinguished by qPCR; however, UGT73C9 may be a pseudogene, and does not glucosylate sapogenins *in vitro* (Augustin *et al.*, 2012), and thus most likely does not contribute to the saponin profiles in *B. vulgaris* when or if expressed in plants.

Phylogenetic analysis of OSCs

Amino acid OSC sequences from *B. vulgaris* and *A. thaliana* were aligned using MUSCLE and used to construct a maximum-likelihood bootstrapped phylogenetic tree based on the Tamura three-parameter model, using MEGA 6.06 (Tamura *et al.*, 2013). Sequences with bootstrap values below 75 were removed from the alignment and a second phylogenetic tree was made. A discrete Gamma distribution was used to model evolutionary rate differences among sites (five categories; +*G*, parameter = 0.9355). The tree is drawn

to scale, with branch lengths measured in the number of substitutions per site. All positions with <95% site coverage were eliminated. Products produced by the OSC are adapted from Augustin *et al.* (2011).

ACKNOWLEDGEMENTS

Dr David Nelson is thanked for naming CYP716A80 and CYP716A81 according to the P450 nomenclature. Drs Takao Kirita and Kiyoshi Ohyama are thanked for providing ursolic aldehyde, betulinic aldehyde and oleanolic aldehyde standards. This work was supported by the Danish Council for Independent Research, Technology and Production Sciences (grant nos. 09-065899 and 1335-00151), the Villum Kann Rasmussen Foundation to Pro-Active Plants, the European Community's Seventh Framework Programme (FP7/2007-2013; grant no. 613692 to TriForC), and by a PhD stipend from the Department of Plant and Environmental Sciences to PØE.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Multiple alignment of the selected OSC sequences from *Barbarea vulgaris* and *Arabidopsis thaliana*.

Figure S2. *Barbarea vulgaris* quantitative trait loci map.

Figure S3. EI-MS fragmentation patterns and retention times of trimethylsilyl (TMS) derivatives of 13 authentic triterpene standards and the five tentatively identified triterpenoids (unknowns 1–5).

Figure S4. LC-MS/MS spectra of tentatively identified saponins produced by expression in *Nicotiana benthamiana*.

Figure S5. Saponins produced by expression of different combinations of *B. vulgaris* genes coding for OSCs, P450s and UGTs in *N. benthamiana* plant leaves.

Table S1. Microsatellite (SSR) markers used for gene mapping and QTL analysis.

Table S2. Quantitative trait loci for G- and P-type saponins and glucosinolates.

Table S3. Triterpenoid saponins produced by the heterologous expression of LUP5G plus CYP716A80 plus UGT73C11 in *Nicotiana benthamiana*.

Table S4. Triterpenoid saponins produced by the heterologous expression of LUP2P plus CYP716A81 plus UGT73C11 in *Nicotiana benthamiana*.

Table S5. Triterpenoid saponins produced by the heterologous expression of LUP2G plus CYP716A80 plus UGT73C11 in *Nicotiana benthamiana.*

Table S6. Triterpenoid saponins produced by the heterologous expression of LUP2P plus CYP716A81 plus UGT73C13 in *Nicotiana benthamiana.*

Table S7. Triterpenoid saponins produced by the heterologous expression of LUP2P plus CYP716A80 plus UGT73C11 in *Nicotiana benthamiana*.

Table S8. Tentative identification of triterpenoid saponin aglycones of unknowns 1–5.

Table S9. Oligonucleotide primers used for cloning and expression analyses.

REFERENCES

Agerbirk, N., Olsen, C.E., Bibby, B.M., Frandsen, H.O., Brown, L.D., Nielsen, J.K. and Renwick, J.A.A. (2003a) A saponin correlated with variable resistance of *Barbarea vulgaris* to the diamondback moth *Plutella xylostella*. *J. Chem. Ecol.* **29**, 1417–1433.

Agerbirk, N., Ørgaard, M. and Nielsen, J.K. (2003b) Glucosinolates, flea beetle resistance, and leaf pubescence as taxonomic characters in the genus Barbarea (Brassicaceae). *Phytochemistry*, **63**, 69–80.

Augustin, J.M., Kuzina, V., Andersen, S.B. and Bak, S. (2011) Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry*, **72**, 435–457.

Augustin, J.M., Drok, S., Shinoda, T. *et al.* (2012) UDP-glycosyltransferases from the UGT73C subfamily in *Barbarea vulgaris* catalyze sapogenin 3-O-glucosylation in saponin-mediated insect resistance. *Plant Physiol.* **160**, 1881–1895.

Badenes-Perez, F.R., Gershenzon, J. and Heckel, D.G. (2014) Insect attraction versus plant defense: young leaves high in glucosinolates stimulate oviposition by a specialist herbivore despite poor larval survival due to high saponin content. *PLoS One*, **9**, e95766.

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S. and Werck-Reichhart, D. (2011) Cytochromes P450. *Arabidopsis Book*, **9**, e0144. doi:10.1199/tab.0144.

Blomstedt, C.K., Gleadow, R.M., O'Donnell, N., Naur, P., Jensen, K., Laursen, T., Olsen, C.E., Stuart, P., Hamill, J.D. and Møller, B.L. (2012) A combined biochemical screen and TILLING approach identifies mutations in *Sorghum bicolor* L. Moench resulting in acyanogenic forage production. *Plant Biotechnol. J.* **10**, 54–66.

Bowles, D., Isayenkova, J., Lim, E.K. and Poppenberger, B. (2005) Glycosyltransferases: managers of small molecules. *Curr. Opin. Plant Biol.* **8**, 254–263.

Carelli, M., Biazzi, E., Panara, F., Tava, A., Scaramelli, L., Porceddu, A., Graham, N., Odoardi, M., Piano, E. and Arcioni, S. (2011) *Medicago truncatula* CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell*, **23**, 3070–3081.

Christensen, S., Heimes, C., Agerbirk, N., Kuzina, V., Olsen, C.E. and Hauser, T.P. (2014) Different geographical distributions of two chemotypes of *Barbarea vulgaris* that differ in resistance to insects and a pathogen. *J. Chem. Ecol.* **40**, 491–501.

Chu, H.Y., Wegel, E. and Osbourn, A. (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* **66**, 66–79.

Field, B. and Osbourn, A.E. (2008) Metabolic diversification - independent assembly of operon-like gene clusters in different plants. *Science*, **320**, 543–547.

Field, B., Fiston-Lavier, A.-S., Kemen, A., Geisler, K., Quesneville, H. and Osbourn, A.E. (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl Acad. Sci. USA*, **108**, 16116–16121.

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B. and Briggs, S.P. (1997) Analysis of a chemical plant defense mechanism in grasses. *Science*, **277**, 696–699.

Fukushima, E.O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., Saito, K. and Muranaka, T. (2011) CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant Cell Physiol.* **52**, 2050–2061.

Hamberger, B. and Bak, S. (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120426.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K. and Shiu, S.-H. (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003.

Khakimov, B., Amigo, J.M., Bak, S. and Engelsen, S.B. (2012) Plant metabolomics: resolution and quantification of elusive peaks in liquid chromatography–mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *J. Chromatogr.* **1266**, 84–94.

Khakimov, B., Motawia, M.S., Bak, S. and Engelsen, S.B. (2013) The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography–mass spectrometry based metabolomics. *Anal. Bioanal. Chem.* **405**, 9193–9205.

Kliebenstein, D.J. and Osbourn, A. (2012) Making new molecules - evolution of pathways for novel metabolites in plants. *Curr. Opin. Plant Biol.* **15**, 415–423.

Kondrashov, F.A. (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* **279**, 5048–5057.

Kristensen, C., Morant, M., Olsen, C.E., Ekstrøm, C.T., Galbraith, D.W., Lindberg Møller, B. and Bak, S. (2005) Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. *Proc. Natl Acad. Sci. USA*, **102**, 1779–1784.

Kuzina, V., Ekstrom, C.T., Andersen, S.B., Nielsen, J.K., Olsen, C.E. and Bak, S. (2009) Identification of defense compounds in *Barbarea vulgaris* against the herbivore *Phyllotreta nemorum* by an ecometabolomic approach. *Plant Physiol.* **151**, 1977–1990.

Kuzina, V., Nielsen, J.K., Augustin, J.M., Torp, A.M., Bak, S. and Andersen, S.B. (2011) *Barbarea vulgaris* linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore *Phyllotreta nemorum*. *Phytochemistry*, **72**, 188–198.

McCullagh, P., Nelder, J.A. and McCullagh, P. (1989) *Generalized Linear Models*. London: Chapman and Hall London.

Moses, T., Pollier, J., Almagro, L., Buyst, D., Van Montagu, M., Pedreno, M.A., Martins, J.C., Thevelein, J.M. and Goossens, A. (2014) Combinatorial biosynthesis of sapogenins and saponins in *Saccharomyces cerevisiae* using a C-16alpha hydroxylase from *Bupleurum falcatum*. *Proc. Natl Acad. Sci. USA*, **111**, 1634–1639.

Moses, T., Pollier, J., Faizal, A., Apers, S., Pieters, L., Thevelein, J.M., Geelen, D. and Goossens, A. (2015) Unraveling the triterpenoid saponin biosynthesis of the African shrub *Maesa lanceolata*. *Mol. Plant*, **8**, 122–135.

Mugford, S.T., Louveau, T., Melton, R., Qi, X., Bakht, S., Hill, L., Tsurushima, T., Honkanen, S., Rosser, S.J. and Lomonossoff, G.P. (2013) Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. *Plant Cell*, **25**, 1078–1092.

Nielsen, J., Nagao, T., Okabe, H. and Shinoda, T. (2010a) Resistance in the plant, *Barbarea vulgaris*, and counter-adaptations in flea beetles mediated by saponins. *J. Chem. Ecol.* **36**, 277–285.

Nielsen, N.J., Nielsen, J. and Staerk, D. (2010b) New resistance-correlated saponins from the insect-resistant crucifer *Barbarea vulgaris*. *J. Agric. Food Chem.* **58**, 5509–5514.

Nützmann, H.W. and Osbourn, A. (2014) Gene clustering in plant specialized metabolism. *Curr Opin Biotechnol.*, **26**, 91–99. Review. Erratum in: Curr Opin Biotechnol. 2014 Apr;26:199. PubMed PMID: 24679264.

Ørgaard, M. and Linde-Laursen, I.B. (2007) Cytogenetics of Danish species of Barbarea (Brassicaceae): chromocentres, chromosomes and rDNA sites. *Hereditas*, **144**, 159–170.

Osbourn, A. (1996) Saponins and plant defence – a soap story. *Trends Plant Sci.* **1**, 4–9.

Osbourn, A. (2010) Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiol.* **154**, 531–535.

Paquette, S.M., Bak, S. and Feyereisen, R. (2000) Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* **19**, 307–317.

Phillips, D.R., Rasbery, J.M., Bartel, B. and Matsuda, S.P. (2006) Biosynthetic diversity in plant triterpene cyclization. *Curr. Opin. Plant Biol.* **9**, 305–314.

Pollier, J., Moses, T., Gonzalez-Guzman, M. *et al.* (2013) The protein quality control system manages plant defence compound synthesis. *Nature*, **504**, 148–152.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R. and Osbourn, A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl Acad. Sci. USA*, **101**, 8233–8238.

Renwick, J.A.A. (2002) The chemical world of crucivores: lures, treats and traps. *Entomol. Exp. Appl.* **104**, 35–42.

Sainsbury, F. and Lomonossoff, G.P. (2008) Extremely high-level and rapid transient protein production in plants without the use of viral replication. *Plant Physiol.* **148**, 1212–1218.

Sainsbury, F., Saxena, P., Geisler, K., Osbourn, A. and Lomonossoff, G.P. (2012) Using a virus-derived system to manipulate plant natural product biosynthetic pathways. *Methods Enzymol.* **517**, 185–202.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K. and Muranaka, T. (2008) Licorice β-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. *Proc. Natl Acad. Sci. USA*, **105**, 14204–14209.

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., Fukushima, E.O., Akashi, T., Aoki, T. and Saito, K. (2011) Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *Plant Cell*, **23**, 4112–4123.

Seki, H., Tamura, K. and Muranaka, T. (2015) P450s and UGTs: key players in the structural diversity of triterpenoid saponins. *Plant Cell Physiol.* **56**, 1463–1471.

Shinoda, T., Nagao, T., Nakayama, M., Serizawa, H., Koshioka, M., Okabe, H. and Kawai, A. (2002) Identification of a triterpenoid saponin from a crucifer, *Barbarea vulgaris*, as a feeding deterrent to the diamondback moth, *Plutella xylostella*. *J. Chem. Ecol.* **28**, 587–599.

Takos, A.M. and Rook, F. (2012) Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.* **17**, 383–388.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. and Osbourn, A. (2014) Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* **65**, 225–257.

Wei, X., Zhang, X., Shen, D., Wang, H., Wu, Q., Lu, P., Qiu, Y., Song, J., Zhang, Y. and Li, X. (2013) Transcriptome analysis of *Barbarea vulgaris* infested with Diamondback Moth (*Plutella xylostella*) larvae. *PLoS One*, **8**, e64481.

Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., ÓMáille, P., Osbourn, A. and Qi, X. (2012) Divergent evolution of oxidosqualene cyclases in plants. *New Phytol.* **193**, 1022–1038.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.