RUNNING HEAD: Integrated Metabolomics Identifies Saponin P450s

*Corresponding authors:

Lloyd W. Sumner

University of Missouri

Department of Biochemistry

Bond Life Sciences Center

1201 Rollins Street

Columbia, MO 65211

Phone: 573-882-5486

sumnerlw@missouri.edu

Vered Tzin

Jacob Blaustein Institutes for Desert Research

Ben-Gurion University of the Negev

Sede Boqer Campus, 849900, Israel

Phone: +972-(0)8-6596749

vtzin@bgu.ac.il

Journal Research Area: Biochemistry and Metabolism

Secondary Research Area: Breakthrough Technologies

TITLE: Integrated Metabolomics Identifies CYP72A67 and CYP72A68 Oxidases in the Biosynthesis of *Medicago truncatula* Oleanate Sapogenins

Authors: Vered Tzin*[1,2], John H. Snyder*[1,3,4], Dong Sik Yang[1,5], David V. Huhman[1], Bonnie S. Watson[1], Stacy N. Allen[1], Yuhong Tang[1], Karel Miettinen[6,7], Philipp Arendt[6,7], Jacob Pollier[6,7], Alain Goossens[6,7], Lloyd W. Sumner[1,8]†

*These authors have contributed equally to this report.

[1]Plant Biology Division, Noble Research Institute, Ardmore, Okla. 73401, USA
[2]Present address: French Associates Institute for Agriculture and Biotechnology of Drylands, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus, Israel
[3]Department of Plant Biology, Cornell University, Ithaca, NY 14850, USA
[4]Present address: National Institute of Biological Sciences, Beijing, China
[5]Present address: Biomaterials Research Center, Samsung Advanced Institute of Technology, Suwon, South Korea
[6] Ghent University, Department of Plant Biotechnology and Bioinformatics, 9052 Gent, Belgium
[7] VIB Center for Plant Systems Biology, 9052 Gent, Belgium
[8]University of Missouri, Department of Biochemistry, Interdisciplinary Plant Group, Bond Life Sciences Center, Columbia, MO 65211

†Corresponding authors: Lloyd W. Sumner (email sumnerlw@missouri.edu) and Vered Tzin (vtzin@bgu.ac.il)

ABSTRACT (150-250 words)

Triterpene saponins are important bioactive plant natural products found in many plant families including the Leguminosae. Here we characterize two *Medicago truncatula* cytochrome P450 enzymes, MtCYP72A67 and MtCYP72A68, involved in saponin biosynthesis including both *in vitro* and *in planta* evidence. UHPLC-(-)ESI-QToF-MS was used to profile saponin accumulation across a collection of 106 *M. truncatula* ecotypes. The profiling results identified numerous ecotypes with high and low saponin accumulation in root and aerial tissues. Four ecotypes with significant differential saponin content in the root and/or aerial tissues were selected, and correlated gene expression profiling was performed. Correlation analyses between gene expression and saponin accumulation revealed high correlations between saponin content with gene expression of *β-amyrin synthase*, *MtCYP716A12*, and two cytochromes P450 genes, *MtCYP72A67* and *MtCYP72A68*. *In vivo* and *in vitro* biochemical assays using yeast microsomes containing MtCYP72A67 revealed hydroxylase activity for carbon 2 of oleanolic acid and hederagenin. This finding was supported by functional characterization of *MtCYP72A67* using RNAi-mediated gene silencing in *M. truncatula* hairy roots, which revealed a significant reduction of 2β-hydroxylated sapogenins. *In vivo* and *in vitro* assays with MtCYP72A68 produced in yeast showed multifunctional oxidase activity for carbon 23 of oleanolic acid and hederagenin. These findings were supported by overexpression of *MtCYP72A68* in *M. truncatula* hairy roots, which revealed significant increases of oleanolic acid, 2β-hydroxyoleanolic acid, hederagenin and total saponin levels. The cumulative data support that MtCYP72A68 is a multisubstrate, multifunctional oxidase and MtCYP72A67 is a 2β-hydroxylase, both of which function during the early steps of triterpene-oleanate sapogenin biosynthesis.

Keywords: Saponin, sapogenin, cytochrome P450, CYP72A67, CYP72A68, *Medicago truncatula*, integrated metabolomics

INTRODUCTION

Saponins are steroidal, steroidal alkaloid or triterpenoid metabolites that are typically conjugated with sugars and present in numerous plant species, including members of the genus *Medicago* (Augustin et al. 2011; Avato et al. 2006; Bialy et al. 1999; Gholami et al. 2014; Huhman et al. 2005; Huhman and Sumner 2002; Pollier et al. 2011; Tava et al. 2011). Many triterpene saponin aglycones (saponins without sugars, also known as sapogenins) are oxidized at various positions on the aglycone (Figure 1). These oxidized positions are often further conjugated with varying numbers of sugars to yield a multitude of saponins. Saponins possess diverse biological activities and plant beneficial properties, which include antifungal, antibacterial, antiviral, antitumor, molluscicidal, insecticidal and antifeedant activities (Augustin et al. 2011; Avato et al. 2006; Dixon and Sumner 2003; Klita et al. 1996; Lu and Jorgensen 1987; Sparg et al. 2004; Yan et al. 2013). In addition, they also affect plant development, including seed germination, vegetative growth and differentiation, fruiting and nodulation (Moses, Papadopoulou, et al. 2014). The pharmacological properties of saponins have been exploited in herbal medicines and, more recently, evaluated for their anticholesterolemic, anticancer, and adjuvant properties (Haridas et al. 2001; Kirk et al. 2004; Kuljanabhagavad et al. 2008; Shibata 2001).

Triterpene saponins constitutively accumulate in plants. However, the saponin biosynthetic pathway and additional accumulation of saponins are further induced during wounding, herbivory and by methyl jasmonate, which is a signaling compound associated with the induction of many defense-responsive plant metabolites (Broeckling et al. 2005; Gholami et al. 2014; Naoumkina et al. 2007; Suzuki et al. 2005). Although saponins are beneficial plant defense compounds, saponins in legume forages, such as alfalfa (*Medicago sativa*), are of particular and substantial economic importance because they result in impaired digestion and reduced weight gain in ruminant animals (Lu and Jorgensen 1987; Sen et al. 1998). Thus, saponins are considered antifeedants in premiere forages such as alfalfa. A detailed molecular and biochemical understanding of saponin biosynthesis would enable future metabolic engineering of crops with increased defense properties resulting in improved fitness and field productivity, and decreased anti-nutrient properties that would result in enhanced livestock weight gain performance. Metabolically engineered legumes with improved performance and

nutritive value would have substantial commercial value that would advance the plant biotechnology industry.

Recent studies in the *Medicago* genus have focused on elucidating the relationship between the biological activities of saponins and their chemical structures. The aglycone type, along with the nature and position of the sugar moieties, appear to correlate with different biological properties (Gholami et al. 2014; Tava et al. 2011). Two distinct classes of sapogenins can be differentiated in *Medicago* spp. based upon the position and the degree of oxidation: (i) sapogenins possessing a hydroxyl group at the C-24 position, without any substituent at the C-28 position atom (i.e., soyasapogenols; A, B and E); and (ii) sapogenins possessing a carboxyl group at the C-28 position that often also contain different oxidized states at the C-23 position (i.e., oleanate sapogenins with H, OH, CHO, or COOH at the C-23 position) (Carelli et al. 2011; Fukushima et al. 2011; Gholami et al. 2014). Some saponins possess hemolytic activity that results from their affinity for membranes, and this activity is related to the nature of the aglycone moiety (Augustin et al. 2011). No hemolytic activity was observed for soyasapogenols (Yoshiki et al. 1998), while oleanate derived sapogenins possessed high (hederagenin and medicagenic acid glycosides) to moderate (zanhic acid glycosides) hemolytic activities (Oleszek 1996). Recently, ectopic accumulation of bioactive monoglycosylated saponins was suggested to affect the integrity of *M. truncatula* roots themselves; hence, saponin producing plants need to develop self-protection mechanisms to allow accumulation of saponins (Pollier et al. 2013).

The first committed step in triterpene saponin biosynthesis is the cyclization of 2,3-oxidosqualene. This reaction is catalyzed by a specific oxidosqualene cyclase (*e.g., β-amyrin synthase; βAS*) which has been functionally characterized in many plant species (Inagaki et al. 2011; Iturbe-Ormaetxe et al. 2003; Sawai and Saito 2011; Suzuki et al. 2002; Thimmappa et al. 2014). Subsequent modifications that impart functional properties and diversify the basic triterpene backbone include the addition of small functional groups such as hydroxy, keto, aldehyde and carboxy moieties which are often followed by glycosylation (Augustin et al. 2012; Miettinen et al. 2018; Moses, Papadopoulou, et al. 2014; Thimmappa et al. 2014).The oxidative reactions prior to glycosylation are catalyzed by cytochrome P450-dependent monooxygenases (P450s). To date, several P450s that utilize β-amyrin as a substrate have been identified in dicotyledonous plants, whereas just one (CYP51H10) has been identified in monocots (Geisler et

al. 2013; Kunii et al. 2012). In oat (*Avena strigosa*) AsCYP51H10 (*Sad2*) is a multifunctional P450 that catalyzes the oxidation of *β*-amyrin on both the C and D rings to give 12,13β-epoxy-16β-hydroxy-*β*-amyrin, an intermediate of root saponin biosynthesis (Geisler et al. 2013). Currently, a significant number of other saponin biosynthetic cytochrome P450s have been characterized and *in planta* activity inferred via heterologous expression predominantly in yeast and to a lesser extent in *Nicotiana benthamiana*. In dicots, GuCYP88D6 in licorice (*Glycyrrhiza uralensis*, Fabaceae) catalyzes the C-11 oxidation of *β*-amyrin in glycyrrhizin biosynthesis (Seki et al. 2008). Members of the CYP93E subfamily of P450s have been shown to catalyze the C-24 hydroxylation of *β*-amyrin in soyasapogenol biosynthesis (Fukushima et al. 2011; Moses, Thevelein, et al. 2014; Seki et al. 2008; Shibuya et al. 2006) CYP87D16 catalyzes the C-16α hydroxylation of *β*-amyrin in the biosynthesis of maesasaponins (Moses et al. 2015). Members of the CYP716 subfamily of P450s catalyze various oxidations on the *β*-amyrin backbone, including the three-step oxidation of *β*-amyrin at the C-28 position to yield oleanolic acid by CYP716A12 (Carelli et al. 2011; Miettinen et al. 2017), the C-16α hydroxylation of *β*-amyrin by CYP716Y1 (Moses, Pollier, et al. 2014), and the C-3 oxidation by CYP716A14 (Moses et al. 2015). In addition, CYP72A154 oxidizes *β*-amyrin at the C-30 position (Seki et al. 2011) and several P450s have been identified that further modify oxidation products of *β*-amyrin, such as MtCYP72A61 and GmCYP72A69 (Sundaramoorthy et al. 2018; Yano et al. 2016) that hydroxylate 24-hydroxy-*β*-amyrin at the C-22 and C-21 positions, respectively, in soyasaponin biosynthesis, MtCYP72A68 that oxidizes C-23 of oleanolic acid, and MtCYP72A67 that catalyzes oxidation at C-2 position (Biazzi et al. 2015; Fukushima et al. 2013). However, several P450s and glycosyltransferases involved in saponin biosynthesis still remain uncharacterized and *in planta* evidence for tentatively identified genes is minimal.

Correlated gene expression analysis has emerged as a powerful tool for predicting gene function as correlation is suggestive of related biological processes (Hirai et al. 2005; Hirai and Saito 2004). Such correlations are facilitated by the availability of large quantities of public gene expression data which enable the calculation of gene coexpression correlation scores across thousands of samples (Usadel et al. 2009). Naoumkina et al. (2010) described a set of coexpressed *M. truncatula* genes based on comprehensive clustering of methyl jasmonate-induced transcript expression patterns along with chromosomal location analysis (Naoumkina et

al. 2010). However, the identification of the specific P450 enzymes responsible for the production of particular metabolites is still a difficult task due to the large numbers and significant diversity within the P450 multigene family (Augustin et al. 2011).

We support that an integrated approach that includes genomic, transcript, and metabolite profiling along with ectopic expression offers a more productive strategy to identify and functionally characterize new biosynthetic genes (Seki et al. 2011). Such a combined approach has been successfully used to identify several glycosyltransferase genes involved in the biosynthesis of triterpene saponins in *M. truncatula* (Achnine et al. 2005; Naoumkina et al. 2010). In the Naoumkina et al., 2010 report, the putative roles of MtCYP72A67 and MtCYP72A68 in saponin biosynthesis were proposed based upon correlated gene expression with functionally characterized genes such as those encoding β-amyrin synthase, MtCYP716A12, MtUGT73F3 and MtCYP93E2 (Carelli et al. 2011; Fukushima et al. 2013; Naoumkina et al. 2010; Seki et al. 2011). Accordingly, *MtCYP72A67* and *MtCYP72A68* expression was found to be regulated by the transcription factor TRITERPENE SAPONIN BIOSYNTHESIS ACTIVATING REGULATOR2 (TSAR2), the regulator of hemolytic triterpene saponin metabolism in *M. truncatula* (Mertens et al. 2016)). Since then, we have also presented our *in vitro* biochemical assays of recombinant MtCYP72A67 and MtCYP72A68 heterologously expressed in yeast (Sumner et al. 2012; Tzin et al. 2012a, 2012b). Similar *in vivo* enzymatic activities have been reported in engineered yeast strains for MtCYP72A61, MtCYP93E2, MtCYP72A67 MtCYP72A68 and MtCYP716A12 (Biazzi et al. 2015; Fukushima et al. 2013). However, evidence obtained through studies in heterologous microbial systems does not always equal *in planta* function. For example, Biazzi *et al*, 2015 associated MtCYP72A68 with medicagenic acid biosynthesis based upon *in vitro* yeast assays. However, we report here that MtCYP72A68 is a multi-functional oxidase responsible for hederagenin, gypsogenin and gypsogenic acid biosynthesis based upon *in vivo* and *in planta* evidence. In addition, *in planta* triterpene engineering has been hampered by a lack of knowledge about the regulatory mechanisms controlling gene expression (Sawai and Saito 2011). Hence, a challenge for future triterpenoid research will be to identify the transcription or other regulatory factors that steer their biosynthesis (Biazzi et al. 2015; Moses et al. 2013)

OBJECTIVES

In this report, we describe a highly productive approach for the discovery of triterpene biosynthetic genes and provide both *in vitro* and *in planta* characterization of two CYP72 family genes. More specifically, *MtCYP72A67* and *MtCYP72A68* were identified based upon large-scale, correlated metabolite accumulation and gene expression. Highly correlated genes were then functionally characterized as multisubstrate, mono and multifunctional oxidases in triterpene saponin biosynthesis using heterologous *in vitro and in vivo* yeast expression assays, heterologous *in vivo* tobacco expression assays, and *in planta* using *M. truncatula* hairy root cultures and Tnt1 mutants.

METHODS

Germplasm plant materials

Seeds for the *Medicago truncatula* ecotype collection were obtained from Jean-Marie Prosperi at L'Institut National De La Recherche Agronomique (INRA; http://www.international.inra.fr/the_institute). The ecotype collection used in the present study was described previously (Ronfort et al. 2006). Single seed descent lines for all of the INRA ecotypes were developed on site at the Noble Foundation.

Plant growth conditions

Plants were grown in a root cone system (Stewe and Sons, OR) with Turface MVP medium (Profile Products, Buffalo Grove, IL) in a Conviron TCR180 walk-in growth chamber maintained at 90% humidity and at an average temperature of 24ºC day (16 h) and 20ºC night (8 h). Plants were fertilized daily with 15 ppm nitrogen (20-10-20 Peat-Lite Special; The Scotts Company). Five-week-old plants were harvested, and the Turface was washed quickly from the roots. Plants were dissected into roots and aerial parts, which were flash frozen in liquid nitrogen and stored at -80 ºC.

Metabolomics analyses by UHPLC-(-)ESI-QToF-MS and GC-MS

Lyophilized tissues were ground into a fine powder using a mortar and pestle, and 10 mg of powder was extracted with 1 ml of 80% methanol in a one-dram vial for two hours on an orbital shaker. An internal standard containing 18 μg/ml umbelliferone was used in all samples.

Extracted samples were centrifuged for 30 min at 2900 x g at 4°C, and supernatants were transferred to UHPLC-MS autosampler vials. The HPLC-(-)ESI-QToF-MS analyses were performed with a Waters Acquity UHPLC system coupled to a Waters Premier hybrid quadrupole time-of-flight (QTOF) mass spectrometer (http://www.waters.com). A reverse-phase, UPLC BEH 1.7 µm C18, 2.1 mm x 150 mm column (Waters) was used for separations. The mobile phase consisted of eluent A (0.1% [v/v] acetic acid/water) and eluent B (100% acetonitrile). Separations were achieved using a linear gradient of 95 % to 30 % A over 30 min, 30 % to 5 % A over 3 min, and 5 % to 95 % A over 3 min. The flow rate was 0.56 mL/min, and the column temperature was maintained at 60 °C. Mass-to-charge ratios (*m/z*) of the eluted compounds were determined in the negative ESI mode from *m/z* 50 to 2,000. The Waters QTOF Premier was operated using the following instrument parameters: desolvation temperature of 385 °C; desolvation nitrogen gas flow of 850 L/h; capillary voltage of 2.9 kV; cone voltage of 48 eV; and collision energy of 10 eV. The MS system was calibrated using sodium formate, and raffinose was used as the lockmass compound. *β*-amyrin, erythrodiol, and cycloartenol assays were extracted twice with 500 µl of ethyl acetate, dried under nitrogen gas, dissolved in 100 µl pyridine, *N*-Methyl-*N*-(trimethylsilyl) trifluoroacetamide (MSTFA)-derivitized, and analyzed by GC-MS as described previously (Broeckling et al. 2005). GC-MS of these compounds were performed because they do not ionize well by negative ESI due to a lower number of hydroxyl substituents.

Data Processing

Raw UHPLC-(-)ESI-QToF-MS data files were annotated and quantified using MarkerLynx XS (Waters; www.waters.com) or converted to CDF file format, followed by metabolite data extraction, alignment, and export using MET-IDEA software (Broeckling et al. 2006; Lei et al. 2012). A target ion list containing 143 known and putative triterpene saponin ions of interest was used for the targeted saponin analyses. This ion list was selected and annotated based on authentic standards and previous MS and MS/MS analyses of triterpene saponins conducted internally in the lab (Huhman et al. 2005; Huhman and Sumner 2002). In addition to the targeted analyses of saponin content, non-targeted analyses of all samples were performed using Waters MarkerLynx software. The spectral abundance values for all metabolites in a separation were normalized to the internal standard of 18 µg/ml umbelliferone. Descriptive statistics were

performed in Microsoft Excel and JMP (SAS Institute Inc; http://www.jmp.com). Correlation coefficients were calculated using a custom MATLAB script (http://www.mathworks.com/).

RNA extraction, quantitative real-time PCR and Medicago Genome Array

Total RNA was isolated using modified cetyl-trimethyl-ammonium bromide (CTAB) extraction as described previously (Pang et al. 2007) or RNeasy Plant Mini Kit (Qiagen, Valencia, CA). Total RNA was purified and concentrated using the RNeasy MiniElute Cleanup Kit (Qiagen, Valencia, CA), and then treated with DNase I (Invitrogen, Carlsbad, CA). RNA concentration and quality were determined with a Nanodrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). First-strand cDNA was synthesized from 2 µg total RNA in a total volume of 20 µL using SuperScript III reverse transcriptase (Invitrogen). Primers for quantitative real-time PCR were designed using Primer3 software (http://www.frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Each primer pair was confirmed to give a single PCR product. All primers for PCR amplification are listed in Supplemental Table S7. The parameters and analysis of the qRT-PCR were as described previously (Pang et al. 2007). All reactions were performed with three technical replicates. Data were analyzed using the SDS 2.2.1 software (Applied Biosystems). Five hundred nanograms of purified RNA for each of the three biological replicates were used for probe synthesis using a GeneChip3' IVT express kit, according to manufacturer's instructions (Affymetrix). Hybridization of probes to Affymetrix GeneChip *Medicago* genome arrays and scanning of arrays was carried out as described (Benedito et al. 2008). Raw data were normalized by robust multichip averaging (RMA), as previously described (Irizarry et al. 2003).

Expression of MtCYP72A67 and MtCYP72A68 genes in WAT11 yeast system

Coding sequences information for *MtCYP72A67* and *MtCYP72A68* (Li et al. 2007) were obtained from NCBI genebank: DQ335782 and DQ335780, respectively. Primers were designed for the coding sequence by using Primer3 software (Rozen and Skaletsky 2000). The upstream cloning primer for *MtCYP72A67* and *MtCYP72A68* included both a BamHI restriction site and a kozak yeast translation initiation sequence, where the downstream cloning primer included an EcoRI cut site. *MtCYP72A67* and *MtCYP72A68* were amplified from cv. Jemalong A17 aerial tissue cDNA template using Platinum Hi-Fi Taq polymerase (Invitrogen, Carlsbad, CA). The *CYP72A68* PCR product was cloned into the pGEM-T easy vector (Promega, WI), then

sequenced using M13 forward and reverse primers. *CYP72A67* and *CYP72A68* were excised from the p-GEM easy vector via a BamHI and EcoRI restriction digest and sub-cloned into the pYeDP60 vector (Pompon et al. 1996) and sequenced using the GAL10 promoter. WAT11 yeast cells were transformed as previously reported (Greenhagen et al. 2003; Urban et al. 1997) and confirmed by PCR.

Recombinant expression and microsomal preparations for enzymatic assays

WAT11 yeast cells were transformed with pYeDP60-MtCYP72A68, pYeDP60-MtCYP72A67 or empty pYeDP60 vector, and microsomes were prepared as previously described (Greenhagen et al. 2003). For initial *in vitro* studies, 100 µg of total microsomal protein (quantified via Bradford assay) (Bradford 1976) was incubated for 2 h at 30°C in a 500 µl reaction volume of 50 mM potassium phosphate buffer (pH 7.25) containing 1 mM NADPH and 40 µM purified authentic substrate of either *β*-amyrin, cycloartenol, erythrodiol, oleanolic acid or hederagenin. An NADPH generation system (3.3 mM glucose-6-phosphate, 1.3 mM of NADPH, 3.3 mM magnesium chloride and 0.4 mM glucose-6-phosphate dehydrogenase) (Mene-Saffrane and Dellapenna, 2009) was also used. All enzymatic assays were performed in triplicate. All reaction assays were analyzed using UHPLC-(-)ESI-QToF-MS as described above, or dissolved in 100µl pyridine, MSTFA-derivitized, and analyzed by GC-MS as previously described (Broeckling et al. 2005).

Generation and cultivation of TM3-derived strains

For expression in TM3-derived yeast strains, *MtCYP72A67* and *MtCYP72A68* were amplified from *M. truncatula* cDNA with primers P1+P2 and P3+P4, respectively, cloned into pDONR221, sequence verified and gateway recombined in the yeast expression vector pAG423GAL-ccdB (Addgene plasmid 14149; (Alberti et al. 2007)). Primer sequences for P1 through P8 are listed in Supplemental Table S8. The construct encoding the self-processing polyprotein with the *M. truncatula* cytochrome P450 reductase 1 (*MtCPR1*; Medtr3g100160) and *CYP716A12* was created by amplifying *MtCPR1* without a stop codon and with a 3′-overhang of the partial T2A sequence with primers P5+P6 and *CYP716A12* with a 5′-overhang of the partial T2A sequence with primers P7+P8. Subsequently the 2 fragments were fused by PCR with primers P5+P8, cloned into pDONR221, sequence verified and recombined in the yeast expression vector pAG425GAL-ccdB (Addgene plasmid 14153; (Alberti et al. 2007)). The yeast

strains (Supplemental Table S8) were generated from strain TM3 and cultivated as described (Moses, Pollier, et al. 2014). Briefly, yeast precultures were grown with agitation in synthetic defined (SD) medium containing glucose with appropriate dropout (DO) supplements (Clontech) at 30 °C for 18 to 20 h. Gene expression was induced by inoculating washed precultures into synthetic defined Gal/Raf medium containing galactose and raffinose with appropriate dropout supplements (Clontech) to a starting OD600 of 0.25 on day 1. The induced cultures were incubated for 24 h, and on day 2, methionine and methylated β-cyclodextrins (MβCD) were added to 1 and 5 mM, respectively. After a further 24 h incubation, MβCD was added once again to 5 mM on day 3, and on day 4 all cultures were extracted for metabolite analyses. For organic extracts of the spent medium, 1 mL of the yeast culture was extracted twice with 0.5 mL of hexane and once with 0.5 mL of ethyl acetate. The organic extracts were pooled, vaporized to dryness and trimethylsilylated for GC-MS analysis. GC-MS analysis was carried out as described (Moses, Pollier, et al. 2014) .

Nicotiana benthamiana leaf infiltration

For transient expression in *N. benthamiana* leaves, the coding sequence of *GgβAS*, *MtCYP716A12*, *MtCYP72A67*, and *MtCYP72A68* were Gateway recombined from their pDONR221 entry vectors into the binary vector pK7WG2D (Karimi et al. 2002). The resulting constructs were individually introduced into the *A. tumefaciens* strain C58C1, carrying the helper plasmid pMP90. *Agrobacterium* strains were grown for 2 d in a shaking incubator (150 rpm) at 28 °C in 5 mL yeast extract broth medium, supplemented with 100 µg/mL kanamycin, 100 µg/mL spectinomycin, and 20 µg/mL gentamycin. After incubation, 0.5 mL of bacterial culture was used to inoculate 9.5 mL of yeast extract broth medium supplemented with antibiotics and containing 10 mM MES (pH 5.7) and 20 mM acetosyringone. After an additional overnight incubation (150 rpm, 28 °C), strains for transient coexpression were mixed, collected via centrifugation, and resuspended in 5 mL of infiltration buffer (100 mM acetosyringone, 10 mM $MgCl_2$, and 10 mM MES, pH 5.7). The amount of bacteria harvested for each construct was adjusted to a final OD_{600} of 0.3 after resuspension in the infiltration buffer. After 2 to 3 h incubation at 150 rpm and 28 °C, the bacteria mixtures were infiltrated to the abaxial side of fully expanded leaves of 3- to 4-week-old *N. benthamiana* plants grown at 25 °C in a 14-h/10-h light/dark regime. The infiltrated plants were incubated under normal growth conditions for 5 d prior to metabolite analysis. *Nicotiana benthamiana* infiltrated leaves were harvested and ground

to a fine powder in liquid nitrogen for metabolite analyses. Then, 0.4 g of ground leaf material was extracted with 1 mL of methanol for 10 min and centrifuged for 5 min at 20,800 x g. The resulting organic extract was evaporated to dryness under vacuum and subsequently resuspended in 0.5 mL of water and 0.5 mL of ethyl acetate. After centrifuging again for 5 min at 20,800 x g, the organic phase was removed, vaporized to dryness, and trimethylsilylated for GC-MS analysis which was carried out as described previously (Moses et al. 2015).

Ectopic expression of MtCYP72A67 and MtCYP72A68 in *Medicago truncatula* hairy roots

The coding sequence of *MtCYP72A67* and *MtCYP72A68* were amplified from cDNA synthesized from *M. truncatula* (cv. Jemalong A17) aerial tissue using Platinum Hi-Fi Taq polymerase (Invitrogen, Carlsbad, CA). The primer sequences used for amplification are listed in Sup. Table S9. PCR products were cloned into the entry vector pENTR/D/TOPO (Invitrogen) and sequenced. The entry vectors were recombined into a destination vector, pK7WG2D for overexpression or pK7GWIWG2D(II) for RNAi (a double-stranded hairpin RNA), by using the LR clonase reaction (Invitrogen). The vectors were transformed into *Agrobacterium rhizogenes* (strain ARqua1) by electroporation (Quandt et al. 1993). Transformed colonies were grown on LB-agar medium at 28°C, with spectinomycin and streptomycin for vector selection. After confirmation by PCR, transformed agrobacteria were used to transform leaves of *M. truncatula* (cv. Jemalong A17) and generate hairy roots (Verdier et al. 2012).

Tnt1 mutant identification of cyp72a68 lines

The MtCYP72A68 coding sequence was used for *in silico* blast searches against the Noble Foundation Tnt1 flanking sequence database https://medicago-mutant.noble.org/mutant/database.php, which yielded an insertion event, NF1698 insertion 4 in R108 ecotype. The Tnt1 insertion in *cyp72a68* was confirmed via cloning and sequencing of the truncated PCR product using gene-specific and Tnt1 border primers amplified from a cyp72a68/*cyp72a68* heterozygous NF1698-4 plant, position 503bp. Generation of the *M. truncatula* Tnt1 insertion mutant population and growth of R1 seeds were performed as described (Tadege et al. 2008). Reverse genetic screening for Tnt1 retrotransposon insertions in *MtCYP72A68* was performed by using a nested PCR approach (Cheng et al. 2011) and cyp72a68-forward: 5'-GCACGAGGAAAACATTTCACAC-'3. PCR products from target mutant NF12169 line were purified with QIAquick PCR purification kit (Qiagen) and sequenced

by using Tnt1 primers to confirm insertions in *MtCYP72A68* at position. Insertions in *Mtcyp72a68* were found at position 503 bp in mutant line NF1698-4 and at 453 bp in line NF12169 which were validated by PCR (both calculated from start codon and located on exon). All mutants were genotyped however no homozygous Tnt1-*Mtcyp72a67* insertion plants were identified.

RESULTS

Metabolite profiling of a *M. truncatula* ecotype collection reveals substantial chemical diversity and identifies high and low saponin-accumulating ecotypes

Species-specific germplasm collections (ecotypes; natural genetic variants) are a powerful resource for exploring the natural chemical variation in triterpene saponin content (Branca et al. 2011; Ronfort et al. 2006). In this study, 106 ecotypes were analyzed using high-resolution biochemical profiling to characterize the chemical diversity in triterpene saponin content within a large *M. truncatula* germplasm collection (collection provided by Dr. Jean-Marie Prosperi and the French National Institute for Agriculture Research; INRA). The triterpene saponin content in each of the above lines was analyzed separately in root and aerial tissues using UHPLC- coupled to a hybrid quadrupole time-of-flight mass spectrometer operated in the negative electrospray ionization mode (UHPLC-(-)ESI-QToF-MS). Metadata compliant with the Metabolomics Standards Initiative (Fiehn et al. 2007; L. W. Sumner et al. 2007) are summarized in Supplemental Table S1. Overall, 143 putative and identified triterpene saponins were measured based upon unique ion mass-to-charge ratio (*m/z*) and chromatographic retention time pairs. Seventeen saponins were rigorously identified based upon co-characterization with authentic standards (e.g., 3-Glc-28-Glc-medicagenic acid standard); 53 saponins were tentatively identified based upon mass spectral (accurate mass, in-source fragmentation and/or MS/MS) and literature data (Huhman et al. 2005; Huhman and Sumner 2002; Pollier et al. 2011); 28 saponins had partial annotation based solely on spectral features resulting from probable source fragmentation (e.g. possibly Glc-Glc-bayogenin); and the remainder were unknowns. The latter unknown saponins were differentiated based upon unique *m/z* values and retention times in the same manner as the known and putatively identified saponins. Total saponin accumulation values were determined for both aerial and root tissues of each ecotype by summing the peak area for each of the saponin ion/RT pairs (Supplemental Table S2 and S3). A scatter plot of the total saponin

content in the aerial tissue versus the root is present in Figure 2. The average value for the total saponin chromatogram peak areas in 106 ecotypes was 78.32 normalized relative instrument response to the internal standard (nrir) in the aerial tissue and 285.35 nrir in the roots. Ecotype ESP105 had the lowest relative content of total saponins in aerial tissue (3.4 nrir), but very high total saponin accumulation in root tissue (394.1 nrir). In contrast, ecotype GRC43 had the lowest total saponin accumulation in the root (44.7 nrir) but very high total accumulation in aerial tissues (131.3 nrir). Thus, these two most diverse ecotypes were selected for further comparative gene expression analyses.

Correlated metabolite and gene expression analyses identify *MtCYP72A67* and *MtCYP72A68* as cytochrome P450s potentially involved in saponin biosynthesis

Correlations between gene-to-gene expression and gene-to-metabolite accumulation have been shown to be a powerful tool for the identification of novel natural product biosynthetic genes (Goossens 2015; Hirai et al. 2005, 2010; Hirai and Saito 2004). Here, correlation analyses between saponin accumulation and gene expression were performed using selected ecotypes with high saponin diversity: *M. truncatula* ESP105 and GRC43 (Figure 2, and Supplemental Tables S2 and S3). These ecotypes were chosen based upon their substantial differential and tissue-specific accumulation of saponins as described above. We also selected two reference ecotypes: A17 that was used for genomic sequencing (Young et al. 2011) and R108 that been used in the generation of a Tnt1 retrotransposon insertion mutant population (Tadege et al. 2008). Affymetrix GeneChip-Medicago Genome Arrays were used for gene expression analyses. Qualitative and relative quantitative analyses of saponin levels were performed using UHPLC-(-)ESI-QToF-MS and based upon a unique ion/RT pair list as described above. Saponins were then grouped according to their triterpene aglycone structures. This grouping of sapogenin-specific accumulation values was performed for eight different sapogenin aglycones, including oleanolic acid, hederagenin, bayogenin, medicagenic acid, polygalagenin (putative identification), zanhic acid, soyasapogenol E and soyasapogenol B (see structures in Figure 1). In addition, two parameters for total saponin calculations were used: i) total known = sum of the known sapogenins and ii) total aglycones = sum of total known saponins and unknown aglycones. Gene-to-gene and gene-to-metabolite Pearson's correlation coefficients (r) were calculated and clustered using gene expression levels for 23 P450 probe sets implicated in a previous report

(Naoumkina et al. 2010), *β-amyrin synthase* (*βAS*), nine other known genes related to *M. truncatula* terpene biosynthesis, known sapogenin aglycones and total saponin levels (Figure 3).The probe sets of four genes, *MtCYP72A67, MtCYP72A68*, *MtCYP716A12* and *MtβAS*, were positively correlated to each other (Pearson's $r \geq 0.5$, upper left triangle) and significant (*P* value <0.01, lower right triangle) as highlighted in the Figure 3 heat map. These probe sets were also highly correlated with the total saponin content and medicagenic acid. The expression levels of *MtCYP72A67, MtCYP72A68*, *MtCYP716A12* and *MtβAS* with the sapogenin aglycones level are presented in Supplemental Figure S1 and the qRT-PCR verification of the gene expression levels in Supplemental Figure S2 along with the full list of transcriptome and metabolome data for the selected *M. truncatula* ecotypes in Supplemental Table S4.

The correlation data reported here were validated with previous studies which reported that both *βAS* (the first enzymatic step of triterpene saponin biosynthesis) and MtCYP716A12 (enzyme associated with oleanate sapogenin biosynthesis) are key enzymatic steps in triterpene saponin biosynthesis in *M. truncatula* (Carelli et al. 2011; Fukushima et al. 2011; Suzuki et al. 2002). The data also highlighted two more important genes, *MtCYP72A67*, a P450 suspected to be involved in *Medicago* saponin biosynthesis (Biazzi et al. 2015; Fukushima et al. 2013; Naoumkina et al. 2010) and *MtCYP72A68*, which was reported to catalyze the three-step oxidation of oleanolic acid in a heterologous yeast system (Fukushima et al. 2013).

Additional correlation analyses and gene expression clustering of the implicated P450 genes were performed using data from the *M. truncatula* Gene Expression Atlas (Benedito et al. 2008). The expression data used for the correlation data are provided in Supplemental Table S5. These analyses revealed similar clustering of the uncharacterized *MtCYP72A67* and partially characterized *MtCYP72A68* genes with the previously characterized *MtβAS*, *MtCYP716A12* and *MtCYP93E2* P450s (Supplemental Figure S3). *MtCYP72A67* and *MtCYP72A68* are also highly correlated with the accumulation of the oleanate sapogenin medicagenic acid. Therefore, the cumulative correlation data strongly support that both *MtCYP72A67* and *MtCYP72A68* have a high potential as putative genes/enzymes involved in triterpene sapogenin biosynthesis.

Heterologous expression of *MtCYP72A67* in yeast

The potential oleanate sapogenin oxidase activity of MtCYP72A67 was tested using heterologous expression in yeast and *in vitro* biochemical assays using yeast microsomes. A

recent article reports on the identification of CYP72A67 (Fukushima et al. 2013); however, upon co-expression of the *βAS*, cytochrome P450 reductase (CPR), CYP716A12, and CYP72A67 genes in yeast, they could not demonstrate CYP72A67 activity (Fukushima et al. 2013). In this study, we used microsomes from yeast WAT11 cells expressing *MtCYP72A67 in vitro* with a variety of triterpene sapogenin substrates, and the products were analyzed by UHPLC-(-)ESI-QToF-MS. NADPH was also added to these assays as a P450 cofactor. When oleanolic acid was used as a substrate, 2β-hydroxyoleanolic acid was detected as a product in the MtCYP72A67 (+) NADPH microsomal samples, but not detected in the MtCYP72A67 (-) NADPH or empty vector control samples (Table 1 and Supplemental Figure S4A). When hederagenin was used as a substrate, 2β-hydroxyhederagenin (*e.g*. bayogenin) was detected as a product in the (+) NADPH samples but not detected in the MtCYP72A67 (-) NADPH and empty vector control samples (Table 1B and Supplemental Figure S4B). Both the empty vector control and assays without NADPH resulted in no P450 activity (Supplemental Figure S4B). In addition, *β*-amyrin and erythrodiol were used as substrates in MtCYP72A67 assays, but no products were detected using gas chromatography-mass spectrometry (GC-MS; data not shown; a summary of all substrates that were used for MtCYP72A67 yeast assay is listed in Supplemental Table S6). It was concluded that microsomes containing recombinant CYP72A67 protein possess multi-substrate C-2β-hydroxylase activity for oleanolic acid and hederagenin, yielding the C-2 β-hydroxy derivatives.

Encouraged by the *in vitro* assays that clearly indicate the oleanolic acid oxidase activity of CYP72A67, additional yeast strains KM1 and KM2 were created from a sterol engineered *β*-amyrin producing yeast strain TM3 (Moses, Pollier, et al. 2014). All yeast strains generated in this study are listed in Supplemental Table S8. Both strains express *CYP716A12* and *M. truncatula* cytochrome P450 reductase (*MtCPR1*) from a high-copy number plasmid to produce a self-processing polyprotein in which the P450 reductase and the P450 are linked via a 2A oligopeptide (de Felipe et al., 2006). When cultivated in the presence of methylated β-cyclodextrins (MβCD), both strains produce high levels of oleanolic acid (Figure 4A). Yeast strain KM2 also expresses MtCYP72A67 from a high-copy number plasmid, whereas KM1 does not. Comparison of the GC chromatograms of extracts from the spent medium of KM1 and KM2 cultured with MβCD showed a single unique peak in strain KM2, but not in strain KM1, that

corresponds to 2β-hydroxyoleanolic acid (Figure 4A). In summary, MtCYP72A67 was shown to catalyze the C-2β-oxidation of oleanolic acid, using *in vitro* assays with yeast microsomes and engineered yeast strains.

Ectopic expression and characterization of *MtCYP72A67 in planta*

To investigate if MtCYP72A67 functions as a 2β-hydroxylase *in planta*, *M. truncatula* hairy roots were generated following *Agrobacterium rhizogenes*-mediated transformation with the *Mtcyp72a67* RNAi, a double-stranded hairpin RNA that triggers post-transcriptional gene silencing, or with *MtCYP72A67*- overexpressing the *MtCYP72A67* full coding sequence. Quantitative RT-PCR analysis of the *MtCYP72A67* transcript levels resulted in an average gene expression reduction of 46% in *Mtcyp72a67*-RNAi hairy roots compared to control. UHPLC-(-)ESI-QToF-MS was used to compare the saponin content in transformed hairy roots relative to an empty vector control. Figure 5 summarizes the changes in saponin content which were measured in *Mtcyp72a67* RNAi hairy roots. Deduced enzymatic products of CYP72A67 and related downstream saponins, including 2β-hydroxyoleanolic acid, bayogenin, polygalagenin (putative identification), medicagenic acid and zanhic acid were significantly reduced while substrates oleanolic acid, hederagenin and gypsogenin (putative identification) were significantly increased in the *Mtcyp72a67* RNAi hairy roots compared to the control. In Figure 5, the oleanate sapogenins were altered in the *Mtcyp72a67* RNAi hairy roots where the downstream C-2 oxidative derivatives of oleanolic acid were decreased while the non-C-2 oxidative pathway metabolites were increased.

MtCYP72A67 was overexpressed in hairy roots and *MtCYP72A67* transcript levels showed an average 2.24-fold induction by qRT-PCR compared to the control. Saponin analyses were performed by UHPLC-(-)ESI-QToF-MS for the *MtCYP72A67*-overexpressing hairy roots and compared to hairy roots transformed with an empty vector. *MtCYP72A67*-overexpressing hairy roots had significantly increased levels of several aglycones, including 2β-hydroxyoleanolic acid, polygalagenin (putative identification) and zanhic acid (Supplemental Figure S5). Unexpectedly, the level of oleanolic acid was also significantly increased which may indicate a more complex regulatory function of this gene. Ectopic expression of *MtCYP72A67 in planta* is in agreement with the results from yeast and thus support the conclusion that MtCYP72A67 possesses multi-substrate 2β-hydroxylase activity for oleanolic acid and

hederagenin, yielding the C-2 alcohols in the oleanate sapogenin branch of the triterpene saponins.

In addition, *G. glabra βAS* and *MtCYP716A12* were transiently expressed with or without *MtCYP72A67* in *Nicotiana benthamiana* leaves using *Agrobacterium tumefaciens*-mediated infiltration. Similar to yeast, comparison of the GC-MS chromatograms of organic extracts from leaves three days after co-infiltration revealed the presence of 2β-hydroxyoleanolic acid in leaves that were co-infiltrated with MtCYP72A67 (Figure 4B). Taken together, MtCYP72A67 was shown to catalyze the 2β-oxidation of oleanolic acid in heterologous tobacco bioassays.

Heterologous expression of *MtCYP72A68* in yeast

The potential oleanate sapogenin oxidase activity of MtCYP72A68 was first tested using heterologous expression and *in vitro* enzymatic assays. Microsomes of WAT11 yeast cells expressing *MtCYP72A68* were tested for *in vitro* enzymatic activity with various substrates, and products analyzed with UHPLC-(-)ESI-QToF-MS or GC-MS as described above. Protein activity was measured with oleanolic acid or hederagenin as the substrate and with or without NADPH as a P450 cofactor. As shown in Table 2A, when oleanolic acid was used as a substrate, hederagenin, gypsogenin, and gypsogenic acid accumulated. None of these products were detected in the empty vector control samples (see also Supplemental Figure S6A). The detected anion at *m/z* 469.35 at retention time (RT) of 24.91 min was tentatively identified in this study as gypsogenin or 3β-hydroxy-23-oxo-olean-12-en-28-oic acid based upon literature information including accurate mass, aglycone anion at *m/z* 469 and a predicted molecular formula of $C_{30}H_{46}O_4$ (Supplemental Figure S6; http://www.chemspider.com).

Another detected anion at *m/z* 485.35 and RT of 21.88 min was previously identified as gypsogenic acid or 2β,3β-dihydroxy-23-oxo-olean-12-en-28-oic acid based upon fragmentation of its aglycone anion at *m/z* 485 and a predicted molecular formula of $C_{30}H_{46}O_5$ (Supplemental Figure S6; http://www.chemspider.com and (Pollier et al. 2011). In addition, the amount of oleanolic acid detected was lower in the MtCYP72A68 (+) NADPH assay, indicating its consumption as a substrate. When hederagenin was used as a substrate (Table 2B), large quantities of gypsogenin and gypsogenic acid were detected as products in the MtCYP72A68 assay and not detected in the empty vector control samples (see also Supplemental Figure S6B). The amount of hederagenin detected was lower in the MtCYP72A68 (+) NADPH assay,

indicating its consumption as a substrate. β-amyrin and erythrodiol were also tested as substrates, but no products were detected via GC-MS (data not shown; a summary of all substrates that were used for MtCYP72A68 yeast assays is listed in Supplemental Table S6). Taken together, these results indicate that microsomes containing recombinant MtCYP72A68 possess the ability to catalyze the sequential oxidation of C-23 on oleanolic acid, yielding the C-23 alcohol, aldehyde and carboxylic acid derivatives. The data also indicated that recombinant MtCYP72A68 will not accept sapogenins with C-28 methyl or C-28 hydroxyl groups as substrates.

Next, yeast strains KM1 and KM3 were created from a sterol engineered β-amyrin producing yeast strain TM3 (Moses, Pollier, et al. 2014). Yeast strain KM3 expresses a self-processing polyprotein in which CYP716A12 and MtCPR1 were linked via a 2A oligopeptide. KM3 also expresses *MtCYP72A68* from a high-copy number plasmid, whereas KM1 does not. Comparison of the GC-MS chromatograms of extracts from the spent medium of KM1 and KM3 cultured with MβCD showed three unique peak in strain KM3 that correspond to hederagenin, gypsogenin and gypsogenic acid (Figure 6A) (Fukushima et al. 2013).

Ectopic expression and characterization of *MtCYP72A68 planta*

M. truncatula hairy roots were transformed and *MtCYP72A68* overexpressed to further substantiate the role of *MtCYP72A68* in triterpene saponin biosynthesis. The *MtCYP72A68* transcript levels in hairy roots were quantified by qRT-PCR, and an average 1.51-fold induction in *MtCYP72A68* expression was observed compared to the control. Saponin analyses were performed by UHPLC-(-)ESI-QToF-MS on the *MtCYP72A68-OE* (overexpression) hairy roots and compared to hairy roots transformed with an empty vector. The fold changes in saponin content that were detected in *MtCYP72A68-OE* hairy roots are presented in Figure 7, and the data revealed significantly increased levels of several aglycones, including oleanolic acid, hederagenin, polygalagenin (putative identification), soyasapogenin E and total saponins. None of the metabolites were decreased. Unexpectedly, the level of 2β-hydroxy oleanolic acid was also induced. As shown in Figure 7, both oleanate sapogenins and soyasapogenols branches were altered in the *MtCYP72A68-OE* hairy roots, which affected the total accumulation of saponins by 1.53-fold. *M. truncatula* hairy roots were also transformed with *Mtcyp72a68* RNAi, a double-stranded RNA (hairpin RNA). However, transcript levels of *Mtcyp72a68*-RNAi were checked by

qRT-PCR and showed only a minor average decrease in gene expression of 16% and no significant metabolic changes. *Glycyrrhiza glabra βAS*, *MtCPR1* and *MtCYP716A12* were also transiently expressed with or without *MtCYP72A68* in *Nicotiana benthamiana* leaves using *Agrobacterium tumefaciens*-mediated infiltration. Like in yeast, comparison of the chromatograms of organic extracts from leaves three days after co-infiltration revealed three unique peaks corresponding to hederagenin, gypsogenin and gypsogenic acid in leaves that were co-infiltrated with *MtCYP72A68* (Figure 6B). Hence, heterologous expression of *MtCYP72A68* in yeast and tobacco all point towards the sequential oxidation of C-23 of oleanolic acid by *MtCYP72A68* and confirm similar results obtained by (Fukushima et al. 2013).

We also measured the saponin levels extracted from root tip tissues from Tnt1 retrotransposon mutants (Tadege et al. 2008) in the *MtCYP72A68* gene. Two independent mutant lines with retrotransposon insertions in the *MtCYP72A68* gene were identified through PCR screening (Tadege et al. 2008). Tnt1 insertion lines NF1698-4, NF12169, and R108 (control) were germinated and the root tips (approximately 3 mm) from 25 plants (a mixed population of heterozygous and wild type, 2:1) were collected. The mRNA levels revealed that both Tnt1 mutant lines possessed lower *MtCYP72A68* mRNA levels, 22 % lower for NF1698-4 and 20% lower for NF12169, compared to wildtype R108 control (Supplemental Figure S7A). This was followed by UHPLC-(-)ESI-QToF-MS analyses of sapogenin aglycones (Supplemental Figure S7B). These analyses showed reduction of medicagenic acid and zanhic acid of NF12169 line.

DISCUSSION

Genomic and coexpression analyses identify genes involved in triterpene saponin biosynthesis

Currently, several P450s from multiple P450 families have been reported in relation to saponin biosynthesis (Augustin et al. 2011; Miettinen et al. 2018; Seki et al. 2015). Diversity of P450s are involved in triterpene saponin biosynthesis across many species, hence the prediction of specific substrate and enzymatic activity based on sequence alone is complex (Nelson and Werck-Reichhart 2011). Thus, there is a need for large-scale data approaches to identify and prioritize candidate P450s and other gene candidates involved in triterpene biosynthesis. Germplasm collections are powerful resources for exploring the natural variation for any number of phenotypes (Ronfort et al. 2006), including saponin content. We measured the total saponin

content for 106 of *M. truncatula* ecotypes and revealed substantial differential accumulation of these specialized metabolites. Four diverse ecotypes with strong differential accumulation of saponins in root and aerial tissues were selected for further transcriptome analyses. Correlation analyses between gene-gene expression and gene-metabolite accumulation were then performed to identify genes that are likely involved in triterpene saponin biosynthesis. The correlation coefficients between microarray gene expression levels for putative P450s, *βAS*, known sapogenins and total saponins revealed a significantly correlated cluster of genes, including *MtCYP72A67, MtCYP72A68, MtCYP716A12* and *MtβAS*. Saponin levels, including total known saponins, total aglycones and medicagenic acid, were also clustered (Figure 3 and Supplemental Figure S3). These correlations represent discovery events that identify putative saponin biosynthetic genes. To further assess the CYP72 candidate genes, correlation coefficients for gene-gene expression values of the P450s were calculated using the *M. truncatula* Gene Expression Atlas (Supplemental Figure S3) (Benedito et al. 2008; He et al. 2009). This revealed similar clustering of *MtCYP72A67* and *MtCYP72A68* genes with *MtβAS* and *MtCYP716A12* (Figure 3). Both *MtβAS* and *MtCYP716A12* genes have been previously reported as saponin biosynthetic enzymes (Supplemental Figure S3) (Carelli et al. 2011; Fukushima et al. 2011; Miettinen et al. 2017). *MtCYP72A61*, *MtCYP72A67* and *MtCYP72A68* were also implicated based upon their coexpression with *β*-amyrin synthase in methyl jasmonate-elicited cell culture data (Naoumkina et al. 2010) and their expression is under control of the regulator of hemolytic saponin biosynthesis, the transcription factor TSAR2 (Mertens et al. 2016). Thus, *MtCYP72A67* and *MtCYP72A68* were prioritized for functional characterization.

MtCYP72A67 is a C-2 hydroxylase involved in the biosynthesis of oleanate sapogenins

Heterologous combinatorial biosynthesis is a method that establishes novel enzyme-substrate combinations *in vivo* (Pollier et al. 2011). However, heterologous combinatorial expression of *β*AS/CPR/CYP716A12/CYP72A67 genes in yeast strains did not demonstrate CYP72A67 activity (Fukushima et al. 2011), whereas in this study, we were able to demonstrate the 2β-hydroxylase activity of MtCYP72A67 on oleanolic acid. This apparent discrepancy could be due to the difference in the yeast strains used or due to different culturing conditions. In the study of Fukushima et al. (2013), the non-engineered yeast strain INVSc1 was used, whereas for this study we used the sterol-engineered yeast strain TM3. In strain TM3, the native yeast gene

ERG7 is under control of a methionine-repressible promoter allowing for increased accumulation of 2,3-oxidosqualene, the substrate for *β*-amyrin synthase. In addition, strain TM3 also expresses a truncated feedback-insensitive copy of isoform 1 of the *S. cerevisiae* 3-hydroxy-3-methylglutaryl-CoA reductase (tHMG1) gene, allowing for an increased accumulation of 2,3-oxidosqualene (Kirby et al. 2008; Moses, Pollier, et al. 2014). Furthermore, the use of MβCD in the cultivation process improves the catalytic efficiency of the P450s, likely due to removal of feedback inhibition on the P450 activity or removal of toxicity due to lower intracellular accumulation of sapogenins (Moses, Pollier, et al. 2014).

Heterologous MtCYP72A67 catalyzed the oxidation of C-2 of both oleanolic acid and hederagenin, yielding the products 2β-hydroxy oleanolic acid and bayogenin, respectively (Table 1). The hydroxylation of hederagenin indicates that compounds with C-23 hydroxyl substitution are also substrates for MtCYP72A67-mediated C-2 oxidation. MtCYP72A67 P450-mediated C-2 oxidation activity is also supported by the lack of product accumulation in assays deficient in NADPH (Supplemental Figure S4). Lack of product accumulation in these assays indicates that the cytochrome P450 reductase/MtCYP72A67 requires NADPH as an electron donor for activity (Liu et al. 2003; Seki et al. 2008). No products were detected when MtCYP72A67 was assayed with *β*-amyrin or erythrodiol, which implies that compounds with a C-28 methyl (*β*-amyrin) or C-28 hydroxyl group (erythrodiol) are not suitable substrates for MtCYP72A67-mediated C-2 oxidation. Functional genomics studies of *Mtcyp72a67* RNAi and *MtCYP72A67* overexpression in *M. truncatula* hairy roots further demonstrated that MtCYP72A67 possessed *in planta* oxidation activity for C-2 of oleanolic acid and hederagenin, yielding the C-2 alcohols in catalyzed substrates of the oleanate sapogenin branch of the triterpene saponins (Figure 5 and Supplemental Figure S5).

MtCYP72A68 is a multifunctional oxidase involved in the oleanate sapogeninin biosynthesis

Previous studies have shown that an individual cytochrome P450 can catalyze the oxidation of a given carbon, yielding the hydroxyl, carbonyl and carboxylic acid products related to diterpenes in loblolly pine and Arabidopsis (He et al. 2009; Ro et al. 2005). More recently, a multifunctional oxidase involved in triterpene saponin biosynthesis (MtCYP716A12) has also been identified in *M. truncatula* (Carelli et al. 2011). Both heterologous expression assays in

yeast and tobacco demonstrate that MtCYP72A68 catalyzes the initial oxidation of C-23 of oleanolic acid, yielding the C-23 hydroxyl product hederagenin (Table 2, Supplemental Figure S6). Additional products with *m/z* values of 469.35 and 485.35 were also detected in the MtCYP72A68 oleanolic acid assays. These have been tentatively identified as gypsogenin, the C-23 aldehyde derivative of oleanolic acid, and gypsogenic acid, the C-23 carboxy derivative of oleanolic acid, based upon accurate mass and literature information. Assays of yeast expressing *MtCYP72A68* tested with hederagenin also showed production of gypsogenin and gypsogenic acid products. No products were detected when MtCYP72A68 was assayed with β-amyrin or erythrodiol, which implies that compounds with a methyl C-28 (β-amyrin) or C-28 hydroxyl group (erythrodiol) are not suitable substrates for MtCYP72A68-mediated C-23 oxidation.

The *in planta* study of *MtCYP72A68* overexpressed in *M. truncatula* hairy roots supported the oxidase activity of CYP72A68 for oleanolic acid, yielding the C-23 alcohol in catalyzed substrates of the oleanate sapogenin branch of the triterpene saponins (Figures 8). Tnt1 mutant lines with an insertion in the *MtCYP72A68* gene accumulated high levels of 2β-hydroxy oleanolic acid due to the reduced level of *MtCYP72A68* (Supplemental Figure S7). However, the reason for high induction of 2β-hydroxy oleanolic acid in *MtCYP72A68* overexpressing lines is not clear and may be due to more complex regulation of the pathway or regulatory function of the *CYP72A68* gene. Taken together, these results indicate that MtCYP72A68 catalyzes the oxidation of C-23 of oleanate sapogenins, yielding the alcohol (hederagenin), and likely also catalyzes the further oxidation towards the aldehyde (gypsogenin) and carboxy acid (gypsogenic acid).

Multisubstrate enzymes enhance chemical diversity

MtCYP72A67 and *MtCYP72A68* are novel enzymes that catalyze the sapogenin biosynthesis in *M. truncatula*. MtCYP72A67 is a multisubstrate C-2 oxidase yielding 2β-hydroxy oleanolic acid and bayogenin, and MtCYP72A68 is a multisubstrate C-23 multifunctional oxidase yielding hederagenin, gypsogenin and gypsogenic acid. Although gypsogenin was not detected *in planta*, it was detected in yeast heterologously producing MtCYP72A68 and supplied *in vivo* or *in vitro* with oleanolic acid and hederagenin acid (Table 2, Figures 7-8 and Supplemental Figure S6-S7). We suggest that *MtCYP72A68* has the potential to convert the C-23 alcohol of hederagenin to a carbonyl group leading to gypsogenin. Together these genes have the potential to synthesize at

least four different sapogenins from a common precursor, oleanolic acid. Biosynthesis of bayogenin, for instance, requires both genes for oxidation of C-2 and C-23 (Figure 1). However, the order of these reactions has not yet been determined. Evidence for cytochrome P450 enzymes as multisubstrate enzymes in biochemistry has been accumulating for some time (Carelli et al. 2011)Siminszky et al., 1999; Schmidt et al., 2003; Ro et al., 2005). Enzymes with broad substrate tolerance are also commonly found in natural product biosynthesis. Enzymes that can act on more than one substrate to give multiple products is a mechanism that generates chemical diversity, and, as long as one of the products enhances the fitness of the producer, the genes coding for the overall process will be favored by selection, and chemical diversity will be retained (Firn and Jones, 2003; Weissman and Leadlay, 2005; Gershenzon and Dudareva, 2007).

CONCLUSIONS

We exploited the genetic and biochemical diversity of a *M. truncatula* population and used integrated metabolomics and transcriptomics to identify novel genes involved in saponin biosynthesis. This was achieved through UHPLC-(-)ESI-QToF-MS metabolite profiling of a diverse collection of *M. truncatula* ecotypes, which further resulted in the identification of four specific ecotypes with substantial differential saponin accumulation. Correlated gene-to-gene expression and gene-to-metabolite accumulation data identified *MtCYP72A67* and *MtCYP72A68* as potential oxidative enzymes associated with saponin biosynthesis. These genes were then functionally characterized using traditional *in vitro, in vivo* and *in planta* biochemical assays along with genetic approaches to prove function. The data provide evidence that MtCYP72A67 is a C-2 β-hydroxylase of oleanolic acid and bayogenin, and that MtCYP72A68 is a multisubstrate, C-23 multifunctional oxidase of hederagenin, gypsogenin, and gypsogenic acid. The successful integration of metabolomics and transcriptomics illustrated here provides additional evidence of the value of these exciting new technologies in the discovery and characterization of novel specialized metabolism genes.

ACKNOWLEDGMENTS

We thank Dr. Jean-Marie Prosperi and the L'Institut National De La Recherche Agronomique (INRA; http://institut.inra.fr/en) for providing the *M. truncatula* germplasm collection. We also thank Joseph Chappell for providing the WAT11 yeast strain. We thank Dr. Qiao Zhao for

constructive comments on this manuscript. The Sumner lab has been graciously supported by several entities over the years for the development of natural products profiling and plant metabolomics. This project was financially supported by the Oklahoma Center for the Advancement of Science and Technology (OCAST Award#PSB10-027), the National Science Foundation Molecular and Cellular Biosciences Award#1024976, The Samuel Roberts Noble Foundation, Oklahoma EPSCoR Research Experience for Undergraduates subaward #EPSCoR-2011-15, the VIB International PhD Fellowship Program (predoctoral fellowship to P.A.), the Research Foundation Flanders (postdoctoral fellowship to J.P.), and the European Union Seventh Framework Program FP7/2007–2013 under grant agreement number 613692–TriForC.

SUPPLEMENTAL DATA

Supplemental Table S1: Metabolomics Standards Initiative Compliant Metadata supporting the experiments reported here (Fiehn et al. 2007).

Supplemental Table S2: Full ecotype UHPLC-ESI(-)-QTofMS data of *Medicago truncatula* aerial tissue

Supplemental Table S3: Full ecotype UHPLC-ESI(-)-QTof-MS data for *Medicago truncatula* root tissue

Supplemental Table S4: Relative saponin accumulation levels and gene expression data for selected *M. truncatula* lines (Log10).

Supplemental Table S5: Expression values from Medicago gene expression atlas of terpene P450s genes, unknown P450s and other genes from the terpenes biosynthesis which were used for CE heat map.

Supplemental Table S6: Substrates that were tested in CYP72A67 and CYP72A68 activity assay

Supplemental Table S7: Primer list used for qRT-PCR.

Supplemental Table S8: Yeast strains generated in this study and primers used for cloning.

Supplemental Table S9: Primer list of *MtCYP72A67* and *MtCYP72A68* cloning to hairy roots transformation.

Supplemental Figure S1. Relative saponin accumulation relative to gene expression levels of *MtCYP72A67, MtCYP72A68, MtCYP716A12* and *βAS* genes. Root and aerial tissues were collected from four selected ecotypes: (A) A17; (B) ESP105; (C) GRC43; and (D) R108. Microarray analyses and UHPLC-(-)ESI-QToF-MS were performed using three biological replicates. Expression values are reported as Log10 of shoots (black bars) and root (white bars).

Supplemental Figure S2. Gene expression of *MtCYP72A67, MtCYP72A68, MtCYP716A12* and *βAS*. Root and aerial tissues were collected from 5-week-old *M. truncatula* ecotypes, including ESP105, GRC43, A17, and R108. Relative gene expression was measured and reported for two methods: A) qRT-PCR; and B) microarray analyses. In both methods, three biological replicates were used.

Supplemental Figure S3. Correlation coefficient (Pearson's r) heat map of known, unknown P450s and genes from the terpene biosynthesis using expression values from Gene Expression

Atlas. Transcript levels were measured in the different tissues (microarray data were obtained from the *M. truncatula* Gene Expression Atlas database version 2, MtGEAv2, https://mtgea.noble.org/v2/). The P450s involved in triterpene biosynthesis were selected according to a previous paper (Naoumkina et al. 2010). Upper left triangle matrix presents probe set correlation coefficients using Pearson's correlation (scale -0.5 to 1), and lower triangle matrix presents *P* value of the correlation coefficient test (light blue present *P* value < 0.01 and brown present non-significant). The black-bordered squares illustrated a highly correlated set of genes, including *MtCYP72A67, MtCYP72A68, MtCYP716A12, βAS, MtCYP93E2, isopentenyl pyrophosphate isomerase, phosphomevalonate kinase and mevalonate kinase.*

Supplemental Figure S4. UHPLC-(-)ESI-QToF-MS chromatograms of the MtCYP72A67 product generated yeast *in vitro* enzymatic assays. (A) Chromatogram of activity assay using oleanolic acid and (B) hederagenin. The full-length MtCYP72A67 was tested with cofactor (MtCYP72A67 (+) NADPH), without cofactor (MtCYP72A67 (-) NADPH) and empty vector with cofactor (empty vector (+) NADPH). The results illustrate that MtCYP72A67 hydroxylates the C-2 position of oleanolic acid and hederagenin to produce 2-hydroxyoleanolic acid and bayogenin.

Supplemental Figure S5. Results of overexpression (OE) of *MtCYP72A67* full gene in *M. truncatula* hairy roots. (A) Proposed biosynthetic pathway of the sapogenins with observed metabolite fold changes in *M. truncatula* hairy roots. The sapogenin names are marked in different colors according to the fold changes: red – saponin significant fold increase; black – no change; and gray – not detected. The observed saponin fold changes are noted in brackets. *Gypsogenin and polygypsogenin were putatively identified using tandem mass (no authentic standard).

Supplemental Figure S6. *In vitro* enzymatic assays of recombinant MtCYP72A68 in yeast WAT11 cells. (A) UHPLC-(-)ESI-QToF-MS chromatograms of activity assays using oleanolic acid and (B) hederagenin as substrates. The full-length CYP72A68 was tested with cofactor (CYP72A68 (+) NADPH) and empty vector with cofactor (empty vector (+) NADPH). Gypsogenin (3-hydroxy-23-oxoolean-12-en-28-oic acid) was tentatively identified here based upon literature information including accurate mass, aglycone anion at *m/z* 469 and a predicted molecular formula of $C_{30}H_{46}O_4$. Gypsogenic acid, ((3β)-3-hydroxyolean-12-ene-23,28-dioic

acid) was previously identified from fragmentation of saponins yielding an aglycone anion at *m/z* 485 and a predicted molecular formula of $C_{30}H_{46}O_5$ (http://www.chemspider.com and (Pollier et al. 2011)).

Supplemental Figure S7. Metabolite accumulation and qRT-PCR levels in *cyp72a68* Tnt1 mutant lines. Metabolites and mRNA were extracted from 3mm root tips of 3-day-old seedlings with three biological repeats from two Tnt1 mutant lines, NF1698-4 and NF12169. (A) The relative level of *CYP72A68* gene expression detected by qRT-PCR in NF1698-4 and NF12169. Expression values were normalized relative to the endogenous ubiquitin control gene. (B) Fold changes in sapogenin content observed in the two mutant lines NF1698-4 and NF12169. In bold, Student's *t*-test with P value < 0.05.

FIGURES and LEGENDS

Fig. 1

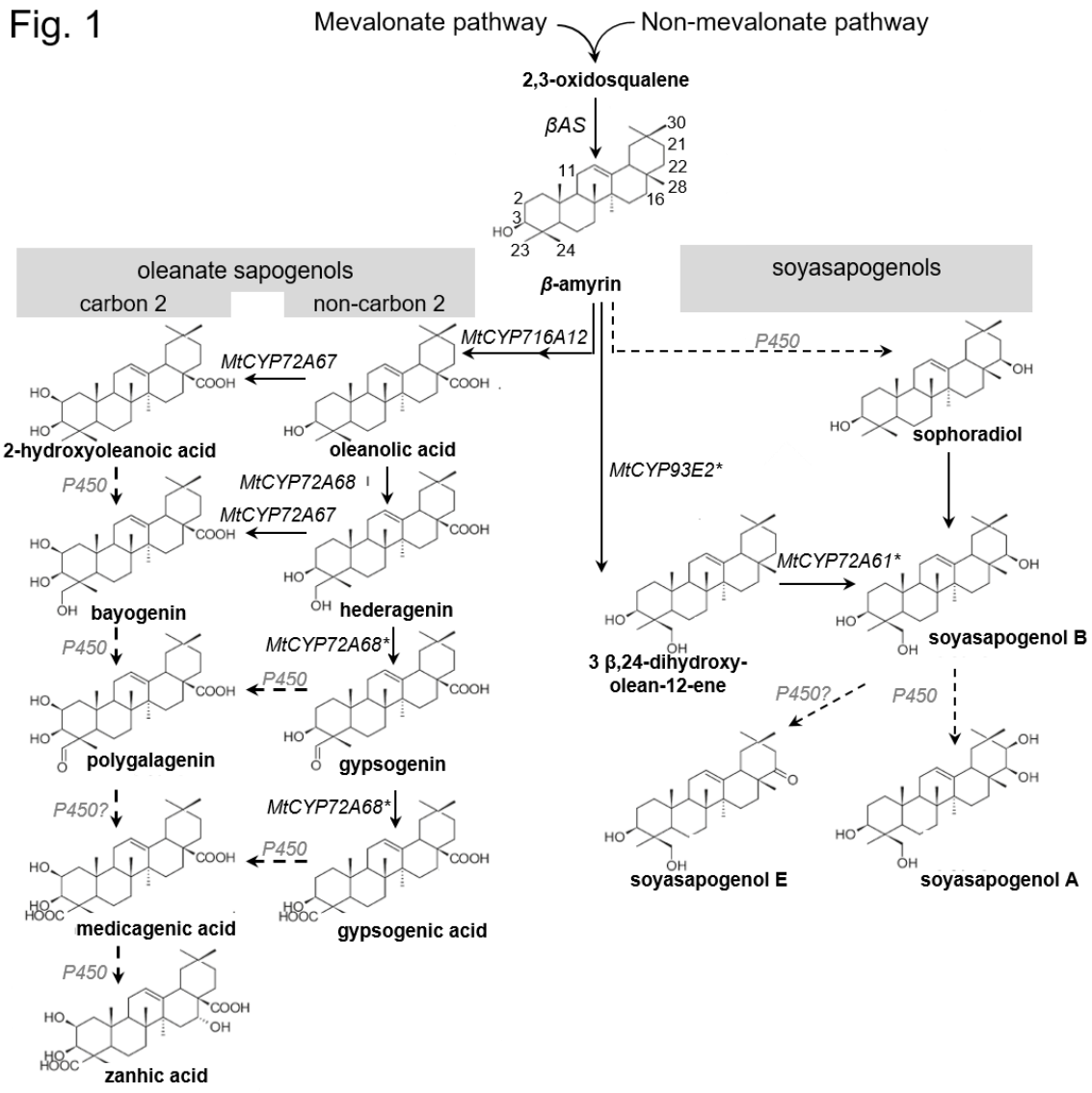

Figure 1. Proposed biosynthetic pathway of triterpene sapogenins reported in *Medicago* spp. (modified from Pollier et al., (2011) *J. Natural Products*). Solid arrows mark the enzymatic steps which have been characterized: *MtCYP72A68* and *MtCYP72A67*in the current paper in addition to *MtβAS* (Suzuki et al. 2002), *MtCYP716A12* (Carelli et al. 2011)(Fukushima et al. 2011)(; Seki et al., 2011), *MtCYP93E2* (Fukushima et al. 2011), *MtCYP72A68** (Fukushima et al. 2013). Dashed arrows mark the unknown enzymes. Asterisk marks enzyme activity that has been tested in yeast system only (not *in planta*).

Fig. 2

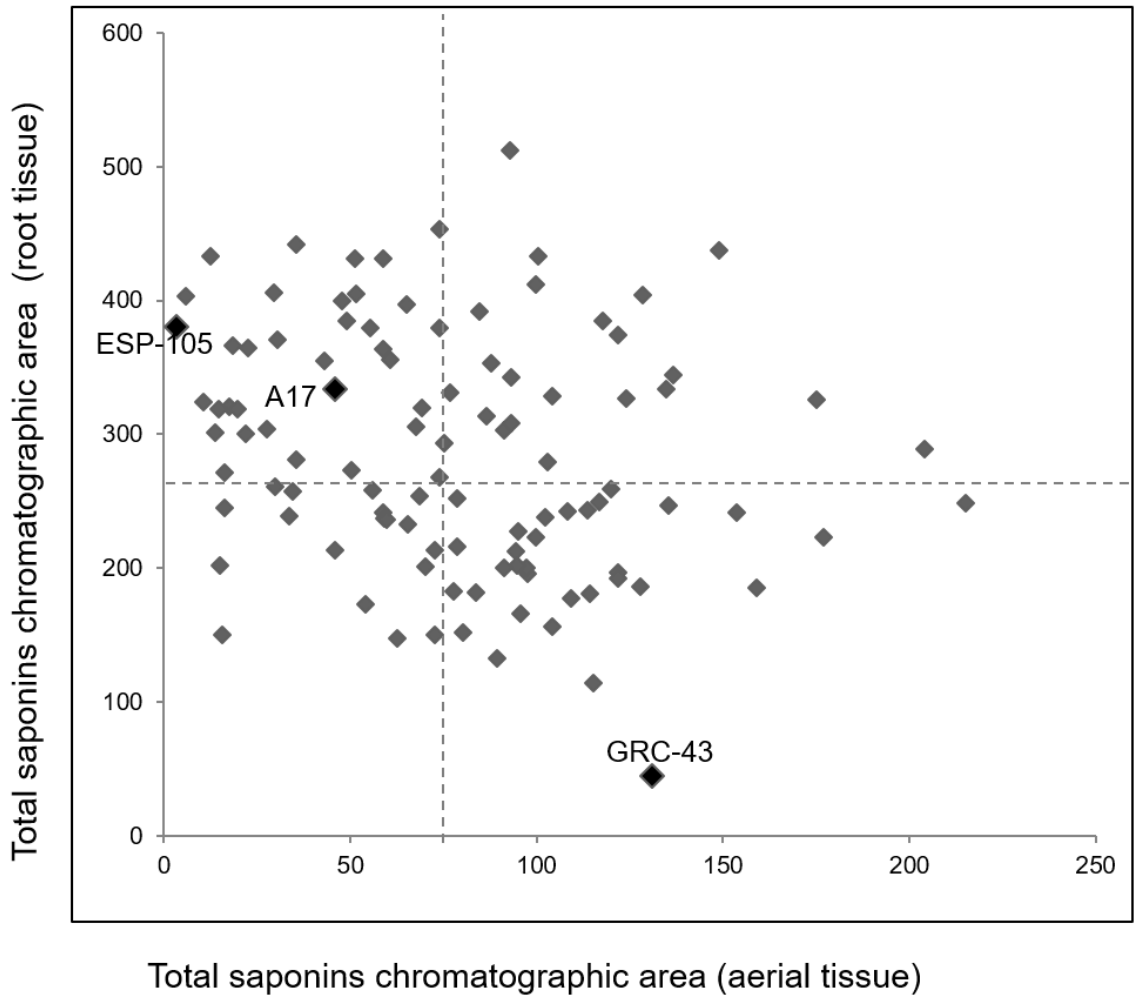

Figure 2. Scatter plot of total saponin content in root and aerial tissue measured for a germplasm collection of 106 *M. truncatula* ecotypes. Saponin content was quantified using UHPLC-(-)ESI-QToF-MS total chromatographic peak area and measured for 5-week-old root and aerial tissues (three to four biological replicates for each ecotype tissue). Dashed line, average values of the X = 78.32 and Y = 285.35 total saponin chromatogram peak areas. ESP-105 had relatively high levels of saponins in roots but low levels in aerial tissues, and vice versa for GRC-43. Ecotype A-17 was used as a reference line.

Fig. 3

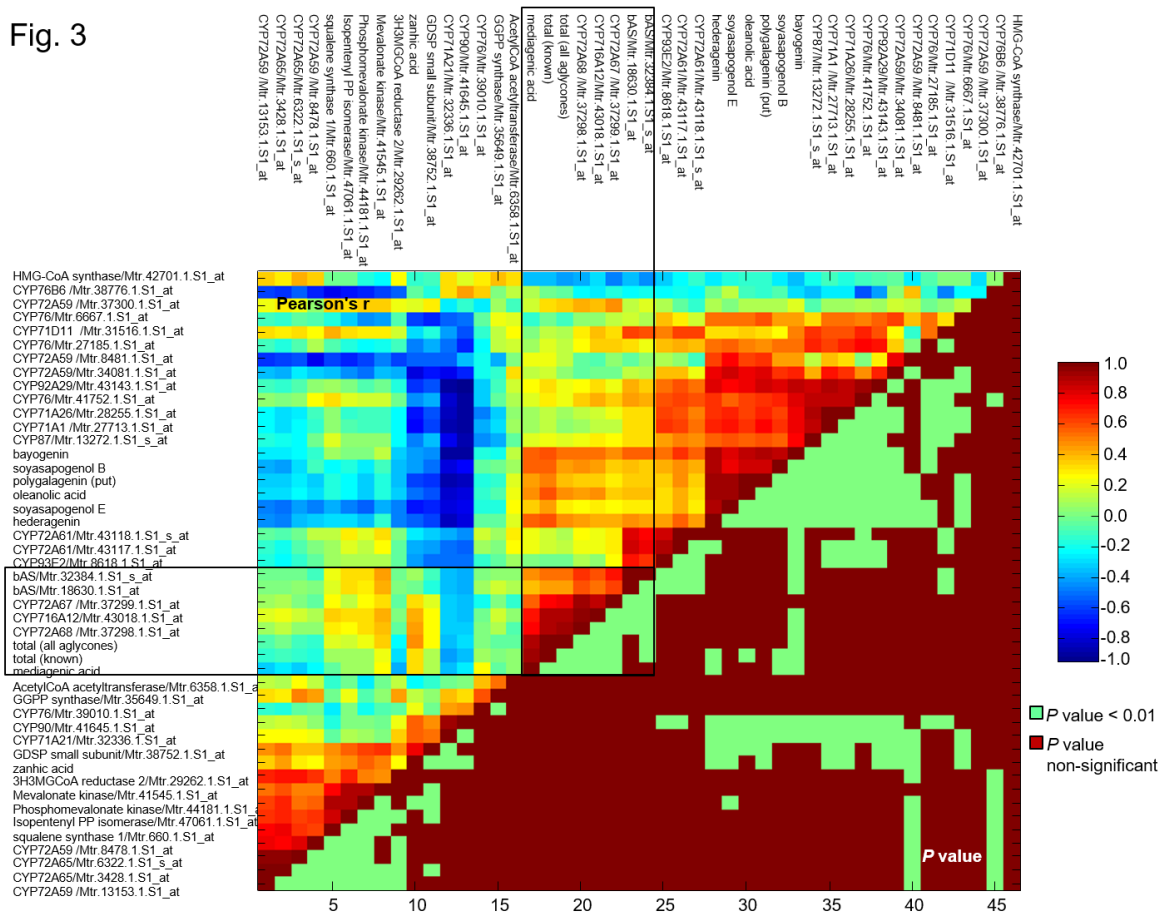

Figure 3. Correlation coefficient Pearson's r heat map of sapogenin metabolites and putative triterpene P450 transcripts. Metabolites and transcript levels were measured in root and aerial tissues of four selected *M. truncatula* ecotypes, A17, ESP105, GRC43 and R108 (three biological replicates each). Characterized and putative P450s involved in terpene biosynthesis were selected according to a previous report (Naoumkina et al. 2010) in addition to βAS and nine other known *M. truncatula* genes involved in early terpene biosynthesis. Upper left triangle matrix presents correlation coefficient Pearson's r as determined using Pearson's correlation (scale -0.8 to 1), and lower right triangle matrix presents *P*-values of the correlation coefficient test (light green highlights *P*-values <0.01, and brown highlights non-significant correlation). The black-bordered boxes encapsulate highly correlated probe sets including CYP72A67, CYPA68, CYP716A12 and *βAS*, and metabolites medicagenic acid, total saponins, and total aglycones.

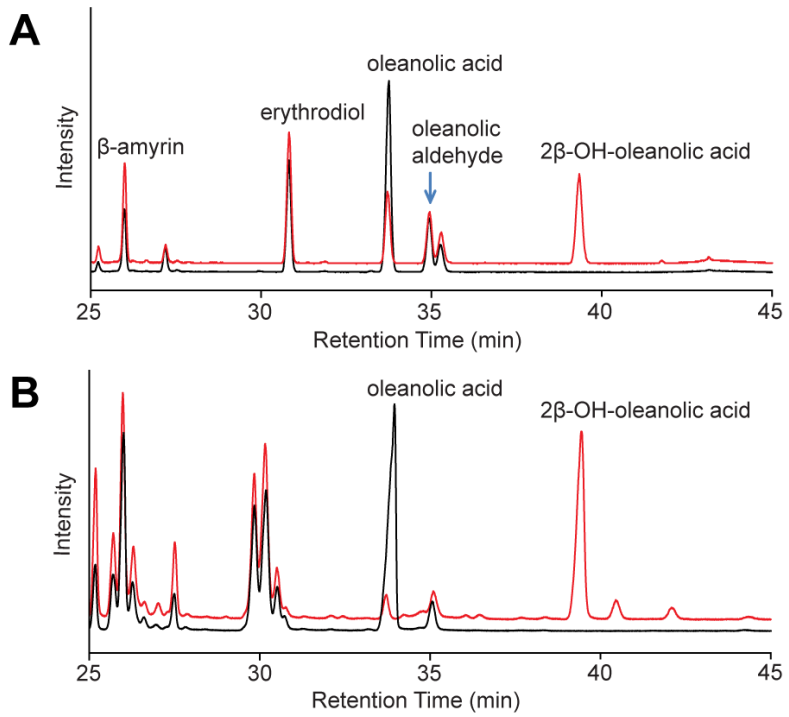

Figure 4. *In vivo* functional analysis of CYP72A67. Overlay of GC-MS chromatograms showing accumulation of trimethylsilylated enzymatic products. (A) *S. cerevisiae* strain KM1 expressing GgβAS, MtCPR1 and CYP716A12 (black) and *S. cerevisiae* strain KM2 expressing GgBAS , MtCPR1, CYP716A12 and CYP72A67 (red). (B) *N. benthamiana* co-infiltrated with *A. tumefaciens* strains armed with gene silencing suppressor p19, GgβAS and CYP716A12 (black) and *N. benthamiana* co-infiltrated with *A. tumefaciens* strains armed with gene silencing suppressor p19, GgβAS, CYP716A12 and CYP72A67 (red). CYP72A67v2 (version 2) was used.

Fig. 5

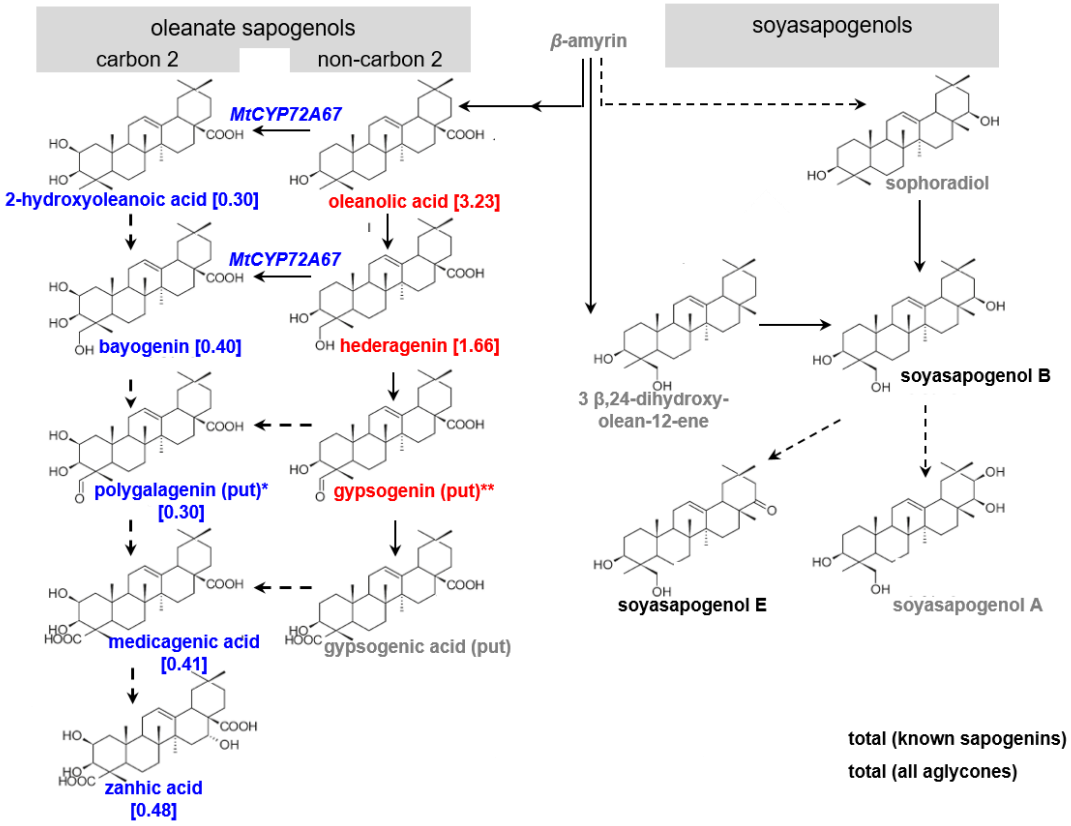

Figure 5. Characterization of *Mtcyp72a67*-RNAi in *M. truncatula* hairy roots. (A) Proposed sapogenin biosynthesis pathway in *M. truncatula* hairy root and observed fold changes noted in brackets. The sapogenin names are marked in different colors, according to the fold changes: red – saponin significant fold increase; blue – significant fold decrease; black – no change; and gray – not detected. Sapogenins were identified based upon authentic standards, except gypsogenin, gypsogenic acid and polygalagenin which were putatively identified using tandem mass. Gypsogenin (put) was detected only in the Mtcyp72a67-RNAi hairy root transgenic lines but not in the wild type.

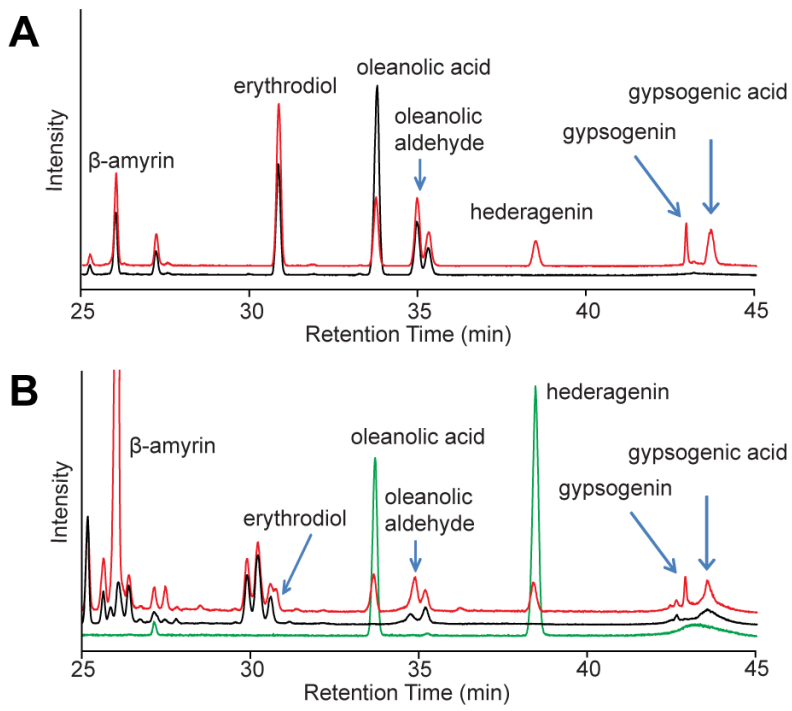

Figure **6**. In vivo functional analysis of CYP72A68. Overlay of GC-MS chromatograms showing accumulation of trimethylsilylated enzymatic products in (A) S. cerevisiae strains KM1 expressing GgβAS , MtCPR1 and CYP716A12 (black) and KM3 expressing GgβAS, MtCPR1, CYP716A12 and CYP72A68 (red). (B) N. benthamiana infiltrated with an *A. tumefaciens* strain armed with gene silencing suppressor p19 (black) and N. benthamiana co-infiltrated with *A. tumefaciens* strains armed with gene silencing suppressor p19, GgβAS, CYP716A12 and MtCYP72A68. Also shown are oleanolic acid and hederagenin standards (green). CYP72A68v2 (version 2) was used.

Fig. 7

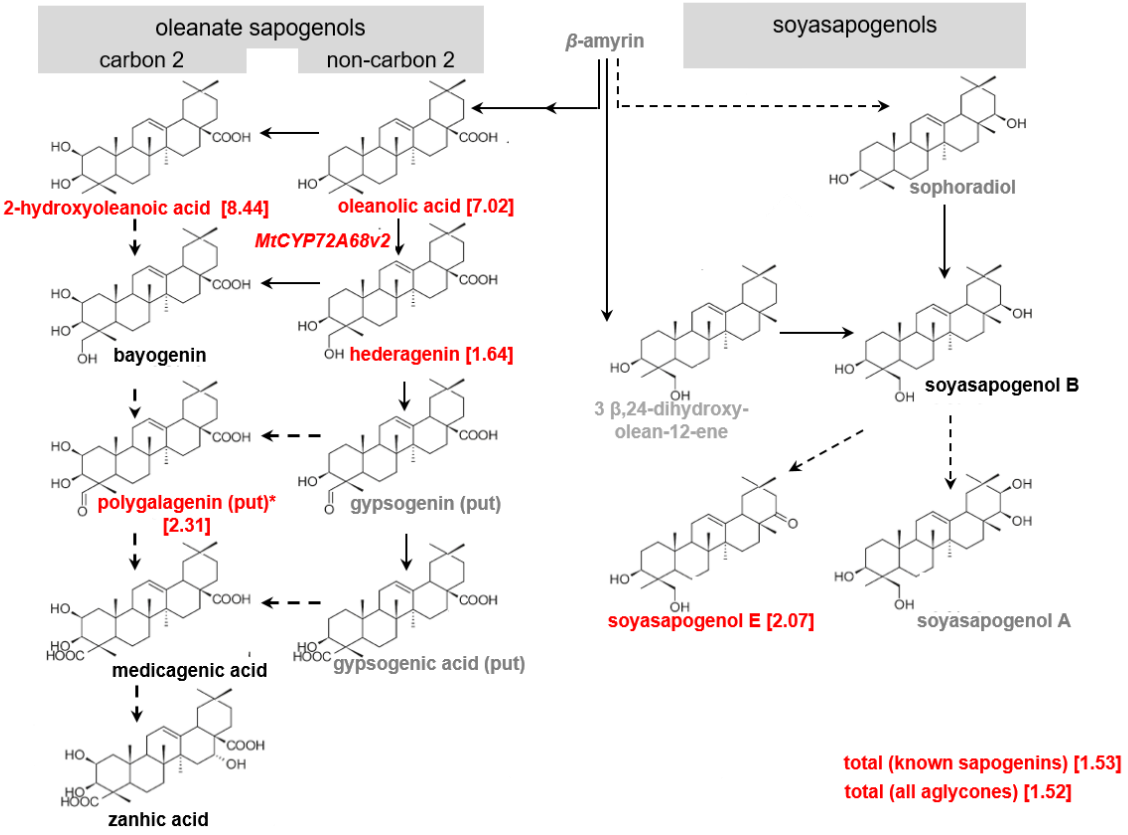

Figure 7. Characterization of *MtCYP72A68* overexpression in *M. truncatula* hairy roots. (A) Proposed sapogenin biosynthesis pathway in *M. truncatula* hairy root and observed fold changes noted in brackets. The sapogenin names are marked in different colors, according to the fold changes: red – saponin significant fold increase; black – no change; and gray – not detected. Sapogenins were identified based upon authentic standards, except gypsogenin, gypsogenic acid, and polygalagenin which were putatively identified using tandem mass.

Table 1. *In vitro* enzymatic assays of CYP72A67. Values in the table represent the mean, normalized chromatogram peak areas and standard error for each of the metabolites (three biological replicates per assay condition). The substrates that were used for CYP72A67 protein were (A) oleanolic acid and (B) hederagenin.

A) Substrate: oleanolic acid

Name	RT	*m/z*	CYP72A67 (+) NADPH			empty vector			fold change CYP72A67 / empty vector	P value
oleanolic acid	28.63	455.35	44866.7	±	2321.2	49566.7	±	1026.9	0.91	1.4E-01
2-hydroxyoleanoic	26.36	471.35	97000.0	±	3507.6	N.D.	±		high	

B) Substrate: hederagenin

Name	RT	*m/z*	CYP72A67 (+) NADPH			empty vector			fold change CYP72A67 / empty vector	P value
hederagenin	22.76	471.35	42033.3	±	1299.1	47966.7	±	1026.9	0.88	2.4E-02
bayogenin	20.28	487.34	45433.3	±	448.5	N.D.	±		high	

Table 2. *In vitro* enzymatic assay of CYP72A68 protein. Values in the table represent the mean, normalized chromatogram peak areas and standard error for each of the metabolites (three biological replicates per assay condition). The substrates that were used for CYP72A68 protein were (A) oleanolic acid and (B) hederagenin.

A) Substrate: oleanolic acid

Name	RT	*m/z*	CYP72A68 (+) NADPH			empty vector			fold change CYP72A68 / empty vector	P value
oleanolic acid	28.80	455.35	218673.0	±	12616.4	328189.4	±	6112.1	0.67	1.7E-04
hederagenin	22.96	471.35	175971.8	±	10435.2	N.D.			high	N.D.
gypsogenin (put)	24.91	469.33	258803.0	±	18852.2	N.D.			high	N.D.
gypsogenic acid (put)	21.88	485.33	156.8	±	39.1	N.D.			high	N.D.

B) Substrate: hederagenin

Name	RT	*m/z*	CYP72A68 (+) NADPH			empty vector			fold change CYP72A68 / empty vector	P value
hederagenin	22.96	471.35	227765.9	±	12260.6	250615.2	±	4084.8	0.91	3.8E-02
gypsogenin (put)	24.91	469.33	119954.8	±	16918.1	N.D.			high	N.D.
gypsogenic acid (put)	21.89	485.55	762.9	±	221.0	N.D.			high	N.D.

Compliance with Ethical Standards: All authors including Vered Tzin, John H. Snyder, Dong Sik Yang, David V. Huhman, Bonnie S. Watson, Stacy N. Allen, Yuhong Tang, Karel Miettinen, Philipp Arendt, Jacob Pollier, Alain Goossens, and Lloyd W. Sumner declare no conflict of interest that could have direct or potential influence or impart bias on the reported work. All authors also declare that no animal or human test subjects were involved in research reported here. Thus, no informed consent is necessary. This project was financially supported by the Oklahoma Center for the Advancement of Science and Technology (OCAST Award#PSB10-027), the National Science Foundation Molecular and Cellular Biosciences Award#1024976, The Samuel Roberts Noble Foundation, Oklahoma EPSCoR Research Experience for Undergraduates subaward #EPSCoR-2011-15, the VIB International PhD Fellowship Program (predoctoral fellowship to P.A.), the Research Foundation Flanders (postdoctoral fellowship to J.P.), and the European Union Seventh Framework Program FP7/2007–2013 under grant agreement number 613692–TriForC.

Literature Cited:

Achnine, L., Huhman, D. V, Farag, M. A., Sumner, L. W., Blount, J. W., & Dixon, R. A. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula. *The Plant Journal*, *41*(6), 875–887. doi:10.1111/j.1365-313X.2005.02344.x

Alberti, S., Gitler, A. D., & Lindquist, S. (2007). A suite of Gateway® cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, *24*(10), 913–919. doi:10.1002/yea.1502

Augustin, J. M., Drok, S., Shinoda, T., Sanmiya, K., Nielsen, J. K., Khakimov, B., et al. (2012). UDP-Glycosyltransferases from the UGT73C Subfamily in Barbarea vulgaris Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance. *Plant Physiology*, *160*(4), 1881 LP-1895. http://www.plantphysiol.org/content/160/4/1881.abstract

Augustin, J. M., Kuzina, V., Andersen, S. B., & Bak, S. (2011). Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry*, *72*(6), 435–457. doi:10.1016/j.phytochem.2011.01.015

Avato, P., Bucci, R., Tava, A., Vitali, C., Rosato, A., Bialy, Z., & Jurzysta, M. (2006). Antimicrobial activity of saponins from Medicago sp.: structure-activity relationship. *Phytotherapy Research*, *20*(6), 454–457. doi:10.1002/ptr.1876

Benedito, V. A., Torres-Jerez, I., Murray, J. D., Andriankaja, A., Allen, S., Kakar, K., et al. (2008). A gene expression atlas of the model legume Medicago truncatula. *Plant Journal*, *55*(3), 504–513. doi:10.1111/j.1365-313X.2008.03519.x

Bialy, Z., Jurzysta, M., Oleszek, W., Piacente, S., & Pizza, C. (1999). Saponins in alfalfa (Medicago sativa L.) root and their structural elucidation. *Journal of Agricultural and Food Chemistry*, *47*(8), 3185–3192. doi:10.1021/jf9901237

Biazzi, E., Carelli, M., Tava, A., Abbruscato, P., Losini, I., Avato, P., et al. (2015). CYP72A67 catalyzes a key oxidative step in Medicago truncatula hemolytic saponin biosynthesis. *Molecular Plant*, *8*(10), 1493–1506. doi:https://doi.org/10.1016/j.molp.2015.06.003

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*(1), 248–254. doi:https://doi.org/10.1016/0003-2697(76)90527-3

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., et al. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. *Proceedings of the National Academy of Sciences*, *108*(42), E864 LP-E870. http://www.pnas.org/content/108/42/E864.abstract

Broeckling, C. D., Huhman, D. V, Farag, M. A., Smith, J. T., May, G. D., Mendes, P., et al. (2005). Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *Journal of Experimental Botany*, *56*(410), 323–336.

http://dx.doi.org/10.1093/jxb/eri058

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., & Sumner, L. W. (2006). MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Analytical Chemistry*, *78*(13), 4334–4341. doi:10.1021/ac0521596

Carelli, M., Biazzi, E., Panara, F., Tava, A., Scaramelli, L., Porceddu, A., et al. (2011). Medicago truncatula CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *The Plant cell*, *23*(8), 3070–81. doi:10.1105/tpc.111.087312

Cheng, X., Wen, J., Tadege, M., Ratet, P., & Mysore, K. S. (2011). Reverse genetics in Medicago truncatula using Tnt1 insertion mutants BT - plant reverse genetics: methods and protocols. In A. Pereira (Ed.), (pp. 179–190). Totowa, NJ: Humana Press. doi:10.1007/978-1-60761-682-5_13

Dixon, R. A., & Sumner, L. W. (2003). Legume natural products: understanding and manipulating complex pathways for human and animal health. *Plant Physiology*, *131*(3), 878 LP-885. http://www.plantphysiol.org/content/131/3/878.abstract

Fiehn, O., Sumner, L. W., Rhee, S. Y., Ward, J., Dickerson, J., Lange, B. M., et al. (2007). Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics*, *3*(3), 195–201. doi:10.1007/s11306-007-0068-0

Fukushima, E. O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., et al. (2011). CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant and Cell Physiology*, *52*(12), 2050–2061. doi:10.1093/pcp/pcr146

Fukushima, E. O., Seki, H., Sawai, S., Suzuki, M., Ohyama, K., Saito, K., & Muranaka, T. (2013). Combinatorial biosynthesis of legume natural and rare triterpenoids in engineered yeast. *Plant and Cell Physiology*, *54*(5), 740–749. doi:10.1093/pcp/pct015

Geisler, K., Hughes, R. K., Sainsbury, F., Lomonossoff, G. P., Rejzek, M., Fairhurst, S., et al. (2013). Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. *Proceedings of the National Academy of Sciences*, *110*(35), 3360–3367. http://www.pnas.org/content/110/35/E3360.abstract

Gholami, A., De Geyter, N., Pollier, J., Goormachtig, S., & Goossens, A. (2014). Natural product biosynthesis in Medicago species. *Natural Product Reports*, *31*(3), 356–380. doi:10.1039/C3NP70104B

Goossens, A. (2015). It is easy to get huge candidate gene lists for plant metabolism now, but how to get beyond? *Molecular Plant*, *8*(1), 2–5. doi:https://doi.org/10.1016/j.molp.2014.08.001

Greenhagen, B. T., Griggs, P., Takahashi, S., Ralston, L., & Chappell, J. (2003). Probing sesquiterpene hydroxylase activities in a coupled assay with terpene synthases. *Archives of Biochemistry and Biophysics*, *409*(2), 385–394. doi:10.1016/S0003-9861(02)00613-6

Haridas, V., Higuchi, M., Jayatilake, G. S., Bailey, D., Mujoo, K., Blake, M. E., et al. (2001).

Avicins : Triterpenoid saponins from Acacia victoriae (Bentham) induce apoptosis by mitochondrial perturbation. *Proceedings of the National Academy of Sciences*, *98*(10), 5821–5826.

He, J., Benedito, V. A., Wang, M., Murray, J. D., Zhao, P. X., Tang, Y., & Udvardi, M. K. (2009). The Medicago truncatula gene expression atlas web server. *BMC bioinformatics*, *10*(1), 441. doi:10.1186/1471-2105-10-441

Hirai, M. Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D. B., Yamazaki, Y., et al. (2005). Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *The Journal of biological chemistry*, *280*(27), 25590–5. doi:10.1074/jbc.M502332200

Hirai, M. Y., & Saito, K. (2004). Post-genomics approaches for the elucidation of plant adaptive mechanisms to sulphur deficiency. *Journal of experimental botany*, *55*(404), 1871–9. doi:10.1093/jxb/erh184

Hirai, M. Y., Sawada, Y., Kanaya, S., Kuromori, T., Kobayashi, M., Klausnitzer, R., et al. (2010). Toward genome-wide metabolotyping and elucidation of metabolic system: metabolic profiling of large-scale bioresources. *Journal of Plant Research*, *123*(3), 291–298. doi:10.1007/s10265-010-0337-2

Huhman, D. V, Berhow, M. A., & Sumner, L. W. (2005). Quantification of saponins in aerial and subterranean tissues of Medicago truncatula. *Journal of Agricultural and Food Chemistry*, *53*(6), 1914–1920. doi:10.1021/jf0482663

Huhman, D. V, & Sumner, L. W. (2002). Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry*, *59*(3), 347–360. doi:https://doi.org/10.1016/S0031-9422(01)00432-0

Inagaki, Y.-S., Etherington, G., Geisler, K., Field, B., Dokarry, M., Ikeda, K., et al. (2011). Investigation of the potential for triterpene synthesis in rice through genome mining and metabolic engineering. *New Phytologist*, *191*(2), 432–448. doi:10.1111/j.1469-8137.2011.03712.x

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, *31*(4), e15–e15. https://www.ncbi.nlm.nih.gov/pubmed/12582260

Iturbe-Ormaetxe, I., Haralampidis, K., Papadopoulou, K., & Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from Medicago truncatula and Lotus japonicus. *Plant Molecular Biology*, *51*(5), 731–743. doi:10.1023/A:1022519709298

Karimi, M., Inze, D., & Depicker, a. (2002). GATEWAY vectors for 0RW1S34RfeSDcfkexd09rT2Agrobacterium1RW1S34RfeSDcfkexd09rT2-mediated plant transformation. *Trends in plant science*, *7*(5), 193–195. doi:10.1016/S1360-1385(02)02251-3

Kirby, J., Romanini, D. W., Paradise, E. M., & Keasling, J. D. (2008). Engineering triterpene

production in Saccharomyces cerevisiae–β-amyrin synthase from Artemisia annua. *The FEBS Journal*, *275*(8), 1852–1859. doi:10.1111/j.1742-4658.2008.06343.x

Kirk, D. D., Rempel, R., Pinkhasov, J., & Walmsley, A. M. (2004). Application of Quillaja saponaria extracts as oral adjuvants for plant-made vaccines. *Expert Opinion on Biological Therapy*, *4*(6), 947–958. doi:10.1517/14712598.4.6.947

Klita, P. ., Mathison, G., Fenton, T., & Hardin, R. (1996). Effects of alfalfa root saponins on digestive function in sheep. *Journal of Animal Science*, *74*, 1144–1156.

Kuljanabhagavad, T., Thongphasuk, P., Chamulitrat, W., & Wink, M. (2008). Triterpene saponins from Chenopodium quinoa Willd. *Phytochemistry*, *69*(9), 1919–1926. doi:https://doi.org/10.1016/j.phytochem.2008.03.001

Kunii, M., Kitahama, Y., Fukushima, E. O., Seki, H., Muranaka, T., Yoshida, Y., & Aoyama, Y. (2012). b-Amyrin oxidation by oat CYP51H10 expressed heterologously in yeast cells: the first example of CYP51-dependent metabolism other than the 14-demethylation of sterol precursors. *Biological and Pharmaceutical Bulletin*, *35*(5), 801–804. doi:10.1248/bpb.35.801

Lei, Z., Li, H., Chang, J., Zhao, P., & Sumner, L. (2012). MET-IDEA version 2.06; improved efficiency and additional functions for mass spectrometry-based metabolomics data processing. *Metabolomics*, *8*, 105–110.

Li, L., Cheng, H., Gai, J., & Yu, D. (2007). Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume Medicago truncatula. *Planta*, *226*(1), 109–123. doi:10.1007/s00425-006-0473-z

Liu, C.-J., Huhman, D., Sumner, L. W., & Dixon, R. A. (2003). Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from Medicago truncatula. *The Plant Journal*, *36*(4), 471–484. doi:10.1046/j.1365-313X.2003.01893.x

Lu, C. D., & Jorgensen, N. A. (1987). Alfalfa saponins affect site and extent of nutrient digestion in Ruminants. *The Journal of Nutrition*, *117*(5), 919–927. http://dx.doi.org/10.1093/jn/117.5.919

Mertens, J., Pollier, J., Vanden Bossche, R., Lopez-Vidriero, I., Franco-Zorrilla, J. M., & Goossens, A. (2016). The bHLH Transcription Factors TSAR1 and TSAR2 Regulate Triterpene Saponin Biosynthesis in Medicago truncatula. *Plant Physiology*, *170*(1), 194 LP-210. doi:10.1104/pp.15.01645

Miettinen, K., Iñigo, S., Kreft, L., Pollier, J., De Bo, C., Botzki, A., et al. (2018). The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis. *Nucleic Acids Research*, *46*(D1), D586–D594. http://dx.doi.org/10.1093/nar/gkx925

Miettinen, K., Pollier, J., Buyst, D., Arendt, P., Csuk, R., Sommerwerk, S., et al. (2017). The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nature Communications*, *8*, 14153. https://doi.org/10.1038/ncomms14153

Moses, T., Papadopoulou, K. K., & Osbourn, A. (2014). Metabolic and functional diversity of

saponins, biosynthetic intermediates and semi-synthetic derivatives. *Critical Reviews in Biochemistry and Molecular Biology*, *49*(6), 439–462. doi:10.3109/10409238.2014.953628

Moses, T., Pollier, J., Almagro, L., Buyst, D., Van Montagu, M., Pedreño, M. A., et al. (2014). Combinatorial biosynthesis of sapogenins and saponins in Saccharomyces cerevisiae using a C-16α hydroxylase from Bupleurum falcatum. *Proceedings of the National Academy of Sciences*, *111*(4), 1634 LP-1639. http://www.pnas.org/content/111/4/1634.abstract

Moses, T., Pollier, J., Shen, Q., Soetaert, S., Reed, J., Erffelinck, M.-L., et al. (2015). OSC2 and CYP716A14v2 catalyze the biosynthesis of triterpenoids for the cuticle of aerial organs of Artemisia annua. *The Plant Cell*, *27*(1), 286 LP-301. http://www.plantcell.org/content/27/1/286.abstract

Moses, T., Pollier, J., Thevelein, J. M., & Goossens, A. (2013). Bioengineering of plant (tri)terpenoids: from metabolic engineering of plants to synthetic biology in vivo and in vitro. *New Phytologist*, *200*(1), 27–43. doi:10.1111/nph.12325

Moses, T., Thevelein, J. M., Goossens, A., & Pollier, J. (2014). Comparative analysis of CYP93E proteins for improved microbial synthesis of plant triterpenoids. *Phytochemistry*, *108*, 47–56. doi:https://doi.org/10.1016/j.phytochem.2014.10.002

Naoumkina, M. a, Modolo, L. V, Huhman, D. V, Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., & Dixon, R. a. (2010). Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in Medicago truncatula. *The Plant cell*, *22*(March), 850–866. doi:10.1105/tpc.109.073270

Naoumkina, M., Farag, M. A., Sumner, L. W., Tang, Y., Liu, C.-J., & Dixon, R. A. (2007). Different mechanisms for phytoalexin induction by pathogen and wound signals in Medicago truncatula. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17909–17915. doi:10.1073/pnas.0708697104

Nelson, D., & Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *The Plant Journal*, *66*(1), 194–211. doi:10.1111/j.1365-313X.2011.04529.x

Oleszek, W. (1996). Alfalfa saponins: structure, biological activity, and chemotaxonomy BT - saponins used in food and agriculture. In G. R. Waller & K. Yamasaki (Eds.), (pp. 155–170). Boston, MA: Springer US. doi:10.1007/978-1-4613-0413-5_13

Pang, Y., Peel, G. J., Wright, E., Wang, Z., & Dixon, R. A. (2007). Early steps in proanthocyanidin biosynthesis in the model legume Medicago truncatula. *Plant Physiology*, *145*(3), 601 LP-615. http://www.plantphysiol.org/content/145/3/601.abstract

Pollier, J., Morreel, K., Geelen, D., & Goossens, A. (2011). Metabolite profiling of triterpene saponins in Medicago truncatula hairy roots by liquid chromatography fourier transform ion cyclotron resonance mass spectrometry, 1462–1476.

Pollier, J., Moses, T., González-Guzmán, M., De Geyter, N., Lippens, S., Bossche, R. Vanden, et al. (2013). The protein quality control system manages plant defence compound synthesis. *Nature*, *504*, 148. https://doi.org/10.1038/nature12685

Pompon, D., Louerat, B., Bronine, A., & Urban, P. (1996). Yeast expression of animal and plant P450s in optimized redox environments. In E. F. Johnson & M. R. B. T.-M. in E. Waterman (Eds.), *Cytochrome P450, Part B* (Vol. 272, pp. 51–64). Academic Press. doi:https://doi.org/10.1016/S0076-6879(96)72008-6

Quandt, H. J., Puhler, A., & Broer, I. (1993). Transgenic root nodules of Vicia hirsuta: a fast and efficient system for the study of gene expression in indeterminate-type nodules. *Molecular plant-microbe interactions, v. 6*(6), 699-706–1993 v.6 no.6. doi:10.1094/MPMI-6-699

Ro, D.-K., Arimura, G.-I., Lau, S. Y. W., Piers, E., & Bohlmann, J. (2005). Loblolly pine abietadienol/abietadienal oxidase PtAO(CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. *Proceedings of the National Academy of Sciences*, *102*(22), 8060 LP-8065. http://www.pnas.org/content/102/22/8060.abstract

Ronfort, J., Bataillon, T., Santoni, S., Delalande, M., David, J. L., & Prosperi, J.-M. (2006). Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. *BMC Plant Biology*, *6*(1), 28. doi:10.1186/1471-2229-6-28

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, *132*, 365–386.

Sawai, S., & Saito, K. (2011). Triterpenoid biosynthesis and engineering in plants. *Frontiers in Plant Science*. https://www.frontiersin.org/article/10.3389/fpls.2011.00025

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., et al. (2008). Licorice β-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. *Proceedings of the National Academy of Sciences*, *105*(37), 14204 LP-14209. http://www.pnas.org/content/105/37/14204.abstract

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., et al. (2011). Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *The Plant cell*, *23*(11), 4112–23. doi:10.1105/tpc.110.082685

Seki, H., Tamura, K., & Muranaka, T. (2015). P450s and UGTs: key players in the structural diversity of triterpenoid saponins. *Plant and Cell Physiology*, *56*(8), 1463–1471. http://dx.doi.org/10.1093/pcp/pcv062

Sen, S., Makkar, H. P. S., & Becker, K. (1998). Alfalfa saponins and their implication in animal nutrition. *Journal of Agricultural and Food Chemistry*, *46*(1), 131–140. doi:10.1021/jf970389i

Shibata, S. (2001). Chemistry and cancer preventing activities of ginseng saponins and some related triterpenoid compounds. *Journal of Korean medical science*, *16 Suppl*(Suppl), S28–S37. doi:10.3346/jkms.2001.16.S.S28

Shibuya, M., Hoshino, M., Katsube, Y., Hayashi, H., Kushiro, T., & Ebizuka, Y. (2006). Identification of β-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. *The FEBS Journal*, *273*(5), 948–959.

doi:10.1111/j.1742-4658.2006.05120.x

Sparg, S. G., Light, M. E., & van Staden, J. (2004). Biological activities and distribution of plant saponins. *Journal of Ethnopharmacology*, *94*(2), 219–243. doi:https://doi.org/10.1016/j.jep.2004.05.016

Sumner, L., Snyder, J., Yang, D., Tzin, V., Huhman, D., Allen, S., & Tang, Y. (2012). Deciphering triterpene saponin biosynthesis using natural diversity, metabolomics and gene expression profiling. In *In American Society for Mass Spectrometry Conference, Springer, Vancouver, Canada* (p. 65).

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society*, *3*(3), 211–221. doi:10.1007/s11306-007-0082-2

Sundaramoorthy, J., Park, G. T., Mukaiyama, K., Tsukamoto, C., Chang, J. H., Lee, J.-D., et al. (2018). Molecular elucidation of a new allelic variation at the Sg-5 gene associated with the absence of group A saponins in wild soybean. *PLOS ONE*, *13*(1), e0192150. https://doi.org/10.1371/journal.pone.0192150

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., & Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in Medicago truncatula. *The Plant Journal*, *32*(6), 1033–1048. doi:10.1046/j.1365-313X.2002.01497.x

Suzuki, H., Reddy, M. S. S., Naoumkina, M., Aziz, N., May, G. D., Huhman, D. V, et al. (2005). Methyl jasmonate and yeast elicitor induce differential transcriptional and metabolic re-programming in cell suspension cultures of the model legume Medicago truncatula. *Planta*, *220*(5), 696–707. http://www.jstor.org/stable/23388822

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., et al. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume Medicago truncatula. *The Plant Journal*, *54*(2), 335–347. doi:10.1111/j.1365-313X.2008.03418.x

Tava, A., Scotti, C., & Avato, P. (2011). Biosynthesis of saponins in the genus Medicago. *Phytochemistry Reviews*, *10*(4), 459–469. doi:10.1007/s11101-010-9169-x

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., & Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annual Review of Plant Biology*, *65*(1), 225–257. doi:10.1146/annurev-arplant-050312-120229

Tzin, V., Snyder, J. H., Yang, D. S., David V. Huhman, B. S. W., Allen, S. N., Tang, Y., & Lloyd W. Sumner. (2012a). Exploiting chemical diversity for triterpene saponin gene discovery in Medicago truncatula. In *In 8th Annual International Conference of the Metabolomics Society, Washington, DC, USA* (p. 34).

Tzin, V., Snyder, J. H., Yang, D. S., David V. Huhman, B. S. W., Allen, S. N., Tang, Y., & Lloyd W. Sumner. (2012b). Integrated metabolomics yields novel gene discoveries related to triterpene saponin biosynthesis in Medicago truncatula. In *In Gordon Research*

Conference on Plant Molecular Biology Holderness, New Hampshire, USA.

Urban, P., Mignotte, C., & Pompon, D. (1997). Cloning, yeast expression, and characterization of the coupling of two distantly related Arabidopsis thaliana NADPH-cytochrome P450 reductases with P450 CYP73A5. *J Biol Chem*, *272*(31), 19176–19186.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., & Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. *Plant, Cell and Environment*, *32*(9), 1211–1229. doi:10.1111/j.1365-3040.2009.01978.x

Verdier, J., Zhao, J., Torres-Jerez, I., Ge, S., Liu, C., He, X., et al. (2012). MtPAR MYB transcription factor acts as an on switch for proanthocyanidin biosynthesis in Medicago truncatula. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1766–71. doi:10.1073/pnas.1120916109

Yan, M., Zhu, Y., Zhang, H. J., Jiao, W. H., Han, B. N., Liu, Z. X., et al. (2013). Anti-inflammatory secondary metabolites from the leaves of Rosa laevigata. *Bioorganic and Medicinal Chemistry*, *21*(11), 3290–3297. doi:10.1016/j.bmc.2013.03.018

Yano, R., Takagi, K., Takada, Y., Mukaiyama, K., Tsukamoto, C., Sayama, T., et al. (2016). Metabolic switching of astringent and beneficial triterpenoid saponins in soybean is achieved by a loss-of-function mutation in cytochrome P450 72A69. *The Plant Journal*, *89*(3), 527–539. doi:10.1111/tpj.13403

Yoshiki, Y., Kudou, S., & Okubo, K. (1998). Relationship between chemical structures and biological activities of triterpenoid saponins from soybean. *Bioscience, Biotechnology, and Biochemistry*, *62*(12), 2291–2299. doi:10.1271/bbb.62.2291

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, *480*, 520. https://doi.org/10.1038/nature10625