EXPLOITING GERMPLASM DIVERSITY FOR TRITERPENE SAPONIN BIOSYNTHETIC GENE DISCOVERY USING INTEGRATED METABOLOMICS AND TRANSCRIPTOMICS

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
John Hugh Snyder
May 2012

EXPLOITING GERMPLASM DIVERSITY FOR TRITERPENE SAPONIN BIOSYNTHETIC GENE DISCOVERY USING INTEGRATED METABOLOMICS AND TRANSCRIPTOMICS

John Hugh Snyder, Ph.D.

Cornell University 2012

Triterpene saponins are a class of structurally diverse plant natural products with a wide range of demonstrated bioactivities. Individual triterpene saponins have been demonstrated to possess allelopathic, anti-fungal, anti-bacterial, anti-insect, anti-feedant, and anti-cancer activities. The biosynthesis of triterpene saponins is poorly characterized. The model legume *Medicago truncatula* is known to accumulate a large variety of triterpene saponin compounds, resulting from the differential glycosylation of at least seven triterpene aglycone structures. In this project, UPLC-ESI-qTOF-MS analysis was used to profile the accumulation of triterpene saponin metabolites in a collection of 100 *M. truncatula* ecotypes (germplasm accessions). Analyses of both aerial and root organs were performed. These metabolomics analyses revealed interesting trends in differential spatial and structural accumulation patterns between the various ecotypes, and between the organs. The high-resolution “biochemical phenotyping" data for the whole ecotype collection enabled an informed selection of hypo- and hyper- accumulating ecotypes for subsequent transcriptomic analyses via Affymetrix Medicago GeneChips®. Integrated analyses of saponin accumulation phenotypes with transcript expression data led to the identification of a number of biosynthetic and regulatory gene candidates. Seven cytochrome P450 gene candidates were cloned and introduced to Wat11 yeast cells, enabling microsomal isolation and detailed in vitro characterization of enzyme function. The cytochrome P450 enzyme CYP72A68 showed sequential oxidase activity for carbon 23 of oleanolic

acid and several structurally related compounds in the triterpene sapogenin biosynthesis pathway. CYP72A67 showed monooxygenase activity at carbon 2 of oleanolic acid and hederagenin, yielding 2-hydroxyoleanolic acid and bayogenin.

Biographical Sketch

John Hugh Snyder is currently a Ph.D. candidate in the Field of Plant Biology at Cornell University. He began his undergraduate career at Reed College, where he studied Mandarin. He went on to graduate with Honors and Distinction from The Schreyer Honors College at Pennsylvania State University with a B.S. in Agroecology/Plant Science. During his studies at Cornell, he worked as a visiting student at the Samuel Roberts Noble Foundation for three years. John's work has been presented at several conferences. He won the NIH Competitive Travel grant for an oral presentation delivered at the *Metabolomics 2010* conference as well as Cornell University's Department of Plant Biology's "Outstanding Teaching Assistant Award". In addition, he has secured a number of fellowships and scholarships.

Currently he lives with his wife and daughter in Beijing, where he is slated to begin postdoctoral work related to his dissertation.

To Ni

Acknowledgments

I would like to thank the members of my degree committee for their encouragement and support of my professional development. I am grateful to Peter Davies for all his support. He mentored me during my teaching assistantship for introductory plant physiology, provided a thorough introduction to the field of plant hormone physiology, and generally supported my various efforts throughout my degree program. Klaas van Wijk provided stalwart support and administration of my unorthodox graduate fellowship program. Robert Raguso gave me extremely helpful suggestions for rhetorical and statistical methods. Frank Schoeder helped me with opinions/admonishments regarding the intellectual danger of an overemphasis on chemical structures (as their own justification) in the field of metabolomics. Finally, I'd like to thank Lloyd Sumner for agreeing to support me as a student, for providing an introduction to traditional phytochemistry methods through pairing me to work with Wensheng Li on the prenylated isoflavonoid project, and for providing an introduction to modern metabolomics methods.

At Cornell, my discussion with Yi Yi regarding the development and commercial release of GM crops were consistently enlightening. Tom Owens provided a thorough introduction to Plant Physiology and Plant Biochemistry, while

my conversations with Robert Turgeon improved my appreciation for the importance of temporal dynamics in biochemical studies. Finally, Steven Tanksley strongly encouraged me to follow my focused goals and obtain research funding to an off-campus site.

At the Noble Foundation, a number of people helped make my transition and work there a pleasant experience. Richard Dixon provided interesting information on the plant natural product research community and the cost of deployment of GM crops. EwaUrbanczyk-Wochniak helped prepare me for the sorts of GC-MS work I would be doing and gave me some much welcomed encouragement in my professional development. Li Wensheng furthered my education by training me in traditional phytochemical protocols, as did Pang Yongzhen with our discussions discussions regarding use of the Affymetrix transcriptomics data.

Throughout my project, there were many others at Noble who were instrumental. In the troubleshooting department, ShenGuoan helped with troubleshooting protocols in molecular biology and enzymology (along with the estimable Changjun Liu). Loachine Achnine gave me several useful suggestions about genomics resources for *Medicago truncatula.* Wang Guodong and Zhou Rui were quite forthcoming with information regarding cloning and sub-cloning protocols.

Bennie Bench, David Human, and Mohamed Bedair all provided help and encouragement for all things analytical chemistry and metabolomics. Bonnie Watson was also useful with regard to plant and plant cell protocols, but perhaps even more edifying was her comments and advice regarding graduate education. Jiangqi Wen and Xiaofei Cheng both helped with identification of TNT1 insertion mutants, with Jiangqi providing additional information regarding genotyping and Xiaofei filling in the gaps in my knowledge regarding troubleshooting PCR protocols. Finally, Carla Welch provided exemplary care for the plants used in my work and gave me a useful window into the broader culture of Oklahoma.

Several people outside of those two institutions were invaluable throughout the years of my project. Anne Osbourn and I engaged in several useful discussions regarding metabolic clusters and the importance of allelic diversity among ecotypes. Ning Zheng's discussion and correspondence regarding the inadequacy of the "lock and key" model to accurately describe biomolecular interactions steered me away from some theoretical problems. I would also like to thank Jeanne Rasberry for protocols and advice for the recombinant expression of cytochrome P450 enzymes. Joe Chappelle provided Wat 11 yeast cultures and general encouragement throughout my

degree program. Kazuki Saito also gave me general encouragement as well as helpful information through several discussions regarding lethal mutations.

In the same category of valued outside advisors, David Liscombe and I had interesting discussions regarding both graduate education and the position of plant natural products within the larger plant biology community. Talking with Joerg Bohlmann about the screening hypothesis and methods for cytochrome P450 enzymology research also proved very helpful. Daneel Ferreira gave me some important information on the mechanism of cytochrome P450 mediated bio-oxidation reactions. Qi Xiaoquan's discussions regarding evolutionary implications of the broad substrate tolerance of certain cytochrome P450 enzymes were helpful, and I look forward to working with him in the coming years.

Finally, at my alma mater, Penn State University, Surinder Chopra provided my introduction to professional research. Michael Saunders, Mark Shriver and James Frazier all provided helpful encouragement for my professional and academic development. I am glad to have been able to have a series of enriching discussions regarding the interface of ecology and biochemistry, and thoughts about graduate education with David Mortenson. Mark Mescher shared his insights regarding the nature of undergraduate science education with me, which helped me better

understand that stage of my education. Richard Doyle provided consistent encouragement and enlightening discussions regarding graduate education.

Table of Contents (vi)

Exploiting Germplasm Diversity For Triterpene Saponin Biosynthetic Gene Discovery Using Integrated Metabolomics And Transcriptomics i

Biographical Sketch iii

Dedication iv

Acknowledgments v

Table of Contents x

List of Figures, Tables, and Equations xiii

Main Text 1

Chapter I - A Metabolomics-Based Platform for the Assessment of Triterpene Saponin Biochemical Variation in a *Medicago truncatula* Germplasm Diversity Collection 1

Authors: John H. Snyder, David V. Huhman, Stacey Allen,and Lloyd W. Sumner 1

Summary 1

Abstract 1

Abbreviations 2

Introduction 4

Methods 9

Results and Discussion 13

Additional Information 28

Sources 28

Chapter II - Identification of Candidate Biosynthetic Genes in Triterpene Saponin Metabolism Through Integrated Analysis of Metabolome And Transcriptome Datasets from *Medicago truncatula* Ecotypes with Differential Triterpene Saponin Accumulation Phenotypes 36

Authors: John H. Snyder, David V. Huhman, Yuhong Tang, and Lloyd W. Sumner 36

Summary 36
Abstract.....37
Glossary 38
Introduction 38
Results 44
Discussion.....56
Methods.....64
Additional Information.....73
Sources.....77
Chapter III - Enzymatic Characterization of CYP72A67 and CYP72A68, Two Cytochrome P450 enzymes in the Triterpene Sapogenin Biosynthetic Pathway of *Medicago truncatula.*.....85
Authors: John H. Snyder, David V. Huhman, Bennie J. Bench, and Lloyd W. Sumner. 85
Summary: 85
Abstract:.....85
Introduction 87
Results 91
Discussion.....107
Methods.....117
Additional Information.....123
Sources.....124
Chapter IV - Loci from the cyp88d Subfamily of Cytochrome P450s are Immediately Adjacent to Oxidosqualene Synthase Loci in the Genomes of *Medicago truncatula* and *Lotus japonicus.* 132
Authors: John H. Snyder, David V. Huhman, Lloyd W. Sumner.....132
Summary: 132
Abstract.....132

Introduction 134
Results 136
Discussion.....145
Methods.....148
Additional Information.....158
Sources.....159
Conclusion.....168
Expanded Repertoire of Reaction Pairs 171
Physiology of Seed Development in *M. truncatula* (Lethality and Organ-Specific Expression) 172
Matrix Pathways 172
Appendices.....174
Appendix RII – Gross Phenotype Comparisons.....174
Appendix RII – Hidden Markoff Models.....175
Cytochrome P450 Models 175
Glycosyltransferase Models 175
Appendix RII– Targeted Ion List.....177
Appendix RII – Primers 188
Appendix RIII – Primers.....190
Appendix RIII – Cloned Sequence.....191
Appendix RIV – Primers.....195
Primers Used in Reverse Screen 197
Appendix RIV – Targeted Ions 197
Bibliography.....202
Chapter I References 202
Chapter II References.....204
Chapter III References 208
Chapter IV References 212

List of Figures, Tables, and Equations

Figure 1_R.I Structures of known and probable sapogenin compounds in *M. truncatula*. 7
Figure 2_R.I. Labeled photographs of aerial organs of nine different ecotypes. Note the diversity of anthocyanin speckling. 14
Figure 3_R.I. Comparative base peak intensity chromatograms of *M. truncatula* ecotypes illustrating dramatic differential accumulation of triterpene saponins eluting in the 9 to 21 minutes region. 16
Table 1_R.I. Summary of descriptive statistics for the total saponin accumulation phenotypes for the aerial and root organs. 17
Figure 4_R.I Scatterplot of z scores for the total accumulation phenotypes of all ecotypes from both root (circles) and aerial (triangles) organ samples.. 19
Figure 5A_R.I shows total ion current chromatograms for aerial organ extracts of A17 and ESP_105 samples for both the initial profiling and the replant confirmation experiments. 22
Table 2_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of medicagenic acid in aerial organs along with the ranks for total saponin accumulation. 23
Table 3_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of soyasapogenol B and soyasapogenol E in root organs.. 23
Table 3_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of soyasapogenol B and soyasapogenol E in root organs along with the ranks for total saponin accumulation.....25
Table 4_R.I details the accumulation values for the top 5 ecotypes, in both root and aerial organs, for zanhic acid saponins and soyasapogenol B and E saponins. 27
Table 5_R.I details the accumulation values for the top 5 ecotypes, in both root and aerial organs, for saponins of bayogenin, hederagenin, and medicagenic acid. 28
Supplemental Figure 1_R.I A visualization of the metabolomics data analysis workflow employed in this project.. 34
Supplemental Figure 2_R.I Shows total ion current chromatograms for root organ extracts of A17 and GRC_105 samples for both the initial profiling and the replant confirmation experiments.....35
FIGURE 1_RII. Explanation of the Ecotype/Organ Experimental Matrix.....43
TABLE 1_RII. Cytochrome P450 and Glycosyltransferase Concatenated Annotation List Summary.....45

Equation 1_RII. Gross Phenotype Comparison Ranking Statistic "*f*".....71
Equation 2_RII. Inverse Case for Regulatory Element Probesets of the Gross Phenotype Comparison Ranking Statistic "*g*".....72
.....72
TABLE 2_RII. Top 15 Cytochrome P450 Probesets from the Gross Phenotype Comparison Ranking Process for the Inter-Genotype, Intra-Aerial-Organ Comparison. 46
TABLE 3_RII. Top 15 Glycosyltransferase Probesets from the Gross Phenotype Comparison Ranking Process for the Inter-Genotype, Intra-Root-Organ Comparison. 47
TABLE 4_RII. Pearson Correlation Coefficient Analysis of High Priority Cytochrome P450 Probesets. 49
FIGURE 2_RII. Graphical and Tabulated Summary of Results for the Selection of cyp72a68 as a High Priority Gene Candidate.....51
FIGURE 3_RII. Graphical and Tabulated Summary of Results for the Selection of cyp88d3 as a High Priority Gene Candidate. 52
FIGURE 4_RII. Ecotype Matrix Expression Dynamics for Known Triterpenoid Biosynthetic Pathway Genes Preceding Triterpene Sapogenin Bio-Oxidation.....54
Graphs showing the transcript expression dynamics in both organ types of all genotypes for squalene synthase, squalene epoxidase 1, squalene epoxidase 2, cycloartenol synthase, and β-amyrin synthase. Error bars represent 1 standard error. The squalene synthase accumulation data is from the microarray experiment. Data for the other genes is from qRT-PCR analysis of the same samples, as cycloartenol synthase and β-amyrin synthase are known to co-hybridize to the same microarray probesets (i. e. "shared probeset"). Similarly squalene epoxidase 1 and squalene epoxidase 2 co-hybridize with a number of probesets. 54
FIGURE 5_RII. Ecotype Matrix Expression Dynamics for Known Glycosyltransferases of the Triterpene Saponin Biosynthetic Pathway.....55
FIGURE 1_RIII. CYP72A67-Mediated Biosynthesis of 2-OH Oleanolic Acid from Oleanolic Acid.....92
FIGURE 2_RIII. CYP72A67-Mediated Biosynthesis of Bayogenin from Hederagenin. 93
TABLE 1_RIII. CYP72A67-Mediated Production and Consumption of Diverse OleanateSapogenins from the Aglycone Mixture.....95

FIGURE 3_RIII. CYP72A68-Mediated Biosynthesis of Hederagenin from Oleanolic Acid.....96
FIGURE 4_RIII. CYP72A68-Mediated Biosynthesis of Putative Gypsogenin and Putative Gypsogenic Acid from Hederagenin. 97
FIGURE 5_RIII. CYP72A68-Mediated Biosynthesis of Medicagenic Acid and Putative Polygalagenin from Bayogenin. 98
TABLE 2_RIII. Necessity of NADPH for CYP72A68 Catalytic Function. 99
TABLE 3_RIII. CYP72A68-Mediated Production and Consumption of Diverse OleanateSapogenins from the Aglycone Mixture.....100
FIGURE 6_RIII. Accumulation of an Unknown Compound in CYP72A68 Expanded Time Series Oleanolic Acid Substrate Assay with NADPH Regeneration System.....102
TABLE 4_RIII. *In Silico* and Reverse Genetic Screening Results for *Tnt-1* Insertion Mutants for All Candidate Loci.....105
FIGURE 7_RIII. Expression Dynamics for Transcripts of Candidate Genes in Diverse Plant Organs.....106
FIGURE 8_RIII. Summary Matrix of CYP72A67 and CYP72A68-Mediated Biosynthetic Reactions in the Oleanate Branch of the *M. truncatula* Sapogenin Biosynthesis Pathway.....108
FIGURE 1_RIV. - β-amyrin synthase and cyp88d1 Loci are Adjacent in the Genome of *M. truncatula*.....137
FIGURE 2_RIV. - β-amyrin synthase and cyp88d4 and cyp88d5 Loci are Adjacent in the Genome of Lotus japonicus.....138
FIGURE 3_RIV. - Correlation of Transcript and Total Saponin Metabolite Accumulation for cyp88d2 from Various Ecotypes. 140
FIGURE 4_RIV. - Correlation of Transcript and Total Saponin Metabolite Accumulation for cyp88d3 from Various Ecotypes. 141
FIGURE 5_RIV. - Expression values for Transcripts of cyp88d Genes in Diverse Plant Organs and Biological Treatments from the Medicago Gene Atlas. 143
TABLE 1_RIV. - Reverse Genetic Screening Results for *Tnt-1*Insertion Mutants for the cyp88d2 Locus. 144

Main Text

Chapter I - A Metabolomics-Based Platform for the Assessment of Triterpene Saponin Biochemical Variation in a *Medicago truncatula* Germplasm Diversity Collection

Authors: John H. Snyder, David V. Huhman, Stacey Allen,and Lloyd W. Sumner

Summary

This chapter details the metabolomics analyses of a large *Medicago truncatula* ecotype collection. Key results include the differences in observed spatial accumulation of both total and individual saponin classes and structures within the ecotype collection. Biochemical and ecological implications of the metabolomics profiling results are considered.

Abstract

The model legume *Medicago truncatula* is known to accumulate a large variety of triterpene saponin compounds, resulting from the differential glycosylation of at least six triterpeneaglycone structures. Previous chemical analyses (using FT-ICR-MS) analyses indicate that there may be several hundred saponin compounds in *Medicago*

sp. In this project, UPLC-ESI-qTOF-MS analysis was used to profile the accumulation of triterpene saponin metabolites in a collection of 110 *M. truncatula* ecotypes (germplasm accessions), which possess substantial metabolic diversity in saponin accumulation. Numerous accessions displayed highly differential total saponin accumulation in both aerial and root organs . Differential accumulation of specific saponin structures was also observed. Zanhic acid saponins were detected exclusively in aerial organs, while soyasapogenol B and soyasapogenol E saponins were detected exclusively in root organs. Additionally, medicagenic acid saponins were relatively more abundant in aerial tissues, while bayogenin saponins were more abundant in root tissue suggesting that the oxidation of carbon 23 of ß-amyrin from a hydroxyl (bayogenin) to a carboxylic acid (medicagenic acid) is more likely to occur in aerial tissues. The differential accumulation of saponins in the root and aerial tissues strongly suggests the presence of differentially regulated or biosynthetically distinct branches of the triterpene saponin pathway. Ecotypes of particular interest for subsequent molecular genetics analysis were identified as genetic resources and tested to ensure reproducibility of the observed biochemical phenotypes.

Abbreviations

FT-ICR-MS: Fourier Transform Ion Cyclotron Resonance Mass Spectrometry

UPLC-ESI-qTOF-MS: Ultra High Performance Liquid Chromatography Electrospray Ionization quadropole Time-of-Flight Mass spectrometry

HPLC-ESI-ion trap-MS: High Performance Liquid Chromatography

PCA: Principal Component Analysis

HCA: Hierarchical Cluster Analysis.

m/z: Mass to charge ratio

N.B. for JHS Dissertation

Development of the single seed descent lines was performed by Stacy Allen under the direction of Greg May and Lloyd Sumner several years prior to JHS arrival at The Noble Foundation. Growth, harvest, sample extraction/preparation, and instrumental analysis of the initial 110 ecotypes was performed by David Huhman prior to JHS arrival at Noble Foundation. Post –acquisition data analysis of raw data and statistical analyses for all samples was performed by JHS. Targeted ion list analyses (see Supplemental Figure 1_R.I) performed by JHS employed ion annotation information developed previously by David Huhman and others in the Sumner group. Subsequent regrowth, harvest, sample extraction/preparation, and post-acquisition analysis for confirmation of results and generation of additional sample materials of selected ecotypes was performed by JHS.

Introduction

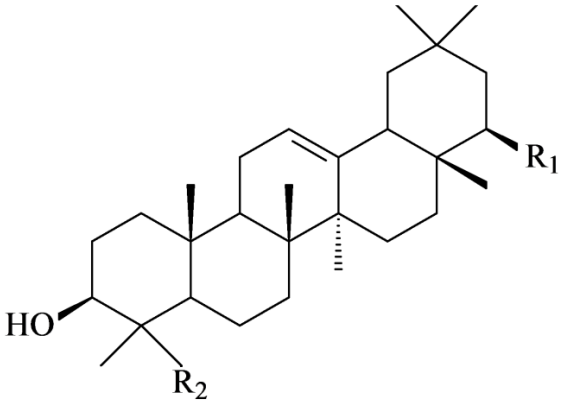
Triterpene saponins are a structurally diverse class of compounds with a wide taxonomic distribution. Although primarily found in Eudicots and especially legumes, triterpene saponins have also been isolated from selected Monocotyledonae plants such as oat and barley (Anne E. Osbourn, 2003; Papadopoulou *et al.*, 1999). Triterpene saponins possess a broad range of biological activities. Particular triterpene saponins have shown deleterious bioactivity against a broad spectrum of organisms such as lepidoptera larvae, aphids, gram-positive bacteria, yeasts, phytopathogenic fungi, human dermatophytic fungi, and other plants (Meesapyodsuk *et al.*, 2007; P. Houghton, 2006; Pedersen *et al.*, 1976). Saponins have also been recently reported to influence nodulation (Confalonieri *et al.*, 2009). Recent studies have reported beneficial pharmacological bioactivities of specific triterpene saponins including anti-inflammatory, anticancer (through induction of apoptosis following mitochondrial perturbation), and cholesterol-lowering properties (Haridas *et al.*, 2001; Kuljanabhagavad *et al.*, 2008). There has been further interest in using triterpene saponins as adjuvants for plant produced vaccines (Kirk *et al.*, 2004). While the above bioactivities are favorable for potential ecological, agrochemical, and pharmacological applications, triterpene saponins also represent the primary antinutritive compounds

in livestock fodder (Lu and Jorgensen, 1987; Lu *et al.*, 1987; Sen *et al.*, 1998). These antinutritive properties restrict the optimum utility of high-protein legumes as livestock feed and limit the ultimate economic potential of forage legumes. High concentrations of triterpene saponins in forages cause a serious reduction in ruminal and total tract forage digestibility through decreased ruminal protozoan populations and increased duodenal N, ultimately resulting in reduced weight gain (Dixon and Sumner, 2003; Lu and Jorgensen, 1987). A sophisticated molecular and biochemical understanding of saponin biosynthesis would enable the metabolic engineering of triterpenoid biosynthesis. For example, specific antimicrobial saponins could be engineered in roots to provide antimicrobial properties. Additionally, engineering a reduction of saponin content in aerial organs would improve nutritional content.

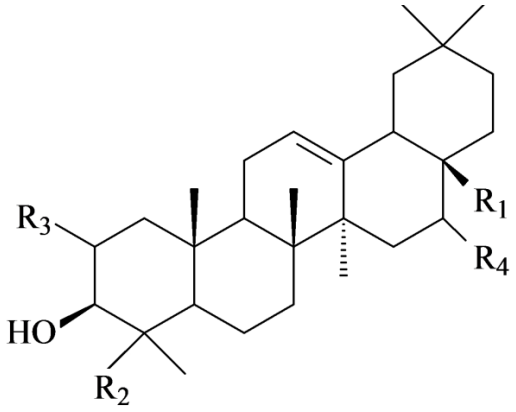
Structurally, triterpene saponins are composed of a lipid-soluble tritepenoid aglycone conjugated with various water soluble sugar residues. Sterol and triterpenoid sapogenin (saponin aglycones) biosyntheses in legumes begin with a common isopentenyl pyrophosphate (IPP) precursor synthesized via the cytosolic mevalonic acid (MVA) and/or plastid localized methylerythritol (MEP) pathways. The progressive condensation of isoprene units leads to various mono, sesqui, di, and triterpenoids. The triterpene oxidosqualene is cyclized by two enzymes resulting in

two branched pathways . Cycloartenol synthase is the first committed step in sterol biosynthesis, whereas ß-amryin synthase is the first committed step in triterpene saponin biosynthesis(Augustin *et al.*, 2011; Pollier *et al.*, 2011). Squalene synthase and squalene epoxidase have been previously characterized in *M. truncatula*(Iturbe-Ormaetxe *et al.*, 2003; Suzuki *et al.*, 2002).

The structural diversity of the triterpene sapogenins in legumes, including the model legume *Medicago truncatula*, has been an active area of phytochemical research for some time(Augustin *et al.*, 2011). Previous FT-ICR-MS studies (unpublished results) indicate that there may be several hundred diverse saponin compounds in *Medicago sp.*, but saponins of only six sapogenins are reported in the phytochemical literature for *M. truncatula* to date (Augustin *et al.*, 2011; Pollier *et al.*, 2011). A straightforward understanding of the diverse structures of the sapogenins of *M. truncatula* can be achieved by examining the sequential oxidation of six different carbons located within the ß-amarin structure. A relatively small number of carbon positions and various degrees of sequential oxidation at those carbon positions describe a very large diversity of chemical structures in this pathway as shown in Figure 1_R.I.



	R_1	R_2
β-amyrin	CH_3	CH_3
sophoradiol	CH_2 - OH	CH_3
olean-12-ene-3β,24-diol	CH_3	CH_2 - OH
soyasapogenol B	CH_2 - OH	CH_2 - OH
soyasapogenol E	CH = O	CH_2 - OH



	R_1	R_2	R_3	R_4
β-amyrin	CH_3	CH_3	H	H
erythrodiol	CH_2 - OH	CH_3	H	H
oleanolic aldehyde	CH = O	CH_3	H	H
oleanolic acid	OH-C = O	CH_3	H	H
2-hydroxyoleanloic acid	OH-C = O	CH_3	OH	H
hederagenin	OH-C = O	CH_2 - OH	H	H
bayogenin	OH-C = O	CH_2 - OH	OH	H
gypsogenin	OH-C = O	CH = O	H	H
polygalagenin	OH-C = O	CH = O	OH	H
gypsogenic acid	OH-C = O	OH-C = O	H	H
medicagenic acid	OH-C = O	OH-C = O	OH	H
zanhic acid	OH-C = O	OH-C = O	OH	OH

Figure 1_R.I Structures of known and probable sapogenin compounds in *M. truncatula*.

Recent progress in the development of robust metabolomics data acquisition and analysis methodologies has enabled novel experimental approaches for high resolution biochemical phenotyping in plant science (Bino *et al.*, 2004; Chan *et al.*, 2010). Several studies have used metabolomics methodologies for high resolution biochemical phenotyping of mutant collections and germplasm diversity panels for gene discovery or QTL analysis (Harjes *et al.*, 2008; Schilmiller *et al.*, 2010). Curated collections of germplasm accessions (ecotypes, natural genetic variants/mutants of a

particular species) are a powerful resource for exploring the natural variation for any number of phenotypes (Ronfort *et al.*, 2006). In this study, high resolution biochemical phenotyping was used to characterize the variation in triterpene saponin content within a large and diverse germplasm collection, i.e. 110 ecotypes. Characterization of the biochemical variation within the collection has enabled the informed selection/identification of particular ecotypes as genetic resources for subsequent molecular genetic analyses. Differences in triterpene saponin accumulation between the ecotypes can ostensibly be explained by either differential gene expression dynamics, or polymorphic alleles for genes that are involved in the biosynthesis, transport, sequestration, catabolism, signaling, or transcriptional activation (among others) for triterpene saponin biosynthesis. The degree of phenotypic differences among the natural variants such as ecotypes are extreme in some cases, but ecotypes are likely to have more complicated genotypic/phenotypic interactions than more traditional mutant collections derived from single genotypes or crosses of controlled and limited pedigree. Nevertheless, identification of ecotypes with extreme enrichments or deficiencies in triterpene saponin accumulation could therefore prove to be valuable as a form of "natural mutants" for studies into the molecular basis of triterpene saponin biosynthsis. This article details the metabolomics analyses of a large

Medicago truncatula ecotype collection. Key results include the differences in observed spatial accumulation of both total and individual saponin classes and structures within the ecotype collection. Biochemical and ecological implications of the metabolomics profiling results are considered.

Methods

Plant Materials

Seeds for the *Medicago truncatula* ecotype collection were obtained from Institut National de la Recherche Agronomique (INRA, http://www.montpellier.inra.fr/ INRA, Montpellier, France). Single seed descent lines for all of the INRA ecotypes were developed on site at The Noble Foundation prior to the plantings for the metabolomics profiling experiments described here.

Plant Growth Conditions

Plants were grown using the D40 H root cone system (Stuewe and Sons, http://www.stuewe.com, Tangent, OR), with Turface MVP medium (Profile Products, Buffalo Grove, IL), in a Conviron TCR180 walk-in growth chamber (http://www.conviron.com/, Winnipeg, Manitoba, Canada) maintained at 90% humidity and at an average temperature of 24 °C day (16 h) and 20 °C night (8 h). Plants were fertilized with 15 ppm nitrogen (Scotts' 20 10 20 Peat-Lite Special,

http://www.scotts.com, Marysville, Ohio) daily in the morning and watered with distilled water in the evening.

Metabolomics Analysis

Plants were harvested at 6 weeks post-germination and dissected into aerial and root organs. Dissected organs from two plants were combined as a single biological replicate, frozen immediately in liquid nitrogen, and lyophilized prior to metabolic profiling. In subsequent experiments, aerial tissues from the youngest 6 metamers (Bucciarelli *et al.*, 2006) of individual plants and whole root organ samples from individual plants were prepared as single biological replicates. Three aerial and root replicates were prepared for all of the ecotypes. Lyophilized tissues were ground into a fine powder using a mortar and pestle. 10.00 ± 0.06 mg of powder was extracted with 1 ml of 80% methanol (containing 0.018 mg/ml umbelliferone as an internal standard) in a dram vial for 2 hours on an orbital shaker. The samples and vials were centrifuged for 30 minutes at 2900g at 4°C. Supernatants were transferred to LC-MS autosampler vials (Agilent, http://www.agilent.com, Santa Clara, CA) and stored at -20°C until analyzed with a Waters Acquity UPLC system coupled to a hybrid quadrupole time-of-flight (QTOF) mass spectrometer (Waters QTofMS Premiere, http://www.waters.com/, Milford, MA). A Waters reverse-phase UPLC BEH, C18,

2.1 x 150 mm column with 1.7-µm particles was used for separations. The mobile phases consisted of eluent A (0.1% [v/v] acetic acid/HPLC grade water) and eluent B (HPLC-grade acetonitrile), and separations were achieved using a linear gradient of 95% to 30% A over 30 min, The flow rate was 0.56 mL/min, and the column temperature was maintained at 60°C. Separated compounds were detected in the negative ESI mode from 50 to 2,000 mass-to-charge ratio. The QTOF Premier mass spectrometer was operated using the following instrumental parameters: desolvation temperature of 400°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.

Targeted ion list and Data Processing

Raw data files were converted to .cdf file format, followed by metabolite data extraction, alignment, and export using MET-IDEA software (Broeckling *et al.*, 2006a). An ion list containing 153 retention time/ion pairs was used for the saponin-targeted metabolomics data analysis of the Ecotype UPLC-ESI(-)-qTOF-MS saponin biochemical phenotypes (Appendix_RI_ion_list). Seventeen of these pairs were determined using validated authentic standards (e.g. 3-Glc-28-Glc-Medicagenic Acid standard), 53 of these pairs were tentatively identified based upon spectral information

(source fragmentation and MS/MS in some cases) as an Aglycone and some combination of sugars(e.g. Hex-Rha-Hex-Hex-Hederagenin), 28 of these pairs have minimal annotation based spectral features resulting from probable source fragmentation (e.g. possibly bayogenin, GlcGlc?), and the remainder are unknowns. The unknown pairs in the ion list were identified with non-targeted MARKERLYNX analysis, and had m/z values and retention times in the same regions as the known and putative pairs and additionally showed statistically significant differential accumulation values among the ecotypes in the collection. In addition to the targeted analysis of saponin content, non-targeted analyses of all samples were performed using Waters MARKERLYNX software. The spectral abundance signals for all metabolites in a separation were normalized to the internal standard (0.018 mg/ml umbelliferone). Descriptive statistics were performed in MS Excel. One-way ANOVA was performed using a custom MATLAB script. Multivariate analyses including principal component analysis (with standardized, z-score values for peak area) and hierarchical clustering were performed using JMP 5.0 software (SAS, http://www.sas.com/). Supplemental Figure 1_R.I presents a visualization of the metabolomics data analysis workflow employed in this project.

Results and Discussion

Morphological Diversity

During the propagation and development of the single seed descent lines of the INRA ecotype collection, an obvious diversity of aerial and reproductive organ morphology among the ecotypes was noted. Photos of all of the ecotypes in the collection can be found at: (http://www.noble.org/medicago/ecotypes.html). The various ecotypes also showed obvious biochemical diversity in the form of varied anthocyanin speckling (Figure 2_R.I) on leaves.

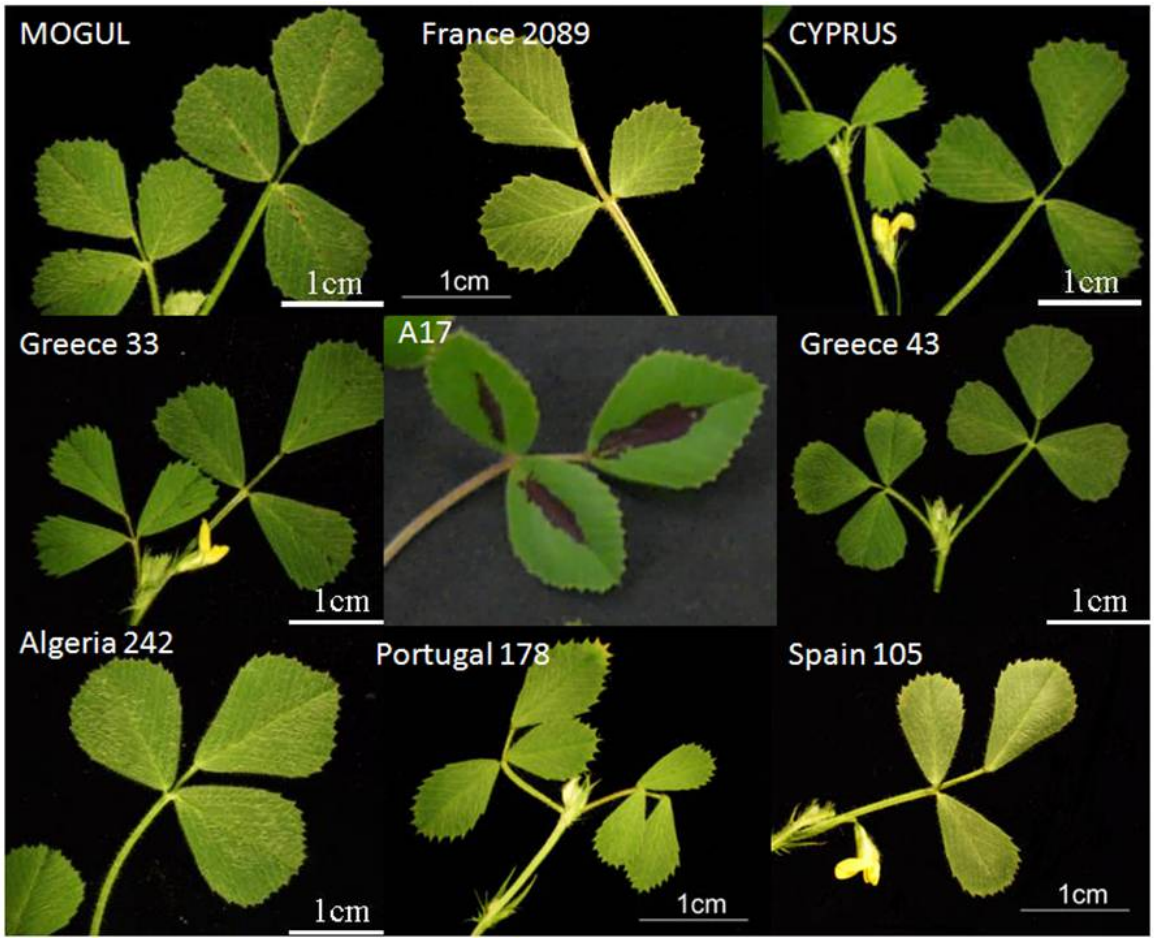


Figure 2_R.I. Labeled photographs of aerial organs of nine different ecotypes. Note the diversity of anthocyanin speckling.

Metabolic profiling

Metabolic profiling was used to determine the saponin content within the 110 *M. truncatula* accessions obtained from INRA. Analyses were performed using a UPLC-ESI-qTOF-MS platform and representative base peak intensity (BPI) chromatograms are presented in Figure 3. Single factor ANOVA of the accumulation of the

normalized peak areas of various ions revealed that 145/153 detected ions were significantly differentially accumulated ($p<0.05$) in root organs among the ecotypes, and that 144/153 detected ions were significantly differentially accumulated ($p<0.05$) in the aerial organ data set (Appendix_RI_ion_list). The metabolic profiling results will be considered at two level of phytochemical resolution: 1.) differences in total saponin accumulation, and 2.) diveristy of saponin structure, as observed between the two organs types and as observed within a particular organ among the various ecotypes.

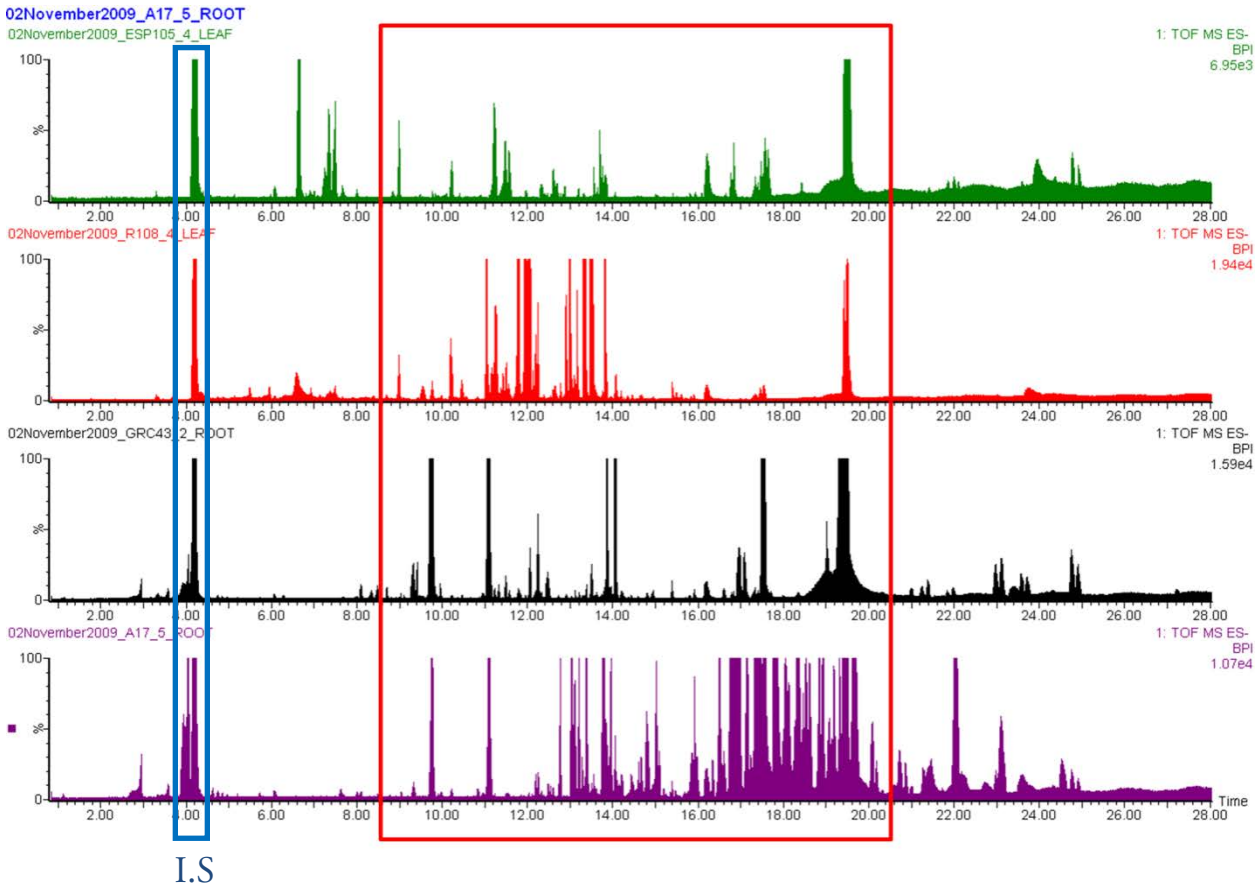


Figure 3_R.I. Comparative base peak intensity chromatograms of *M. truncatula* ecotypes illustrating dramatic differential accumulation of triterpene saponins eluting in the 9 to 21 minutes region (red highlighted box). I.S. indicates umbelliferone internal standard.

Total Saponin Accumulation

Total saponin accumulation phenotype values were obtained for each of the ecotypes by summing the individual accumulation values for each of ion/R_t pairs (Appendix_RI_ion_list). Table 1_R.I presents a summary of descriptive statistics for the total saponin accumulation phenotypes for the aerial and root organs.

A.

Statistic	Normalized Peak Area
Aerial_Mean	9049
Aerial_Standard Deviation	5397
Aerial_Max	32538
Aerial_Min	346
Aerial_Range	32192

B.

Statistic	Normalized Peak Area
Root_Mean	29787
Root_Standard Deviation	9329
Root_Max	52032
Root_Min	7136
Root_Range	44896

C.

	% of ecotypes
Aerial within 1 st dev	69
Aerial within 2 st dev	95
Aerial within 3 st dev	99
Aerial within 4 st dev	100

D.

	% of ecotypes
Root within 1 st dev	60
Root within 2 st dev	96
Root within 3 st dev	100

Table 1_R.I. Summary of descriptive statistics for the total saponin accumulation phenotypes for the aerial and root organs.

Root organs accumulated higher quantities of total saponins than aerial organs. This is consistent with a previous study (Huhman *et al.*, 2005), but the ratio of root to aerial saponins (3.3) is lower than the comparisons reported for root to leaf (approximately 5) and root to stem (approximately 10). This difference may be due to several factors,

including but not limited to the increased depth of coverage in the analytical platform (HPLC-ESI-ion trap-MS compared to UPLC-ESI-qTOF-MS) and additional structural annotations available in the metabolomics informatics workflow, or simply as a consequence of the increased biological variation in the saponin accumulation represented by the much larger (one ecotype vs. 110) representation of germplasm diversity in this study. Indeed, the desire to examine the likely increased biological variation for both total saponin accumulation (as well as variation in saponin structural diversity) within the large ecotype collection was the primary motivation for this study. The variation of total saponin accumulation between the various ecotypes is presented in Figure 4.

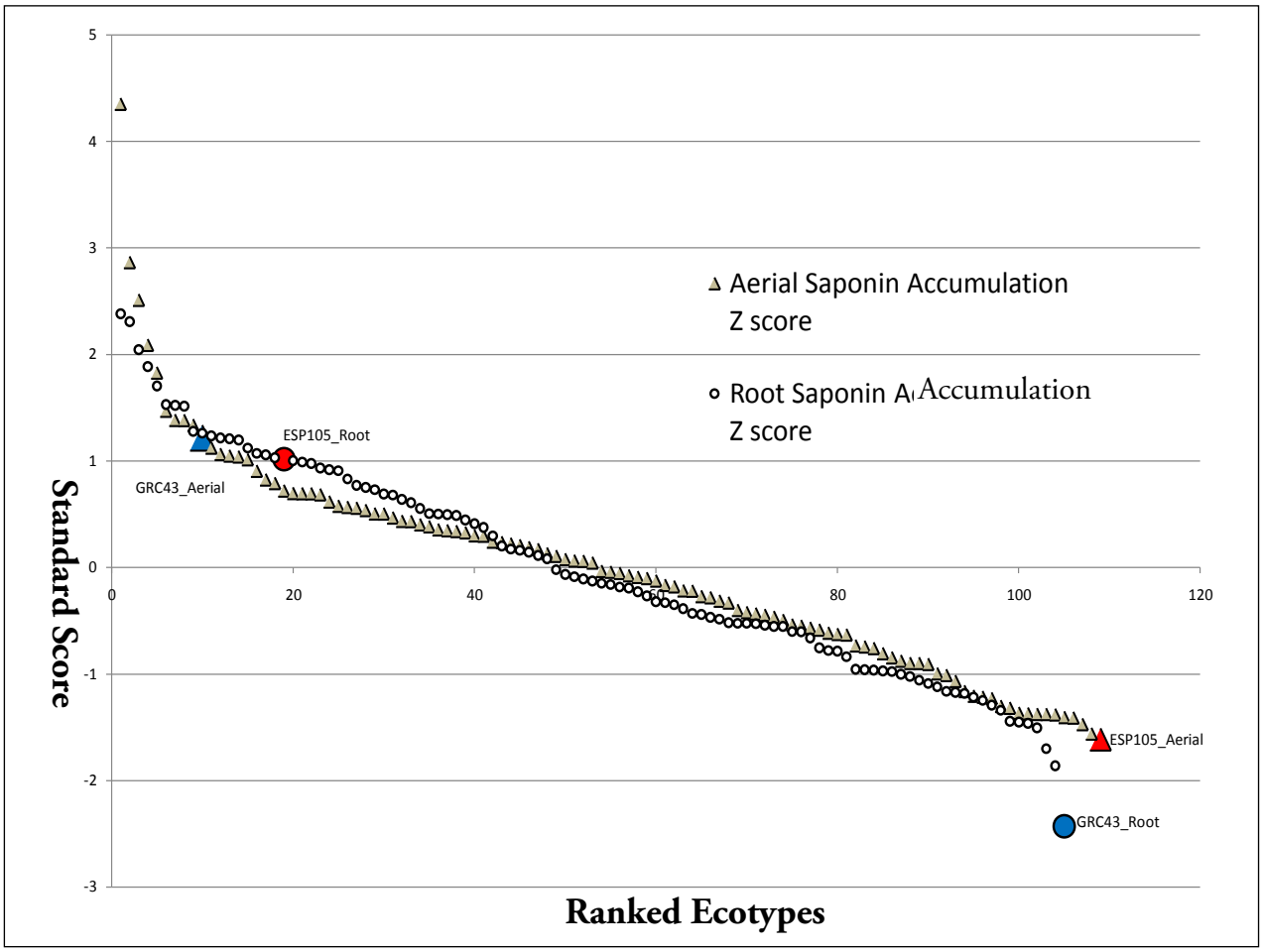


Figure 4_R.I Scatterplot of z scores for the total accumulation phenotypes of all ecotypes from both root (circles) and aerial (triangles) organ samples. ESP_105 samples are labeled and shown enlarged and in red color. GRC_43 samples are labeled and shown enlarged and in blue color.

Outliers in total saponin accumulation values were more pronounced in aerial organs than in root organs as evidenced by the magnitude of the range in the z-score distribution. Perhaps the most interesting aspect of the total saponin accumulation analysis was the observation that a low accumulator in aerial organs may be a high accumulator in root organs and vice versa. For example, the ecotype ESP_105 is the

lowest total saponin accumulator in aerial organs but a very high (top 20) total accumulator in root organs. Likewise, the ecotype GRC_43 is the lowest total accumulator in root organs but a very high (top 10) total accumulator in aerial organs.

Specific ecotypes were identified as genetic resources for potential use in subsequent molecular genetics analyses based on UPLC-ESI(-)-qTOF-MS profiling and metabolomics data analyses. Ecotypes that demonstrated extremes of either hypo- or hyper-total saponin accumulation were considered to be of primary importance. Ecotypes with differential accumulation of saponins of particular sapogenin structures were not prioritized for immediate exploitation. As highlighted previously, a low accumulator in aerial organs may be a high accumulator in root organs and vice versa. This observation was exploited in the experimental design of the molecular genetics analyses that eventually followed from the metabolomics profiling (see JHS_Research_Chapter_II). The ecotype ESP_105 was selected as the lowest total saponin accumulator in aerial organs, but it had potential additional value as a resource because it was also very high (top 20) total accumulator in root organs. Likewise, the ecotype GRC_43 was selected as the lowest total accumulator in root organs but was also an extremely high (top 10) total accumulator in aerial organs. ESP_104 was selected as an additional low total saponin accumulator for aerial organs,

and DZA_242 was selected as an additional low total saponin accumulator for root organs. The popular isolines A17 and R108 were selected as reference ecotypes with relatively high total saponin accumulation in both aerial and root organs, due primarily to their role in the development of genomics and mutant population resources for research in *M. truncatula*. Reproducibility of highly-dimensional metabolomics phenotypes is often difficult to achieve. In order to assure that the saponin phenotypes observed in the initial profiling experiment were reproducible, the selected ecotypes were re-grown, harvested, extracted/processed, and analyzed in the same manner as the initial profiling experiment . Figure 5A_R.I and Supplemental Figure 2_R.I indicate that the low total saponin accumulation phenotypes for ESP_105 aerial organs and GRC_43 root organs compared to A17 are broadly reproducible between experiments, thereby enabling reliable generation of plant material for subsequent phytochemical and molecular genetic studies.

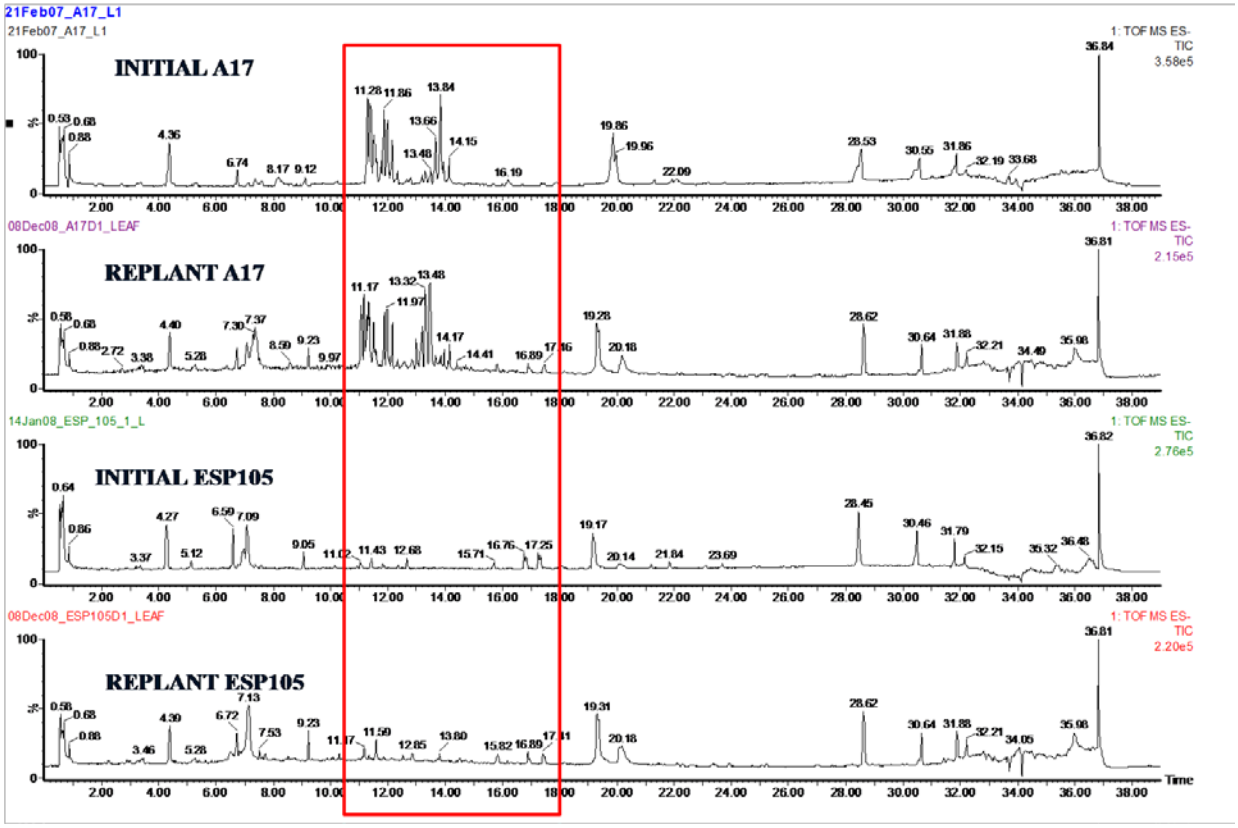


Figure 5A_R.I shows total ion current chromatograms for aerial organ extracts of A17 and ESP_105 samples for both the initial profiling and the replant confirmation experiments.

Diversity of Saponin Structure Observed Between the Two Organ Types

3.) Diversity of saponin structures observed within a particular organ among the various ecotypes.

TOP TEN

Ecotype	Normalized Peak Area	Rank (medicagenic acid)	Rank (total)
F20_89_aerial	3881.7	1	2
GRC_33_aerial	3351.7	2	3
ESP_48_aerial	3232.5	3	24
ESP_50_aerial	2916.8	4	14
DZA_55_aerial	2548.8	5	35
DZA_246_aerial	2484.7	6	11
F20_81_aerial	2403.9	7	19
DZA_45_aerial	2349.0	8	13
F20_58_aerial	2297.3	9	59
DZA_59_aerial	2208.6	10	16

BOTTOM TEN

Ecotype	Normalized Peak Area	Rank (medicagenic acid)	Rank (total)
ESP_162_aerial	47.2	99	106
CALIPH__aerial	46.8	100	78
DZA_58_aerial	45.6	101	104
DZA_61_aerial	42.4	102	94
F34_42_aerial	41.3	103	20
DZA_46_aerial	36.8	104	92
ESP_140_aerial	22.6	105	101
MOGUL__aerial	20.9	106	80
ESP_105_aerial	18.6	107	109
F11_7_aerial	13.0	108	77
HARBINGER__aerial	5.9	109	105

Table 2_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of medicagenic acid in aerial organs along with the ranks for total saponin accumulation.

Table 3_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of soyasapogenol B and soyasapogenol E in root organs. It is worth noting that root organs of the ecotype DZA_46 rank 6th overall for saponins of soyasapogenols B and E but rank 50th in terms of

total saponin accumulation. Root organs of the ecotype GRC_65 rank 102nd for accumulation for saponins of soyasapogenols B and E, while ranking 13th in total saponin accumulation.

TOP TEN

Ecotype	Normalized Peak Area	Rank (combined soy-E & soy-B)	Rank (total)
ESP_105_root	4456.8	1	18
DZA_45_root	4370.0	2	41
ESP_39_root	4209.0	3	17
ESP_104_root	4153.9	4	9
ESP_155_root	3916.2	5	14
DZA_46_root	3894.1	6	50
ESP_96_root	3794.4	7	2
ESP_165_root	3670.6	8	1
ESP_50_root	3592.6	9	22
ESP_74_root	3376.4	10	16

BOTTOM TEN

Ecotype	Normalized Peak Area	Rank (combined soy-E & soy-B)	Rank (total)
GRC_64_root	1383.0	95	98
SALERNES__root	1330.1	96	88
CRE_9_root	1288.9	97	65
DZA_309_root	1272.0	98	96
DZA_242_root	1207.1	99	104
ESP_161_root	1139.9	100	103
CRE_5_root	1092.4	101	90
GRC_65_root	939.9	102	13
HARBINGER__root	926.4	103	31
DZA_221_root	273.4	104	89
GRC_043B_root	116.5	105	105

Table 3_R.I presents the top ten and bottom ten accumulator ecotypes for saponins of soyasapogenol B and soyasapogenol E in root organs along with the ranks for total saponin accumulation.

Zanhic Acid Saponins	Normalized Peak Area
HARBINGER_Root	25
DZA_055_Root	9
DZA_327_Root	9
PRT_180_Root	9
DZA_323_Root	9
F11_012_Aerial	6915
CALIPH_Aerial	3393
F11_007_Aerial	3241
PRT_176_Aerial	3226
SALERNES_Aerial	2600

Soyasapogenol B Saponins	Normalized Peak Area
ESP_104_Root	1919
ESP_165_Root	1821
ESP_162_Root	1547
ESP_171_Root	1087
GRC_052_Root	1001
F11_007_Aerial	29
CRE_009_Aerial	17
MOGUL_Aerial	14
DZA_033_Aerial	12
F11_012_Aerial	12

Soyasapogenol E Saponins	Normalized Peak Area
ESP_105_Root	3916
ESP_039_Root	3731
DZA_046_Root	3721
DZA_045_Root	3594
ESP_155_Root	3351
ESP_096_Aerial	19
ESP_031_Aerial	13
ESP_098A_Aerial	12
ESP_074_Aerial	10
ESP_040_Aerial	9

Table 4_R.I details the accumulation values for the top 5 ecotypes, in both root and aerial organs, for zanhic acid saponins and soyasapogenol B and E saponins.

Bayogenin Saponins	**Normalized Peak Area**
ESP_155_Root	15990
ESP_096_Root	14777
ESP_159_Root	14705
ESP_100_Root	14488
DZA_323_Root	14486
F11_012_Aerial	162
PRT_178_Aerial	148
PRT_176_Aerial	121
PRT_179_Aerial	116
CRE_009_Aerial	89

Hederagenin Saponins	**Normalized Peak Area**
MOGUL_Root	16079
ESP_162_Root	15639
F20_025_Root	14731
ESP_105_Root	12666
DZA_105_Root	12275
MOGUL_Aerial	1648
DZA_241_Aerial	1157
JEMALONG_3_Aerial	1058
A17_Aerial	1032
CALIPH_Aerial	800

Medicagenic Acid Saponins	**Normalized Peak Area**
ESP_041_Root	19776
ESP_165_Root	18597
ESP_074_Root	16373
JEMALONG_Root	16315
ESP_096_Root	14473
F20_089_Aerial	3882
GRC_033_Aerial	3353
ESP_048_Aerial	3233

ESP_050_Aerial	2919
DZA_055_Aerial	2549

Table 5_R.I details the accumulation values for the top 5 ecotypes, in both root and aerial organs, for saponins of bayogenin, hederagenin, and medicagenic acid.

Additional Information

Appendices

Appendix_RI_ion_list

Supplemental Figure 1_R.I

Supplemental Figure 2_R.I

Sources

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin, (2003) Dissecting plant secondary metabolism; constitutive chemical defences in cereals. *New Phytologist* 159, 101-108.

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

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S. and Sumner, L. W. (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9, 418-425.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X. and Sumner, L. W. (2006a) MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. *Anal. Chem.* 78, 4334-4341.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X. and Sumner, L. W. (2006b) MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. *Analytical Chemistry* 78, 4334-4341.

Bucciarelli, B., Hanan, J., Palmquist, D. and Vance, C. P. (2006) A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development 10. 1104/pp. 106. 082594. *Plant Physiol.* 142, 207-219.

Chan, E. K. F., Rowe, H. C., Hansen, B. G. and Kliebenstein, D. J. (2010) The Complex Genetic Architecture of the Metabolome. *PLoS Genet* 6, e1001198.

Confalonieri, M., Cammareri, M., Biazzi, E., Pecchia, P., Fevereiro, M. P. S., Balestrazzi, A., Tava, A. and Conicella, C. (2009) Enhanced triterpene saponin biosynthesis and root nodulation in transgenic barrel medic (*Medicago truncatula* Gaertn.) expressing a novel β-amyrin synthase (AsOXA1) gene. *Plant Biotechnology Journal* 7, 172-182.

Delis, C., Krokida, A., Georgiou, S., Peña-Rodríguez, L. M., Kavroulakis, N., Ioannou, E., Roussis, V., Osbourn, A. E. and Papadopoulou, K. K. Role of

lupeol synthase in Lotus japonicus nodule formation. *New Phytologist* 189, 335-346.

Dixon, R. A. and Sumner, L. W. (2003) Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. *Plant Physiology* 131, 878-885.

Haridas, V., Higuchi, M., Jayatilake, G. S., Bailey, D., Mujoo, K., Blake, M. E., Arntzen, C. J. and Gutterman, J. U. (2001) Avicins: Triterpenoid saponins from Acacia victoriae (Bentham) induce apoptosis by mitochondrial perturbation

10. 1073/pnas. 101619098. *Proceedings of the National Academy of Sciences* 98, 5821-5826.

Harjes, C. E., Rocheford, T. R., Bai, L., Brutnell, T. P., Kandianis, C. B., Sowinski, S. G., Stapleton, A. E., Vallabhaneni, R., Williams, M., Wurtzel, E. T., Yan, J. and Buckler, E. S. (2008) Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification. *Science* 319, 330-333.

Huhman, D. V., Berhow, M. A. and Sumner, L. W. (2005) Quantification of Saponins in Aerial and Subterranean Tissues of *Medicago truncatula*. *Journal of Agricultural and Food Chemistry* 53, 1914-1920.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K. and Osbourn, A. E. (2003) Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. *Plant Molecular Biology* 51, 731-743.

Kirk, D. D., Rempel, R., Pinkhasov, J. and Walmsley, A. M. (2004) Application of Quillaja saponaria extracts as oral adjuvants for plant-made vaccines doi:10. 1517/14712598. 4. 6. 947. *Expert Opinion on Biological Therapy* 4, 947-958.

Kuljanabhagavad, T., Thongphasuk, P., Chamulitrat, W. and Wink, M. (2008) Triterpene saponins from Chenopodium quinoa Willd. *Phytochemistry* 69, 1919-1926.

Lu, C. D. and Jorgensen, N. A. (1987) Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. *The Journal of Nutrition* 117, 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M. and Jorgensen, N. A. (1987) Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. *Journal of dairy science* 70, 799-805.

Meesapyodsuk, D., Balsevich, J., Reed, D. W. and Covello, P. S. (2007) Saponin Biosynthesis in Saponaria vaccaria. cDNAs Encoding beta-Amyrin Synthase and a Triterpene Carboxylic Acid Glucosyltransferase 10. 1104/pp. 106. 088484. *Plant Physiol.* 143, 959-969.

Nielsen, J., Nagao, T., Okabe, H. and Shinoda, T. (2010) Resistance in the Plant Barbarea vulgaris and Counter-Adaptations in Flea Beetles Mediated by Saponins. *Journal of Chemical Ecology* 36, 277-285.

P. Houghton, N. P., M. Jurzysta, Z. Biely, C. Cheung, (2006) Antidermatophyte activity of medicago extracts and contained saponins and their structure-activity relationships. *Phytotherapy Research* 20, 1061-1066.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. and Osbourn, A. E. (1999) Compromised disease resistance in saponin-deficient plants *Proceedings of the National Academy of Sciences of the United States of America* 96, 12923-12928.

Pedersen, M. W., Barnes, D. K., Sorensen, E. L., Griffin, G. D., Nielson, M. W., Hill, R. R., Jr., Frosheiser, F. I., Sonoda, R. M., Hanson, C. H., Hunt, O. J., Peaden, R. N., Elgin, J. H., Jr., Devine, T. E., Anderson, M. J., Goplen, B. P., Elling, L. J. and Howarth, R. E. (1976) Effects of Low and High Saponin Selection in Alfalfa on Agronomic and Pest Resistance Traits and the Interrelationship of these Traits. *Crop Sci* 16, 193-199.

Ronfort, J., Bataillon, T., Santoni, S., Delalande, M., David, J. and Prosperi, J. -M. (2006) Microsatellite diversity and broad scale geographic structure in a model

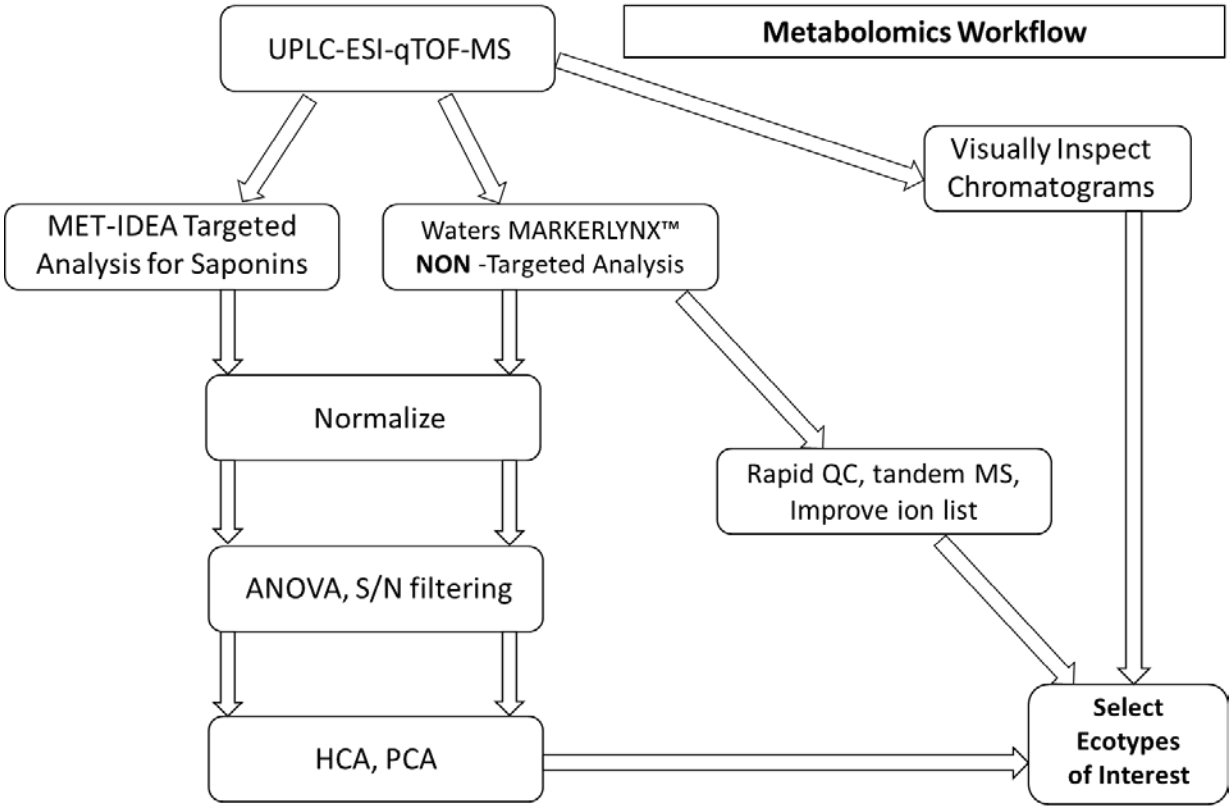
legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*. *BMC Plant Biology* 6, 28.

Schilmiller, A., Shi, F., Kim, J., Charbonneau, A. L., Holmes, D., Daniel Jones, A. and Last, R. L. (2010) Mass spectrometry screening reveals widespread diversity in trichome specialized metabolites of tomato chromosomal substitution lines. *The Plant Journal* 62, 391-403.

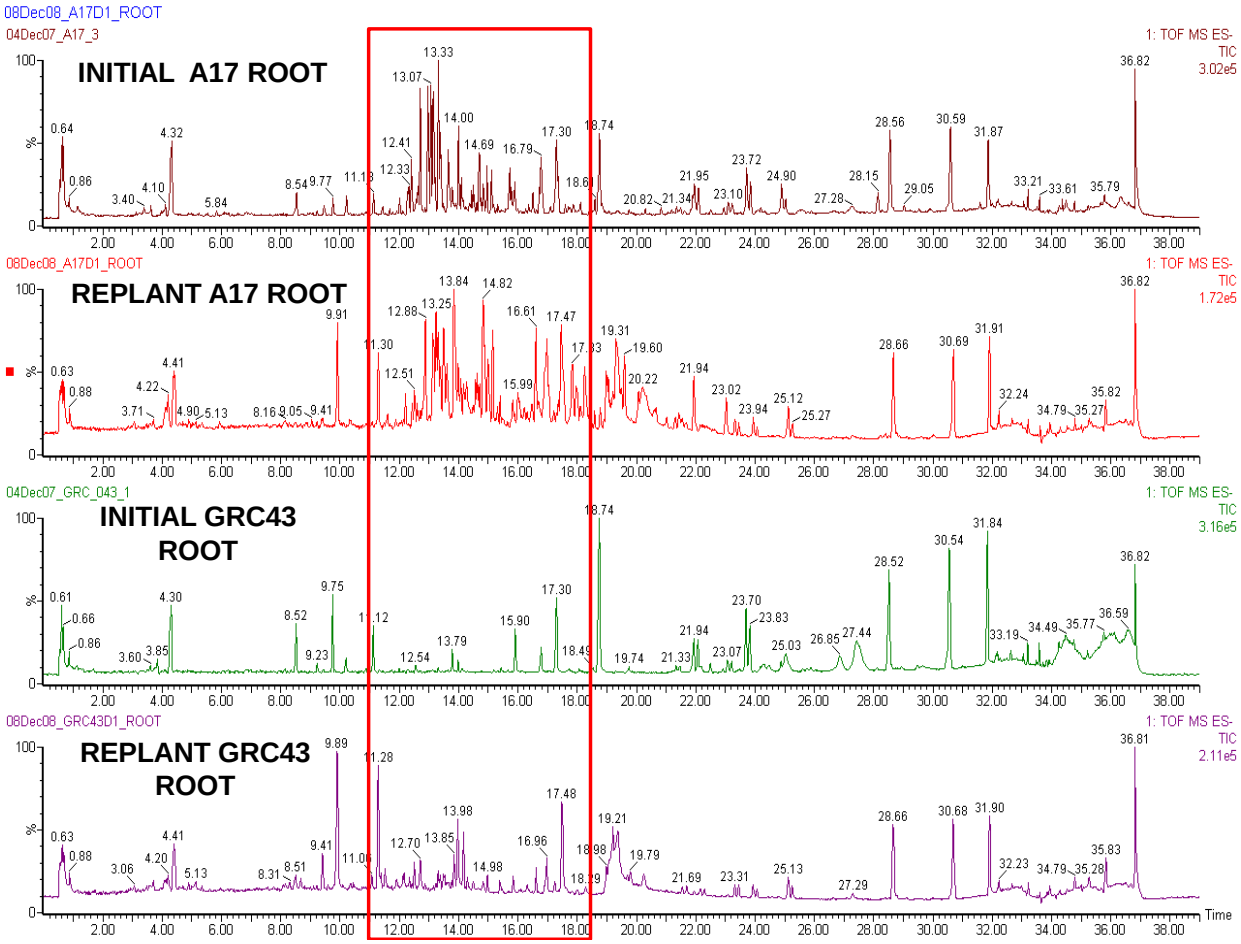
Sen, S., Makkar, H. P. S. and Becker, K. (1998) Alfalfa Saponins and Their Implication in Animal Nutrition. *Journal of Agricultural and Food Chemistry* 46, 131-140.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T. and Dixon, R. A. (2002) A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. *The Plant Journal* 32, 1033-1048.



Supplemental Figure 1_R.I A visualization of the metabolomics data analysis workflow employed in this project. PCA: Principal Component Analysis, HCA: Hierarchical Cluster Analysis.



Supplemental Figure 2_R.I Shows total ion current chromatograms for root organ extracts of A17 and GRC_105 samples for both the initial profiling and the replant confirmation experiments.

Chapter II - Identification of Candidate Biosynthetic Genes in Triterpene Saponin Metabolism Through Integrated Analysis of Metabolome And Transcriptome Datasets from *Medicago truncatula* Ecotypes with Differential Triterpene Saponin Accumulation Phenotypes

Authors: John H. Snyder, David V. Huhman, Yuhong Tang, and Lloyd W. Sumner

Summary

This chapter will detail the large-scale transcriptomics experiment performed with the ecotypes/organs selected in JHS_Research_CHAPTER_I. The focus of this chapter is the process of identification of candidate genes (particularly cytochrome P450, glycosyltransferase, and regulatory element genes) for triterpene saponin biosynthesis. This process is accomplished through the integrated analysis of these separate "omics" datasets through two approaches (gross phenotype comparisons and Pearson correlation coefficient analysis). Additionally, this chapter contains the results of a *de novo* profile hidden Markov model annotation of the tentative consensus sequences used to design the probesets of the Affymetrix Medicago Gene Chip

Abstract

Based on the metabolomics profiling of a large germplasm diversity collection, an experimental matrix of grouped hypo- and hyper-triterpene saponin accumulating ecotypes and organs were selected for transcriptomics analysis. The dramatic differences in total saponin accumulation between the various ecotypes and organs selected for the matrix enabled several approaches for the integration of the metabolomics and transcriptomics datasets for the identification of likely candidate genes for the as yet uncharacterized biosynthetic steps of the triterpene saponin pathway of *M. truncatula*. Identification of likely regulatory element candidate genes which may control saponin biosynthesis was also a goal of the study. A predictive model was developed for the ranking of transcripts which matched the gross saponin accumulation phenotypes from the various ecotypes and organs from the experimental matrix. Pearson correlation coefficient analysis was also performed for a large number of [transcript] vs. [metabolite] and [ecotype] vs. [organ] permutations from the experimental matrix. As the annotations for the tentative consensus sequences used to design the probesets of the Affymetrix Medicago Gene Chip™ showed poor representation for known cytochrome P450 and glycosyltransferase genes, a *de novo* profile hidden Markov

model (HMM) protein domain annotation was performed. This annotation enabled more comprehensive analysis of the transcripts from these protein families. Results from the various integrated analyses of transcriptomics and metabolomics datasets motivated the selection of five cytochrome P450 genes (cyp72a68, cyp72a67, cyp716a12, cyp83g1, and cyp88d3) as likely candidates involved in the bio-oxidation of triterpene sapogenins in the triterpene saponin biosynthesis pathway of *M. truncatula*.

Glossary

Hidden Markov Model (HMM): A statistical model used for analyzing unknown (but probabilistically defined) sequences when useful outputs from those sequences are available.

Probeset: A term for a proprietary technology consisting of "a collection of probes designed to interrogate a given sequence" (Affymetrix, http://www.affymetrix.com/support/help/faqs/mouse_430/faq_8.jsp).

Introduction

Metabolomics methodologies enable researchers to obtain extremely high resolution biochemical phenotypic datafor biological samples (Fiehn 2002). Metabolomics is

now established as an important tool in broader functional genomics (Bino, Hall et al. 2004), and has become a critical component of Systems Biology (Sulpice, Trenkamp et al. in press). Metabolomics biochemical phenotypic data can be seen as a particularly useful in plant functional genomics, as the largely uncharacterized genes/enzymes for the synthesis, modification, degradation and/or transport of exotic metabolites ultimately yield the awesome phytochemical diversity observed in plants (Dixon and Sumner 2003). A number of studies in plants have explored the associations between transcriptomics and metabolomics datasets. Many of these studies have used a single reference genotype, and focused on developmental stages and spatially-resolved tissue types (Krueger, Giavalisco et al. 2011 ; Matsuda, Hirai et al. 2010) or discreet perturbations such as diverse growth conditions (e. g. temperature, day length) or nutritional status (Hirai, Yano et al. 2004; Hannah, Caldana et al. 2010). Other integrated studies have focused on the differential transcript and metabolite accumulation dynamics between transgenic and wild-type plants (Tohge, Nishiyama et al. 2005). Studies which integrate transcriptomics and metabolomics datasets among germplasm diversity panels ("natural mutant collections") have become a more recent focus for plant metabolomics (Tohge and Fernie 2010). These germplasm diversity based studies have focused their integrated transcript and

metabolite models on “major”(i. e. not expressly “biochemical”) phenotypes, such as total biomass (Sulpice, Trenkamp et al. in press), or tomato fruit color (Ballester, Molthoff et al. 2011).

Triterpene saponins are a structurally diverse class of compounds with a wide taxonomic distribution and a broad range of biological activities (Augustin, Kuzina et al. 2011). Although primarily found in dicots and especially legumes, triterpene saponins have also been isolated from selected monocots such as oat and barley (Papadopoulou, Melton et al. 1999; Anne E. Osbourn 2003). Triterpene saponins represent the primary antinutritive compounds in livestock fodder (Lu and Jorgensen 1987; Lu, Tsai et al. 1987). Structurally, triterpene saponins are composed of a lipid-soluble triterpenoid aglycone conjugated with various water soluble sugar residues. Sterol and triterpenoid sapogenin (saponin aglycones) biosyntheses in legumes begin with a common isopentenyl pyrophosphate (IPP) precursor synthesized via the cytosolic mevalonic acid (MVA) and/or plastid localized methylerythritol (MEP) pathways. The progressive condensation of isoprene units leads to various mono, sesqui, di, and triterpenoids. The triterpene oxidosqualene is cyclized by two enzymes resulting in two branched pathways. Cycloartenol synthase is the first committed step in sterol biosynthesis, whereas ß-amryin synthase is the first committed step in

triterpene saponin biosynthesis (Augustin, Kuzina et al. 2011). Squalene synthase and squalene epoxidase have been previously characterized in *M. truncatula* (Suzuki, Achnine et al. 2002; Iturbe-Ormaetxe, Haralampidis et al. 2003). Very little is known about the remaining enzymatic (bio-oxidation by cytochrome P450 enzymes and gylcosylation) steps following ß-amyrin synthase in triterpene saponin biosynthesis, although UGT73K1, UGT71G1, UGT73F3 have recently been characterized as GTs in the triterpene saponin biosynthesis pathway of *M. truncatula*. (Lahoucine Achnine 2005; Naoumkina, Modolo et al. 2010). None of the enzymes which catalyze the bio-oxidation of ß-amyrin in *M. truncatula* have been characterized, and the mechanisms of additional glycosylation steps remain uncharacterized.

It has been observed that the 'guilt by association' phenomenon (co-accumulation dynamics for genes which are co-regulated and thus co-expressed under the control of a shared regulatory system)is particularly pronounced in the case of plant secondary metabolism(Saito, Hirai et al. 2008). The aim of this study was to exploit this strong 'guilt by association' phenomenon by performing transcriptomics analysis of a matrix (Figure 1_RII) of germplasm accessions (ecotypes) with extreme phenotypes for differential triterpene saponin accumulation. In combination with the extreme saponin accumulation phenotypes, the transcript expression data could

potentially identify genes of unknown function that had strong 'guilt by association' with the saponin biosynthetic pathway. As no enzymes which catalyze the bio-oxidation (putatively cytochrome P450s) have been characterized in the saponin biosynthesis pathway of *M. truncatula*, particular emphasis was placed on 'guilt by association' relationships between the phenotypes and expression dynamics for unknown cytochrome P450 genes.

Metabolomics analysis of a large germplasm diversity (ecotype) collection revealed substantial metabolic diversity in triterpene saponin accumulation both within the various ecotypes, and between the root and aerial organs of individual ecotypes (JHS_RESEARCH_CHAPTER_I). The metabolomics phenotyping results enabled the informed selection of specific ecotypes for an experimental matrix of hypo- and hyper- saponin accumulating ecotypes and organs (Figure 1_RII). The ecotype ESP_105 was selected as the lowest total saponin accumulator in aerial organs, but it had potential additional value as a resource because it was also very high (top 20) total accumulator in root organs. Likewise, the ecotype GRC_43 was selected as the lowest total accumulator in root organs but was also an extremely high (top 10) total accumulator in aerial organs. The popular isolines A17 and R108 were selected as reference ecotypes with relatively high total saponin accumulation in both aerial

and root organs, due primarily to their role in the development of genomics and mutant population resources for research in *M. truncatula*. An evaluation of the several possible permutations of comparisons enabled by the experimental matrix is critical for understanding the results of the integrated transcriptomics and metabolomics datasets.

A.

Ecotype:	A17	ESP105	GRC43	R108
AERIAL ORGAN	**HIGH**	**LOWEST**	**VERY HIGH**	**HIGH**
ROOT ORGAN	**HIGH**	**VERY HIGH**	**LOWEST**	**MODERATE**

B.

I. Inter-Genotype, Inter-Organ **(n = 24)**

Genotype:	A17	ESP105	GRC43	R108
AERIAL ORGAN Gross Saponin Phenotype:	***H***	***L***	***H***	***H***
ROOT ORGAN Gross Saponin Phenotype:	***H***	***H***	***L***	***H***

II. Inter-Genotype, Intra-Organ (Aerial) **(n=12)**

Genotype:	ESP105	A17	GRC43	R108
AERIAL ORGAN Gross Saponin Phenotype:	***L***	***H***	***H***	***H***

III. Inter-Genotype, Intra-Organ (Root) **(n=12)**

Genotype:	GRC43	A17	ESP105	R108
ROOT ORGAN Gross Saponin Phenotype:	***L***	***H***	***H***	***H***

IV. Intra-Genotype (ESP105) , Inter-Organ **(n=6)**	ESP105
AERIAL ORGAN Gross Saponin Phenotype:	***L***
ROOT ORGAN Gross Saponin Phenotype:	***H***

V. Intra-Genotype (GRC43) , Inter-Organ **(n=6)**	GRC43
AERIAL ORGAN Gross Saponin Phenotype:	***H***
ROOT ORGAN Gross Saponin Phenotype:	***L***

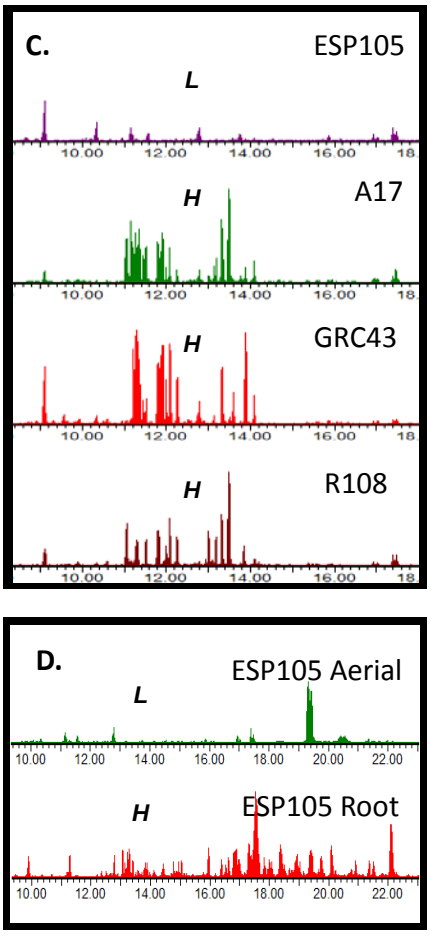


FIGURE 1_RII. Explanation of the Ecotype/Organ Experimental Matrix.

(A) Summary table showing the complete Ecotype/Organ Experimental Matrix. The terms "High," "Lowest" etc. presented in the table refer to the total triterpene saponin accumulation phenotypes for

each of these ecotypes from the large scale biochemical phenotyping (metabolomics) analysis of the 110 ecotypes in the germplasm diversity panel presented in JHS_Research_CHAPTER_I.

(B) Tables presenting the various permutations of comparisons between the samples in the Ecotype/Organ Experimental Matrix. The notation refers to either the "**H**igh" or "**L**ow" gross phenotypic state for each of the samples (see Equation 1_RII for detail). The number of samples used for each of the permutations is noted (e. g. "n=24" in table I).

(C) Base peak intensity chromatograms (8 to 18 min of 39 min UPLC-ESI-qTOF-MS analysis) for the aerial organs of the 4 genotypes of the experimental matrix. This panel offers a visual representation of the inter-genotype, intra-aerial-organ (B. table II) comparison permutation. The gross phenotypic state (e. g. "L") is indicated for each of the samples. Saponins from aerial organs primarily elute in the 8 to 18 min window of the UPLC separation.

(D) Base peak intensity chromatograms (9 to 23 min of 39 min UPLC-ESI-qTOF-MS analysis) for the aerial and root organs of the genotype ESP105 from experimental matrix. This panel offers a visual representation of the intra-genotype, inter-organ (B. table IV) comparison permutation.

Results

Probeset Annotation

Table 1_RII presents the combined results of the profile HMM annotation of the tentative consensus sequences of the 61,278 probes of the Affymetrix Medicago Gene Chip™ and the BLAST-based annotation of the tentative consensus sequences which map to the current draft of the *M. truncatula* genome ||unpublished, ZHAO||. 672 of

the 1349 probesets annotated as likely regulatory elements (Udvardi, Kakar et al. 2007)were found to differ significantly($p < 0.05$) between the ecotypes, based on the associative t-test of the Affymetrix Medicago Gene Chip™ data.

	Concatenated Total	Unique to HMM	Unique to Genome BLAST
Cytochrome P450	344	103	25
Glycosyltransferase	550	405	66

TABLE 1_RII. Cytochrome P450 and Glycosyltransferase Concatenated Annotation List Summary. A table summarizing the number of hits (i. e. probeset annotated as P450 or GT) of the *de novo* profile HMM analysis of the tentative consensus sequences used to design the probesets of the Affymetrix Medicago Gene Chip™, and the BLAST-basedre-annotation of the probesets which mapped to the *M. truncatula* genome sequence.

Gross Phenotype Comparisons for Selection of Candidate Genes of Interest

The differences in total saponin accumulation between the various ecotypes and organs selected for the experimental matrix enabled a relatively simple process of candidate gene identification- here termed “gross phenotype comparisons” (see Methods).

The state filter of the gross phenotype comparison for aerial organs isolated 93 of the 344 cytochrome P450 probesets. These 93 probesets were then ranked based on the f values obtained from application of EQUATION 1_RII. Table 2_RII presents the top 15 probesets resulting from the complete gross phenotype comparison selection

process. Recall that the intra-ecotype (see Figure 1_RII), inter-organ comparisons are also critically useful for selection of candidate genes. As such, the expression values for root organs for ecotype ESP105 are included in Table 2_RII.

Rank	Probeset	BLAST ID (*M.truncatula*)	*f* value	ESP105 Aerial	A17 Aerial	GRC43 Aerial	R108 Aerial		ESP105 Root
1	Mtr.37298.1.S1_at	*M.t*_ cyp72a68	1629	88	2515	3154	1902		3153
2	Mtr.43018.1.S1_at	*M.t*_ cyp716a12	1275	17	1788	5602	3002		6768
3	Msa.1808.1.S1_at	*M.t*_ cyp716a12	560	16	861	2990	1544		3418
4	Mtr.37299.1.S1_at	*M.t*_ cyp72a67	424	28	1225	1332	200		1681
5	Mtr.31199.1.S1_s_at	*M.t*_ cyp716a12	337	9	577	1830	884		2821
6	Mtr.2065.1.S1_at	*M.t*_ cyp88d3	104	63	351	265	355		11
7	Mtr.37358.1.S1_s_at	*M.t*_ cyp83g1	90	71	640	210	791		2790
8	Mtr.37356.1.S1_at	*M.t*_ cyp83g1	75	73	718	238	1026		3538
9	Mtr.42226.1.S1_at	*M.t*_ cyp88d3	24	43	222	144	198		13
10	Mtr.12672.1.S1_at	No sig similarity found	-9	47	138	234	177		869
11	Mtr.49920.1.S1_x_at	AC233070.6 (genome seq)	-11	6	7	8	10		47
12	Mtr.17322.1.S1_x_at	*M.t*_ cyp71d64	-12	6	8	7	8		41
13	Mtr.4753.1.S1_at	*M.t*_ cyp88d2	-13	11	46	42	14		11
14	Mtr.23217.1.S1_at	AC145061.27 (genome seq)	-13	7	8	11	10		45
15	Mtr.5109.1.S1_at	AC152936.21 (genome seq)	-14	7	7	7	9		514

TABLE 2_RII. Top 15 Cytochrome P450 Probesets from the Gross Phenotype Comparison Ranking Process for the Inter-Genotype, Intra-Aerial-Organ Comparison.

Values in the table represent the mean (n = 3) expression (hybridization) level for the probesets indentified via the gross phenotype comparison ranking (equation 1) process for the inter-genotype, intra-aerial-organ (state filter) comparison. The values shaded in gray are for from the ecotype which represents the "Low" state for this comparison (ESP105 Aerial). The expression level for "High" state from the intra-genotype (ESP105 Root), inter-organ comparison is also included. BLAST IDs are from tBLASTn analysis; using a given probeset design sequence as the query against the *M. truncatula* records in the NCBI Nucleotide Collection (nr/nt) database. The annotations presented represent the record with the most significant E value, except in the cases when records with some form of functional annotation beyond simple BAC clone ID or genome records were available.

The second illustrative example of the gross phenotype comparison selection process is from the GT probesets from root organs. The state filter of the gross phenotype comparison for root organs isolated 120 of the 550 GT probesets. These 120 probesets were then ranked (Table 3_RII) based on the *f* values obtained from application of EQUATION 1_RII. Appendix_RII_Gross_Phenotype_ Inverse _Examples presents a similar illustrative example for the inverse state of regulatory probesets from root organs ranked according to *g* values from the application of EQUATION 2_RII, and includes probesets annotated as transcription factors (Chen, Yu et al. ; Kalo, Gleason et al. 2005).

Rank	Probeset	BLAST ID (*M.truncatula*)	*f* value	GRC43 Root	A17 Root	ESP105 Root	R108 Root		GRC43 Aerial
1	Mtr.22118.1.S1_s_at	AC119419.11	1844	277	3264	2515	3532		212
2	Mtr.37250.1.S1_at	No sig similarity found	817	52	946	1309	1588		22
3	**Mtr.12473.1.S1_at**	***M.t*_ UGT73F3**	666	260	1506	1662	2001		16
4	Mtr.1550.1.S1_at	No sig similarity found	444	17	1244	2404	701		21
5	Mtr.41983.1.S1_at	AC142095.11	442	349	2214	2172	1169		75
6	Mtr.9221.1.S1_at	*M.t*_ GT63G	372	214	1323	1436	917		78
7	**Mtr.4547.1.S1_at**	***M.t*_ UGT73K1**	252	32	485	460	305		16
8	Mtr.11212.1.S1_s_at	No sig similarity found	162	9	1092	2180	282		8
9	**Mtr.28421.1.S1_x_at**	***M.t*_ UGT73K1**	42	10	235	483	113		9
10	Mtr.28421.1.S1_at	No sig similarity found	29	9	135	197	46		9
11	Mtr.25168.1.S1_at	CR932040.2	-5	16	20	155	122		16
12	Mtr.50388.1.S1_at	AC140034.14	-6	9	26	81	46		6
13	**Mtr.27374.1.S1_at**	***M.t* _UGT73K1**	-10	35	190	367	136		8
14	Mtr.46668.1.S1_at	BT051872.1	-12	7	13	17	7		6
15	Mtr.37105.1.S1_at	AC136472.40	-13	7	9	9	10		15

TABLE 3_RII. Top 15 Glycosyltransferase Probesets from the Gross Phenotype Comparison Ranking Process for the Inter-Genotype, Intra-Root-Organ Comparison.

Values in the table represent the mean (n = 3) expression (hybridization) level for the probesets identified via the gross phenotype comparison ranking (equation 1) process for the inter-genotype, intra-root-organ (state filter) comparison. The values shaded in gray are for the ecotype which represents the "Low" state for this comparison (GRC43 Root). The expression level for "High" state from the intra-genotype (Aerial), inter-organ comparison is also included. BLAST IDs are from tBLASTn analysis; using a given probeset design sequence as the query against the *M. truncatula* records in the NCBI Nucleotide Collection (nr/nt) database. The annotations presented represent the record with the most significant E value, except in the cases when records with some form of functional annotation beyond simple BAC clone ID or genome records were available. Annotations in bold are for probesets which likely represent genes that have been functionally characterized and established as GTs in the *M. truncatula* triterpene saponin biosynthetic pathway.

Pearson Correlation Coefficients for Transcripts vs. Metabolites

Pearson correlation coefficients for high-priority candidate (from the gross phenotype comparisons) probesets in four different [genotype] vs. [organ] permutations (using the total saponin accumulation values) are presented in Table 4_RII.

Intra-Genotype ESP105, Inter-Organ

Gene of interest	Representative Probeset	Pearson's r (Intra-Genotype ESP105)(n=6)	90 % bootstrap confidence interval (r)	bootstrap standard error of (r)	rank (Intra-Genotype ESP105)
*M.t_* cyp72a68	Mtr.37298.1.S1_at	0.882	0.73 - 1.00	0.165	66
*M.t_* cyp716a12	Mtr.43018.1.S1_at	0.887	0.71 - 1.00	0.202	64
*M.t_* cyp72a67	Mtr.37299.1.S1_at	0.870	0.68 - 1.00	0.188	71
*M.t_* cyp88d3	Mtr.2065.1.S1_at	-0.547	-1.00 - 0.33	0.313	240
*M.t_* cyp83g1	Mtr.37356.1.S1_at	0.893	0.72 - 1.00	0.287	60

Inter-Genotype & Inter-Organ

Gene of interest	Representative Probeset	Pearson's r (Inter-Genotype & Inter-Organ)(n=24)	90 % bootstrap confidence interval (r)	bootstrap standard error of (r)	rank (Inter-Genotype & Inter-Organ)
*M.t_* cyp72a68	Mtr.37298.1.S1_at	0.575	0.28 - 0.76	0.153	4
*M.t_* cyp716a12	Mtr.43018.1.S1_at	0.370	0.06 - 0.61	0.170	90
*M.t_* cyp72a67	Mtr.37299.1.S1_at	0.233	-0.10 - 0.50	0.184	68
*M.t_* cyp88d3	Mtr.2065.1.S1_at	0.236	0.06 - 0.45	0.121	63
*M.t_* cyp83g1	Mtr.37356.1.S1_at	0.091	-0.30 - 0.35	0.200	120

Inter-Genotype, Intra-Aerial-Organ

Gene of interest	Representative Probeset	Pearson's r (Inter-Genotype Aerial)(n=12)	90 % bootstrap confidence interval (r)	bootstrap standard error of (r)	rank (Inter-Genotype, Intra-Aerial-Organ)
*M.t_* cyp72a68	Mtr.37298.1.S1_at	0.668	0.23 - 0.89	0.307	9
*M.t_* cyp716a12	Mtr.43018.1.S1_at	0.501	0.10 - 0.80	0.262	54
*M.t_* cyp72a67	Mtr.37299.1.S1_at	0.357	-0.09- 0.68	0.300	94
*M.t_* cyp88d3	Mtr.2065.1.S1_at	0.370	0.07 - 0.67	0.197	88
*M.t_* cyp83g1	Mtr.37356.1.S1_at	0.572	0.36 - 0.78	0.134	33

Inter-Genotype, Intra-Root-Organ

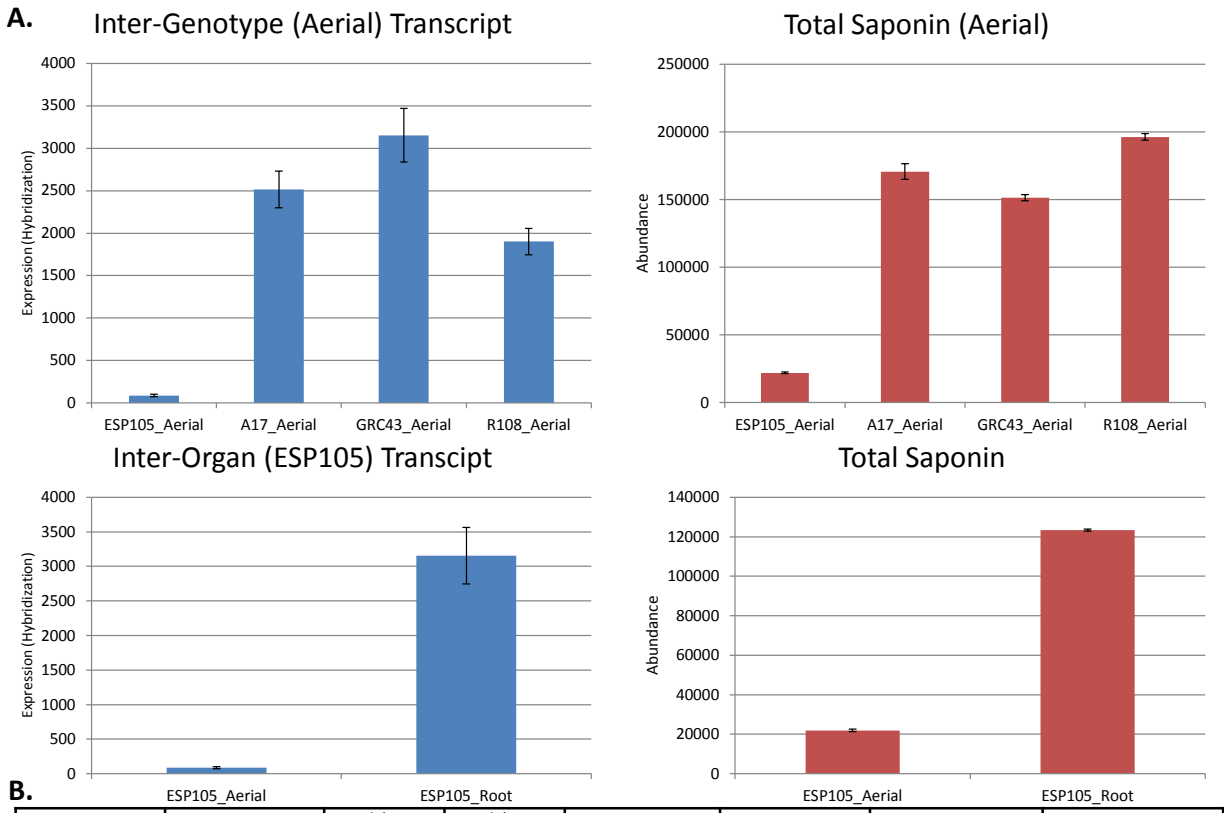
Gene of interest	Representative Probeset	Pearson's r (Inter-Genotype Root)(n=12)	90 % bootstrap confidence interval (r)	bootstrap standard error of (r)	rank (Inter-Genotype, Intra-Root-organ)
*M.t_* cyp72a68	Mtr.37298.1.S1_at	0.569	-0.22 - 0.84	0.312	33
*M.t_* cyp716a12	Mtr.43018.1.S1_at	0.243	-0.23 - 0.73	0.278	25
*M.t_* cyp72a67	Mtr.37299.1.S1_at	-0.012	-0.42 - 0.56	0.283	155
*M.t_* cyp88d3	Mtr.2065.1.S1_at	-0.276	-0.64 - 0.20	0.249	247
*M.t_* cyp83g1	Mtr.37356.1.S1_at	-0.262	-0.58 - 0.27	0.253	242

TABLE 4_RII. Pearson Correlation Coefficient Analysis of High Priority Cytochrome P450 Probesets.

Pearson correlation coefficient (Pearson's r) values from 4 sample permutations from the experimental matrix for [transcript] vs. [total saponin accumulation]. The table also includes: Bootstrapped (5000 iteration) 90% confidence intervals of r, bootstrapped standard errors of r, and the rank of the r value (high to low) within a given permutation among all cytochrome P450 probesets.

Example of Selected Candidate Genes

Figure 2_RII and Figure 3_RII present the combined gross phenotype comparison and Pearson correlation coefficient analyses for two (cyp72a68 and cyp88d3) of the five cytochrome P450 genes selected as high priority candidates. Graphs of both transcript and saponin accumulation are presented in order to visually emphasize the obvious relationships between the two data sets for these candidate genes. The probeset and total saponin accumulation values for cyp72a68 are characteristic of the probeset/metabolite relationships (inter-genotype, intra-aerial-organ, and the intra-genotype, inter-organ comparisons) used in the selection of candidates cyp72a68, cyp72a67, cyp716a12, and cyp83g1. The selection of cyp88d3 as a candidate is based on the relationship of the probeset and metabolite values from both the inter-genotype, intra-aerial-organ and the inter-genotype, inter-organ comparisons. Note the relative lack of cyp88d3 expression in root organs.



B.

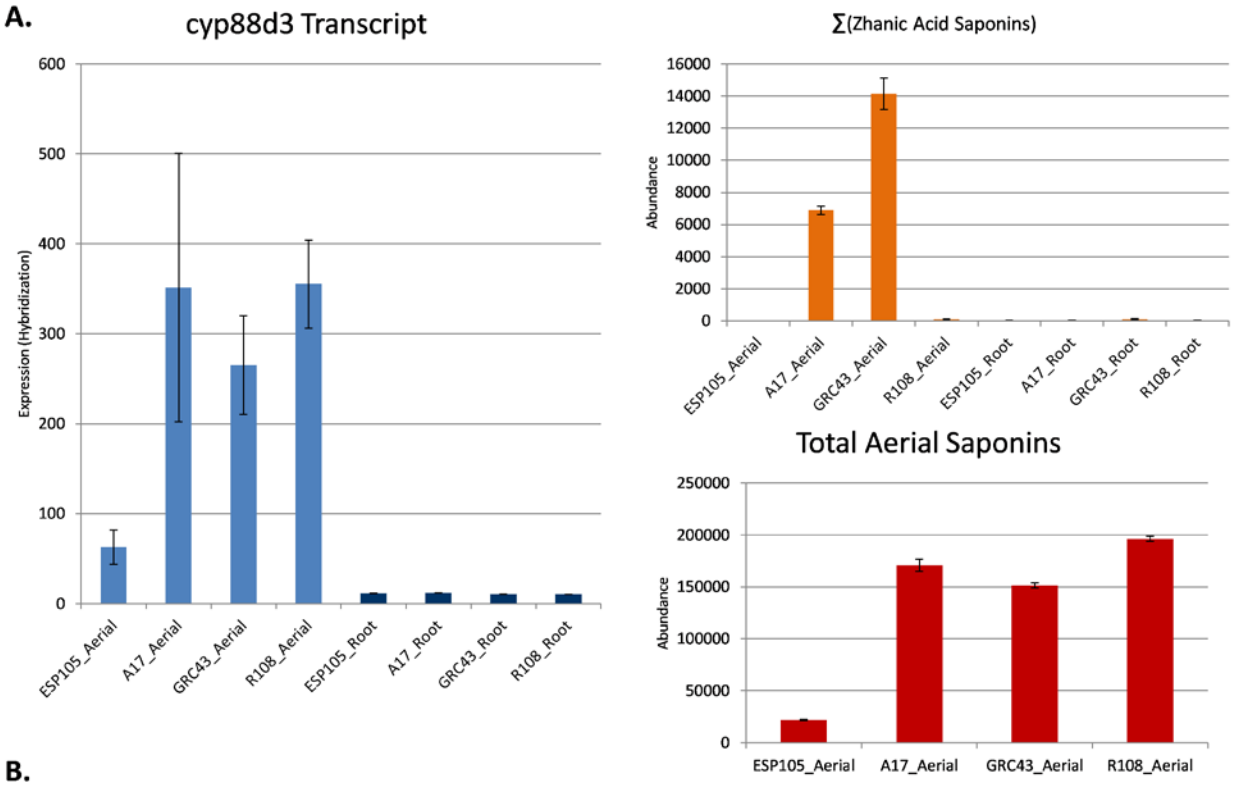
Gene of interest	Representative Probeset	Rank (Gross Phenotype Aerial)	Rank (Gross Phenotype Root)	Rank (Intra-Genotype ESP105)	Rank (Inter-Genotype & Inter-Organ)	Rank (Inter-Genotype Aerial)	Rank (Inter-Genotype Root)
*M.t*_cyp72a68	Mtr.37298.1.S1_at	1	excluded by state filter	66 (r =0.88)	4 (r =0.58)	9 (r =0.67)	33 (r =0.57)

FIGURE 2_RII. Graphical and Tabulated Summary of Results for the Selection of cyp72a68 as a High Priority Gene Candidate.

(A) Graphs showing the transcript expression dynamics (in blue, on left side) for the cyp72a68 probeset "Mtr. 37298. 1. S1_at" shown for the inter-genotype, intra-aerial-organ (top), and intra-ESP105-genotype, intra-organ (bottom) comparative permutations. Total saponin accumulation values are presented for the same comparative permutations (in red, at right).

(B) Summary table of the rankings for the cyp72a68 probeset "Mtr. 37298. 1. S1_at" from both the gross phenotype comparison rankings and the Pearson correlation (r values are included in

parentheses) analysis of total saponin accumulation values with 4 permutations of samples from the experimental matrix.



Gene of interest	Representative Probeset	Rank (Gross Phenotype Aerial)	Rank (Gross Phenotype Root)	Rank (Intra-Genotype ESP105)	Rank (Inter-Genotype & Inter-Organ)	Rank (Inter-Genotype Aerial)	Rank (Inter-Genotype Root)
*M.t*_cyp88d3	Mtr.2065.1.S1_at	6	13	240 (r = -0.55)	63 (r =0.24)	88 (r =0.37)	247 (r = -0.28)

FIGURE 3_RII. Graphical and Tabulated Summary of Results for the Selection of cyp88d3 as a High Priority Gene Candidate.

(A) Graph showing the transcript expression dynamics (in blue, at left) of the cyp88d3 probeset "Mtr. 2065. 1. S1_at" shown for the inter-genotype, inter -organ (in blue at top left) comparative permutation. A graph of the summed accumulation values of all saponin compounds annotated with zhanic acid as the sapogenin aglycone in inter-genotype, inter -organ (in orange at top right)

permutation. Total saponin accumulation values are presented for the inter-genotype, intra-aerial-organ permutation (in red, at bottom).

(B) Summary table of the rankings for the cyp88d3 probeset "Mtr. 2065. 1. S1_at" from both the gross phenotype comparison rankings and the Pearson correlation (r values are included in parentheses) analysis of transcript vs. total saponin accumulation values with 4 permutations of samples from the experimental matrix.

Expression Dynamics for Known Triterpenoid Biosynthetic Genes in *M. truncatula*.

Figure 4_RII presents expression data for previously characterized biosynthetic genes from triterpene metabolism of *M. truncatula*. Figure 5_RII presents expression data for previously characterized glycosyltransferase genes from triterpene saponin metabolism of *M. truncatula*.

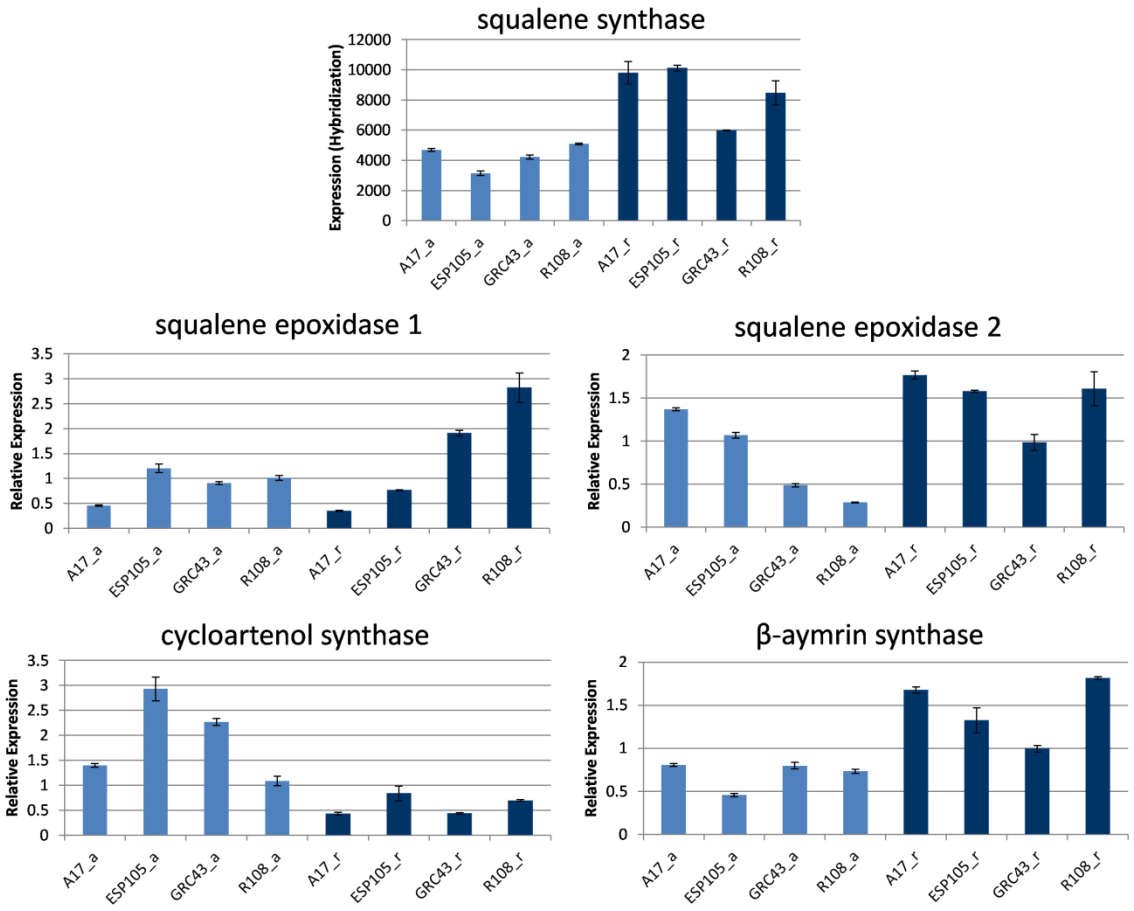


FIGURE 4_RII. Ecotype Matrix Expression Dynamics for Known Triterpenoid Biosynthetic Pathway Genes Preceding Triterpene Sapogenin Bio-Oxidation.

Graphs showing the transcript expression dynamics in both organ types of all genotypes for squalene synthase, squalene epoxidase 1, squalene epoxidase 2, cycloartenol synthase, and β-amyrin synthase. Error bars represent 1 standard error. The squalene synthase accumulation data is from the microarray experiment. Data for the other genes is from qRT-PCR analysis of the same samples, as cycloartenol synthase and β-amyrin synthase are known to co-hybridize to the same microarray probesets (i. e. "shared probeset"). Similarly squalene epoxidase 1 and squalene epoxidase 2 co-hybridize with a number of probesets.

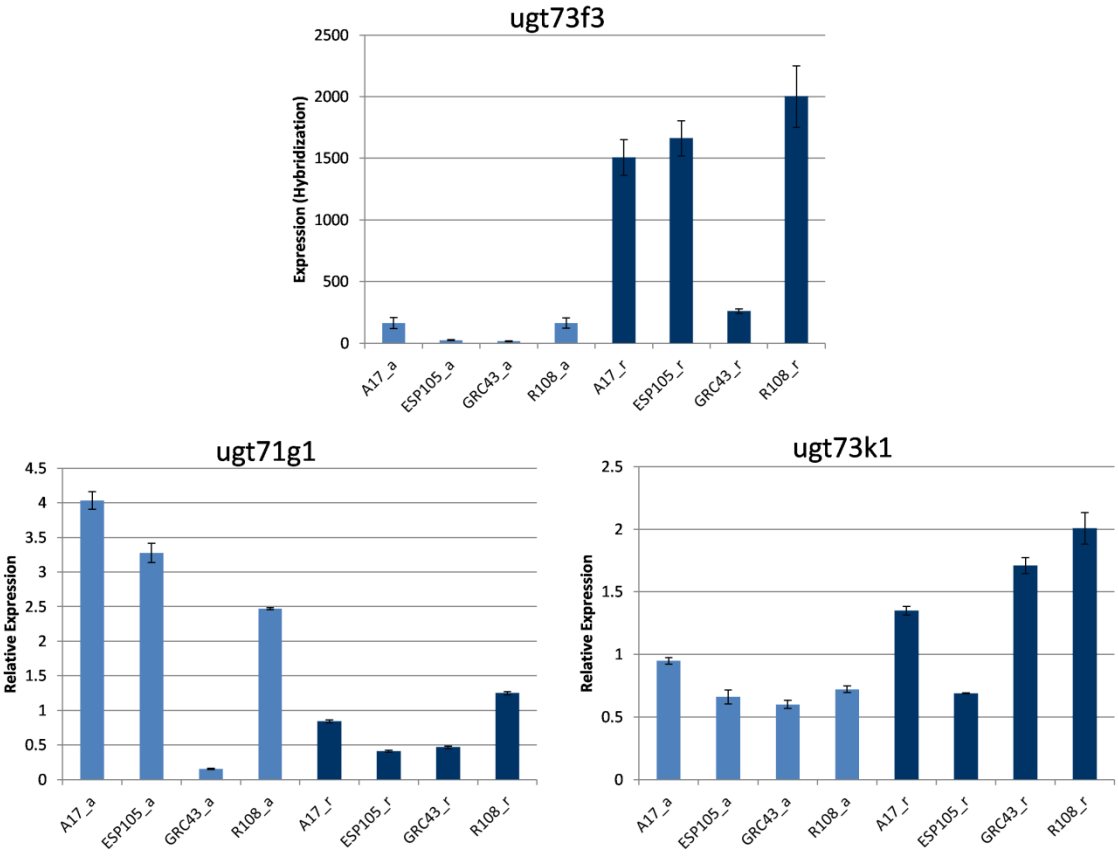


FIGURE 5_RII. Ecotype Matrix Expression Dynamics for Known Glycosyltransferases of the Triterpene Saponin Biosynthetic Pathway.

Graphs showing the transcript expression dynamics in both organ types of all genotypes for ugt73f3, ugt71g1, and ugt73k1. Error bars represent 1 standard error. The ugt73f3 accumulation data is from the microarray experiment. Data for ugt71g1 and ugt73k1 is from qRT-PCR analysis of the same samples, as these transcripts are known to co-hybridize to the same microarray probesets (i.e. "shared probeset").

Discussion

Probeset Annotation

The most comprehensive analysis of the cytochrome P450 gene superfamily in *M. truncatula*(Li, Cheng et al. 2007)identified 151 putative P450 genes, including 135 novel sequences not present in NCBI Genbank (NCBI, http://www.ncbi. nlm. nih. gov/genbank/ , Bethesda, MD) at the time of the analysis. These putative P450 genes were classified into 9 clans and 44 families. Four of the clans and 21 of the families had not been reported previously in legumes. The annotations for the tentative consensus sequences of the 61,278 Affymetrix Medicago Gene Chip™ probe sequences available at the time of the ecotype experiment showed poor coverage for known cytochrome P450 and glycosyltransferase genes (Lahoucine Achnine 2005). Further, the anarchic gene descriptions (Blast results against various plant genomes) made comprehensive analysis of these gene families impossible, as they lacked a consistent signifier. Some probesets had function-based annotations, while others had protein family-based annotations. For example, the probeset ("Mtr. 2065. 1. S1_at") representing cyp88d3 was annotated as "Similar to Ent-kaurenoic acid oxidase," while the probeset ("Mtr. 37298. 1. S1_at") representing cyp72a68 was annotated as "Cytochrome P450." Ent-kaurenoic acid oxidase happens to be a cytochrome P450

enzyme known to function in the gibberellin biosynthetic pathway (Helliwell, Chandler et al. 2001), but this function-based annotation highlights the inability to systematically isolate all of the probesets representing cytochrome P450 genes. Subsequent efforts to ameliorate this problematic dual (function-based and protein family-based) annotation schema have focused exclusively on tentative consensus sequences which map to the *M. truncatula* genome (unpublished, personal communication with Yuhong Tang and Patrick Zhao).

As the tentative consensus sequences used for the design of the Affymetrix Medicago Gene Chip™ were based not on the genome sequence but rather on EST library sequences, the possibility exists that the recent efforts at re-annotation will not comprehensively represent the probesets for a given gene family. Indeed, the *de novo* profile HMM results of the tentative consensus sequences show this to be the case. 103 putative P450 probesets and 405 putative GT probesets were identified in the HMM analysis but not the genome BLAST analysis. Critically, four of the five high priority cytochrome P450 candidate genes identified in this study (cyp72a67, cyp72a68, cyp716a12, and cyp88d3) were correctly annotated in the HMM results but not found in the genome BLAST analysis. Similarly, a probeset representing cyp88d2 (studied in JHS_RESEARCH_CHAPTER_IV) was correctly annotated in

the HMM results but not in the genome BLAST results. Inversely, 25 putative P450 and 66 putative GT probesets were uniquely identified in the genome BLAST analysis. As the primary goal of the HMM annotation was to enable the systematic isolation of P450 and GT probesets, the actual scores of the HMM or genome BLAST results are not critically important. It is possible that some of the probesets included in the concatenated HMM/BLAST P450 and GT lists are not actually members of these families, but the list dramatically improves the comprehensiveness of gene family representation. Obviously, more detailed analysis of the design sequence for particular candidate probesets of interest can identify inaccurate annotations from the list as needed.

Integrated Analysis of Transcriptomics and Metabolomics Datasets

The *f* statistic was developed empirically through successive iterations of simple arithmetic manipulations of the state filtered probeset list for cytochrome P450 probesets from the inter-genotype, intra-aerial-organ comparison, where a "correspondence" or "match" between the probeset values and total saponin phenotypes was starkly obvious to someone familiar with the design of the experimental matrix. In Figure 2_RII A and Figure3_RII A, this is made clear by the fact that the values of the transcripts for expression levels and the values for total

saponin content between the ecotypes are extremely similar, particularly with regard to the large absolute difference between the lowest ecotype and the next lowest. Likewise, in the intra-ecotype comparison, the differences between the organs are extremely similar between expression levels and total saponin content, and the stark absolute magnitude of the difference is apparent.

The first term of Equation 1_RII simply represents the sum of the differences for a probeset's expression value for the three high state ectypes minus the expression value for the probeset from the low state ecotype. The second term of Equation 1_RII (average value the probeset of the three high states minus the value of the probeset from the low state) acts to emphasize the proximity to zero (no expression) of the probeset value from the low state ecotype. This is critical because of the extremes of the saponin accumulation phenotypes among the ecotypes selected for the matrix, and is motivated by the strong 'guilt by association' phenomenon observed in genes of plant secondary metabolism (Saito, Hirai et al. 2008). The final term of Equation 1_RII (subtraction of the maximum value) is included to remove influence of extreme outlier probeset values from any of the three high state ecotypes.

While the *f* statistic was obtained through "training" from the cytochrome P450 probeset data from the inter-genotype, intra-aerial-organ comparison, it was

found to generalize to the other comparative permutations of the experimental matrix. Non-rigorous validation evidence for the generalization of the *f* statistic to other permutations of the matrix is provided by the inclusion in the top 15 GT list (Table 3_RII) from the inter-genotype, intra-root-organ permutation comparison of several probesets with annotations for previously characterized GTs known to function in the *M. truncatula* triterpene saponin biosynthetic pathway (Lahoucine Achnine 2005; Naoumkina, Modolo et al. 2010). As highlighted in the introduction, most of the models used to link transcript and metabolite information from germplasm diversity samples have focused on a "major" phenotype such as biomass or fruit color. Focus on a value external to the transcript or metabolite accumulation values enables considerable flexibility to modelers seeking to elucidate the transcript/metabolite relationships, and development of such models is an active area of research with obvious applications in fields such as plant breeding (Goodacre, Roberts et al. 2007).

The gross phenotype comparisons (state filter plus *f* or *g*statistic) ultimately proved to be a more useful analytical model than Pearson correlation coefficient analysis for selection of high priority candidate genes. Evidence for this conclusion is evident in the summary tables for the example selected candidate genes (Figure 2_RII B., Figure3_RII B.). It is important to note that the initial selection of candidate genes

aimed to identify early genes in the pathway, and that the high resolution of individual compounds from the metabolomics dataset was not exploited to any large degree in this study. Indeed, the high sensitivity of Pearson correlation coefficient analysis to differences among intra-state probeset or saponin accumulation values (e. g. H1, H2, and H3 from inter-genotype, intra-aerial-organ comparison) likely masked the grosser trends that were the primary focus of the integrated transcriptomics and metabolomics dataset analysis. The bootstrapped confidence intervals obtained for the Pearson's r (Table 4_RII) are quite large in many cases, reinforcing the notion that Pearson correlation coefficient analysis is not a strong model for representation of useful trends of transcript/metabolite accumulation, at least for the particular comparative permutations employed in this study to date.

As more information about the of the molecular basis of sapogenin and saponin biosynthesis is obtained, it is likely that Pearson correlation coefficient analysis of probesets vs. precisely selected groups of compounds or organ/ecotype pairings will become more powerful (offer finer resolution) for subsequent identification of candidate genes. For example the significantly higher expression of cyp88d3 in aerial organs suggests that it may be important in the bio-oxidation of carbon 16 of triterpene sapogenins (see JHS_RESEARCH_CHAPTER_I). The structural diversity

of saponins between the ecotypes or organs such as the absence of zhanic acid saponins detected in R108 aerial organs (see Figure 4_RII A.) or the lack of strong cyp88d3 expression in root organs will offer a huge combinatorial space of "sub-matrices" to explore with Pearson correlation coefficient analysis or other more sophisticated comparative transcript/metabolite models.

The UPLC-ESI-qTOF-MS analytical method used for the metabolomics analysis was not optimized for separation and detection of triterpene saponins, but developed to enable simultaneous detection of compounds from several classes of plant secondary metabolites (e. g. flavonoids, isoflavonoids). Thus, the targeted flavonoid accumulation data and non-targeted MARKERLYNX analysis should prove useful to researchers interested in exploring the 'guilt by association' relationship of a very large number of probeset/metabolite combinations from the experimental matrix.

Other Triterpenoid Pathway Genes

Examination of the expression dynamics for known genes preceding the bio-oxidation of β-amyrin from *M. truncatula* triterpenoid metabolism reveals several interesting trends (Figure 4_RII). Expression levels of triterpene biosynthetic genes (squalene synthase and squalene epoxidases) which precede the critical branch point for sapogenin or sterol biosynthesis (cyclization of squalene to β-amyrin or cycloartenol,

respectively) do not share the same patterns of extreme highs and lows for transcript accumulation in the experimental matrix as those observed for the candidate cytochrome P450 genes. Further, the expression of the sterol biosynthesis entry point enzyme (cycloartenol synthase) is highest in the ecotype and organ (ESP105 aerial organ) with the lowest total saponin accumulation value, suggesting that the sink for sterols is most pronounced in this aerial organs of this ecotype. The lowest expression value for β-amyrin synthase is also found in ESP105 aerial organs, which is consistent with the notion that triterpenoid skeletons are preferentially shuttled into sterol rather than triterpene sapogenin biosynthesis in ESP105 aerial organs. However, it is important to note that the transcriptomics and metabolomics datasets from this study are for steady state conditions, so such inferences about biosynthetic flux are ultimately speculative/hypothetical.

Examination of the expression dynamics for known glycosyltransferase genes from the *M. truncatula* biosynthetic pathway also revealed several interesting trends (Figure 5_RII). First and foremost, the expression levels of these previously functionally characterized GTs served a form of validation for the predictive qualities of the *f* statistic from the gross phenotype comparison process. The higher expression levels of ugt73f3 in root organs compared to aerial organs suggests that the

glycosylation products of the UGT73F3 activity may be detected in higher levels in root organs, which could offer useful structural information for the chemical characterization of unknown saponin compounds. Further, the low relative expression of ugt71g1 in the aerial organs of GRC43 aerial organs may offer a "key" for the structural elucidation of unknown saponin compounds which are uniquely present or absent in GRC43 aerial organs when compared to the saponins from aerial organs of A17, ESP105, and R108 (see JHS_RESEARCH_CHAPTER_1). Finally, the expression data for these known GTs should enable the development of more sophisticated predictive models to explore transcript/metabolite relationships that could more effectively identify likely candidate GTs from the datasets from the experimental matrix.

Methods

Plant Growth and Harvest

ESP_105 and GRC_43 seeds used in this study were of the same single seed descent lines developed in [chapter R. I]. A17 and R108 isoline seeds were obtained from the greenhouse manager (David McSweeny) at the Samuel Roberts Noble Foundation. Plants were grown in a root cone system (Stuewe and Sons, http://www.stuewe.com, Tangent, OR) with Turface MVP medium (Profile Products, Buffalo Grove, IL) in a

Conviron TCR180 walk-in growth chamber (http://www.conviron.com/, Winnipeg, Manitoba, Canada) maintained at 90% humidity and at an average temperature of 24 °C day (16 h) and 20 °C night (8 h). Plants were fertilized with 15 ppm nitrogen (Scotts' 20 10 20 Peat-Lite Special, http://www.scotts.com, Marysville, Ohio) daily in the morning and watered with distilled water in the evening. Plants were harvested at 6 weeks post-germination and dissected into aerial and root organs. Aerial tissues from the youngest 6 metamers of individual plants and whole root organ samples from individual plants were prepared as single biological replicates. For both aerial and root samples, three biological replicates were prepared for all of the ecotypes. Samples were frozen immediately in liquid nitrogen, ground using a mortar and pestle, and stored at –80C. The same sample material was used for the metabolomics, microarray, and qRT-PCR analyses.

DNA Preparations, mRNA Isolation, Microarray Analysis

Total RNA was extracted using TRIZOL reagent (Invitrogen, http://www.invitrogen.com/), treated with DNaseI (Ambion, http://www.ambion.com/), and column purified with an RNeasyMinEluteCleanUp Kit (Qiagen, http://www.qiagen.com/). RNA was quantified using a Nanodrop Spectrophotometer ND-100 (NanoDrop Technologies, http://www.nanodrop.com/)

and evaluated for quality with a Bioanalyzer 2100 (Agilent, http://www.home.agilent.com/). The Affymetrix Medicago Gene Chip™ (Affymetrix, http://www.affymetrix.com/) was used for expression analysis. The RNA from three independent biological replicates was analyzed for both root and aerial organs for each of the four ecotypes (a total of 24 chips). Probe labeling using 10 μg RNA, array hybridization, and scanning were performed according to the manufacturer's instructions for eukaryotic RNA, using a one-cycle protocol for cDNA synthesis. For each Affymetrix array hybridized, the resulting . cel file was exported from GeneChip Operating Software Version 1.4 (Affymetrix) and imported into Robust Multiarray Average (Irizarry, Bolstad et al. 2003) for global normalization. Presence/absence call for each probe set was obtained using dCHIP(Parmigiani, Garrett et al. 2003). Gene selections based on an associative t-test (Dozmorov and Centola 2003)were made using Matlab (MathWorks, http://www.mathworks.com/). The complete Affymetrix data set (ID# E-MEXP-2984) is publicly available at ArrayExpress (http://www.ebi. ac. uk/arrayexpress).

cDNA Synthesis and qRT-PCR Analysis

For cDNA synthesis preceding qRT-PCR analysis, 10μg of total RNA (prepared and assessed for quality as above) was primed with oligo(dT)20 and synthesized with

Super Script III according to manufacturer's instructions. qRT-PCR reactions were performed in an optical 384-well plate with an ABI PRISM 7900 HTsequence detection system (Applied Biosystems), using SYBR Green to monitor dsDNA synthesis. Reactions contained 2µl of primer pair (1 µM), 2µl of 1:20 dilution of cDNA, 5µl of 2x power SYBR Green MASTER MIX, and 1µl water. The following standard thermal profile was used for all PCR reactions: 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 s, and 60°C for 1 min. Amplicon dissociation curves were recorded after cycle 40 by heating from 60°C to 95°C with a ramp speed of 1. 9°C/min. Primers (Appendix_RII_Primers) were designed using Primer Express® Software (Applied Biosystems). All reactions were performed with 3 technical replicates for each of 3 biological replicates. Data were analyzed using SDS 2. 2. 1 software (Applied Biosystems). PCR reaction efficiencies were determined using LinReg PCR software(Ruijter, Ramakers et al. 2009). Transcript expression levels were determined relative to two housekeeping genes (ubiquitin and actin), based on modifications (equation below) of formulae presented in (Pfaffl 2001; Czechowski, Stitt et al. 2005). Briefly, the ΔCt terms for the target and reference genes were calculated as the mean Ct of all samples minus the Ct of a given sample, rather than "control" minus "treatment" Ct values.

$$relative\ expression\ ratio = \frac{E_{target}^{\Delta Ct_{target}(\bar{x}-x)}}{E_{ref}^{\Delta Ct_{ref}(\bar{x}-x)}}$$

Extractions and Metabolomics Analysis

Harvested sample material was lyophilized prior to extraction. 10.00 ± 0.06 mg of powder was extracted with 1 ml of 80% Methanol (containing 0.018 mg/ml umbelliferone as an internal standard) in a dram vial for 2 hours on an orbital shaker. Extracted samples were centrifuged for 30 minutes at 2900g at 4°C, and supernatants were transferred to LC-MS sample vials (Agilent, http://www.agilent.com, Santa Clara, CA) and stored at -20°C. They were then analyzed with a Waters Acquity UPLC system coupled to a hybrid quadropole time-of-flight (QTOF) Premier mass spectrometer (Waters, http://www.waters.com/, Milford, MA). A reverse-phase, 1.7-mm UPLC BEH C18, 2.1 3 150 mm column (Waters) was used for separations. The mobile phase consisted of eluent A (0.1% [v/v] acetic acid/water) and eluent B (acetonitrile), and separations were achieved using a linear gradient of 95% to 30% A over 30 min, 30% to 5% A over 3.0 min, and 5% to 95% A over 3.0 min. The flow rate was 0.56 mL min21, and the column temperature was maintained at 60°C. Masses of the eluted compounds were detected in the negative ESI mode from 50 to 2,000 mass-to-charge ratio. The QTOF Premier was operated under the following

instrument parameters: desolvation temperature of 400°C, desolvation nitrogen gas flow of 850 L h21, 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.

Ion List and Metabolomics Data Processing

Waters .raw data files were converted to .cdf file format, followed by metabolite data extraction, alignment, and exported using MET-IDEA software (Broeckling, Reddy et al. 2006). An ion list containing 463 ion/retention time pairs was used for the targeted metabolomics data analysis of the ecotype UPLC-ESI (-)-qTOF-MS biochemical phenotypes (Appendix_RII_Targeted_Ion_List). 133 of these ion/retention time pairs were annotated as sapogenin or saponin compounds. Annotations were based on authentic reference standards or spectral information from either source fragmentation or MS/MS (ESI-q-CID-TOF-MS) experimentation. The unknown pairs included in the ion list were identified with non-targeted MARKERLYNX analysis, and had m/z values and retention times in the same regions as the known and putative pairs and additionally showed statistically significant differential accumulation values among the ecotypes. In addition to the targeted analysis, *de novo* non-targeted analysis of all samples was performed using Waters MARKERLYNX software. Spectral abundance

signals for all metabolites in a separation were normalized to the internal standard (0. 018 mg/ml umbelliferone). Descriptive statistics were performed in Excel. One-way ANOVA was performed using a custom MATLAB script (MathWorks, http://www.mathworks.com/). Multivariate analyses including principal component analysis and hierarchical clustering were performed using JMP 5. 0 software (SAS, http://www.sas.com/).

Gross Phenotype Comparisons for Selection of Candidate Genes of Interest

These gross phenotype comparisons consisted of two separate steps. The first step was the application of a simple "state filter" to identify all probesets with expression values that corresponded to the phenotypic "states" from the experimental matrix. Recall that the matrix had three high "states" (high total saponin accumulator ecotypes) and one low "state" (low total saponin accumulator) for each of the organ types (see Figure 1_RII). For aerial organs, A17, GRC43, and R108 represented the high states, while ESP105 represented the low state. Thus, the state filter for aerial organs was simply the selection of all probesets for which ESP105 had the lowest expression (hybridization) value. In root organs, A17, ESP105, and R108 represented the high states while GRC43 represented the low state. The root organ state filter was therefore the selection of all probesets for which GRC43 had the lowest expression

(hybridization) value. In order to account for the possibility of negative regulation mechanisms of regulatory elements, the inverse of the filter was also applied to the regulatory element probesets for both organs(ESP105 as the high state in aerial organs and GRC43 as the high state in root organs). The second step used in the gross phenotype comparisons is less intuitive than the initial state filter. Briefly, EQUATION 1_RII was used to calculate a Phenotype Comparison Ranking Statistic "*f*" value which was used for the ranking of all state filtered probesets (see Discussion section for details). EQUATION 2_RII was used to calculate the Inverse Case Regulatory Element Gross Phenotype Comparison Ranking Statistic "*g*" for the inversely filtered regulatory element probesets. In addition to preparation of a master list of gross phenotype comparisons for all probesets for each of the organs types, probesets for GT, P450, and regulatory element probesets were isolated and prepared as separate lists. The list representing the inverse state for regulatory element probesets was also prepared as a separate sheet.

$$f = \sum_{i=1}^{3} (H_i - L) - \sum_{i=1}^{3} \frac{H_i}{3} - \max_{i=1,2,3} (H_i)$$

Equation 1_RII. Gross Phenotype Comparison Ranking Statistic "*f*".

Given the three expression values for a probeset (H1, H2, and H3) from the three different "high" state ecotype samples and one expression value for the probeset (L) from the "low" state ecotype from the experimental matrix, Equation 1_RII yields *f*.

$$g = \sum_{i=1}^{3} (L - H_i) - \sum_{i=1}^{3} \frac{H_i}{3} - L$$

Equation 2_RII. Inverse Case for Regulatory Element Probesets of the Gross Phenotype Comparison Ranking Statistic "*g*".

Given the three expression values for a probeset (H1, H2, and H3) from the three different "high" state ecotype samples and one expression value for the probeset (L) from the "low" state ecotype from the experimental matrix, Equation 2_RII yields "*g*". Recall that the inverse of the "state filter" was applied to regulatory element probesets in order to account for the possibility of negative regulation mechanisms, and that the "high" and "low" state designations refer to the original application of the state filter.

Hidden Markov Model Annotation

24 profile HMM models for cytochrome P450 proteins and 24 profile HMM models for glycosyltransferase proteins (Appendix_RII_HMM_models) were used to analyze the 61,278 tentative consensus sequences of the Affymetrix Medicago Gene Chip™ (Affymetrix, http://www.affymetrix.com/) with HMMER software (http://hmmer. janelia. org/). Profile models were obtained from SUPERFAMILY (http://supfam.

org/SUPERFAMILY/index. html). These *de novo* HMM results were concatenated with the cytochrome P450 and GT annotation results of an *M. truncatula* genomic sequence based BLAST annotation ||unpublished, ZHAO||.

Pearson Correlation Coefficients for Transcripts vs. for Metabolites Selection of Candidate Genes of Interest

Pearson correlation coefficient analysis was performed on a series of ([probeset] vs. [metabolite]) and ([genotype] vs. [organ]) permutations.

Bootstrap Analysis of Pearson Correlation Coefficients

A custom MATLAB (MathWorks, http://www.mathworks.com/) script was used to generate 90% bootstrapped confidence intervals and bootstrapped standard errors for Pearson's r for transcript vs. total saponin content correlations for high priority probesets (using 5000 iterations). The bootstrapping algorithm in the script was the "bbcorr" function (http://www.mathworks.com/matlabcentral/), which computes double block bootstrap (Lee and Lai 2009) percentile confidence intervals and bootstrap standard errors.

Additional Information

Accession Numbers

Currently found in Appendix_RII_Primers

Appendices

Appendix_RII_HMM_models - A list of the all P450 and GT profile models used in the HMM analysis of the tentative consensus sequences used to design the probesets of the Affymetrix Medicago Gene Chip.

Appendix_RII_Primers - Primer sequence information for all of the primers used in the qRT-PCR analysis of gene expression.

Appendix_RII_Targeted_Ion_List- The 463 Ion/Retention Time pairs used for the targeted metabolomics data analysis of the ecotype UPLC-ESI (-)-qTOF-MS data.

Appendix_RII_Gross_Phenotype_ Inverse _Examples- A list of the top 15 probesets resulting from the complete gross phenotype comparison selection process using the inverse state filter and EQUATION 2_RII for regulatory element probesets from root organs.

Supplemental Data

SUPP 1_RII - Results of de novo profile HMM analysis of the tentative consensus sequences used to design the probesets of the 61,278 Affymetrix Medicago Gene Chip™ (including scores, E-values, model IDs, and translation frame of tentative consensus sequence for all hits). This file also contains the “comprehensive”

annotation lists for both P450 and GT gene families from the concatenation of HMM and genome BLAST annotation results.

SUPP 2_RII - Data from the microarray analysis, including separated worksheets for 1.)all probes, 2.) probes selected as significantly ($p < 0.05$) different based on the associative t-test analysis of thediverse combinations of ecotypes, 3.) probes annotated as regulatory elements, 4.) GT probes from the concatenated annotation list, and 5.) P450 probes from the concatenated annotation list.

SUPP 3_RII- Data from the metabolomics analysis of the same samples used in the microarray analysis. It includes separated worksheets for 1.) aerial organ non-targeted MARKERLYNX results, 2.)root organ non-targeted MARKERLYNX results, 3.) aerial organ targeted MET-IDEA results with the full ion/retention time pair list, 4.) root organ targeted MET-IDEA results with the full ion/retention time pair list, 5.) aerial organ targeted MET-IDEA results with the saponin-only ion/retention time pairs, and 6.) root organ targeted MET-IDEA results with the saponin-only ion/retention time pairs.

SUPP 4_RII - Ranked results of the gross phenotype comparison selection process applied to aerial organ samples, presented in separate worksheets for 1.) All probesets, 2.)GTs, 3.)P450s, 4.) Regulatory Elements and 5.) inverse Regulatory Elements.

SUPP 5_RII - ranked results of the gross phenotype comparison selection process applied to root organ samples, separated as in SUPP 4_RII.

SUPP 6_RII - Pearson correlation coefficients for [all probesets] vs. [Each of the 462 ion/rt pairs, excluding internal standard] for both the A.) individual case (n = 24), and B.) averaged sample case (n = 8).

SUPP 7_RII - Pearson correlation coefficients for [all probesets] vs. [totaled saponin content] for the individual sample case (n = 24). Correlation coefficients are presented for three [ecotype] vs. [organ] combinations; 1.) [inter-ecotype, inter-organ], 2.)[inter-ecotype, aerial intra-organ], and 3.)[inter-ecotype, root intra-organ]. This file also contains separate worksheets which contain the above information for the "comprehensive" lists of A.) GT probesets, B.)P450 probesets, and C.)regulatory element probesets.

SUPP 8_RII - Pearson correlation coefficients for [all probesets] vs. [totaled saponin content] for the averaged sample case (n = 8), presented as in SUPP 7_RII.

SUPP 9_RII - Pearson correlation coefficients for [all probesets] vs. [the summed value of ion/rt pairs representing a particular sapogenin aglycone] for the individual sample case (n = 24). Summed accumulation values for particular sapogeninaglycones were prepared for hederagenin, bayogenin, zhanic acid, medicagenic acid,

putative_gypsogenin (“new aglycone”), and a combination of soyasapogenols B and E. Correlation coefficients are presented for three [ecotype] vs. [organ] combinations; 1.) [inter-ecotype, inter-organ], 2.)[inter-ecotype, aerial intra-organ], and 3.)[inter-ecotype, root intra-organ].

SUPP 10_RII - Pearson correlation coefficients for [all probesets] vs. [the summed value of ion/rt pairs representing a particular sapogenin aglycone] for the averaged sample case (n = 8), presented as in SUPP 9_RII.

SUPP 11_RII - Results from qRT-PCR analysis of the same samples used in the microarray and metabolomics analyses, presented with standard errors.

Sources

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin,. (2003). Dissecting plant secondary metabolism; constitutive chemical defences in cereals. New Phytologist **159,** 101-108.

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

Ballester, A. -R., Molthoff, J., de Vos, R., Hekkert, B. t. L., Orzaez, D., FernÃ¡ndez-Moreno, J. -P., Tripodi, P., Grandillo, S., Martin, C.,

Heldens, J., Ykema, M., Granell, A., and Bovy, A. (2011). Biochemical and Molecular Analysis of Pink Tomatoes: Deregulated Expression of the Gene Encoding Transcription Factor SlMYB12 Leads to Pink Tomato Fruit Color. Plant Physiology **152,** 71-84.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. Trends in Plant Science **9,** 418-425.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. Anal. Chem. **78,** 4334-4341.

Bucciarelli, B., Hanan, J., Palmquist, D., and Vance, C. P. (2006). A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development

10. 1104/pp. 106. 082594. Plant Physiol. **142,** 207-219.

Chen, J., Yu, J., Ge, L., Wang, H., Berbel, A., Liu, Y., Chen, Y., Li, G., Tadege, M., Wen, J., Cosson, V., Mysore, K. S., Ratet, P., MadueÃ±o, F., Bai, G.,

and Chen, R. Control of dissected leaf morphology by a Cys(2)His(2) zinc finger transcription factor in the model legume *Medicago truncatula*. Proceedings of the National Academy of Sciences **107,** 10754-10759.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W. -R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. Plant Physiol. **139,** 5-17.

Dixon, R. A., and Sumner, L. W. (2003). Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. Plant Physiology **131,** 878-885.

Dozmorov, I., and Centola, M. (2003). An associative analysis of gene expression array data. Bioinformatics **19,** 204-211.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol **48,** 155 - 171.

Goodacre, R., Roberts, L., Ellis, D., Thorogood, D., Reader, S., Ougham, H., and King, I. (2007). From phenotype to genotype: whole tissue profiling for plant breeding. Metabolomics **3,** 489-501.

Hannah, M. A., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A. R., and Willmitzer, L. (2010).combined Transcript and Metabolite Profiling of

Arabidopsis Grown under Widely Variant Growth Conditions Facilitates the Identification of Novel Metabolite-Mediated Regulation of Gene Expression. Plant Physiology **152,** 2120-2129.

Helliwell, C. A., Chandler, P. M., Poole, A., Dennis, E. S., and Peacock, W. J. (2001). The CYP88A cytochrome P450, ent-kaurenoic acid oxidase, catalyzes three steps of the gibberellin biosynthesis pathway. Proceedings of the National Academy of Sciences **98,** 2065-2070.

Hirai, M., Yano, M., Goodenowe, D., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proc Natl Acad of Sci USA **101,** 10205 - 10210.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research **31,** e15.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. Plant Molecular Biology **51,** 731-743.

Kalo, P., Gleason, C., Edwards, A., Marsh, J., Mitra, R. M., Hirsch, S., Jakab, J. l., Sims, S., Long, S. R., Rogers, J., Kiss, G. r. B., Downie, J. A., and Oldroyd, G. E. D. (2005). Nodulation Signaling in Legumes Requires NSP2, a Member of the GRAS Family of Transcriptional Regulators. Science **308,** 1786-1789.

Krueger, S., Giavalisco, P., Krall, L., Steinhauser, M. -C., BÃ¼ssis, D., Usadel, B., FlÃ¼gge, U. -I., Fernie, A. R., Willmitzer, L., and Steinhauser, D. (2011). A Topological Map of the Compartmentalized <italic>Arabidopsis thaliana</italic> Leaf Metabolome. PLoS ONE **6,** e17806.

Lahoucine Achnine, D. V. H., Mohamed A. Farag, Lloyd W. Sumner, Jack W. Blount, Richard A. Dixon,. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. The Plant Journal **41,** 875-887.

Lee, S. M. S., and Lai, P. Y. (2009). Double block bootstrap confidence intervals for dependent data. Biometrika **96,** 427-443.

Li, L., Cheng, H., Gai, J., and Yu, D. (2007). Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*. Planta **226,** 109-123.

Lu, C. D., and Jorgensen, N. A. (1987). Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. The Journal of Nutrition **117,** 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M., and Jorgensen, N. A. (1987). Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. Journal of dairy science **70,** 799-805.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., Sakurai, T., Shimada, Y., and Saito, K. (2010). AtMetExpress Development: A Phytochemical Atlas of Arabidopsis Development. Plant Physiology **152,** 566-578.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., and Dixon, R. A. (2010). Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*. Plant Cell **22,** 850-866.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., and Osbourn, A. E. (1999). Compromised Disease Resistance in Saponin-Deficient Plants. Proceedings of the National Academy of Sciences of the United States of America **96,** 12923-12928.

Parmigiani, G., Garrett, E., Irizarry, R., Zeger, S., Li, C., and Wong, W. (2003). DNA-Chip Analyzer (dChip). In The Analysis of Gene Expression Data, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, and W. Wong, eds (Springer London), pp. 120-141.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RTPCR. Nucleic Acids Research **29,** e45.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M. H., Karlen, Y., Bakker, O., van den Hoff, M. J. B., and Moorman, A. F. M. (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research **37,** e45.

Saito, K., Hirai, M. Y., and Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics - `majority report by precogs'. Trends in Plant Science **13,** 36-43.

Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E. -T., Tschoep, H., Steinhauser, M. C., Guenther, M., Hoehne, M., Rohwer, J. M., Altmann, T., Fernie, A. R., and Stitt, M. (in press). Network Analysis of Enzyme Activities and Metabolite Levels and Their

Relationship to Biomass in a Large Panel of Arabidopsis Accessions. The Plant Cell Online.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., and Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. The Plant Journal **32,** 1033-1048.

Tohge, T., and Fernie, A. R. (2010).combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat. Protocols **5,** 1210-1227.

Tohge, T., Nishiyama, Y., Hirai, M., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D., and Kitayama, M. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J **42,** 218 - 235.

Udvardi, M., Kakar, K., Wandrey, M., Montanari, O., Murray, J., Andriankaja, A., Zhang, J., Benedito, V., Hofer, J., Chueng, F., and Town, C. (2007). Legume transcription factors: global regulators of plant development and response to the environment. Plant Physiol **144,** 538 - 549.

Chapter III - Enzymatic Characterization of CYP72A67 and CYP72A68, Two Cytochrome P450 enzymes in the Triterpene Sapogenin Biosynthetic Pathway of *Medicago truncatula*.

Authors: John H. Snyder, David V. Huhman, Bennie J. Bench, and Lloyd W. Sumner.

Summary:

This chapter will detail characterization experiments for 5 cytochrome P450 gene candidates that emerged from chapters I and II. *In vitro* enzymology and mutant genetics experimental avenues will be covered. Particular emphasis will be placed on the functions of CYP72A67 and CYP72A68, with demonstrated broad substrate tolerance. The content presented in this chapter will be the primary focus of the main publication of my degree project.

Abstract:

The model legume *Medicago truncatula* is known to accumulate a large variety of triterpene saponin compounds, resulting from the differential glycosylation of at least seven triterpene aglycone structures. Previous metabolomics analyses of a large

germplasm diversity (ecotype) collection revealed substantial metabolic diversity in triterpene saponin accumulation both within the various ecotypes and between the root and aerial organs of individual ecotypes. The metabolomics phenotyping results enabled the informed selection of specific ecotypes for an experimental matrix of hypo- and hyper- saponin accumulating ecotypes and organs. The global transcript expression dynamics of the ecotypes and organs in the matrix were profiled with microarrays, and candidate genes for triterpene saponin biosynthesis were chosen based on obvious relationships between saponin content and transcript accumulation. In particular five high priority cytochrome P450 genes (cyp72a67, cyp72a68, cyp83g1, cyp88d3, and cyp716a12) were selected for detailed characterization. Multiple *Tnt-1* insertion mutagenesis lines for each of the five loci were identified. These five genes were cloned and introduced to the *Wat11* yeast strain for recombinant expression and in vitro enzymatic analysis. Results from a series of microsomal assays with a variety of potential substrates demonstrate that CYP72A67 is a multisubstrate oleanate sapogenin carbon 2 oxidase enzyme, and CYP72A68 is a multifunction, multisubstrate oleanate sapogenin carbon 23 sequential oxidase enzyme.

Introduction

Triterpene saponins are a structurally diverse class of compounds with a wide taxonomic distribution and a broad range of biological activities (Augustin, Kuzina et al. 2011). Although primarily found in dicots and especially legumes, triterpene saponins have also been isolated from selected monocots such as oat and barley (Papadopoulou, Melton et al. 1999; Anne E. Osbourn 2003). Triterpene saponins represent the primary antinutritive compounds in livestock fodder (Lu and Jorgensen 1987; Lu, Tsai et al. 1987). These antinutritive properties restrict the optimum utility of high-protein legumes as livestock feed and limit the ultimate economic potential of forage legumes. A sophisticated molecular and biochemical understanding of saponin biosynthesis would enable the metabolic engineering of triterpenoid biosynthesis. For example, specific antimicrobial saponins could be engineered in roots to provide antimicrobial properties while simultaneously reducing saponin content in aerial tissues would improve nutritional content (Dixon and Sumner 2003).

Structurally, triterpene saponins are composed of a lipid-soluble triterpenoid aglycone conjugated with various water soluble sugar residues. Sterol and triterpenoid sapogenin (saponin aglycones) biosyntheses in legumes begin with a common isopentenyl pyrophosphate (IPP) precursor synthesized via the cytosolic mevalonic

acid (MVA) and/or plastid localized methylerythritol (MEP) pathways. The progressive condensation of isoprene units leads to various mono, sesqui, di, and triterpenoids. The triterpene oxidosqualene is cyclized by two enzymes resulting in two branched pathways. Cycloartenol synthase is the first committed step in sterol biosynthesis, whereas, ß-amryin synthase is the first committed step in triterpene saponin biosynthesis (Augustin, Kuzina et al. 2011). Squalene synthase and squalene epoxidase have been previously characterized in *M. truncatula* (Suzuki, Achnine et al. 2002; Iturbe-Ormaetxe, Haralampidis et al. 2003). Very little is known about the remaining enzymatic steps following ß-amyrin synthase in triterpene saponin biosynthesis. Oxidation of six different alkyl carbons (2,16,22,23,24,28) of ß-amyrin yield at least seven empirically determined aglycone structures found in *Medicago sp.* (Augustin, Kuzina et al. 2011). These oxidative reactions are likely catalyzed by cytochrome P450 enzymes. Elucidation of the molecular and biochemical mechanisms for these enzymatic oxdations of sapogenin compounds is the primary focus of this study, as enzymes responsible for these oxidations have not been characterized in *M. truncatula* to date. Oxidation of alkyl carbon 24 and carbon 11 of ß-amyrin have been demonstrated for CYP93E1 from *Glycine max*(Shibuya, Hoshino et al. 2006)and CYP88D6 from *Glycyrrhiza uralensis*(Seki, Ohyama et al. 2008),

respectively. In addition to the oxidation of alkyl carbons of ß-amyrin, a series of glycosyltransferases (GTs) are also necessary to conjugate the diversity of aglycone structures for saponin biosynthesis. Recently, GTs have been characterized in Medicago: UGT73K1 with specificity for hederagenin and soyasapogenols B and E, and UGT71G1 with specificity for medicagenic acid (Lahoucine Achnine 2005; Naoumkina, Modolo et al. 2010). A large number of additional GTs still remain uncharacterized.

Ecotype Metabolomics

Metabolomics analysis of a large germplasm diversity (ecotype) collection revealed substantial metabolic diversity in triterpene saponin accumulation both within the various ecotypes, and between the root and aerial organs of individual ecotypes. The metabolomics phenotyping results enabled the informed selection of specific ecotypes for an experimental matrix of hypo- and hyper- saponin accumulating ecotypes and organs. The ecotype ESP_105 was selected as the lowest total saponin accumulator in aerial organs, but it had potential additional value as a resource because it was also very high (top 20) total accumulator in root organs. Likewise, the ecotype GRC_43 was selected as the lowest total accumulator in root organs but was also an extremely high (top 10) total accumulator in aerial organs. The popular isolines A17 and R108

were selected as reference ecotypes with relatively high total saponin accumulation in both aerial and root organs, due to primarily to their role in the development of genomics and mutant population resources for research in *M. truncatula*.

Ecotype Microarray

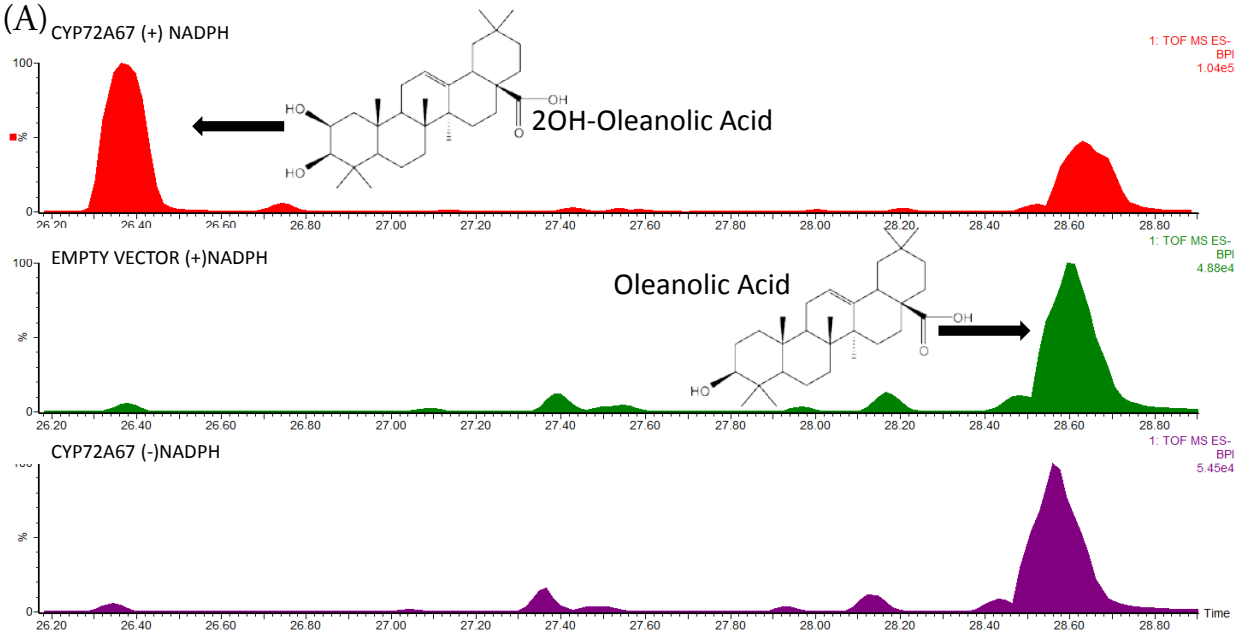
The global transcript expression dynamics of the ecotypes and organs in the matrix were profiled with microarrays, and candidate genes for triterpene saponin biosynthesis based on obvious relationships between saponin content and transcript accumulation. In particular five high priority cytochrome P450 genes (cyp72a67, cyp72a68, cyp83g1, cyp88d3, and cyp716a12) were selected for detailed characterization. As an example, the expression level of the transcript for cyp72a68 in ecotype ESP105 aerial organ samples was 36-fold less than GRC43 aerial, 29-fold less than A17 aerial, and 22-fold less than R108 aerial, which is extremely similar to the total saponin accumulation phenotypes for aerial organs in these ecotypes. Additionally, the expression level of this transcript in ESP105 root organ samples was second highest of all ecotypes, consistent with the observation that ESP105 root organs accumulate very high levels of total saponins. As a second example, the expression level of the transcript for cyp72a67 in ecotype ESP105 aerial organ samples was 47-fold less than GRC43 aerial, 42-fold less than A17 aerial, and 7-fold less than

R108 aerial. The highest expression level of this transcript was found in ESP105 root organ samples. Transcripts for cyp83g1, cyp88d3, and cyp716a12 showed similar expression dynamics, and were therefore prioritized for further molecular and biochemical functional analyses.

Results

In Vitro Enzymatic Assays of CYP72A67 with Standards asSubstrates

When oleanolic acid was used as a substrate (FIGURE 1_RIII), 2-OH oleanolic acid was detected a product in the (+)CYP72A67 microsomal samples and not detected in the empty vector control samples. Additionally, the amount of oleanolic acid detected was higher in the empty vector control samples.



(B)

Ion/Retention Time Pair	Oleanolic Acid		
	CYP72A67 (+) NADPH	EMPTY Vector	CYP72A67 (-) NADPH
Oleanolic Acid (455.35 m/z,28.63 min)	44867 ± 2321	49567 ± 1027	53900 ± 379
2-OH Oleanolic Acid (471.35 m/z, 26.36 min)	97000 ± 3507	67 ± 10	328 ± 90

FIGURE 1_RIII. CYP72A67-Mediated Biosynthesis of 2-OH Oleanolic Acid from Oleanolic Acid.

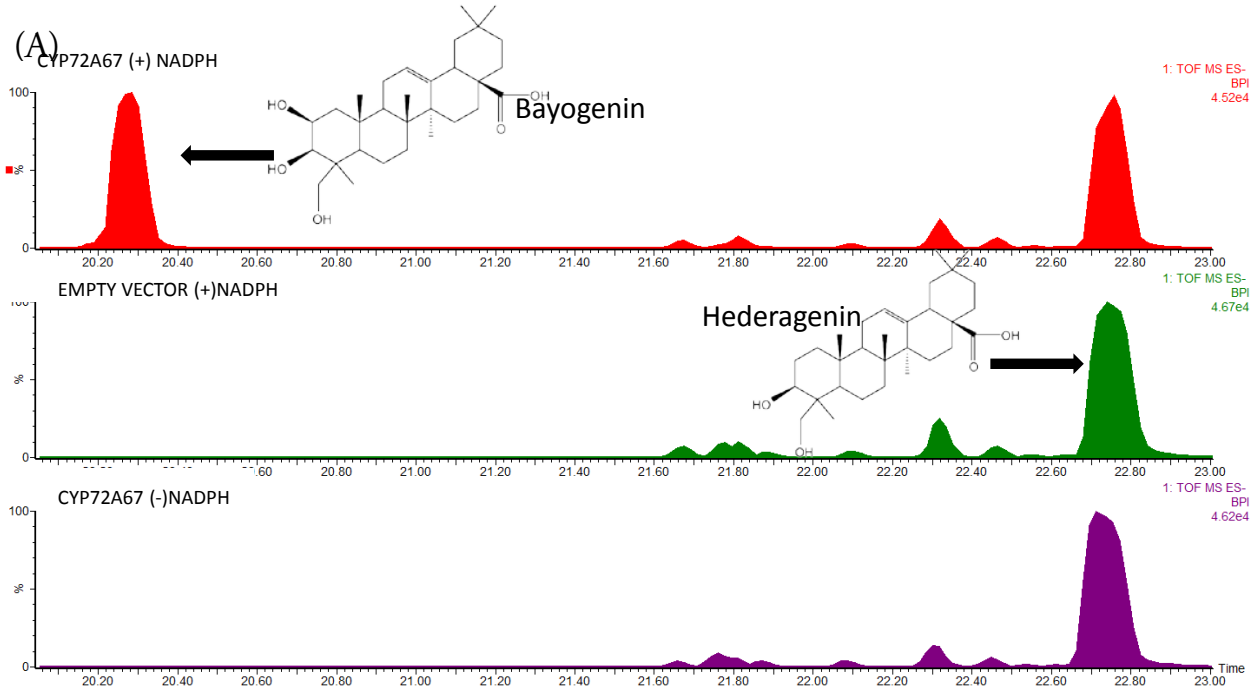
(A) Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for CYP72A67 (+) NADPH, Empty Vector (+) NADPH, and CYP72A67 (-)NADPH samples, with structures for the substrate and product.

(B) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

When hederagenin was used as a substrate (FIGURE 2_RIII), bayogenin was detected a product in the (+)CYP72A67 samples and not detected in the empty vector control

samples. The amount of hederagenin detected was higher in the empty vector control samples.

(A)



(B)

Ion/Retention Time Pair	Hederagenin		
	CYP72A67 (+) NADPH	EMPTY Vector	CYP72A67 (-) NADPH
Hederagenin (471.35 m/z, 22.76 min)	42033 ± 1299	47967± 1027	47266 ± 1168
Bayogenin (487.34 m/z, 20.28 min)	45433 ± 448	326 ± 4	326 ± 9

FIGURE 2_RIII. CYP72A67-Mediated Biosynthesis of Bayogenin from Hederagenin.

(A) Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for CYP72A67 (+) NADPH, Empty Vector (+)NADPH, and CYP72A67 (-)NADPH samples, with structures for the substrate and product.

(B) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

When β-amyrin or erythrodiol were used as substrates in CYP72A67 assays, no products were detected (data not shown).

CYP72A67 (+/-) NADPH Assays

Oleanolic acid substrate assays of CYP72A67 microsomes both with and without NADPH showed accumulation of2-OH oleanolic acid in the (+)NADPH samples, but not the (-)NADPH control samples(FIGURE 1_RIII). Hederagenin substrate assaysof CYP72A67 microsomes both with and without NADPH showed accumulation of bayogenin in the (+)NADPH samples, but not in the (-)NADPH control samples (FIGURE 2_RIII).

CYP72A67 Assays withAglycone Mix as Substrate

A solution of partially purified aglycones obtained through acid hydrolysis of saponin extracts obtained from *M. truncatula* root tissues was also tested as a substrate(mixture) with (+)CYP72A67 and (-)CYP72A67 (empty vector control) assays (TABLE 1_RIII). 2-OH oleanolic acid, bayogenin, putative polygalagenin, and medicagenic acid were detected in higher amounts in the (+)CYP72A67 samples than in the empty vector controls. Oleanolic acid, hederagenin, putative gypsogenin, and putative gypsogenic acid were detected in higher amounts in the empty vector control samples than in the (+)CYP72A67 samples.

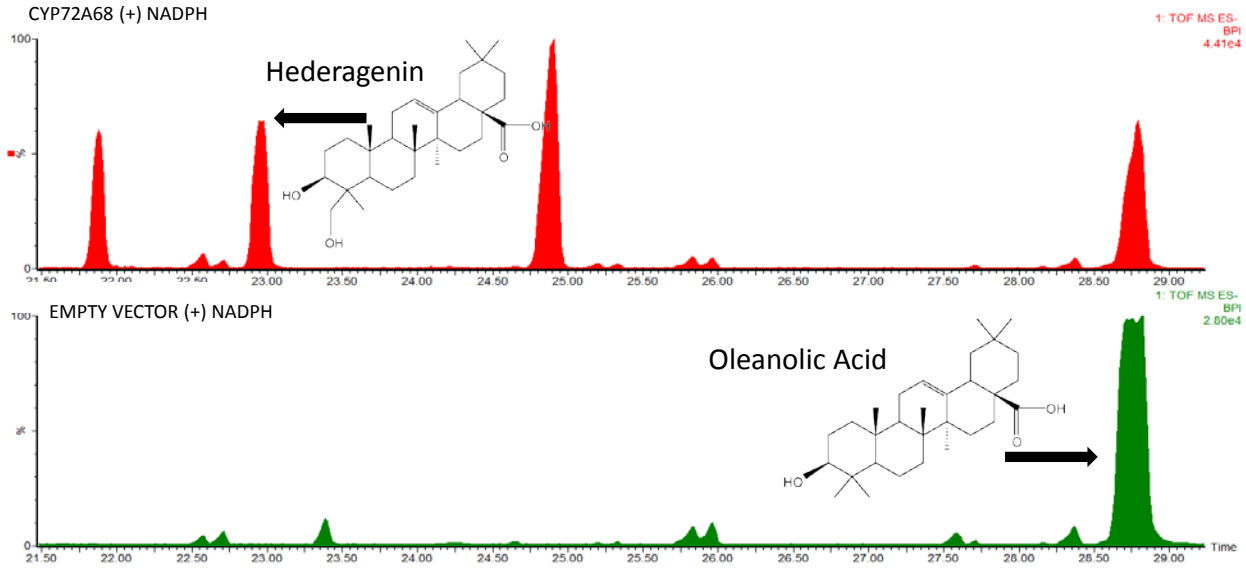
	Aglycone Mixture	
Ion/Retention Time Pair	CYP72A67	Empty vector
Oleanolic Acid (455.35 m/z, 28.47 min)	5977 ± 521	10524 ± 1269
Hederagenin (471.34 m/z,22.68 min)	5182 ± 533	19988 ± 2035
putative_Gypsogenin (469.33 m/z, 24.60 min)	269 ± 20	903 ± 146
putative_Gypsogenic Acid (485.33 m/z, 21.60 min)	287 ± 30	521 ± 222
2OH-Oleanolic Acid (471.34 m/z, 26.22 min)	14268 ± 1513	6326 ± 657
Bayogenin (487.35 m/z, 20.22 min)	30160 ± 3780	18794 ± 2354
putative_Polygalagenin (485.33 m/z, 22.42 min)	11565 ± 1307	9841 ± 1317
Medicagenic Acid (501.32 m/z, 19.43 min)	26816 ± 2737	23665 ± 2685

TABLE 1_RIII. CYP72A67-Mediated Production and Consumption of Diverse OleanateSapogenins from the Aglycone Mixture.

Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards. Shaded values highlight the assay condition with the higher detection value for each of the Ion/Retention Time Pairs.

In Vitro Enzymatic Assays of CYP72A68 with Standards as Substrates

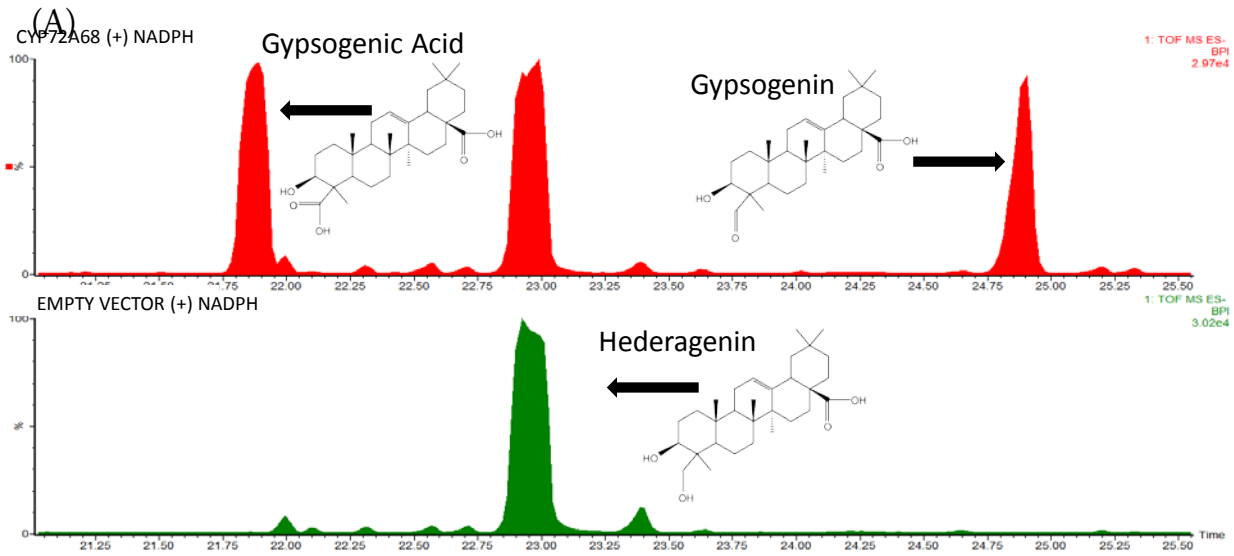
When oleanolic acid was used as a substrate (FIGURE 3_RIII), hederagenin, putative gypsogenin, and putative gypsogenic acid were detected as products in the CYP72A68 assays and not detected in the empty vector control samples. Additionally, the amount of oleanolic acid detected was higher in the empty vector control samples.



Ion/Retention Time Pair	CYP72A68	Empty Vector
Oleanolic Acid (455.35 m/z, 28.8 min)	19457 ± 465	28716 ± 381
Hederagenin (471.35m/z, 22.94 min)	15509 ± 480	n.d.
putative_Gypsogenin (469.332 m/z, 24.91 min)	22753 ± 783	n.d.
putative_Gypsogenic Acid (485.327 m/z, 21.88 min)	11156 ± 610	n.d.

FIGURE 3_RIII. CYP72A68-Mediated Biosynthesis of Hederagenin from Oleanolic Acid.
(A) Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for CYP72A68 and Empty Vector samples, with structures for the substrate and product.
(B) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

When hederagenin was used as a substrate (FIGURE 4_RIII), putative gypsogenin, and putative gypsogenic acid were detected as products in the CYP72A68 assays. They were not detected in the empty vector control samples. The amount of hederagenin detected was higher in the empty vector control samples.



(B) Ion/Retention Time Pair	Hederagenin	
	CYP72A68	Empty Vector
Hederagenin (471.35 m/z, 22.92 min)	19789 ± 538	21530 ± 271
putative_Gypsogenin (469.33 m/z, 24.91 min)	10624 ± 662	n.d.
putative_Gypsogenic Acid (485.33 m/z, 21.89 min)	16247 ± 406	n.d.

FIGURE 4_RIII. CYP72A68-Mediated Biosynthesis of Putative Gypsogenin and Putative Gypsogenic Acid from Hederagenin.
(A) Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for CYP72A68 and Empty Vector samples with structures for the substrate and products.
(B) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

Assays using bayogenin as a substrate (FIGURE 5_RIII) showed accumulation of medicagenic acid and putativepolygalagenin as products.

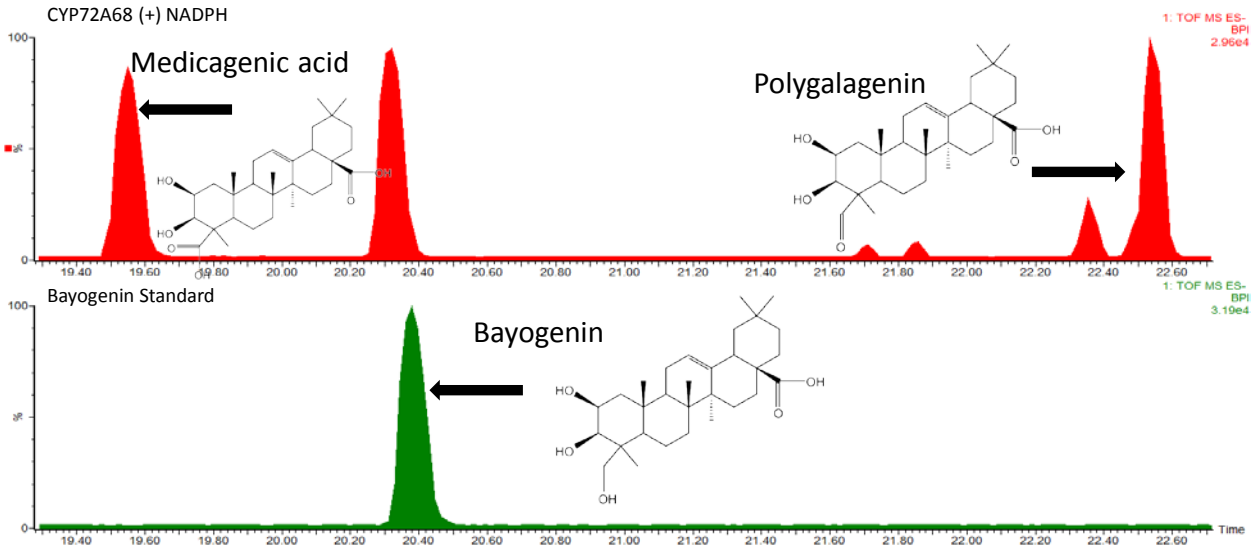


FIGURE 5_RIII. CYP72A68-Mediated Biosynthesis of Medicagenic Acid and Putative Polygalagenin from Bayogenin.

Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for a CYP72A68 sample and an analysis of thebayogenin authentic reference standard , with structures for the substrate and products.

When β-amyrin, cycloartenol, or erythrodiol were used as substrates in CYP72A68 assays, no products were detected. Additionally, genistein, daidzein, formononetin, and 7,4'-Dihydroxyflavone were tested as substrates in CYP72A68 assays, and no products were detected.

CYP72A68 (+/-) NADPH Assays

Oleanolic acid substrate assays of CYP72A68 microsomes with and without NADPH showed accumulation of hederagenin, putative gypsogenin, and putative gypsogenic acid in the (+)NADPH samples but not the (-)NADPH control samples (TABLE 2_RIII).

	Oleanolic Acid (+) & (-) NADPH	
Ion/Retention Time Pair	CYP72A68 (+) NADPH	CYP72A68 (-) NADPH
Oleanolic Acid (455.35 m/z, 28.42 min)	9681 ± 1047	13256 ± 113
Hederagenin (471.35 m/z, 22.63min)	10783 ± 492	n.d.
putative_Gypsogenin (469.33 m/z, 24.53 min)	13413 ± 525	n.d.
putative_Gypsogenic Acid (485.33 m/z, 21.88 min)	6822 ± 891	n.d.

TABLE 2_RIII. Necessity of NADPH for CYP72A68 Catalytic Function.

Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates of CYP72A68 (+) NADPH or CYP72A68 (-) NADPH assay conditions.

CYP72A68Assays with Aglycone Mix as Substrate

(+)CYP72A68 and empty vector control samples were also assayed with the solution of partially purified aglycones from *M. truncatula* roots (TABLE 3_RIII). Putativegypsogenin, putativegypsogenic acid, and medicagenic acid were detected in higher levels in the (+)CYP72A68 samples than in the empty vector controls. Oleanolic acid, 2-OH oleanolic acid, hederagenin, bayogenin, and putative polygalagenin were detected at higher levels in the empty vector control samples than in the (+)CYP72A68 samples.

	Aglycone Mixture	
Ion/Retention Time Pair	CYP72A68	Empty vector
Oleanolic Acid (455.35 m/z, 28.79 min)	4957 ± 120	6448 ± 383
2OH-Oleanolic Acid (471.34 m/z, 26.54 min)	1386 ± 44	3303 ± 314
Hederagenin (471.34 m/z,22.95 min)	7680 ± 112	8155 ± 350
Bayogenin (487.35 m/z, 20.49 min)	2699 ± 43	7489 ± 257
putative_Polygalagenin (485.33 m/z, 22.70 min)	1660 ± 170	4273 ± 408
putative_Gypsogenic Acid (485.33 m/z, 21.85 min)	731 ± 65	50 ± 50
putative_Gypsogenin (469.33 m/z, 24.89 min)	3417 ± 191	352 ± 21
Medicagenic Acid (501.32 m/z, 19.69 min)	8477 ± 318	6989 ± 286

TABLE 3_RIII. CYP72A68-Mediated Production and Consumption of Diverse OleanateSapogenins from the Aglycone Mixture.

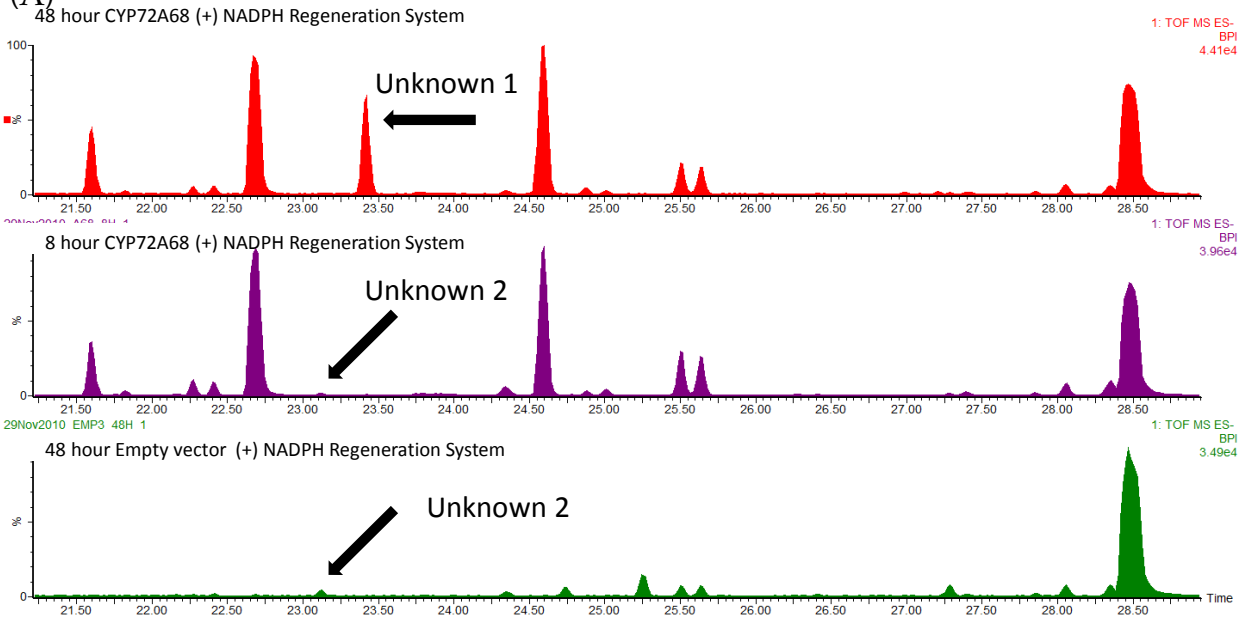
Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates per assay condition. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

Shaded values highlight the assay condition with the higher detection value for each of the Ion/Retention Time Pairs.

CYP72A68 Time Series with NADPH Regeneration System

Oleanolic acid (+) CYP72A68 and empty vector assays with an expanded time domain and an NADPH regeneration system. Microsomal preparations of (+) CYP72A68 and emptyvector control samples were assayed with an expanded time domain and an NADPH regeneration system using oleanolic acid as a substrate (FIGURE 6_RIII). Accumulation of hederagenin, putative gypsogenin, and putative gypsogenic acid were similar to the initial oleanolic acid assays. However, a new product (unknown 1) accumulated in the 48 hour (+)CYP72A68 samples, but not in the 8 hour (+)CYP72A68 or empty vector control samples. An additional compound (unknown 2) was detected in higher amounts in the 8 hour (+)CYP72A68 andempty vector control samples.

(A)



(B)

Ion/Retention Time Pair	48 Hour Oleanolic Acid Assay with NADPH regeneragtion system	
	48 hour CYP72A68	48 hour Empty vector
Oleanolic Acid (455.35 m/z, 28.47 min)	30798 ± 5171	39453 ± 8259
Hederagenin (471.34 m/z, 22.68 min)	25352 ± 3772	105 ± 45
putative_Gypsogenin (469.33 m/z, 24.59 min)	5659 ± 1629	17 ± 4
putative_Gypsogenic Acid (485.33 m/z, 21.6min)	18731 ± 3303	142 ± 31
Unknown 1 (469.33 m/z, 23.42 min)	10994 ± 1253	25 ± 4
Unknown 2 (471.34 m/z, 23.12 min)	266 ± 58	832 ± 185

(C)

Ion/Retention Time Pair	8 and 48 Hour Oleanolic Acid Assay with NADPH regeneragtion system	
	48 hour CYP72A68	8 hour CYP72A68
Unknown 1 (469.33 m/z, 23.42 min)	10994 ± 1253	307 ± 89
Unknown 2 (471.34 m/z, 23.12 min)	266 ± 58	798 ± 285

FIGURE 6_RIII. Accumulation of an Unknown Compound in CYP72A68 Expanded Time Series Oleanolic Acid Substrate Assay with NADPH Regeneration System.

(A) Representative UPLC-ESI-qTOF-MS base peak intensity chromatograms for 48 hour CYP72A68 (+) NADPH Regeneration System, 8 hour CYP72A68 (+) NADPH Regeneration System, and 48 hour Empty Vector (+)NADPH Regeneration System samples, with arrows indicating the unknown peaks of interest.

(B) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates for the 48 hour CYP72A68 and Empty

Vector samples. Ion/Retention Time Pairs in **bold** represent compounds identified via comparison with authenticated reference standards.

(C) Values in the table represent the mean of normalized areas (and 1 standard error) for each of the Ion/Retention Time pairs, from three biological replicates for the 48 hour and 8 hour CYP72A68 samples.

In Vitro Enzymatic Assays of CYP83G1, CYP88D3, and CYP716A12

When β-amyrin, erythrodiol, oleanolic acid, hederagenin, or the partially purified aglycone solution from *M. truncatula*roots were assayed as substrates for CYP83G1, CYP88D3, or CYP716A12,no products were detected (data not shown).

Genomics

Of the five loci investigated in this study, only cyp83g1 (Medtr5g072930. 1,1786 bp) was included in the genome sequence of *M. truncatula* (http://www.medicagohapmap.org/?genome). Amplification of cyp72a68 and cyp72a67, and cyp88d3 from genomic DNA showed loci size of approximately 2. 5, 3, and 1. 5 kb, respectively (data not shown).

Tnt-1 Mutant Collection Screening

Results of both the *Tnt-1* flanking sequence database BLAST screen and the reverse screen of the pooled *Tnt-1*germplasm are presented in TABLE 4_RIII. 138 individual plants for *Tnt-1*-insertion lineNF1698 and 6 individuals of NF12169 (both

representing the cyp72a68 locus) were genotyped, and no homozygous insertion *Tnt-1_cyp72a68* plants were identified. Genomic DNA from a heterozygous cyp72a68/*Tnt-1_cyp72a68* NF1698 plant was used as a template for cloning/confirmation of the insertion. Ten individual plants for *Tnt-1*-insertion line NF5264 and 6 individuals of NF13243 (both representing the cyp72a67 locus) were genotyped, and no homozygous insertion *Tnt-1_cyp72a67* plants were identified. Seven individual plants for *Tnt-1*-insertion lineNF14380 (representing the cyp83g1 locus) were genotyped, and a confirmed homozygous insertion *Tnt-1_cyp83g1* plant was identified. Metabomolics analysis of R108 (wild type) and NF1698 heterozygous cyp72a68/*Tnt-1_cyp72a68* plants did not reveal differences in triterpene saponin accumulation (data not shown).

FST BLAST

Target Locus	Database Hit	NF *Tnt-1* Insertion Line ID
cyp72a67	>NF5264-Insertion-2 NF5264	NF5264
cyp72a68	>NF1698-Insertion-4 NF1698	NF1698

(B) REVERSE SCREEN

Target Locus	Primer Combination	Insertion Site in Locus	NF *Tnt-1* Insertion Line ID

cyp72a67	A67-F + Tnt1-F	at ~ 0. 8 kb	NF13243
cyp72a68	A68-F + Tnt1-F	at ~1. 2kb at intron	NF12169
cyp83g1	CYP83-F + Tnt1-F	at 400 bp	NF10345
	CYP83-F + Tnt1-F	at 467 bp	NF14380
cyp88d3	CYP88-F + Tnt1-R	at 990 bp reverse	NF10938
	CYP88-F + Tnt1-R	at 1090 bp reverse	NF12871
	CYP88-F + Tnt1-R	at ~1170 bp reverse	NF10044
cyp716a12	CYPA12-F + Tnt1-R	at 1340 bp reverse	NF11197
	CYPA12-F + Tnt1-R	at 405 bp	NF3726

TABLE 4_RIII. *In Silico* and Reverse Genetic Screening Results for *Tnt-1* Insertion Mutants for All Candidate Loci.

(A) *In silico* screening results (hits) and *Tnt-1* insertion line identification numbers from the BLAST analysis of candidate gene sequences as a query against the *Tnt-1* flanking sequence tag database.

(B) Successful primer combinations, insertion site in target loci, and *Tnt-1* insertion line identification numbers for the reverse genetic screen of the *Tnt-1* mutant collection.

Molecular Genetics

The expression data forcyp72a67, cyp72a68, cyp83g1,cyp716a12, and cyp88d3 transcripts in a variety of plant organs and developmental stages (FIGURE 7_RIII) were obtained from The Medicago Gene Expression Atlas web server (Benedito, Torres-Jerez et al. 2008; He, Benedito et al. 2009). Note that the highest transcript accumulation for probesets representing cyp72a67, cyp72a68, and cyp716a12 were in

late developing (24 days after pollination) seed organs. Also note that the expression level for these transcripts is approximately 3 fold higher in the 24 days after pollination organ sample compared to the next highest organ sample.

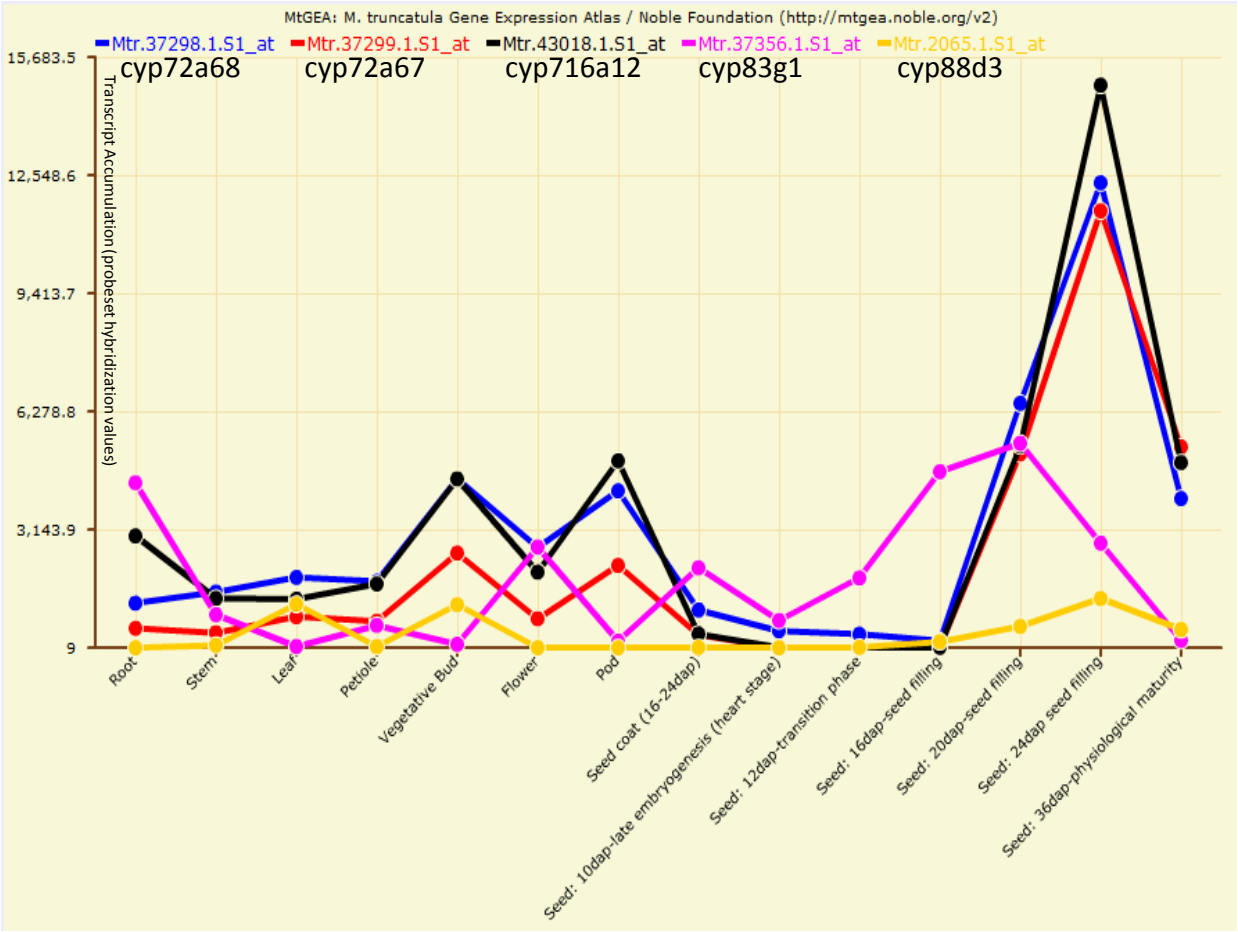


FIGURE 7_RIII. Expression Dynamics for Transcripts of Candidate Genes in Diverse Plant Organs.

Medicago Gene Expression Atlas accumulation data for transcripts of cyp72a68 (blue), cyp72a67 (red), cyp716a12 (black), cyp83g1 (pink), and cyp88d3 (orange), in root, stem, leaf, petiole,

vegetative bud, flower, pod, and seed coat organs, and seed developmental stages of 10, 12, 16, 20, 24, and 36 days following pollination.

Discussion

CYP72A67 Activity

Microsomal assays with reference standards as substrates demonstrate that CYP72A67 catalyzes the oxidation of carbon 2 of both oleanolic acid and hederagenin, yielding the products 2-OH oleanolic acid and bayogenin, respectively. In each of these reactions the abstracted hydrogen atom from oxidized carbon 2 is replaced with a hydroxyl group, and both products show the requisite mass increase of 16 Daltons. Microsomal assays with the partially purified aglycone mix as substrates demonstrate broader substrate tolerance/additional enzymatic activity for CYP72A67. Compounds lacking hydroxyl groups at carbon 2 (oleanolic acid, hederagenin, putative gypsogenin, and putative gypsogenic acid) were all detected in higher amounts in the empty vector control samples, indicating their consumption as substrates. Products with the characteristic 16 Dalton mass increase (2-OH oleanolic acid, bayogenin, putative polygalagenin, and medicagenic acid) were detected in higher amounts in the (+)CYP72A67 samples. These results indicate that compounds with carbon 23 methyl,

carbon 23 hydroxyl, carbon 23 carbonyl, and carbon 23 carboxylic acid substitutions are substrates for CYP72A67-mediated carbon 2 oxidation (See FIGURE 8_RIII).

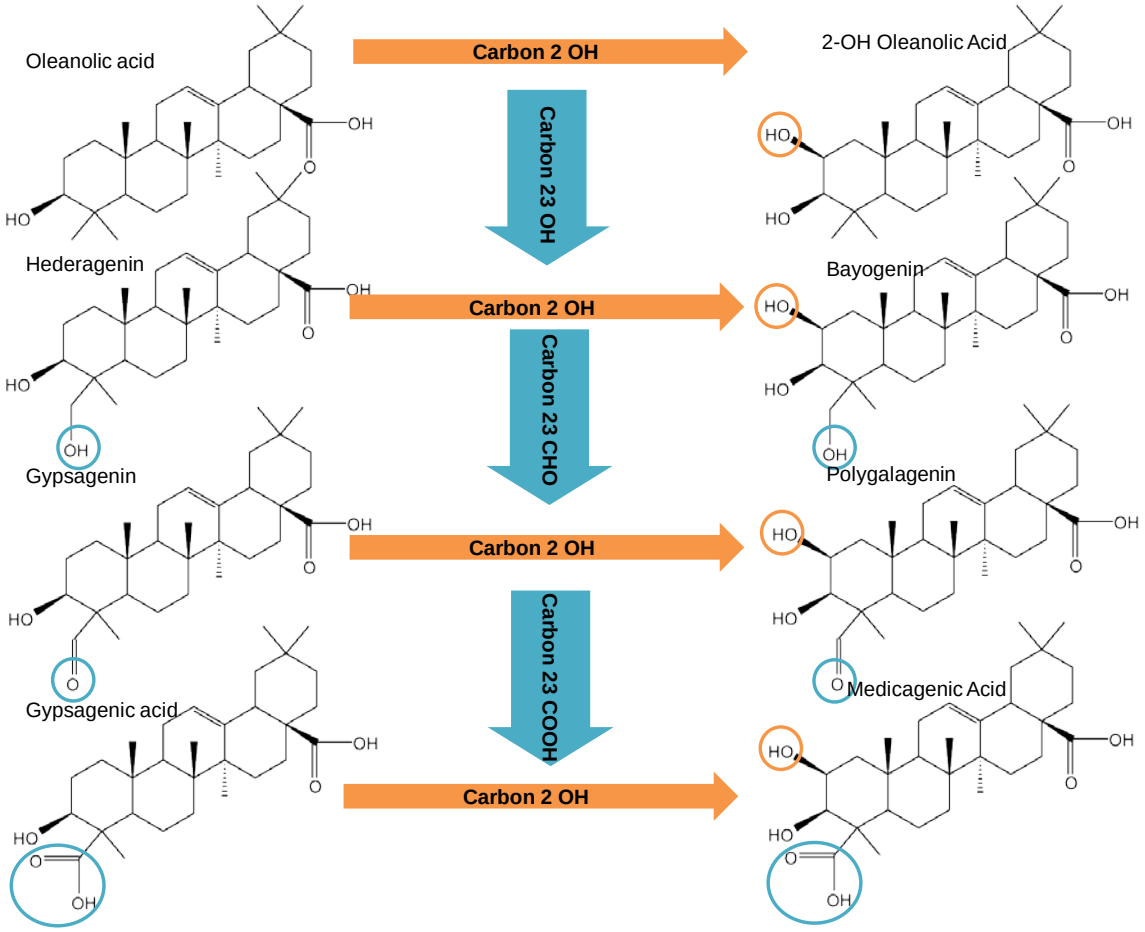


FIGURE 8_RIII. Summary Matrix of CYP72A67 and CYP72A68-Mediated Biosynthetic Reactions in the Oleanate Branch of the *M. truncatula* Sapogenin Biosynthesis Pathway.

Orange arrows pointing from left to right indicate the CYP72A67-Mediated oxidation of carbon 2 of the four substrates on the left hand side.

Blue arrows pointing from top to bottom indicate the multistep CYP72A68-Mediated sequential oxidation of carbon 23 of the various substrates.

Additional support for demonstration of CYP72A67-mediated carbon 2 oxidation activity in these assays is shown by the lack of product accumulation in assays lacking NADPH. Lack of product accumulation in these assays indicates that the CYTOCHROME P450 REDUCTASE/CYP72A67 complex require NADPH as an electron donor for activity (Chang-Jun Liu 2003; Seki, Ohyama et al. 2008). No products were detected when CYP72A67 was assayed with ß-amyrin or erythrodiol. This implies that compounds with a methyl carbon 28(ß-amyrin) or carbon 28 hydroxyl group (erythrodiol) are not substrates for CYP72A67-mediated carbon 2 oxidation as tested in this experimental system.

CYP72A68 Activity

Previous studies have shown that an individual cytochrome p450 enzyme can catalyze the sequential/consecutive oxidation of a given carbon,yielding the hydroxyl(16 Dalton increase), carbonyl (14 Dalton increase), carboxylic acid (30 Dalton increase) products (Helliwell, Poole et al. 1999; Ro, Arimura et al. 2005).

CYP72A68 demonstrates this type of multifunctional activity. Similar to CYP72A67, CYP72A68 also demonstrates an expanded substrate tolerance for oleanate-type sapogenin compounds with and without hydroxyl groups at carbon 2.

Microsomal assays with oleanolic acid as a substrate demonstrate that CYP72A68 catalyzes the initial oxidation of carbon 23 of oleanolic acid, yielding the carbon 23 hydroxyl product, hederagenin. Two additional products were detected in the CYP72A68/oleanolic acid assays. The mass of the first additional product corresponds to oleanolic acid plus 14 Daltons, likely representing a product with carbonyl group. The mass of the second additional product corresponds to oleanolic acid plus 30 Daltons, likely representing a product with a carboxylic acid group. The CYP72A68/hederagenin assays also showed production of the likely carbonyl and carboxylic acid products. The decrease in hederagenin content in the (+)CYP72A68 samples indicates that hederagenin is consumed as a substrate. Taken together, these assay results indicate that CYP72A68 catalyzes the sequential oxidation of carbon 23 of oleanate-type sapogenins, yielding the alcohol (hederagenin), the aldehyde (putative gypsogenin), and the carboxylic acid (putative gypsogenic acid).

Bayogenin and hederagenin both have a hydroxyl group at carbon 23, but bayogenin also has a hydroxyl group at carbon 2. The CYP72A68/bayogenin assays showed accumulation of medicagenic acid (carbon 23 carboxylic acid) and accumulation of a product with a mass characteristic of a hydroxyl to carbonyl oxidation (decrease of 2 Daltons). This second additional product is likely putative

polygalagenin. The elution of the carbon 23 carbonyl products (putative gypsogenin and putative polygalagenin) is consistent, in that they both elute after the carbon 23 carboxylic acid products (putative gypsogenic acid and medicagenic acid) and the carbon 23 hydroxyl products (hederagenin and bayogenin).

The assays with the partially purified aglycone mix as substrates reinforce the conclusions drawn from the assays with reference standards. It is important to note that aglycone mix contained multiple demonstrated CYP72A68 substrates. Because of the presence of multiple substrates and products in the mixture, one would expectdynamic accumulation results that account for the simultaneous production and consumption of compounds such as hederagenin. Briefly, hederagenin is a product in the CYP72A68 catalyzed oxidation of carbon 23 of oleanolic acid, but a substrate in the CYP72A68 catalyzed sequential oxidation of carbon 23 to putative gypsogenin and putative gypsogenic acid. The compounds with carbon 23 methyl groups (oleanolic acid, 2-OH oleanolic acid) were detected in higher levels in the empty vector control samples, indicating their consumption as substrates in the (+)CYP72A68 samples. Compounds with carbon 23 hydroxyl groups (hederagenin and bayogenin) were likewise detected in higher levels in the empty vector control

samples. The (+)CYP72A68 samples showed accumulation of the products with carbon 23 carboxylic acid groups (putative gypsogenic acid and medicagenic acid). As with CYP72A67, the lack of product accumulation in the (+)CYP72A68 assays lacking NADPH provides support for the carbon 23 sequential oxidation activity of CYP72A68.

The results of the (+)CYP72A68 and empty vector control samples assayed with the expanded time domain and an NADPH regeneration system using oleanolic acid as a substrate indicate that CYP72A68 has catalytic capabilities beyond the scope of the defined sapogenin structures and reactions detailed above. The compound unknown 1 (469.33 m/z) is likely the product of a hydroxyl to carbonyl (decrease of 2 Daltons) oxidation of an unknown carbon of compound unknown 2 (471.34 m/z) which is most probably an impurity in the oleanolic acid reference standard. The elution of unknowns 1 and 2 is consistent with the observations for the other alcohol and aldehydes investigated, in that unknown 1 elutes after unknown 2.

No products were detected when CYP72A68 was assayed with ß-amyrin or erythrodiol. This implies that compounds with a methyl carbon 28(ß-amyrin) or carbon 28 hydroxyl group (erythrodiol) are not substrates for CYP72A68-mediated carbon 23 sequential oxidation as tested in this experimental system.

Combined CYP72A67 and CYP72A68 Reactions

FIGURE 8_RIII presents the demonstrated *in vitro* activities for CYP72A67 and CYP72A68 in the context of the oleanate branch(Augustin, Kuzina et al. 2011)of the *M. truncatula* triterpene sapogenin biosynthetic pathway. CYP72A67 and CYP72A68 are the first two demonstrated enzymes in the *M. truncatula* sapogenin biosynthetic pathway. CYP72A67 is a multisubstrate oleanate sapogenin carbon 2 oxidase enzyme. CYP72A68 is a multifunction, multisubstrate oleanate sapogenin carbon 23 sequential oxidase enzyme. CYP93E1 (Shibuya, Hoshino et al. 2006) from the soyasapogenol pathway of *Glycine max*, CYP88D6 (Seki, Ohyama et al. 2008) CYP72A67 from the glycyrrhizin pathway of *Glycyrrhiza uralensis*, and now CYP72A67 and CYP72A68 of the oleanate sapogenin pathway of *M. truncatula*remain the only plant triterpenesapogenin oxidase enzymes demonstrate to date.

Evidence for cytochrome P450 enzymes as multifunction, multisubstrate enzymes in biochemistry has been accumulating for some time (Siminszky, Corbin et al. 1999; Morant, Bak et al. 2003; Schmidt, Sunyaev et al. 2003). The phenomenon of broad substrate tolerance (multisubstrate) of cytochrome P450 enzymes in plant biochemistry lends support to the contention that exotic and structurally diverse natural product "pathways" in plant metabolism are better conceptualized as matrices

(Firn and Jones 2003; Peñuelas and Llusià 2004; Owen and Peñuelas 2005). The phenomenon of multifunctionality (here, sequential oxidation of alkyl carbons) by a single cytochrome p450 enzyme is helpful in thinking about generation of the awesome chemical diversity in the plant kingdom (Dixon 2001), as fewer proteins can catalyze a larger number of reactions. The case for a reasonably simple/parsimonious mechanism for the biosynthesis of extreme plant metabolic diversity is bolstered by the combined phenomena of multifunctionality and broad substrate tolerance by plant cytochrome P450 enzymes.

CYP83G1, CYP88D3, CYP716A12 *In Vitro*

The lack of detected product formation in the ß-amyrin, erythrodiol, oleanolic acid, hederagenin, or the partially purified aglycone assays for CYP83G1, CYP88D3, or CYP716A12 does not preclude the possibility that these proteins may catalyze reactions in the *M. truncatula* sapogenin biosynthesis pathway. It is possible that the recombinant expression system and/or assay conditions employed in this study may have been inappropriate for proper protein folding/assembly/modification or catalytic function for these proteins. CYP716A12 in particular remains extremely tantalizing as a likely enzyme in the biosynthesis of sapogenins, owing to the similarities of cyp72a67, cyp72a68, and cyp716a12 transcript expression levels in both the ecotype

transcriptomics experiment from JHS_CHAPTER_RII and the expression data from the Medicago Gene Expression Atlas.

Genomics

As four of the 5 genes studied were not in current draft of the *M. truncatula* genome sequence, it is reasonable to raise concerns about the comprehensiveness the *M. truncatula* genome (unpublished). Lack of genomic loci was particularly frustrating as it prevented the type of genomic proximity/gene cluster analysis presented in JHS_Research_CHAPTER_IV. The 1.5 kb size of the cyp88d3 amplification product from a genomic DNA template indicates that the locus likely lacks introns of any significant length, and would be interesting to investigate in light of genomic locus size of homologs ((Seki, Ohyama et al. 2008) JHS_Research_CHAPTER_IV) and recent results that demonstrate the retroposition and neofunctionalization of a cytochrome P450 gene in a pollen-specific branch of phenylpropanoid metabolism in *Arabidopsis thaliana* (Matsuno, Compagnon et al. 2009).

Mutant/Molecular Genetics

The failure to identify homozygous intersertional mutant individuals in either of the two *cyp72a67* or two *cyp72a68Tnt-1* mutagenesis lines strongly suggests that these genes/gene products are critical for plant growth and function. It seems likely that

homozygous intersertional mutant individuals for either of these loci result in a lethal phenotype. The cyp72a67 and cyp72a68 transcript expression data for various plant organs obtained from the Medicago Gene Expression Atlas showed that these transcripts were accumulated in the highest levels in late developing seeds. It is reasonable to infer that high expression level of these transcripts in late developing seeds may be related to putative lethal phenotype of homozygous intersertional mutant individuals for these loci. Further, one could postulate that the accumulation of the functional gene products of these highly accumulated transcripts may be critical in the metabolism of seed development. Finally, as these gene products have been shown to function as biosynthetic enzymes in the *M. truncatula* triterpene sapogenin biosynthetic pathway, one could warrant that triterpene sapogenins function critically in the metabolism of *M. truncatula* seed developmental physiology.

The biotechnological limitations of reverse mutant screening methodologies for characterization of gene function is highlighted if in fact the homozygous intersertional mutant state for these loci does indeed result in a lethal phenotype. The use of germplasm diversity collections for dissection of gene function seems particularly suited as a means to avoid the limitations of reverse mutant screening

methodologies, as the less extreme yet significantly different phenotypes evinced in the ecotype collection enable experimentation with viable plants.

The cyp72a67 and cyp72a68 transcript expression data from the Medicago Gene Expression Atlas showed that these transcripts were expressed in both root and various aerial and reproductive organs. When considered in combination with the demonstrated catalytic functions for these gene products, the spatially resolved expression data for the various organs suggests that sapogenin biosynthesis is likely to occur throughout the plant, and not through a mechanism of centralized biosynthesis and subsequent translocation.

Methods

Plant Material

A17 and R108 isoline seeds were obtained from the greenhouse manager (David McSweeny) at the Samuel Roberts Noble Foundation. Seeds for the *Tnt-1* insertion mutagenesis lines (Tadege, Wen et al. 2008)were obtained from the curator of biological materials (Dr. Jiangqi Wen)at the Samuel Roberts Noble Foundation.

Plant Growth Conditions

Plants were grown in a root cone system (Stuewe and Sons, http://www.stuewe.com, Tangent, OR) with Turface MVP medium (Profile Products, Buffalo Grove, IL) in a

Conviron TCR180 walk-in growth chamber (http://www.conviron.com/, Winnipeg, Manitoba, Canada) maintained at 90% humidity and at an average temperature of 24 °C day (16 h) and 20 °C night (8 h). Plants were fertilized with 15 ppm nitrogen (Scotts' 20 10 20 Peat-Lite Special, http://www.scotts.com, Marysville, Ohio) daily in the morning and watered with distilled water in the evening.

DNA preparations, RNA isolation, cDNA synthesis

All genomic DNA isolations were performed as previously described (Sambrook, Russell et al. 2001). For all RNA isolations, harvested material was frozen immediately in liquid nitrogen and stored at -80°C prior to RNA isolation. Total RNA was extracted using TRIZOL reagent (Life Technologies, http://www.lifetechnologies.com/, Carlsbad, California), treated with DNaseI (Ambion, http://www.ambion.com/), and column purified with a RNeasyMinEluteCleanUp Kit (Qiagen, http://www.qiagen.com/). RNA was quantified using a Nanodrop Spectrophotometer ND-100 (NanoDrop Technologies, http://www.nanodrop.com/) and evaluated for quality with a Bioanalyzer 2100 (Agilent, http://www.home. agilent.com/). For cDNA synthesis, 10µg of total RNA (prepared and assessed for quality as above) was primed with oligo(dT)20 and

synthesized with Super Script III according to manufacturer's instructions (Life Technologies, http://www.lifetechnologies.com/, Carlsbad, California)

Cloning

All primer sequences and NCBI Genbank (NCBI, http://www.ncbi.nlm.nih.gov/genbank/, Bethesda, MD) accession numbers are presented in Appendix_R.III._Primers. Coding sequences for cyp72a67, cyp72a68, cyp83g1, cyp88d3, and cyp716a12 were obtained from NCBI Genbank. All cloning primers were designed using primer3 (Rozen and Skaletsky 1999). The forward primer for each target included both a BamHI restriction site and a kozak yeast translation initiation sequence, while each reverse primer included an EcoRI cut site. Targets were amplified from cDNA prepared from aerial organs from the isoline A17 using Plantium Hi-Fi Taq polymerase (Life Technologies, http://www.lifetechnologies.com/home. html, Carlsbad, California). Amplified products were cloned into the pGEM-easy vector (Promega, http://www.promega.com/, Madison, WI) and sequenced using M13 forward and reverse primers. The targets were excised from the p-GEM easy vector via BamHI and

EcoRI restriction digest, sub-cloned into the *pYeDP60* vector (Pompon, Louerat et al. 1996; Urban, Mignotte et al. 1997) and sequenced using the *gal10* promoter primer (Appendix_RIII_cloned_sequence). *Wat11* yeast cells were transformed as previously reported (Greenhagen, Griggs et al. 2003). Transformation of yeast was confirmed via colony PCR with gene specific primers. cyp72a67 and cyp72a68 genomic loci were amplified from genomic DNA and cloned into the pGEM-easy vector. The *Tnt-1-_cyp72a68* allele was amplified from genomic DNA of the *Tnt-1* insertion mutagenesis line NF 1698 and cloned into the pGEM-easy vector.

Recombinant expression and microsomal preparations of CYP72A68 enzymatic assays

The potential catalytic activities of CYP72A68, CYP72A67, CYP83G1, CYP88D3, and CYP716A12 were tested using *in vitro* assays with a variety of triterpene sapogenin substrates. *Wat11* cells containing*pYeDP60*+target or empty *pYeDP60* were grown and microsomes were prepared as previously described (Greenhagen, Griggs et al. 2003). All assays were performed in triplicate. For assays with purified reference standards as substrates, 100µg of total microsomal protein (quantified via Bradford assay) (Seki, Ohyama et al. 2008) was assayed for 2 hours at 30°C in a 500µl reaction volume of 50mM potassium phosphate buffer (pH 7.25) containing 1mM NADPH,

and 40μM purified substrate. ß-amyrin, erythrodiol, and oleanolic acid were obtained from Sigma-Aldrich(Sigma-Aldrich, http://www.sigmaaldrich.com/, St. Louis, MO). Hederagenin and cycloartenol were obtained from Chromadex (Chromadex, http://www.chromadex.com, Irvine, CA). Bayogenin was obtained from PhytoLab (PhytoLab, http://www.phytolab.com, Vestenbergsgreuth, Germany). A mixture of partially purified aglycones obtained through acid hydrolysis of saponin extracts obtained from *Medicago truncatula* root tissues (Huhman and Sumner 2002) were at assayed at an approximate 80μM concentration. The expanded time-series (8 and 48 hour) CYP72A68 aglycone mix substrate assays contained an NADPH generation system (3.3 mM glucose-6-phosphate, 1.3 mM of NADPH, 3.3 mM magnesium chloride, and 0.4 U/ml glucose-6-phosphate dehydrogenase) (Yu, Shin et al. 2003). Glucose-6-phosphate dehydrogenase was obtained from Sigma-Aldrich.

Extraction and Instrumental Analysis

Assay reaction mixtures were extracted 2 times with 500 μl of ethyl acetate, and dried under nitrogen gas. Oleanolic acid, hederagenin, bayogenin, and aglycone mix assay contents were resolubilized in 250μl of 80% Methanol (containing 0.018 mg/ml umbelliferone as an internal standard) and analyzed with a Waters Acquity UPLC system coupled to a hybrid quadrupole time-of-flight (QTOF) Premier mass

spectrometer (Waters, http://www.waters.com/, Milford, MA). A reverse-phase, 1.7-mm UPLC BEH C18, 2.1 3 150 mm column (Waters) was used for separations. The mobile phase consisted of eluent A (0.1% [v/v] acetic acid/water) and eluent B (acetonitrile), and separations were achieved using a linear gradient of 95% to 30% A over 30 min, 30% to 5% A over 3.0 min, and 5% to 95% A over 3.0 min. The flow rate was 0.56 mL min 21, and the column temperature was maintained at 60°C. Masses of the eluted compounds were detected in the negative ESI mode from 50 to 2,000 mass-to-charge ratio. The QTOF Premier was operated under the following instrument parameters: desolvation temperature of 400°C, desolvation nitrogen gas flow of 850 L h21, 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. ß-amyrin, erythrodiol, and cycloartenol assays were extracted 2 times with 500 µl of ethyl acetate, dried under nitrogen gas, dissolved in 100µl pyridine, MSTFA-derivitized, and analyzed by GC-MS as described previously (Broeckling, Huhman et al. 2005).

Data Processing

Raw data files were converted to .cdf file format, followed by metabolite data extraction, alignment, and export using MET-IDEA software (Broeckling, Reddy et al.

2006) or Waters MARKERLYNX software. The spectral abundance signals for all metabolites in a separation were normalized to the internal standard (0.018 mg/ml umbelliferone). Descriptive statistics were performed in Microsoft Excel.

Screening the *M. truncatula Tnt-1*Retrotransposon Insertion Population for Identification transposon insertion mutants

The *M. truncatula* R108 *Tnt-1* population (Million Tadege 2008)was screened for insertions in cyp72a67, cyp72a68, cyp83g1, cyp716a12, cyp88d3 loci (APPENDIX_*Tnt-1*_PRIMERS) as previously described (Pang, Wenger et al. 2009). BLAST analysis (Altschul, Gish et al. 1990) was performed for all target loci against the Noble Foundation *Tnt-1* flanking sequence database (http://bioinfo4.noble.org/mutant/).

Additional Information

Accession Numbers

Currently found in Appendix_RIII_Primers

Supplemental Data

Appendix_RIII_Primers-Cloning primers used in RIII

Appendix_RIII_cloned_sequence-Nucleotide sequence of cloned target genes

APPENDIX_Tnt-1_PRIMERS-Primers used in reverse screen R III

Sources

Altschul, S., Gish, W., Miller, W., Meyers, E., and Lipman, D. (1990). Basic Local Alignment Search Tool. J Mol Biol **215,** 403 - 410.

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin,. (2003). Dissecting plant secondary metabolism; constitutive chemical defences in cereals. New Phytologist **159,** 101-108.

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

Benedito, V., Torres-Jerez, I., Murray, J., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P., Tang, Y., and Udvardi, M. (2008). Affymetrix GeneChip Medicago Genome Array

A gene expression atlas of the model legume *Medicago truncatula*. Plant J **55,** 504 - 513.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. Anal. Chem. **78,** 4334-4341.

Broeckling, C. D., Huhman, D. V., Farag, M. A., Smith, J. T., May, G. D., Mendes, P., Dixon, R. A., and Sumner, L. W. (2005). Metabolic profiling of. Journal of Experimental Botany **56,** 323-336.

Chang-Jun Liu, D. H., Lloyd W. Sumner, Richard A. Dixon,. (2003). Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from *Medicago truncatula*. The Plant Journal **36,** 471-484.

Dixon, R. A. (2001). Natural products and plant disease resistance. Nature **411,** 843-847.

Dixon, R. A., and Sumner, L. W. (2003). Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. Plant Physiology **131,** 878-885.

Firn, R. D., and Jones, C. G. (2003). Natural products - a simple model to explain chemical diversity. Natural Product Reports **20,** 382-391.

Greenhagen, B. T., Griggs, P., Takahashi, S., Ralston, L., and Chappell, J. (2003). Probing sesquiterpene hydroxylase activities in a coupled assay with terpene synthases. Archives of Biochemistry and Biophysics **409,** 385-394.

He, J., Benedito, V., Wang, M., Murray, J., Zhao, P., Tang, Y., and Udvardi, M. (2009). The *Medicago truncatula* gene expression atlas web server. BMC Bioinformatics **10,** 441.

Helliwell, C. A., Poole, A., James Peacock, W., and Dennis, E. S. (1999). Arabidopsis ent-Kaurene Oxidase Catalyzes Three Steps of Gibberellin Biosynthesis. Plant Physiology **119,** 507-510.

Huhman, D. V., and 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,** 347-360.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. Plant Molecular Biology **51,** 731-743.

Lahoucine Achnine, D. V. H., Mohamed A. Farag, Lloyd W. Sumner, Jack W. Blount, Richard A. Dixon,. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. The Plant Journal **41,** 875-887.

Lu, C. D., and Jorgensen, N. A. (1987). Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. The Journal of Nutrition **117,** 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M., and Jorgensen, N. A. (1987). Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. Journal of dairy science **70,** 799-805.

Matsuno, Michiyo, Compagnon, V., Schoch, G. A., Schmitt, M., Debayle, D., Bassard, J. -E., Pollet, B., Hehn, A., Heintz, D., Ullmann, P., Lapierre, C., Bernier, F. o., Ehlting, J. r., and Werck-Reichhart, D. l. (2009). Evolution of a Novel Phenolic Pathway for Pollen Development. Science **325,** 1688-1692.

Million Tadege, J. W., Ji He, Haidi Tu, Younsig Kwak, Alexis Eschstruth, Anne Cayrel, Gabriella Endre, Patrick X. Zhao, Mireille Chabaud, Pascal Ratet, Kirankumar S. Mysore,. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. The Plant Journal **54,** 335-347.

Morant, M., Bak, S., Møller, B. L., and Werck-Reichhart, D. (2003). Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Current Opinion in Biotechnology **14,** 151-162.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., and Dixon, R. A. (2010). Genomic and

Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*. Plant Cell **22,** 850-866.

Owen, S. M., and Peñuelas, J. (2005). Opportunistic emissions of volatile isoprenoids. Trends in Plant Science **10,** 420-426.

Pang, Y., Wenger, J. P., Saathoff, K., Peel, G. J., Wen, J., Huhman, D., Allen, S. N., Tang, Y., Cheng, X., Tadege, M., Ratet, P., Mysore, K. S., Sumner, L. W., Marks, M. D., and Dixon, R. A. (2009). A WD40 Repeat Protein from *Medicago truncatula* Is Necessary for Tissue-Specific Anthocyanin and Proanthocyanidin Biosynthesis But Not for Trichome Development. Plant Physiol. **151,** 1114-1129.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., and Osbourn, A. E. (1999).compromised disease resistance in saponin-deficient plants Proceedings of the National Academy of Sciences of the United States of America **96,** 12923-12928.

Peñuelas, J., and Llusià, J. (2004). Plant VOC emissions: making use of the unavoidable. Trends in ecology & evolution (Personal edition) **19,** 402-404.

Pompon, D., Louerat, B., Bronine, A., Urban, P., Eric, F. J., and Michael, R. W. (1996). [6] Yeast expression of animal and plant P450s in optimized redox environments. In Methods in Enzymology (Academic Press), pp. 51-64.

Ro, D. -K., Arimura, G. -I., Lau, S. Y. W., Piers, E., and Bohlmann, J. r. (2005). Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. Proceedings of the National Academy of Sciences of the United States of America **102,** 8060-8065.

Rozen, S., and Skaletsky, H. (1999). Primer3 on the WWW for General Users and for Biologist Programmers, pp. 365-386.

Sambrook, J., Russell, D. W., and Cold Spring Harbor, L. (2001). Molecular cloning : a laboratory manual / Joseph Sambrook, David W. Russell. (Cold Spring Harbor, N. Y. :: Cold Spring Harbor Laboratory).

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? Trends in Biochemical Sciences **28,** 336-341.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K., and Muranaka, T. (2008). Licorice Î²-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,** 14204-14209.

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

Siminszky, B., Corbin, F. T., Ward, E. R., Fleischmann, T. J., and Dewey, R. E. (1999). Expression of a soybean cytochrome P450 monooxygenase cDNA in yeast and tobacco enhances the metabolism of phenylurea herbicides. Proceedings of the National Academy of Sciences **96,** 1750-1755.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., and Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. The Plant Journal **32,** 1033-1048.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., Cayrel, A., Endre, G., Zhao, P. X., Chabaud, M., Ratet, P., and Mysore, K. S. (2008). Large-

scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. The Plant Journal **54,** 335-347.

Urban, P., Mignotte, C., Kazmaier, M., Delorme, F., and Pompon, D. (1997). Cloning, Yeast Expression, and Characterization of the Coupling of Two Distantly Related Arabidopsis thaliana NADPH-Cytochrome P450 Reductases with P450 CYP73A5

10. 1074/jbc. 272. 31. 19176. J. Biol. Chem. **272,** 19176-19186.

Yu, C., Shin, Y. G., Kosmeder, J. W., Pezzuto, J. M., and van Breemen, R. B. (2003). Liquid chromatography/tandem mass spectrometric determination of inhibition of human cytochrome P450 isozymes by resveratrol and resveratrol-3-sulfate. Rapid Communications in Mass Spectrometry **17,** 307-313.

Chapter IV - Loci from the cyp88d Subfamily of Cytochrome P450s are Immediately Adjacent to Oxidosqualene Synthase Loci in the Genomes of *Medicago truncatula* and *Lotus japonicus.*

Authors: John H. Snyder, David V. Huhman, Lloyd W. Sumner

Summary:

This chapter will detail my attempts to test the hypothesis that there may be gene clusters of biosynthetically-related genes from triterpenoid metabolism in the *Medicago truncatula* genome. Promising early Ecotype qRT-PCR and *M. truncatula* genomics analysis will be contrasted with inconclusive results from*in vitro* enzymology, cell culture, and mutant analysis for genes in the cyp88d subfamily.

Abstract

Five plant secondary metabolic gene clusters have been discovered to date. The oxidosqualene synthase enzyme β-amyrin synthase is the entry point enzyme in the triterpene saponin pathway of *Medicago truncatula*. The enzymes which catalyze the bio-oxidation of β-amyrin in the triterpene sapogenin biosynthetic pathway of *M. truncatula* have not been characterized, but CYP88D6 from

Glycyrrhiza uralensis has been shown bio-oxidation activity for carbon 11 of ß-amyrin. In light of the recent discovery of secondary metabolic gene clusters in plant genomes, the genome of *M. truncatula* and L. japonicus were analyzed in this study in order to identify potential clusters which included oxidosqualene synthase genes. In *M. truncatula*, a cyp88d1 locus was identified immediately adjacent to a β-amyrin synthase locus. In *L. japonicus*, cyp88d4 and cyp88d5 loci were immediately adjacent to a β-amyrin synthase locus. These exciting findings motivated our efforts to characterize the function of cyp88d1-3 from *M. truncatula*. Experimental approaches in the characterization effort included in vitro enzymatic assays, *in planta* integrated transcript/metabolomics analyses from a root cell suspension culture methyl jasmonate elicitation time series, *in planta* integrated transcript/metabolomics analyses from a collection of germplasm diversity accessions showing differential triterpene saponin accumulation dynamics, and metabolomics analyses of transposon insertion mutants. The *in planta* integrated transcript/metabolomics analyses from a collection of germlasm diversity accessions showed strong correlation values for cyp88d2 and cyp88d3 transcripts vs. total saponin accumulation. Recombinant expression and *in vitro* enzymatic assay analysis of these 3 proteins did not show

activity for any of the substrates tested (β-amyrin, oleanolic acid, hederagenin, or a mixture of *M. truncatula* sapogenin compounds. Analysis of saponin accumulation phenotypes of two independent *cyp88d2* mutants did not reveal significant differences in the saponin phenotypes between the wild-type and *cyp88d2* mutants.

Introduction

Five plant secondary metabolic gene clusters have been discovered in higher plants to date (Chu, Wegel et al. in press). These numerous intriguing examples of clusters of functionally related but non-homologous genes from plant defense compound pathways are proving useful in functional genomics efforts for the characterization of genes of unknown function. More broadly, the discovery of these clusters is enabling powerful new methodologies for the investigation of adaptive evolution and genome plasticity in plants (Osbourn 2010).

The oxidosqualene synthase enzyme β-amyrin synthase is the entry point enzyme in the triterpene saponin pathway of *Medicago truncatula* (Suzuki, Achnine et al. 2002; Iturbe-Ormaetxe, Haralampidis et al. 2003). The enzymes which catalyze the bio-oxidation of β-amyrin in the triterpene sapogenin biosynthetic pathway of *M. truncatula* have not been characterized. Bio-oxidation of alkyl carbon 24 and carbon

11 of ß-amyrin have been demonstrated for CYP93E1 from *Glycine max* (Shibuya, Hoshino et al. 2006) and CYP88D6 from *Glycyrrhiza uralensis* (Seki, Ohyama et al. 2008), respectively. The Seki et al. study which characterized CYP88D6 also contained full length coding sequence for other cyp88d family members from *M. truncatula* (cyp88d1-3) and *L. japonicas* (cyp88d4-5).

In light of the recent discovery of secondary metabolic gene clusters in plant genomes, the genome of *M. truncatula* and L. japonicus were analyzed in this study in order to identify potential clusters which included oxidosqualene synthase genes. In *M. truncatula*, a cyp88d1 locus was identified immediately adjacent to a β-amyrin synthase locus. In *L. japonicus*, cyp88d4 and cyp88d5 loci were immediately adjacent to a β-amyrin synthase locus. These exciting findings motivated our efforts to characterize the function of cyp88d1-3 from *M. truncatula*. Experimental approaches in the characterization effort included in vitro enzymatic assays, *in planta* integrated transcript/metabolomics analyses from a root cell suspension culture methyl jasmonate elicitation time series, *in planta* integrated transcript/metabolomics analyses from a collection of germplasm diversity accessions showing differential triterpene saponin accumulation dynamics, and metabolomics analyses of transposon insertion mutants.

Results

Genomics

Medicago truncatula genome analysis was performed to identify cytochrome P450 and gylcosyltransferase loci in the immediate genomic vicinity of β-amyrin synthase and other oxidosquanlene cyclase loci in order to explore the possibility of "operon-like gene clusters" (Field and Osbourn 2008) related to triterpene metabolism. All genome regions of interest are presented in Supp 1_RIV. Figure 1_RIV shows a detail from chromosome 4 of *M. truncatula* where a β-amyrin synthase locus is adjacent to a cytochrome P450 locus from a subfamily which has been previously demonstrated to bio-oxidize β-amyrin (Seki, Ohyama et al. 2008).

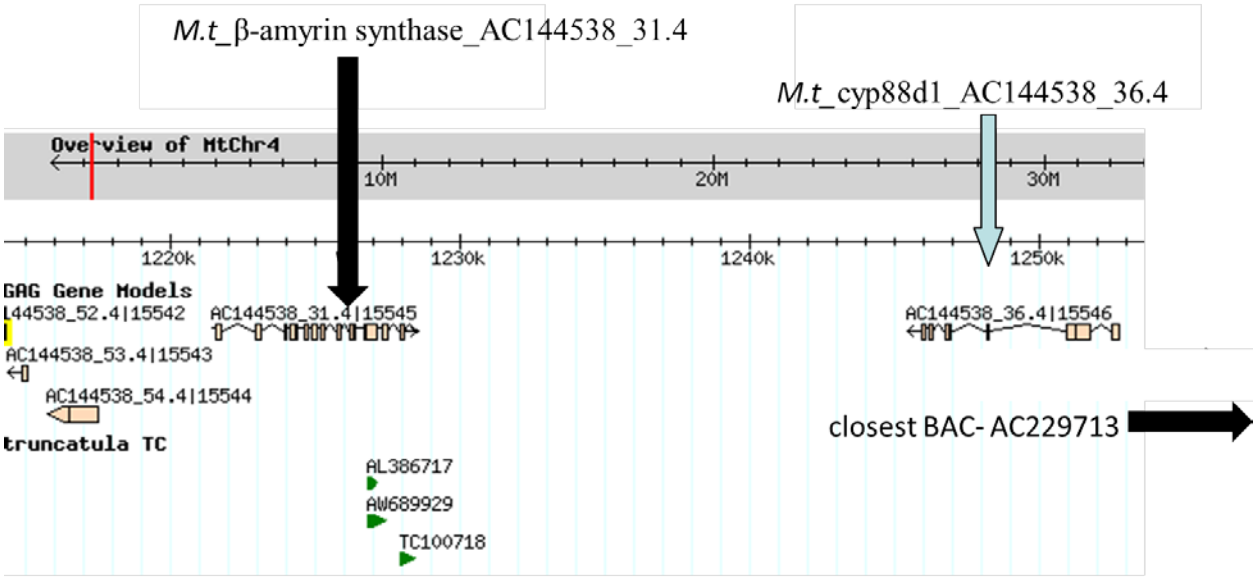


FIGURE 1_RIV. - β-amyrin synthase and cyp88d1 Loci are Adjacent in the Genome of *M. truncatula*.

Details of the terminal (centromeric) end of BAC 144538 showing the uninterrupted genomic proximity of predicted gene models for AC144538_31(β-amyrin synthase) and AC144538_36(cyp88d1), prepared using the Medicago genome browser program (http://bioinfo4.noble.org/cgi-bin/gbrowse/gbrowse/medicago).

Full length coding sequence of two cyp88d family members (cyp88d4, cyp88d5) from *Lotus japonicas* were identified in a previous study (Seki, Ohyama et al. 2008). In order to assess the possibility of related cyp88d-β-amyrin synthase loci proximity, BLAST analysis was performed using cyp88d4 and cyp88d5 coding sequences as queries against the cDNA models mapped to the genome sequence of *L. japonicus.* Figure 2_RIV shows a detail from chromosome 3 of *L. japonicus* where a β-amyrin synthase locus is immediately adjacent to cDNA models for cyp88d4 and cyp88d5.

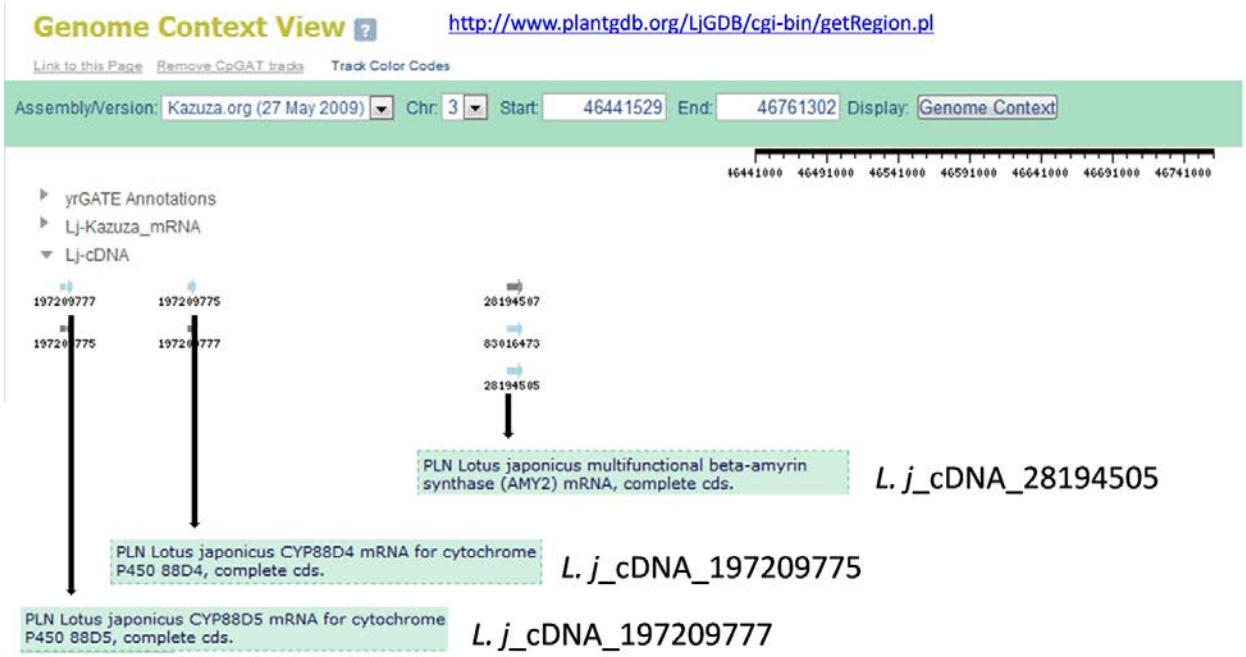


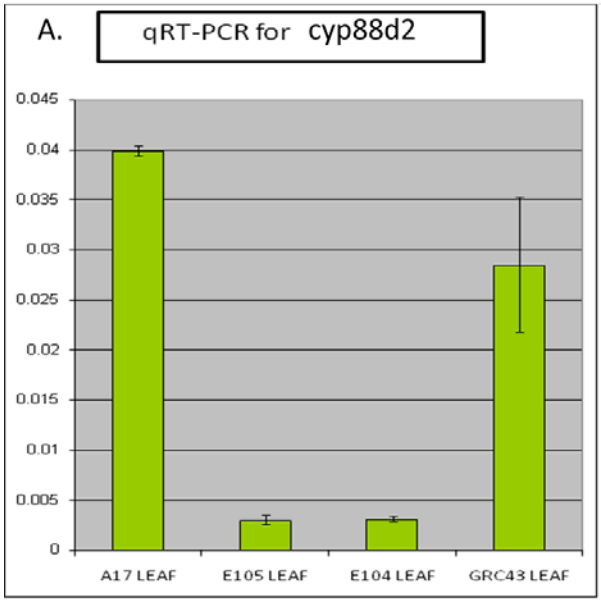
FIGURE 2_RIV. - β-amyrin synthase and cyp88d4 and cyp88d5 Loci are Adjacent in the Genome of Lotus japonicus.

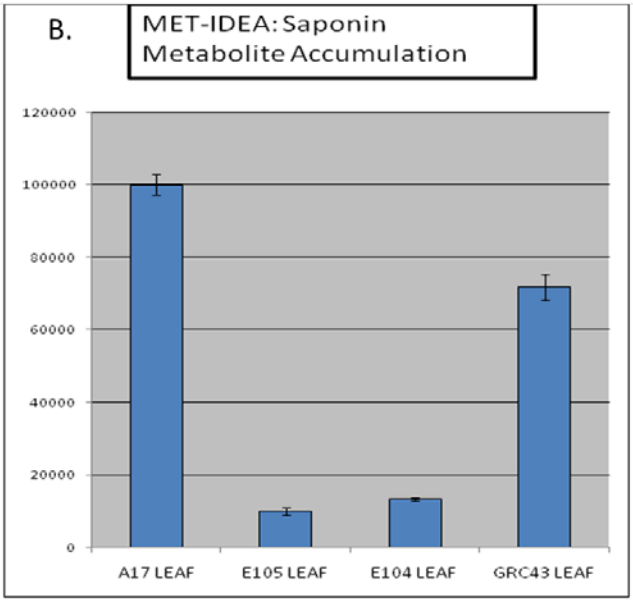
L. japonicus chromosome 3,bases 46441529 to 46761302, showing the uninterrupted genomic proximity of the cDNA gene models for β-amyrin synthase, cyp88d4, and cyp88d5,prepared using the genome browser tools available at http://www.plantgdb.org/LjGDB.

Of the three loci investigated in this study, only cyp88d1 (AC144538_36, 6.78 kb) was included in the genome sequence of *M. truncatula* (http://www.medicagohapmap.org/?genome). Amplification of cyp88d2 and cyp88d3 from genomic DNA of isoline A17 showed loci size of approximately 6kb and 1.5 kb, respectively (data not shown).

Integrated Analysis of Relative Transcript Expression and Metabolomics Datasets

Complete data from the qRT-PCR and metabolomics analyses of the selected ecotypes/organs are presented in SUPP 2_RIV and SUPP 3_RIV. Complete data from the qRT-PCR and metabolomics analyses of the root cell suspension culture MeJA-elicitation time series experiment are presented in SUPP 4_RIV and SUPP 5_RIV. Detailed descriptions of these files can be found in the Supplementary Data section of this manuscript. Figure 3_RIV(cyp88d2) and Figure 4_RIV(cyp88d3) present Pearson correlation values as well as transcript and total saponin accumulation results for the inter-ecotype, intra-aerial-organ comparison.





C.

cyp88d2 expression vs total saponin accumulation	
Inter-Ecotype, Intra-Aerial-Organ_individual case (n=12)	
Pearson's r	0.8522
90% Bootstrap Confidence Interval of r	0.65 - 0.98
Bootstrap Standard Error of r	0.1155

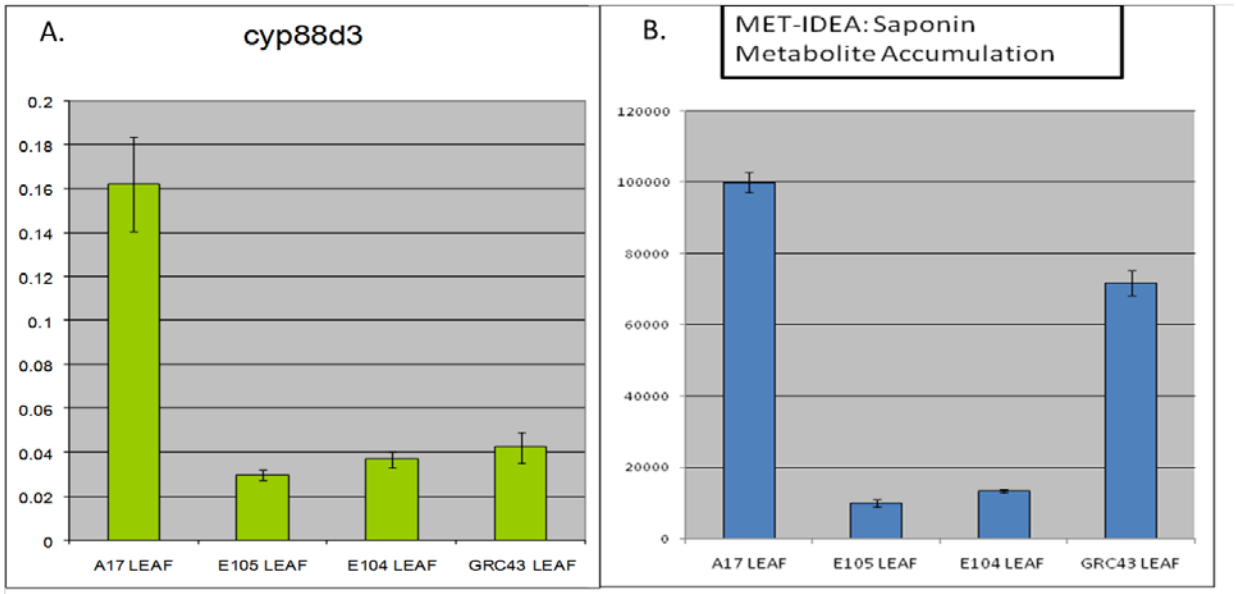
Inter-Ecotype, Intra-Aerial-Organ_averaged case (n= 4)	
Pearson's r	0.9684
95% Bootstrap Confidence Interval of r	0.9367 - 0.9989
Bootstrap Standard Error of r	Not a Number (n too small)

FIGURE 3_RIV. - Correlation of Transcript and Total Saponin Metabolite Accumulation for cyp88d2 from Various Ecotypes.

(A) Graph showing the relative transcript expression levels of cyp88d2 (green) for an inter-genotype (A17, ESP105, ESP104, and GRC43), intra-aerial-organ comparison. Error bars represent 1 standard error.

(B) Graph showing the total saponin accumulation values from the metabolomics analysis of the inter-genotype (A17, ESP105, ESP104, and GRC43), intra-aerial-organ comparison. Error bars represent 1 standard error.

(C) Table of Pearson correlation coefficient (Pearson's r) values for [cyp88d2 relative transcript expression] vs. [total saponin accumulation] in the inter-ecotype, intra-aerial-organ comparison permutation, for both the individual (n= 12) and averaged (n= 4) cases. The table also includes: Bootstrapped (5000 iteration) confidence intervals (90% for individual case, 95% for averaged case) of r, and bootstrapped standard errors of r.



C.

cyp88d3 expression vs total saponin accumulation	
Inter-Ecotype, Intra-Aerial-Organ_individual case (n=12)	
Pearson's r	0.6593
90% Bootstrap Confidence Interval of r	0.3761 - 0.8624
Bootstrap Standard Error of r	0.1805

Inter-Ecotype, Intra-Aerial-Organ_averaged case (n= 4)	
Pearson's r	0.7661
85% Bootstrap Confidence Interval of r	0.1791 - 1.0000
Bootstrap Standard Error of r	Not a Number (n too small)

FIGURE 4_RIV. - Correlation of Transcript and Total Saponin Metabolite Accumulation for cyp88d3 from Various Ecotypes.

(A) Graph showing the relative transcript expression levels of cyp88d3 (green) for an inter-genotype (A17, ESP105, ESP104, and GRC43), intra-aerial-organ comparison. Error bars represent 1 standard error.

(B) Graph showing the total saponin accumulation values from the metabolomics analysis of the inter-genotype (A17, ESP105, ESP104, and GRC43), intra-aerial-organ comparison. Error bars represent 1 standard error.

(C) Table of Pearson correlation coefficient (Pearson's r) values for[cyp88d3 relative transcript expression] vs. [total saponin accumulation] in the inter-ecotype, intra-aerial-organ comparison permutation, for both the individual (n= 12) and averaged (n= 4) cases. The table also includes: Bootstrapped (5000 iteration) confidence intervals (90% for individual case, 85% for averaged case) of r, and bootstrapped standard errors of r.

Molecular Genetics

The expression data for cyp88d1, cypp88d2, and cyp88d3 transcripts in a variety of plant organs and developmental stages (Figure 5_RIV) were obtained from The MedicagoGene Expression Atlas web server (Benedito, Torres-Jerez et al. 2008; He, Benedito et al. 2009). Note that the highest transcript accumulation value for the probeset representing cyp88d1 was found in roots involved in mycorrhizal symbiosis.

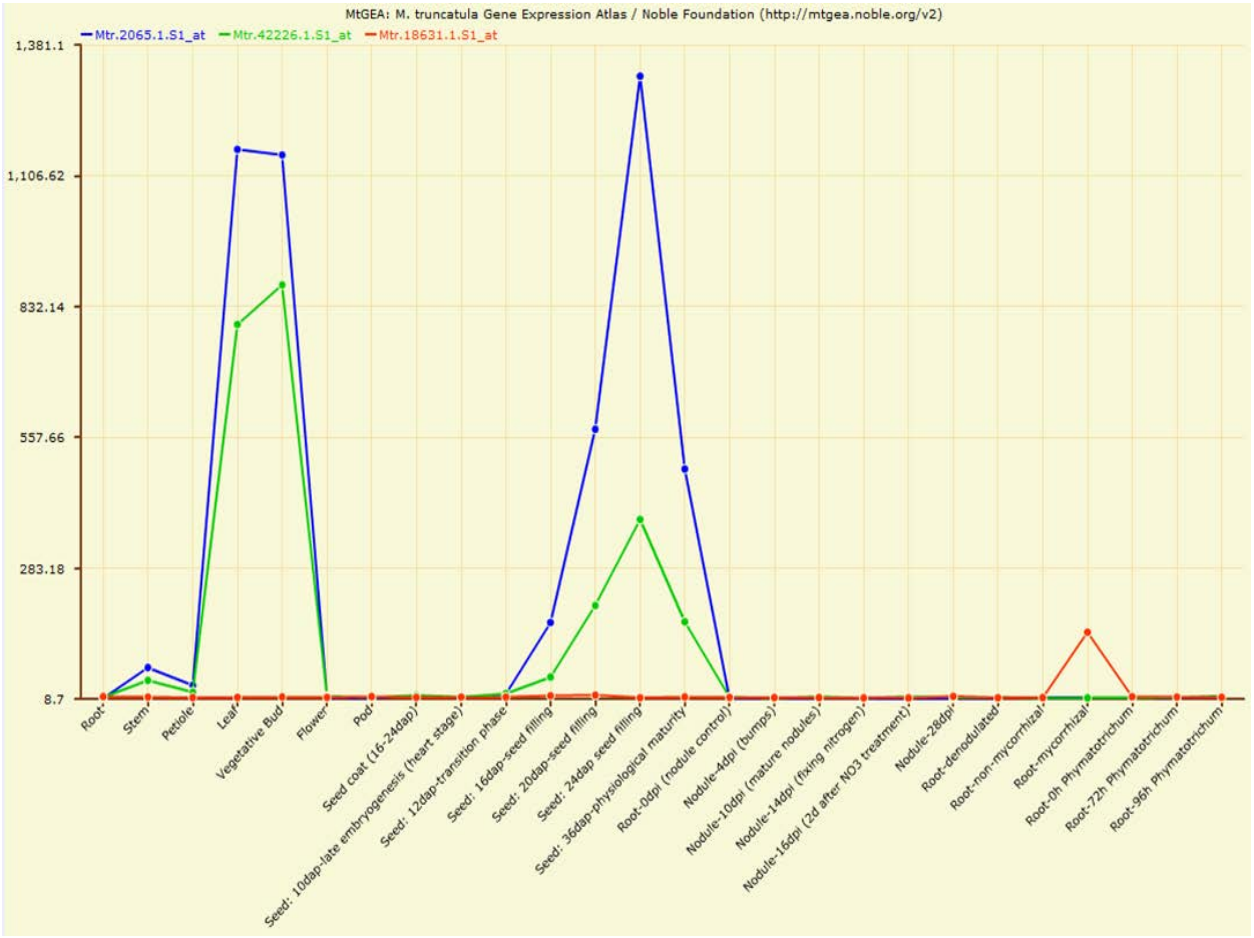


FIGURE 5_RIV. - Expression values for Transcripts of cyp88d Genes in Diverse Plant Organs and Biological Treatments from the Medicago Gene Atlas.

Medicago Gene Expression Atlas accumulation data for transcripts of cyp88d1 (red), cyp88d2 (green), and cyp88d3 (blue), in root, stem, leaf, petiole, vegetative bud, flower, pod, and seed coat organs, seed developmental stages of 10, 12, 16, 20, 24, and 36 days following pollination, root at rhizobial inoculation (control), 4 days after rhizobial inoculation (lumps), 10 days after rhizobial inoculation (immature nodules), 14 days after rhizobial inoculation (N2 fixing), 16 days after rhizobial inoculation (2 days after NO3 treatment), 28 days after rhizobial inoculation (reference organ), 28 day-old denodulated roots, 6 week-old uninoculated roots (control for mycorrhization), 6

week-old (30dpi) mycorrhizal roots, 0 hour Phymatotrichum root rot infection (time-course), and 72 hour Phymatotrichum root rot infection (time-course).

Tnt-1 Mutant Collection Screening and Metabolomics Analysis

Results of the reverse screen of the pooled *Tnt-1* germplasm are presented in TABLE 1_RIV. Results for the reverse screen for cyp88d3 are presented in JHS_RESEARCH_CHAPTER_III. Multiple homozygous *Tnt-1_cyp88d2* plants were identified in the NF8050 and NF5409 lines. Metabolomics analysis of R108 (wild type) and homozygous *Tnt-1_cyp88d2* from NF8050 or NF5409 plants did not reveal differences in triterpene saponin accumulation (data not shown).

Reverse Screen			
Target Locus	Primer Combination	Insertion Site in Locus	NF *Tnt-1* Insertion Line ID
cyp88d2	cyp88d2 F + Tnt1-R	at base 695 of coding sequence	NF5409
	cyp88d2 F2 + Tnt1-R	intron	NF8050

TABLE 1_RIV. - Reverse Genetic Screening Results for *Tnt-1*Insertion Mutants for the cyp88d2 Locus.

Successful primer combinations, insertion site in target loci, and *Tnt-1* insertion line identification numbers for the reverse genetic screen of the *Tnt-1* mutant collection.

In Vitro Enzymatic Assays of CYP88D1, CYP88D2, and CYP88D3

When β-amyrin, erythrodiol, oleanolic acid, hederagenin, or the partially purified aglycone solution from *M. truncatula* roots were assayed as substrates for CYP88D1, CYP88D2, and CYP88D3, no products were detected (data not shown).

Discussion

Genomics

In addition to the terpenoid synthases or tryptophan synthase homolog "signature" enzymes (Chu, Wegel et al. in press), Cytochrome P450 genes are included in the gene clusters for biosynthesis of the triterpenoid thalianol from *Arabidopsis thaliana* (Field and Osbourn 2008), biosynthesis of the triterpenoid avenacin from *Avena strigosa* (Qin, Eagles et al.; Qi, Bakht et al. 2004; Mylona, Owatworakit et al. 2008; Mugford, Qi et al. 2009), biosynthesis of the diterpenoidsmomilactone and phytocassane from *Oryza sativa* (Sakamoto, Miura et al. 2004; Wilderman, Xu et al. 2004; Shimura, Okada et al. 2007; Swaminathan, Morrone et al. 2009), and biosynthesis of the benzoxazinoid, 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one(DIMBOA) from *Zea mays* (Frey, Schullehner et al.; Gierl and Frey 2001; Frey, Huber et al. 2003; Jonczyk, Schmidt et al. 2008). It is therefore extremely interesting that β-amyrin synthase lociare immediately adjacent tocytochrome P450 loci from a subfamily which contains a member with demonstrated β-amyrin bio-oxidation activity. Further, it is exceedingly unlikely that the co-occurrence of immediately adjacent β-amyrin synthase and cyp88d family loci in the genomes of *M. truncatula* and L. japonicus is an evolutionary coincidence. It is unfortunate that the BAC

sequence which contained the immediately adjacent β-amyrin synthase and cyp88d1 loci was an "island" in the assembly of chromosome 3 of the *M. truncatula* genome, as additional cyp88d loci may be in the same vicinity. Several attempts to obtain additional sequence information (BAC end sequence, adjacent BACs etc.) were made in 2009 and 2010.

Integrated Analysis of Relative Transcript Expression and Metabolomics Datasets

The strong transcript vs. total saponin accumulation correlation values for cyp88d2 and cyp88d3 in the selected ecotypes were initially perceived as very promising examples of a likely 'guilt by association' phenomenon. Indeed, the strong correlation values motivated the initial emphasis on Pearson correlation coefficient analysis of the microarray dataset (see JHS_RESEARCH_CHAPTER_II). The metabolomics results from the root cell culture MeJA elicitation time series experiment did not reproduce the previous findings (Naoumkina, Modolo et al. 2010) for strong MeJA-induction of triterpene saponin biosynthesis, so the integrated analysis of relative transcript expression and metabolomics datasets from the time series is fundamentally problematic. Difficulty in reproducing the results of plant cell suspension culture experiments is a frequent and serious problem (Miguel and Marum 2011, Richard Dixon, personal communication).

Tnt-1 Mutant Collection Screening

Unlike the case for *Tnt-1* mutants for cyp72a67 and cyp72a68 (JHS_RESEARCH_CHPATER_III), homozygous *Tnt-1_cyp88d2* insertion mutant plants were identified in the NF8050 and NF5409 lines *Tnt-1* lines. If functional copies of cyp72a67 or cyp72a68 gene products are indeed required in seed developmental metabolism, the survival of viable *Tnt-1_cyp88d2* insertion mutant seeds/plants may indicate that cyp88d2 does not function critically in this biological process. As metabolomics analysis of R108 (wild type) and homozygous *Tnt-1_cyp88d2* individuals from NF8050 or NF5409 plants did not reveal differences in triterpene saponin accumulation, it is likely that cyp88d2 either does not function in triterpene sapogenin biosynthesis, or that gene products from other loci can replace/substitute for CYP88D2 function.

Molecular Genetics

The probeset representing cyp88d1 was found to express above baseline (little or no expression) only in roots involved in mycorrhizal symbiosis. Researchers studying mycorrhizal symbiosis may benefit from additional characterization of CYP88D1 enzymatic function, as there may be mycorrhizal symbiosis-specific triterpenoid compounds that are critical for establishment or physiological function in this

symbiotic relationship. The expression profiles for cyp88d2 and cyp88d3 are largely aerial and seed organ specific. The highest absolute expression values for these two transcripts is approximately 10-13 fold lower than the absolute expression values for cyp72a68, cyp72a68, and cyp716a12 in similar tissues (JHS_RESEARCH_CHAPTER_III).

In Vitro Enzymatic Assays of CYP88D1, CYP88D2, and CYP88D3

The lack of detected product formation in the ß-amyrin, erythrodiol, oleanolic acid, hederagenin, or the partially purified aglycone assays for CYP88D1, CYP88D2, or CYP88D3 does not preclude the possibility that these proteins may catalyze reactions in the *M. truncatula* sapogenin biosynthesis pathway. It is possible that the recombinant expression system and/or assay conditions employed in this study may have been inappropriate for proper protein folding/assembly/modification or catalytic function for these proteins.

Methods

Genomic Loci Proximity Analyses

BLAST and genome browser analysis tools for the *M. truncatula* (http://medicago.org/genome/IMGAG/) and *L. japonicas* (http://www.plantgdb. org/LjGDB) were used for the genomic loci proximity analyses.

Growth and Harvest of Biological Materials

ESP105, ESP104 and GRC_43 seeds used in this study were of the same single seed descent lines developed in JHS_REEARCH_CHAPTER_I. A17 and R108 isoline seeds were obtained from the greenhouse manager (David McSweeny) at the Samuel Roberts Noble Foundation. Plants were grown in a root cone system(Stuewe and Sons, http://www.stuewe.com, Tangent, OR) with Turface MVP medium(Profile Products, Buffalo Grove, IL) in a Conviron TCR180 walk-in growth chamber (http://www.conviron.com/, Winnipeg, Manitoba, Canada) maintained at 90% humidity and at an average temperature of 24°C day (16 h) and 20°C night (8 h). Plants were fertilized with 15 ppm nitrogen (Scotts' 20 10 20 Peat-Lite Special, http://www.scotts.com, Marysville, Ohio) daily in the morning and watered with distilled water in the evening. Plants were harvested at 6 weeks post-germination and dissected into aerial and root organs. Aerial tissues from the youngest 6 metamers of individual plants (Bucciarelli, Hanan et al. 2006) and whole root organ samples from individual plants were prepared as single biological replicates. For both aerial and root samples, three biological replicates were prepared for all of the ecotypes. Samples were frozen immediately in liquid nitrogen, ground using a mortar and pestle, and stored at

–80C. The same plant sample material was used for the metabolomics and qRT-PCR analyses.

The *M. truncatula* cell culture line was the sub-cultured descendant of the cell line from previously reported experimentation (Broeckling, Huhman et al. 2005; Naoumkina, Modolo et al. 2010) at the Noble Foundation. The methyl jasmonate (MeJA) induced time series (0,24, and 48 hour) treatment was initiated 5 days after subculturing of the cell suspension cultures. For each sample, 2 X 40 ml of cells were added to 160ml SH medium (Schenk and Hildebrandt 1972) in a 500ml Ehrlenmeyer flask on to yield an approximate volume of 250ml culture. A 50mM MeJA stock solution (in ethanol) was used at a 1:100 dilution (2. 5ml) for a final concentration of 500uM. 2.5ml of ethanol was added to the control flasks. Samples were harvested at the 0,24, and 48 hour time points by vacuum filtration through a 300u Nitex nylon membrane in a Büchnerfunnel using an Ehrlenmeyer side-arm flask and house vacuum. The cells were rinsed in the funnel under vacuum with 50ml of 25% strength MS salts (GibcoBRL Murashige & Skoog salt mixture, Invitrogen, http://www.invitrogen.com). Three biological replicates were prepared for the both control and(+)MeJA groups for all of the time points. Samples were frozen

immediately in liquid nitrogen, ground using a mortar and pestle, and stored at –80C. The same sample material was used for the metabolomics and qRT-PCR analyses.

DNA Preparations, mRNA Isolation, cDNA synthesis, and qRT-PCR Analysis

All genomic DNA isolations were performed as previously described (Sambrook, Russell et al. 2001). For mRNA isolation, total RNA was extracted using TRIZOL reagent (Invitrogen, http://www.invitrogen.com/), treated with DNaseI (Ambion, http://www.ambion.com/), and column purified with an RNeasyMinEluteCleanUp Kit (Qiagen, http://www.qiagen.com/). RNA was quantified using a Nanodrop Spectrophotometer ND-100 (NanoDrop Technologies, http://www.nanodrop.com/) and evaluated for quality with a Bioanalyzer 2100 (Agilent, http://www.home.agilent.com/). For cDNA synthesis preceding qRT-PCR analysis, 10µg of total RNA (prepared and assessed for quality as above) was primed with oligo(dT)20 and synthesized with Super Script III according to manufacturer's instructions. qRT-PCR reactions were performed in an optical 384-well plate with an ABI PRISM 7900 HTsequence detection system (Applied Biosystems), using SYBR Green to monitor dsDNA synthesis. Reactions contained 2µl of primer pair (1 µM), 2µl of 1:20 dilution of cDNA, 5µl of 2x power SYBR Green MASTER MIX, and 1µl water. The following standard thermal profile was used for all PCR reactions: 50°C for 2 min,

95°C for 10 min, 40 cycles of 95°C for 15 s, and 60°C for 1 min. Amplicon dissociation curves were recorded after cycle 40 by heating from 60°C to 95°C with a ramp speed of 1.9°C/min. Primers (Appendix_RIV_Primers) were designed using Primer Express® Software (Applied Biosystems). All reactions were performed with 3 technical replicates for each of 3 biological replicates. Data were analyzed using SDS 2. 2.1 software (Applied Biosystems). PCR reaction efficiencies were determined using LinReg PCR software (Ruijter, Ramakers et al. 2009). Transcript expression levels were determined relative to two housekeeping genes (ubiquitin and actin), based on modifications (equation below) of formulae presented in (Pfaffl 2001; Czechowski, Stitt et al. 2005). Briefly, the ΔCt terms for the target and reference genes were calculated as 41(cycle) minus the Ct vale of a given sample ("x"), rather than "control" minus "treatment" Ct values.

$$relative\ expression\ ratio = \frac{E_{target}^{\Delta Ct_{target}(41-x)}}{E_{ref}^{\Delta Ct_{ref}(41-x)}}$$

Extractions and Metabolomics Analysis

Harvested plant sample material was lyophilized prior to extraction. 10.00 ± 0.06 mg of powder was extracted with 1 ml of 80% Methanol (containing 0.018 mg/ml umbelliferone as an internal standard) in a dram vial for 2 hours on an orbital shaker.

Microsomal enzymatic assay reaction mixtures were extracted 2 times with 500 µl of ethyl acetate, and dried under nitrogen gas. Oleanolic acid, hederagenin, bayogenin, and aglycone mix assay contents were resolubilized in 250µl of 80% Methanol (containing 0. 018 mg/ml umbelliferone as an internal standard). Extracted samples were centrifuged for 30 minutes at 2900g at 4°C, and supernatants were transferred to LC-MS sample vials(Agilent, http://www.agilent.com, Santa Clara, CA) and stored at -20°C. They were then analyzed with a Waters Acquity UPLC system coupled to a hybrid quadropole time-of-flight (QTOF) Premier mass spectrometer (Waters, http://www.waters.com/, Milford, MA). A reverse-phase, 1. 7-mm UPLC BEH C18, column (Waters) was used for separations. The mobile phase consisted of eluent A (0. 1% [v/v] acetic acid/water) and eluent B (acetonitrile), and separations were achieved using a linear gradient of 95% to 30% A over 30 min, 30% to 5% A over 3.0 min, and 5% to 95% A over 3.0 min. The flow rate was 0.56 mL min21, and the column temperature was maintained at 60°C. Masses of the eluted compounds were detected in the negative ESI mode from 50 to 2,000 mass-to-charge ratio. The QTOF Premier was operated under the following instrument parameters: desolvation temperature of 400°C, desolvation nitrogen gas flow of 850 L h21, 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.

Ion List and Metabolomics Data Processing

Waters .raw data files were converted to .cdf file format, followed by metabolite data extraction, alignment, and they were exported using MET-IDEA software (Broeckling, Reddy et al. 2006). An ion list containing 377 retention time/ion pairs was used for the targeted metabolomics data analysis of the saponin biochemical phenotypes (APPENDIX_TARGETED_ION_RIV). An ion list containing 151 retention time/ion pairs was used for the saponin-targeted metabolomics data analysis. 17 of these pairs are validated authentic standards (e.g. 3-Glc-28-Glc-Medicagenic Acid standard), 53 of these pairs are tentatively identified via spectral information (source fragmentation and MS/MS in some cases) as an Aglycone and some combination of sugars (e.g. Hex-Rha-Hex-Hex-Hederagenin), 28 of these pairs have minimal annotation based spectral features resulting from probable source fragmentation (e.g. possibly bayogenin, GlcGlc?), and the remainder are unknowns. The unknown pairs in the ion list were identified with non-targeted MARKERLYNX analysis and had m/z values and retention times in the same regions as the known and putative pairs. In addition to the targeted analysis, *de novo* non-targeted analysis of all samples was

performed using Waters MARKERLYNX software. Spectral abundance signals for all metabolites in a separation were normalized to the internal standard (0.018 mg/ml umbelliferone). Descriptive statistics were performed in Excel. One-way ANOVA was performed using a custom MATLAB script (MathWorks, http://www.mathworks.com/). Multivariate analyses including principal component analysis and hierarchical clustering were performed using JMP 5.0 software (SAS, http://www.sas.com/).

Cloning

All primer sequences and NCBI Genbank (NCBI, http://www.ncbi.nlm.nih.gov/genbank/, Bethesda, MD) accession numbers are presented in Appendix_RIV_Primers. Coding sequences for cyp88d2 and cyp88d3 were obtained from NCBI Genbank, cyp88d1 coding sequence was obtained from the genome sequence of *M. truncatula* (http://www.medicagohapmap.org/?genome). All cloning primers were designed using primer3 (Rozen and Skaletsky 1999). The forward primer for each target included both a BamHI restriction site and a kozak yeast translation initiation sequence, while each reverse primer included an EcoRI cut site. Targets were amplified from cDNA prepared from aerial organs from the isoline A17 using Plantium Hi-Fi Taq polymerase (Life Technologies,

http://www.lifetechnologies.com/home.html, Carlsbad, California). Amplified products were cloned into the pGEM-easy vector (Promega, http://www.promega.com/, Madison, WI), and sequenced using M13 forward and reverse primers. The targets were excised from the p-GEM easy vector via BamHI and EcoRI restriction digest, sub-cloned into the *pYeDP60* vector (Pompon, Louerat et al. 1996; Urban, Mignotte et al. 1997) and sequenced using the *gal10* promoter primer (Appendix_RIV_cloned_sequence). *Wat11* yeast cells were transformed as previously reported (Greenhagen, Griggs et al. 2003). Transformation of yeast was confirmed via colony PCR with gene specific primers.

Recombinant expression and microsomal preparations of CYP72A68 enzymatic assays

The potential catalytic activities of CYP88D1, CYP88D2, and CYP88D3 were tested using *in vitro* assays with a variety of triterpene sapogenin substrates. *Wat11* cells containing *pYeDP60*+target or empty *pYeDP60* were grown and microsomes were prepared as previously described (Greenhagen, Griggs et al. 2003). All assays were performed in triplicate. For assays with purified reference standards as substrates, 100µg of total microsomal protein (quantified via Bradford assay) (Seki, Ohyama et al. 2008) was assayed for 2 hours at 30°C in a 500µl reaction volume of 50mM

potassium phosphate buffer (pH 7.25) containing 1mM NADPH, and 40μM purified substrate. ß-amyrin, erythrodiol, and oleanolic acid were obtained from Sigma-Aldrich (Sigma-Aldrich, http://www.sigmaaldrich.com/, St. Louis, MO). Hederagenin and cycloartenol were obtained from Chromadex (Chromadex, http://www.chromadex.com, Irvine, CA). Bayogenin was obtained from PhytoLab (PhytoLab, http://www.phytolab.com, Vestenbergsgreuth, Germany). A mixture of partially purified aglycones obtained through acid hydrolysis of saponin extracts obtained from *Medicago truncatula* root tissues (Huhman and Sumner 2002) were at assayed at an approximate 80μM concentration.

Pearson Correlation Coefficients for Transcripts vs. Metabolites

Pearson correlation coefficient analysis was performed for [gene of unknown function] vs. [total saponin accumulation] for the inter-ecotype, intra-aerial-organ comparison permutation for both the averaged (n=4) and individual (n=12) cases. A custom MATLAB (MathWorks, http://www.mathworks.com/) script was used to generate bootstrapped confidence intervals and bootstrapped standard errors for Pearson's r for transcript vs. total saponin content correlations for high priority transcripts(using 5000 iterations). The bootstrapping algorithm in the script was the "bbcorr" function (http://www.mathworks.com/matlabcentral/), which computes double block

bootstrap (Lee and Lai 2009) percentile confidence intervals and bootstrap standard errors.

Screening the *M. truncatula Tnt-1* Retrotransposon Insertion Population for Identification transposon insertion mutants

The *M. truncatula* R108 *Tnt-1* population (Million Tadege 2008) was screened for insertions in cyp88d2 and cyp88d3 loci (Appendix_RIV_Primers) as previously described (Pang, Wenger et al. 2009).

Additional Information

Accession Numbers

Currently found in Appendix_RIV_Primers

Appendices

Appendix_RIV_Primers-Primer sequence information for all of the primers used in the qRT-PCR analysis of gene expression, cloning, and *Tnt-1* reverse mutant screening.

APPENDIX_TARGETED_ION_RIV-The 377 Ion/Retention Time pairs used for the targeted metabolomics data analysis of the ecotype UPLC-ESI(-)-qTOF-MS data, including the 151 saponin-target pairs prepared as a separate list.

Appendix_RIV_cloned_sequence-Nucleotide sequence of cloned target genes

Sources

Benedito, V., Torres-Jerez, I., Murray, J., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P., Tang, Y., and Udvardi, M. (2008). Affymetrix GeneChip Medicago Genome Array

A gene expression atlas of the model legume *Medicago truncatula*. Plant J **55,** 504 - 513.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. Anal. Chem. **78,** 4334-4341.

Broeckling, C. D., Huhman, D. V., Farag, M. A., Smith, J. T., May, G. D., Mendes, P., Dixon, R. A., and Sumner, L. W. (2005). Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism

10. 1093/jxb/eri058. J. Exp. Bot. **56,** 323-336.

Bucciarelli, B., Hanan, J., Palmquist, D., and Vance, C. P. (2006). A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development

10. 1104/pp. 106. 082594. Plant Physiol. **142,** 207-219.

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

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W. -R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. Plant Physiol. **139,** 5-17.

Field, B., and Osbourn, A. E. (2008a). Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Plants

10. 1126/science. 1154990. Science, 1154990.

Field, B., and Osbourn, A. E. (2008b). Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Different Plants. Science **320,** 543-547.

Frey, M., Schullehner, K., Dick, R., Fiesselmann, A., and Gierl, A. Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. Phytochemistry **70,** 1645-1651.

Frey, M., Huber, K., Park, W. J., Sicker, D., Lindberg, P., Meeley, R. B., Simmons, C. R., Yalpani, N., and Gierl, A. (2003). A 2-oxoglutarate-

dependent dioxygenase is integrated in DIMBOA-biosynthesis. Phytochemistry **62,** 371-376.

Gierl, A., and Frey, M. (2001). Evolution of benzoxazinone biosynthesis and indole production in maize. Planta **213,** 493-498.

Greenhagen, B. T., Griggs, P., Takahashi, S., Ralston, L., and Chappell, J. (2003). Probing sesquiterpene hydroxylase activities in a coupled assay with terpene synthases. Archives of Biochemistry and Biophysics **409,** 385-394.

He, J., Benedito, V., Wang, M., Murray, J., Zhao, P., Tang, Y., and Udvardi, M. (2009). The *Medicago truncatula* gene expression atlas web server. BMC Bioinformatics **10,** 441.

Huhman, D. V., and 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,** 347-360.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. Plant Molecular Biology **51,** 731-743.

Jonczyk, R., Schmidt, H., Osterrieder, A., Fiesselmann, A., Schullehner, K., Haslbeck, M., Sicker, D., Hofmann, D., Yalpani, N., Simmons, C., Frey,

M., and Gierl, A. (2008). Elucidation of the Final Reactions of DIMBOA-Glucoside Biosynthesis in Maize: Characterization of Bx6 and Bx7. Plant Physiology **146,** 1053-1063.

Lee, S. M. S., and Lai, P. Y. (2009). Double block bootstrap confidence intervals for dependent data. Biometrika **96,** 427-443.

Miguel, C. l., and Marum, L. (2011). An epigenetic view of plant cells cultured in vitro: somaclonal variation and beyond. Journal of Experimental Botany.

Million Tadege, J. W., Ji He, Haidi Tu, Younsig Kwak, Alexis Eschstruth, Anne Cayrel, Gabriella Endre, Patrick X. Zhao, Mireille Chabaud, Pascal Ratet, Kirankumar S. Mysore,. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. The Plant Journal **54,** 335-347.

Mugford, S. T., Qi, X., Bakht, S., Hill, L., Wegel, E., Hughes, R. K., Papadopoulou, K., Melton, R., Philo, M., Sainsbury, F., Lomonossoff, G. P., Roy, A. D., Goss, R. J. M., and Osbourn, A. (2009). A Serine Carboxypeptidase-Like Acyltransferase Is Required for Synthesis of Antimicrobial Compounds and Disease Resistance in Oats. The Plant Cell Online **21,** 2473-2484.

Mylona, P., Owatworakit, A., Papadopoulou, K., Jenner, H., Qin, B., Findlay, K., Hill, L., Qi, X., Bakht, S., Melton, R., and Osbourn, A. (2008). Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat

10. 1105/tpc. 107. 056531. Plant Cell **20,** 201-212.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., and Dixon, R. A. (2010). Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*. Plant Cell **22,** 850-866.

Osbourn, A. (2010). Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. Plant Physiology **154,** 531-535.

Pang, Y., Wenger, J. P., Saathoff, K., Peel, G. J., Wen, J., Huhman, D., Allen, S. N., Tang, Y., Cheng, X., Tadege, M., Ratet, P., Mysore, K. S., Sumner, L. W., Marks, M. D., and Dixon, R. A. (2009). A WD40 Repeat Protein from *Medicago truncatula* Is Necessary for Tissue-Specific Anthocyanin and Proanthocyanidin Biosynthesis But Not for Trichome Development. Plant Physiol. **151,** 1114-1129.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RTPCR. Nucleic Acids Research **29,** e45.

Pompon, D., Louerat, B., Bronine, A., Urban, P., Eric, F. J., and Michael, R. W. (1996). [6] Yeast expression of animal and plant P450s in optimized redox environments. In Methods in Enzymology (Academic Press), pp. 51-64.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., and Osbourn, A. (2004). A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants

10. 1073/pnas. 0401301101. Proceedings of the National Academy of Sciences of the United States of America **101,** 8233-8238.

Qin, B., Eagles, J., Mellon, F. A., Mylona, P., Peña-Rodriguez, L., and Osbourn, A. E. High throughput screening of mutants of oat that are defective in triterpene synthesis. Phytochemistry **71,** 1245-1252.

Rozen, S., and Skaletsky, H. (1999). Primer3 on the WWW for General Users and for Biologist Programmers, pp. 365-386.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M. H., Karlen, Y., Bakker, O., van den Hoff, M. J. B., and Moorman, A. F. M. (2009). Amplification

efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research **37,** e45.

Sakamoto, T., Miura, K., Itoh, H., Tatsumi, T., Ueguchi-Tanaka, M., Ishiyama, K., Kobayashi, M., Agrawal, G. K., Takeda, S., Abe, K., Miyao, A., Hirochika, H., Kitano, H., Ashikari, M., and Matsuoka, M. (2004). An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice. Plant Physiology **134,** 1642-1653.

Sambrook, J., Russell, D. W., and Cold Spring Harbor, L. (2001). Molecular cloning : a laboratory manual / Joseph Sambrook, David W. Russell. (Cold Spring Harbor, N. Y. :: Cold Spring Harbor Laboratory).

Schenk, R. U., and Hildebrandt, A. C. (1972). Medium and techniques for induction and growth of monocotyledonous and dicotyledonous plant cell cultures. Canadian Journal of Botany **50,** 199-204.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K., and Muranaka, T. (2008a). Licorice Î²-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,** 14204-14209.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K., and Muranaka, T. (2008b). Licorice beta-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,** 14204-14209.

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

Shimura, K., Okada, A., Okada, K., Jikumaru, Y., Ko, K. -W., Toyomasu, T., Sassa, T., Hasegawa, M., Kodama, O., Shibuya, N., Koga, J., Nojiri, H., and Yamane, H. (2007). Identification of a Biosynthetic Gene Cluster in Rice for Momilactones. Journal of Biological Chemistry **282,** 34013-34018.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., and Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. The Plant Journal **32,** 1033-1048.

Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B., and Peters, R. J. (2009). CYP76M7 Is an ent-Cassadiene C11Î±-Hydroxylase Defining a Second Multifunctional Diterpenoid Biosynthetic Gene Cluster in Rice. The Plant Cell Online **21,** 3315-3325.

Urban, P., Mignotte, C., Kazmaier, M., Delorme, F., and Pompon, D. (1997). Cloning, Yeast Expression, and Characterization of the Coupling of Two Distantly Related Arabidopsis thaliana NADPH-Cytochrome P450 Reductases with P450 CYP73A5

10. 1074/jbc. 272. 31. 19176. J. Biol. Chem. **272,** 19176-19186.

Wilderman, P. R., Xu, M., Jin, Y., Coates, R. M., and Peters, R. J. (2004). Identification of Syn-Pimara-7,15-Diene Synthase Reveals Functional Clustering of Terpene Synthases Involved in Rice Phytoalexin/Allelochemical Biosynthesis. Plant Physiology **135,** 2098-2105.

Conclusion

The research detailed in this dissertation adds to several areas of knowledge within the fields of Plant Biology, Plant Natural Product Biochemistry, and Cytochrome P450 Enzymology.

The research indicates that it appears unlikely that triterpene sapogenin biosynthesis does not take place at a single "source" tissue and subsequently transport sapogenins or saponins. Rather, the observation that saponins of soyasapogenols B and E were not detected in the aerial organs of any of the ecotypes lends additional support to the conjecture that the bio-oxidation of carbon 22 and carbon 24 of ß-amyrin does not occur in aerial organs. Similarly, the observation that saponins zanhic acid were not detected in the root organs of any of the ecotypes indicates that bio-oxidation of carbon 16 is unlikely to occur in root organs. As was highlighted extensively in Chapter I, the metabolomics profiles for the ecotypes ESP_105 and GRC_43 indicate that a low total saponin accumulator for aerial organs may be a high accumulator in root organs, and vice versa. Taken together with the organ-specific accumulation for soyasapogenol and zanhic acid saponins, these observations of varied spatial accumulation patterns offer a fundament of data to address questions about the physiological location of triterpene saponin biosynthesis.

The results of the metabolomics profiling are also of potential utility to other researchers in related fields. Indeed, metabolomics "phenotyping" of germplasm on the scale presented in Chapter I is just now emerging as a viable experimental complement to genome-scale investigations. As such, it is difficult to assess the ultimate utility of studies at this scale, and exciting novel applications of such data may emerge as more researchers become aware of and come to understand the content, promise and limitations of these high-resolution metabolomics phenotypes. The metabolomics profiling data for the germplasm diversity collection could also be paired with modern genomics-level information and methodologies such as genome wide association studies (GWAS) to associate the observed phenotypic diversity with genomic and allelic variation from the same germplasm. In addition to genomics-level investigations, ecologists interested in plant-insect or plant-pathogen interactions could use the aglycone-specific saponin accumulation data to select ecotypes with the appropriate chemical profiles to investigate hypotheses concerning the role of specific compounds in ecological interactions. Similarly, physiologists interested in abiotic stresses could use the profiling data to select appropriate ecotypes for structure-specific investigations.

Researchers interested in cytochrome P450 and glycosyltransferase enzymes from *Medicago truncatula* should benefit from the concatenated lists of cytochrome P450 and glycosyltransferase annotations for the probe design sequences of the Affymetrix Medicago Gene Chip, as this list represents the most comprehensive such resource available to date.

The success of the experimental matrix of the selected ecotypes and organs in identifying gene candidates for the target trait (triterpene saponin biosynthesis, as indicated by the successful in vitro characterization of CYP72A67 and CYP72A68) offers a form of validation of several aspects of the integrated transcript and metabolite methodology. The combined analyses of the total saponin accumulation data with the transcript expression levels lent support to the predicted supposition that there is a strong relationship between these discrete properties/levels of biological organization. The supposition that low accumulation of total saponin content would be associated with low expression levels for transcripts of biosynthetic genes from the target pathway was bolstered by the data, and offers validation for the notion of selected hypo-accumulator ecotypes as an operational form of "knock-down mutant" in transcriptomics studies. If one accepts the successful cases of cyp72a67 and cyp72a68 as validation of the relationship between transcript expression levels and saponin

accumulation, the formalization of the relationship (Equations 1 & 2) between the observed crude phenotypes ("high vs. low total saponin content") and transcript expression levels from two permutations of the experimental matrix implicates numerous additional gene candidates for involvement in the biosynthesis and/or regulation of the target pathway. Finally, the observed relationship between transcript expression levels and saponin accumulation data among the two different organ types lends further support to the case for localized (i.e. non-centralized) biosynthesis of triterpene saponin compounds.

Expanded Repertoire of Reaction Pairs

The biochemical results offer new knowledge about the molecular basis of the bio-oxidation reactions of the triterpene sapogenin biosynthetic pathway. Characterization of enzymatic activities for CYP72A67 and CYP72A68 contributes to the larger efforts of gene discovery and functional annotation in the model legume *M. truncatula* specifically and plant functional genomics generally. The *in vitro* enzymology results provide new structure and direction to the bio-oxidation pathway for triterpene sapogenins. Specifically, the results show that CYP72A67 and CYP72A68 accept substrates with bio-oxidation at carbon 28, and do not accept substrates lacking bio-oxidation at carbon 28. The broad substrate tolerance demonstrated for CYP72A67

and CYP72A68 adds to the growing evidence of broad substrate tolerance as a frequently observed property of cytochrome P450 enzymes which function in plant natural product metabolism. Further, the multifunctionality demonstrated for CYP72A68 provides yet another example of the sequential bio-oxidation activities of cytochrome P450 enzymes which function in plant metabolism.

Physiology of Seed Development in *M. truncatula* (Lethality and Organ-Specific Expression)

The lethal mutations observed for insertion mutants of cyp72a67 and cyp72a68 loci, in combination with the noted high levels of expression of cyp72a67 and cyp72a68 transcripts in developing seed organs raises a tantalizing prospect of a critical physiochemical role for triterpene sapogenin compounds in the seed developmental physiology of *M. truncatula.*

Matrix Pathways

The use of conceptual "matrix" pathways in the place of more traditional "linear" pathways was shown to be profitable in the case of the triterpene sapogenin biosynthetic pathway in *M. truncatula.* The hypothetical model of the matrix pathway predicted the likely presence of several minor compounds (gypsogenin, gypsogenic

acid, polygaligenin) that were subsequently identified (putatively) in extracts of both plant material and enzymatic reactions. The dramatic reduction in dimensionality afforded through the imposition of a hypothetical matrix pathway was useful in that it greatly reduced the number of probable intermediate compounds to account for when analyzing complex analytical results.

Appendices

Appendix RII – Gross Phenotype Comparisons

Gross Phenotype Comparison Rank	Probeset	Sequence source for probeset design	BLAST ID (*M. truncatula*)	Value from Equation 2	GRC43 Root	A17 Root	ESP105 Root	R108 Root		GRC43 Aerial
1	Mtr.50075.1.S1_s_at	IMGAG|986. m00012	AC140721.15	10109	13214	56	447	11737		13
2	Mtr.40021.1.S1_at	TC106322	AC183304.14	5070	6730	853	5140	300		13
3	Mtr.8678.1.S1_at	TC100985	CU468275.4	3391	4130	489	2434	729		1873
4	Mtr.43645.1.S1_at	TC95765	CR962121. 2	1009	2400	709	1107	1028		131
5	Mtr.12441.1.S1_at	TC94806	CU468276.4	918	3364	752	2085	1520		1755
6	Mtr.11646.1.S1_at	TC110595	CR962122. 2	908	1183	338	327	429		1024
7	Mtr.43644.1.S1_at	TC95764	*Medicago truncatula* class III HD-Zip protein CNA2	838	886	268	237	196		29
8	Mtr.9513.1.S1_at	TC103568	MYB transcription factor MYB52	664	529	52	117	125		311
9	Mtr.42879.1.S1_at	TC94010	AC135413.43	644	1877	425	1045	862		713
10	Mtr.35791.1.S1_s_at	TC94009	AC137603.16	642	1600	354	855	710		609
11	Mtr.49779.1.S1_at	IMGAG|1198.m00032	AC148242.14	581	692	272	32	299		626
12	Mtr.50733.1.S1_at	IMGAG|1042.m00006	nsp2 gene for GRAS family protein ||320||	440	273	21	23	36		8
13	Mtr.1885.1.S1_at	BE239880	No sig similarity found	410	2547	774	1981	758		775
14	Mtr.23572.1.S1_at	1681.m00026	zinc finger transcription factor ||319|| palmate-like pentafoliatea1	335	948	152	608	412		25
15	Mtr.5719.	BF646019	No sig	283	1007	142	705	450		1113

	1.S1_at		similarity found							

Appendix RII – Hidden Markoff Models

Cytochrome P450 Models

http://supfam.org/SUPERFAMILY/cgi-bin/models_list.cgi?sf=48264;listtype=sf

Model ID	No. of seqs	Build date	Seed sequence	Family of seed
54912	10095	5/31/2010	d3czha1	Cytochrome P450
44092	4008	9/8/2005	d1tqna_	Cytochrome P450
41802	3250	9/8/2005	d1odoa_	Cytochrome P450
39101	2672	9/8/2005	d1io7a_	Cytochrome P450
36587	3890	9/8/2005	d1dt6a_	Cytochrome P450
44304	2788	9/8/2005	d1ueda_	Cytochrome P450
42400	3466	9/8/2005	d1po5a_	Cytochrome P450
40482	2863	9/8/2005	d1lfka_	Cytochrome P450
36859	3920	9/8/2005	d1e9xa_	Cytochrome P450
49296	9115	9/10/2008	d1s1fa_	Cytochrome P450
42621	3260	9/8/2005	d1q5da_	Cytochrome P450
41273	3727	9/8/2005	d1n97a_	Cytochrome P450
38307	3037	9/8/2005	d1gwia_	Cytochrome P450
36106	3408	9/8/2005	d1cpta_	Cytochrome P450
44298	3271	9/8/2005	d1ue8a_	Cytochrome P450
41837	3864	9/8/2005	d1og2a_	Cytochrome P450
39299	3007	9/8/2005	d1izoa_	Cytochrome P450
36682	2010	9/8/2005	d1dz4a_	Cytochrome P450
46634	5531	9/10/2008	d1eupa_	Cytochrome P450
42421	3878	9/8/2005	d1pq2a_	Cytochrome P450
41213	2523	9/8/2005	d1n40a_	Cytochrome P450
37346	2629	9/8/2005	d1f24a_	Cytochrome P450
35710	3862	9/8/2005	d1bu7a_	Cytochrome P450

Glycosyltransferase Models

http://supfam.org/SUPERFAMILY/cgi-bin/models_list.cgi?sf=53756

Model ID	No. of seqs	Build date	Seed sequence	Family of seed
35835	2	9/8/2005	d1c3ja_	beta-Glucosyltransferase (DNA-modifying)
46669	450	9/10/2008	d1f0ka_	Peptidoglycan biosynthesis glycosyltransferase MurG
46716	797	9/10/2008	d1f6da_	UDP-N-acetylglucosamine 2-epimerase
48500	497	9/10/2008	d1o6ca_	UDP-N-acetylglucosamine 2-epimerase

44616	216	9/8/2005	d1v4va_	UDP-N-acetylglucosamine 2-epimerase
52910	1192	9/10/2008	d2gj4a1	Oligosaccharide phosphorylase
48983	881	9/10/2008	d1qm5a_	Oligosaccharide phosphorylase
51304	497	9/10/2008	d1ygpa_	Oligosaccharide phosphorylase
46587	843	9/10/2008	d1em6a_	Oligosaccharide phosphorylase
49238	694	9/10/2008	d1rrva_	Gtf glycosyltransferase
42392	283	9/8/2005	d1pn3a_	Gtf glycosyltransferase
47375	1298	9/10/2008	d1iira_	Gtf glycosyltransferase
38341	280	9/8/2005	d1gz5a_	Trehalose-6-phosphate synthase, OtsA
48814	1572	9/10/2008	d1pswa_	ADP-heptose LPS heptosyltransferase II
51981	6238	9/10/2008	d2bisa1	Glycosyl transferases group 1
52623	10289	9/10/2008	d2f9fa1	Glycosyl transferases group 1
53131	4324	9/10/2008	d2iw1a1	Glycosyl transferases group 1
43452	1301	9/8/2005	d1rzua_	Glycosyl transferases group 1
52533	17	9/10/2008	d2ex0a1	Sialyltransferase-like
54538	3989	5/31/2010	d2pq6a1	UDPGT-like
53693	3513	5/31/2010	d2acva1	UDPGT-like
54711	3669	5/31/2010	d2vcha1	UDPGT-like
53762	3891	5/31/2010	d2c1xa1	UDPGT-like
54312	610	5/31/2010	d2nzwa1	FucT-like

Appendix RII– Targeted Ion List

Authentic Standards David Huhman

161.0239 (4.45, Umbelliferone)
1087.4955 (13.728, 3-Glc-Glc-28-Ara-Rha-Xyl Medicagenic Acid)
1383.6057 (11.26, O Zhan Stand)
1545.6609 (11.39, O Zhan Stand)
1545.6595 (11.56, O Zhan)
941.5112 (17.955, Rha-Gal-GlcA-SoyB)
1073.5208 (14.15, 3-Glc-28-Ara-Rha-Xyl Medicagenic Acid)
351.0705 (2.4, ChlorogenicAcid)
269.0441 (10.189, Apigenin)
269.0441 (2.449, pel_3_O)
267.0671 (12.53, 7-hydroxy-3-methoxyflavone)
415.1044 (4.17, Daidzin)
415.1044 (4.42, Daidzin)
163.0386 (4.99, m-coumaric acid)
445.079 (6.51, quercitrin)
431.0962 (7.18, apigenin-7-O-glu)
269.248 (32.39, 6-hydroxy genistein)
285.0389 (1.96, CY_3_O)
175.0378 (6.35, 4MethylUmbelliferone)
271.0606 (6.83, Narigenin-7-O-glucoside - aglycone)
269.0816 (14.11, medicarpin)
431.0989 (5.97, Genistin)
193.0482 (4.9, Ferulic acid)
433.1123 (6.83, Narigenin-7-O-glucoside)
381.0609 (4.84, scopoletin dimer)
315.0123 (6.95, myricetin)
299.0182 (8.81, Quercetin)
447.0925 (6.1, luteolin-5'- 7-O-glucoside)
283.0606 (15.173, Biochanin)
283.061 (15.37, 7-methoxy apigenin)
179.0367 (2.78, caffeic acid)
269.045 (10.06, Genistein)
163.0382 (6.12, o-CoumaricAcid)
593.1554 (6.99, kaempferol-7-neohesperidoside)
299.023 (8.64, Quercetin)
503.1612 (0.64, Cellotrise)
447.0938 (7.05, kaempferol-7-O-glucoside)
447.0924 (7.1, luteolin4'- 7-O-glucoside)

593.1505 (6.25, kempferol-3-O-rutinoside)
461.0737 (5.43, CY_3_O)
299.0566 (10.894, chrysoecin)
299.057 (11.07, diosmetin)
267.0663 (8.47, Formononetin_7_OG)
267.0656 (8.64, Ferulic acid)
193.0492 (5.37, 3-Hydroxy-4-methoxycinnamic acid)
579.175 (7, Naringin)
285.0502 (8.36, eriodictyol)
459.0567 (5.38, scutellarein-&-0-glucuronide)
283.0602 (10.92, Sissotrin)
301.0335 (5.08, taxifolin)
447.0597 (6.69, luteolin-4-O-glu)
607.1699 (7.478, diosmin)
117.0188 (0.89, succinic acid)
609.1505 (5.14, luteolin-3-7-di-O-glu)
579.1342 (5.95, luteolin 7-O-glucoside)
285.0407 (8.68, Luteolin)
325.0919 (4.25, o-Coumaric acid-B-glucoside)
299.0565 (15.26, kaempferide)
153.0178 (1.76, gentisic acid)
167.0346 (4.9, 5-Methoxysalicylic acid)
283.0614 (10.68, sissotrin)
255.0669 (7.894, Liquiritin)
431.098 (5.54, genistoside)
431.0971 (5.56, Vitaxin)
283.0257 (8.96, Luteolin)
271.0607 (9.88, Naringenin)
271.0598 (10.07, Naringenin)
289.0726 (3.22, Epicatechin)
135.1213 (2.07, 1-Hydroxybenzotriazole)
301.0353 (4.862, DHQ)
267.0667 (12.35, 7-hydroxy-2-methoxyflavone)
255.0667 (11.7, isoliquiritigenin)
577.1594 (7.135, rhoifolin)
284.0313 (11.05, scutellarein)
417.1193 (19.03, gardenin A)
253.0468 (14.76, chrysin)
163.0385 (3.92, p-Coumaric acid)
451.1235 (3.09, EpicatechinGlucoside)
285.0396 (10.34, Kaempferol)

255.0662 (8.13, Liquiritigenin)
433.1139 (4.81, naringenin 4'-O-glucoside)
285.0398 (9.11, Luteolin)
237.0546 (17.3, 3Hydroxyflavone)
299.0914 (14.03, faureral)
415.104 (3.45, Puerarin)
191.0344 (4.9, Scopoletin)
177.0158 (2.81, Esculetin)
447.0931 (5.4, luteolin3'- 7-O-glucoside)
253.0481 (8.25, Daidzein)
283.0607 (9.08, Glycitein)
144.0464 (3.69, alpha-Cyano-3-hydroxycinnamicacid fragment)
431.0971 (6.451, Genisitin)
663.3777 (17.53, 3-Glc-MedicagenicAcid)
285.0398 (10.61, Kaempferol)
445.112 (10.92, sissotrin)
358.0234 (0.84, 50ngSinigrin_MW397_46 Indofine)
349.0707 (6.35, 4MethylUmbelliferone - Dimer)
1677.7001 (11.45, O Zhan Stand)
313.0713 (15.84, irisolidone)
237.0551 (18.72, 5-hydroxyflavone)
301.0332 (8.81, Quercetin)
289.0712 (2.4, Catechin)
957.4825 (17.4, Glc-Gal-GlcA-SoyB)
957.5084 (17.414, Soy Mix Stand)
287.0545 (2.4, Catechin fragment)
593.1508 (4.45, saponarin)
151.0382 (4, isoVanillin)
237.0545 (11.95, 7Hydroxyflavone)
465.1035 (1.96, CY_3_O)
237.0546 (11.64, 4Hydroxyflavone)
609.1814 (7.64, hesperidin)
593.1312 (9.53, tiliroside)
449.108 (2.449, pel_3_O)
315.0515 (9.28, 6-methoxyluteolin)
315.087 (12.62, eucomol)
227.0721 (7.12, resveratrol)
151.02 (1.76, gentisic acid fragment)
267.0669 (11.95, 4-hydroxy-7-methoxy flavone)
267.0295 (6.77, orionin)
237.0557 (13.2, 6Hydroxyflavone)

267.0285 (12.25, Baicalein)
267.0666 (12.07, 7-hydroxy-4-methoxyflavone)
237.0542 (13.37, 2Hydroxyflavone)
149.0602 (8.4, HydroCinnamicAcid)
607.1701 (7.82, NeoDiosmin)
609.1821 (8, neohesperidin)
237.0541 (12, 5-hydroxy-flavone)
301.0716 (10.56, homo-eriodictyol)
1007.3277 (0.64, Cellotrise - Dimer)
911.5005 (18.303, Rha-Ara-GlcA-SoyB)
473.1038 (13.2, 6Hydroxyflavone - Dimer)
577.1187 (4.1, lucenin B)
473.1036 (11.95, 7Hydroxyflavone - Dimer)
343.0826 (15.48, nevadensin)
577.1584 (6.76, isorhoifolin)
265.0492 (5.51, tectochrysin)
408.0443 (1.82, 50ngBenzylglucosinolate_C14H18O9NCH3_4Canada)
147.044 (8.92, trans-cinnamic acid)
422.0589 (2.95, 50ngPhenylEthylglucosinolate_Chromadex)
473.1023 (18.97, 5-Hydroxyflavone - Dimer)
471.3469 (23.17, hederagenin)
269.0446 (7.3, 6,7,4-Trihydroxyflavone)
285.04 (4.89, 3',4''7'8-tetrahydroxyflavone)
285.0407 (5.039, 7,3,4,5-tetrahydroxyflavone)
285.0418 (6.84, 3,6,2,4-tetrahydroxyflavone)
285.0423 (7.25, 3,3,3,4-tetrahydroxyflavone)
153.0168 (1.33, 3,5 dihydroxybenzoic acid)
283.0602 (15.31, 4,5-dihydroxy-7-methoxy isoflavone)
283.06 (15.375, 3',4'-O-methoxyflavone)
179.035 (2.79, 5,6,7-OH flavone)
301.0353 (6.989, 5,7,3,4,5-penthahydroxyflavone)
609.1476 (5.4, luteolin-3,7-O-glu)
285.0413 (8.4, 3,6,2,3-tetrahydroxyflavone)
253.0477 (10, 4',6-dihydroxy aurone)
253.0475 (10.06, 4,6,OH-aurone)
195.0645 (5.53, 3,5,Dimethoxy4Hydroxyacetophenone)
253.0491 (8.05, 7,4-OH-flavone)
297.0791 (8.85, 4,6-dimethoxyisoflavone-7-O-beta-D-glucoparanoside)
311.0909 (13.12, 3',4'-methoxy-7-hydroxy-flavone)
341.0981 (15.21, 2-OH-5,7-dimethoxy-isoflavone)
301.0373 (5.022, 3,7,3,4,5-pentahydroxyflavone)

299.0946 (12.15, 5,7-dimethoxyapigenin)
341.1035 (16.22, 7,4-dimethoxyflavone)
253.0475 (11.75, 3,4-dihydroxy aurone)
151.0347 (6.89, 2,6,Dihydroxyacetphenone)
151.0025 (1.51, 3,4DiHydroxyBenzoicAcid)
151.0399 (5.55, 2,4,Dihydroxyacetophenone)
151.039 (4.81, 2,5,Dihydroxyacetophenone)
223.0607 (5.15, 3,5-dimethoxy-4-hydroxy oxycinnamic acid)
285.0761 (14.28, 5,7-dihydroxy4'-methoxy-flavone)
341.1031 (14.71, ?5,4-dimethoxyflavone)
297.0778 (17.81, 7,4-dimethoxy-3hydroxyflavone)
167.0346 (5.01, 2,4,6,Trihydroxyacetophenone)
297.0763 (19.72, 4,7-dimethoxy apigenin)
267.0285 (12.18, 5,6,7-trihydroxy-flavone)
237.0549 (12.71, 3',7-hydoroxy-flavone)
297.0761 (12.51, 3,3',4'-methoxy-phenylo-7-OH-Coumarin)
371.1106 (18.4, 5,7,4-trimethoxy aurone)
327.0863 (21.03, Kaempferol-3,7,4-trimethyl ether)
401.1241 (16.62, 7,8,3,4-tetramethoxy aurone)
224.046 (17.71, 1,8,9-Anthracenetriol)

Authentic Standards John Snyder

469.33181 (24.73, GLYCYRRHETINIC ACID, 18Beta)
471.34746 (24.08, COROSOLIC ACID)
471.34746 (23.08, HEDERAGENIN)
471.34746 (23.15, PYGENIC ACID A)
471.3469 (23.17, hederagenin)
487.342375 (18.44, ASIATIC ACID)
455.352545 (22.81, Oleanolic acid)

Literature and Empirical Validation David Huhman

462.0923 (7.66, 7-Methylthio-n-heptyl-glucosin)
494.0787 (6.83, 4-Benzoyloy-n-butyl-glucosino)
524.0737 (10.924, 6'Benzoyl-4-methyl-sulfinyl-bu)
402.0891 (4.92, 4-Methylpentyl-glucosio)
448.0772 (5.549, 6-Methylthio-n-heyl-gluc)
434.0621 (3.6, 5-Methylthio-n-pentyl-glucosinolate)
480.0628 (5.351, 3-Benzoyloy-ethyl-glucosinola)
406.03 (1.56, 3-Methylthio-n-propyl-glucosinolate)
436.039 (0.86, 4-Methylsulfinyl-n-butyl-gluco)
447.0516 (2.613, Indol-3-ylmethyl-glucosinolat)

463.0471 (2.613, Methoyindol3-ylmethyl glucosinolae)
430.1207 (9.971, iso-Heptyl glucosinolate from Armoracia lapathifolia)
416.1057 (7.205, iso-Hexylglucosinolate from Armoracia lapathifolia)
376.0383 (0.7, 3-Hydroy-n-propyl-gluc)
464.0729 (1.349, 6-Methylsulfinyl-n-heyl-glucosinolate)
492.1024 (3.22, 8-Methylsulfinyl-n-octyl glucosinolate)
420.0447 (2.14, 4-Methylthio-n-butyl-glucosinolate)
376.0371 (0.86, 3-Hydroypropyl glucosinolate)
390.0516 (0.86, 4-Hydroy-n-butyl-glucosinolate)
476.108 (9.99, 8-Methylthio-n-octyl-gluc)
372.0416 (1.69, 3-Butenyl-glucosinolate)
450.0562 (0.97, 5-Methylsulfinyl-n-pentyl-glucosinolate)
478.0874 (2.02, 7-Methylsulfinyl-n-heptyl-glu)

Putative Identification Based on MS/MS analysis John Snyder

1067.5469 (19.58, soyasapogenol B_2x Rha, Hex)
1383.6111 (11.69, Gypsogenin_Arab/xyl,)
1235.5293 (11.316, zanhic acid_2x GlcA, Arab/xyl)
1251.5713 (11.963, zanhic acid?)
1397.5726 (11.186, zanhic acid_2x arab/xyl, GlcA, Hex)
1221.5436 (11.926, Gypsogenin_Arab/xyl)
1221.562 (11.963, Gypsogenin_Arab/xyl,)
1103.5247 (11.39, zanhic acid_GlcA, more)
1265.5645 (11.154, zanhic acid_?)
1265.5499 (11.315, zanhic acid_ ?)
1205.5675 (13.756, Gypsogenin_?)
1205.5549 (13.83, many possible_3x Arab/xyl, 2x Hex?)
1161.5352 (13.793, ?)
1089.5249 (12.06, zanhic acid_HEX, Arab/xyl)
1089.5249 (12.167, ?)
1089.5249 (11.389, zanhic acid_?)
1089.5249 (19.617, gypsogenic acid_?)
1089.5249 (19.525, gypsogenic acid_3x Rha, hex?)
1089.5249 (12.203, bayogenin_3x Hex, Rha)
1089.5249 (13.424, ?)
1089.5249 (14.089, hederagenin_ara/xyl, 2x HEX)
1089.5249 (14.699, Gypsogenin?_3x Hex)
1089.5249 (11.389, bayogenin_3x Hex, Rha)
469.335 (24.96, possible_Gypsogenin)
469.3321 (22.01, possible_Gypsogenin)
469.332 (23.77, possible_Gypsogenin)

Putative Identification based on m/z and/or fragmentation David Huhman

1087.4974 (13.81, Mt Leaf)
1067.5479 (19.89, Mt Leaf)
1383.6086 (11.84, Mt Leaf)
1383.608 (11.97, Mt Leaf)
1235.5402 (11.61, Mt Leaf)
1251.5621 (12.138, Mt Leaf)
1397.5889 (11.29, Mt Leaf)
1221.554 (12.059, Mt Leaf)
1103.4962 (11.77, Mt Leaf)
1205.5613 (13.93, Mt Peak)
1161.5363 (13.9, Mt Peak)
1089.5117 (12.225, Mt Leaf)
1251.5675 (11.5, Mt Leaf)
941.5107 (17.877, Ara-Rha-GlcA-Bayogenin)
941.5099 (17.6709, 3-Rha-Gal-GlcA-Soyasapogenol B)
1367.6135 (13.63, 3Glc-Glc-28-Ara-Rha-Xyl-Api-Med)
1083.5422 (19.89, Mt Leaf)
1235.5725 (13.8147, Mt Peak)
1235.5703 (13.827, 3-GlcA-28-Ara-Rha-Xyl Medicagenic Acid)
1235.5771 (13.84, Mt Peak Leaf)
793.5449 (33.93, Rha-?)
957.507 (13.65, Hex-Hex-Hex-Hederagenin)
957.5068 (13.85, Hex-Hex-Hex-Hederagenin)
957.4758 (13.979, Glc-Gal-GlcA-SoyB)
1307.5969 (13.562, Mt Leaf)
939.498 (19.535, 3-Rha-Xyl-GlcA)
971.4856 (14.16, Mt peak)
1119.564 (10.18, Mt Leaf)
1085.5544 (16.206, Leaf)
809.4349 (17.945, Hex-HexA-Hederagenin)
823.4152 (14.05, Hex-HexA-New Aglycone)
955.495 (14.93, Rha-Hex-?)
1089.5494 (10.25, Mt Leaf)
973.5025 (12.84, Hex-Hex-Hex-Bayogenin)
1413.6189 (11.609, Mt Leaf)
987.4836 (12.536, Mt Root)
987.4818 (12.52, Mt Root)
823.4148 (14.09, GlcA-Glc-NewAglycone)
809.4337 (13.816, Hex-HexA-Hed)

955.4971 (14.376, Mt peak-?)
927.497 (15.1, Hex-Hex-Pent-Hederagenin)
985.4704 (12.928, Mt Root)
1119.5602 (11.57, Mt Root)
1145.5433 (13.81, Mt Leaf)
911.4344 (14.06, 3-Glc-28-Glc-Malonyl-Med)
793.4389 (16.23, HexA-Hex-Soy E)
795.4543 (17.945, Gal-GlcA-SoyB)
941.514 (16.97, Rha-Hex-Hex-Hederagenin)
811.4469 (12.441, GlcGlcBayogenin)
811.4475 (12.47, Hex-Hex-Bayogenin-?)
925.4822 (19.55, Hex-Hex-Rha-SoyE)
955.4933 (15.26,)
955.4926 (15.356, GlcA-?)
793.4408 (19.62, Hex-HexA-455 ?)
811.4481 (16.92, Hex-Hex-Bayogenin)
925.5173 (14.86, Rha-Hex-Hex-SoyE may be related to 1087)
987.4858 (12.466, Hex-Hex-Hex-Med Scotts work)
1043.5476 (17.45, ?)
795.4542 (18.357, Hex-Hex-Hederagenin?)
957.5092 (16.05, Rha-Hex-Hex-Bayogenin)
809.4334 (18.78, Glc-Glc-hed?)
941.5093 (17.96, Leaf)
1057.4866 (13.65, Mt Peak Root)
1101.5518 (13.55, Glc)
795.4526 (19.819, Hex-Hex-Hederagenin)
989.4892 (8.257, Mt Root)
825.4666 (15.77, GlcGlcMed?)
825.4643 (15.95, GlcGlc?)
649.3969 (17.47, Hex-Bayogenin)
649.394 (17.44, HexA-Bayogenin)
663.3756 (15.49, Hex-Medicagenic Acid)
663.3762 (17.48, Hex-Medicagenic Acid)
1087.5739 (14.86, Hex-Rha-Hex-Hex-SoyE)
647.3831 (14.14, Hex-New Aglycone)
1119.5729 (7.845, GlcRha?)
809.4335 (15.29, Hex-HexA-Hederagenin)
749.4464 (15.57, Hex-Pent-Soyasapogenol E)
765.4424 (19.41, Hex-Hex-Hederagenin)
989.4884 (9.54, Glc)
1119.5544 (12.3889, Rha-Hex-Hex-Hex-Bayogenin)

1103.517 (10.07, Mt Leaf)
809.4333 (15.77, HexA-Hex-Hederagenin)
1159.4995 (14.143, Mt peak-?)
647.3798 (18.78, Hex-New Aglycone)
501.3228 (19.881, Mediagenic Acid)
795.4525 (14.99, Hex-Hex-Hederagenin)
1085.5581 (16.231, Saponin?)
1117.5343 (10.661, Mt Root)
825.4304 (16.998, 3-Glc-Glc-MedicagenicAcid?)
647.4343 (20.997, GlcA-Hederagenin)
765.4431 (18.552, Ara-GlcA-SoyB)
809.4341 (19.05, Unknown - Hed)
1105.5775 (10.337, Glc)
957.507 (9.305, Mt Root)
647.3817 (21.57, GlcA-Hederagenin)
809.4313 (19.243, Unknown)
631.3854 (22.029, ?)
1073.5581 (17.02, ?)
765.4431 (18.994, GlcAHed)
925.5151 (19.028, ?)
705.3849 (18.28, 3-Glc-Malonyl-MedicagenicAcid)
855.4741 (15.86, GlcAGlcHed?)
1113.5566 (19.29, Unknown)
989.5104 (8.928, Glc)
487.3421 (20.637, Bayogenin)
1367.5752 (11.666, Mt Leaf)
971.4877 (16.129, HexHex)
633.4041 (20.989, Hex-Hederagenin)
987.4865 (10.766, GlcGlcGlc)
515.3385 (23.54, Zhanic Acid Aglycone?)
485.3254 (22.89, New Aglycone)
897.4828 (18.159, 3-Ara-Glc-Ara-Hederagenin)
617.4049 (22.001, Hex-SoyE?)
749.4512 (17.25, Pent-Hex-SoyE)
853.4593 (18.357, ?)
1027.5154 (16.62, Saponin V-?)
1129.5471 (15.56, Leaf)
1057.5605 (23.05, Unknown)
515.3362 (25.46, Zhanic Acid Aglycone?)
1113.5524 (18.45, Unknown)
1073.5574 (17.339, GlcA-?)

Putative Identification based on m/z and/or fragmentation Mohamed Bedair

1073.5175 (14.1311, 3-Glc-28-Ara-Rha-Xyl Medicagenic Acid)
269.0452 (10.4704, Apigenin)
957.506 (14.031, Hex-Hex-Hex-Hederagenin)
939.4945 (19.2447, Dehydrosoyasaponin)
1119.5579 (10.137, 3-Hex-Hex-Hex-28-Hex-Echinocystic acid)
175.039 (6.2289, 4-Methyl Umbelliferone)
973.4991 (11.7007, Hex-Hex-Hex-Bayogenin)
957.5018 (12.4346, Hex-Hex-Rha-Bayogenin)
973.5004 (13.5744, Hex-Hex-Hex-Bayogenin)
825.4275 (13.4499, GlcA-Glc-Bayogenin)
939.4936 (13.5333, Dehydrosoyasaponin)
957.5032 (11.5434, Hex-Hex-Rha-Bayogenin)
973.4997 (10.2012, Hex-Hex-Hex-Bayogenin)
433.1121 (3.824, Naringenin chalchone 4-O-glucoside)
987.4791 (13.5266, GlcA-Glc-Glc--Bayogenin)
255.0649 (11.9383, Isoliquiritigenin)
925.511 (14.827, Rha-hex-hex-Soyasapogenol E fragment of 1087)
941.51 (13.2652, Rha-Hex-Hex-Hederagenin fragment of 1103-162)
285.076 (9.5444, 4',5-Dihydroxy-7-methoxyflavonone or 7,2'-Dihydroxy-4'-methoxyisoflavanone)
633.3975 (14.2205, Hex-Herdeagenin)
253.0501 (6.376, Daidzein)
1103.564 (14.0897, Hex-Rha-Hex-Hex-Hederagenin)
793.433 (11.442, Fragment hexA-hex-soyasapogenol E)
1884.0123 (17.6683, [2M-1] of m/z 941.509 3-Rha-Gal-GlcA-SoyB)
455.3538 (28.766, Soyasapogenol E)
811.4449 (12.8105, Glu Glu Bayogenin Fragment to 973-162)
1117.5419 (13.6447, Rha-Hex-Hex-Hex-Quillaic acid)
1117.5395 (11.0018, Rha-Hex-Hex-Hex-Gypsogenic acid)
973.4977 (14.5488, Hex-Hex-Hex-Bayogenin)
1073.5138 (14.7251, 3-Glc-28-Ara-Rha-Xyl Medicagenic Acid)
475.1244 (4.7971, Luteolin 7,3'-dimethyl ether 5-glucoside or any isomer C23H24O11)
607.1312 (6.3976, Kaempferol 3-rhamnoside-7-galacturonide or any isomer C27H28O16)
617.4051 (23.6318, Hex-Soyasapogenol E)
647.3793 (16.8011, Hex-Quillaic acid)
647.3766 (14.5771, Hex-Gypsogenic acid)
1189.5656 (13.7904, Pen-Pen-Pen-Rha-GlcA-Echinocystic acid)
471.3486 (26.543, Aglycone triterpene C30H48O4)
811.4447 (11.4075, Fragment hex-hex-bayogenin)
485.3257 (22.7271, Gypsogenic acid)

649.3932 (16.6029, Hex-Bayogenin)
1027.5117 (15.5922, GlcA-Rha-Glc-malonyl-Soyasapogenol B)
471.3457 (27.0986, Aglycone triterpene C30H48O4)
471.3468 (25.8634, Echinocystic acid)
485.327 (26.3936, Quillaic acid)
471.3485 (23.0424, Hederagenin)

Putative Identification based on m/z and/or fragmentation Dongsik Yang

277.2173 (28.606, Linolenic acid)
279.2367 (30.6349, Linoleic acid)
255.232 (31.8631, Palmitic acid)
1235.536 (11.3357, 3-GlcA-28-Ara-Rha-Xyl Medicagenic Acid)
421.2078 (9.1802, Epicatechin Pentose)
283.2629 (33.1727, Stearic acid)
1367.6123 (13.6662, 3Glc-Glc-28-Ara-Rha-Xyl-Api-Med)
607.1294 (5.1396, Kaempferol Glucuronide Rhamnose)
461.1079 (7.5346, Leteolin 3'-methyl ether 7-glucoside)
227.2007 (28.7351, Myristic acid)
431.0992 (6.9725, Genistein 7-O-b-D-glucoside)
253.2166 (29.2798, Palmitoleic acid)
607.1297 (5.4053, Kaempferol Glucuronide Rhamnose)
1103.5603 (11.2146, Hex-Rha-Hex-Hex-Hederagenin)
241.2167 (30.5534, Pentadecanoic acid)
447.2224 (10.648, Linalool glucoside Pentose)
283.0218 (5.876, Lucernol)
141.0163 (23.9785, 18-Hydroxy-9-octadecenoic acid)
245.0429 (9.8794, Isopimpinellin)
461.0723 (5.3592, Kaempferol Glucuronide)
447.0931 (6.4753, Kaempferol-3-O-glucoside)
255.027 (8.8021, Purpurin)
477.1047 (6.1909, Kaempferol Hexose)
283.0603 (7.3836, Biochanin A)
241.2164 (27.0009, Pentadecanoic acid)
227.2009 (24.7706, Myristic acid)
447.0919 (6.7914, Kaempferol Hexose)
607.1086 (9.0041, Biochanin A b-D-diglucoside)
607.1077 (8.6031, Biochanin A b-D-diglucoside)
593.1288 (10.0838, Kaempferol Coumaroyl Hexose)
367.3582 (35.8388, Lignoceric acid)
593.1301 (9.6433, Genistein b-D-di-glucoside)
433.2587 (20.7929, Rasfonin)

339.0724 (9.5337, Cichoriin)
299.2586 (31.0314, Octadecene-1,9,18-triol ()
447.2738 (24.4337, Isolinaritriol triacetate)
321.2055 (21.0978, Rapanone)
321.2064 (20.6406, Rapanone)

Appendix RII – Primers

Primer Set 8(cyp88d3)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d3 (BAG68926)
Forward Primer Sequence: AAGGAAACCTTCTTCATCTCTTTCAA
Reverse Primer Sequence: AGGACATTGCAATCAATTCGTTAG

Primer Set 4 (cyp88d2)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d2 (BAG68925)
Forward Primer Sequence: ACGGCGACCAGATGAGAAATA
Reverse Primer Sequence: CAATTTCCACTACCTCCTGGTGAT

Primer Set 3 (cyp88d1)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d1 (BAG68925)
Forward Primer Sequence: TGATATGGCGTATTGTTTCATCAA
Reverse Primer Sequence: GCCAAGGAAGAGCAAGAAGGA

Primer Set 23 (GT2-1 R)

Used for qRT-PCR transcript expression level analysis
Gene Target (Seki, 2008): triterpene udp-glucosyl transferase ugt73k1
Forward Primer Sequence: ACGAAATGAGCAGCCATGTG
Reverse Primer Sequence: TTTCGCTGCTTCCGATAACC

Primer Set 22 (GT2-2 R)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): triterpene udp-glucosyl transferase ugt71g1 (AAW56091)
Forward Primer Sequence: TAGTCCACTCTCAGTCCCAAACC
Reverse Primer Sequence: ATGCAGAACAACAGCTTAATGCTT

Primer Set 19 (CAS-1)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cycloartenol synthase (CAA75588)
Forward Primer Sequence: GGATTCGGGCTAAATGAAGTTTG
Reverse Primer Sequence: GATAGCGCGTTGGGTTGAAG

Primer Set 18 (BAS1-2)

Used for qRT-PCR transcript expression level analysis

Gene Target (NCBI Protein accession ID): beta-aymrin synthase (CAD23247)
Forward Primer Sequence: CCAAGGGAGGCATGAAAAATAG
Reverse Primer Sequence: GCAAACCAGTGATGGCCATT

Primer Set 16 (SE2-2)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): squalene monooxygenase 2 (CAD23248)
Forward Primer Sequence: CCCAAGTGTATGAGCCAAAGC
Reverse Primer Sequence: CGGTGATGCTGATGTTATCATTG

Primer Set 14 (SE1-2)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): squalene monooxygenase 1 (CAD23249)
Forward Primer Sequence: AAAGGAAATTGTAGAGTGCAGCAA
Reverse Primer Sequence: CGGTTTCGGGTGGATCAC

Primer Set 43 (qRT_72A68_1429)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp72a68 (ABC59077. 1)
Forward Primer Sequence: GTTTGGAGCGGGTCCTAGAAT
Reverse Primer Sequence: TCTTTGCTTCCAACAGGGAAA

Primer Set 44 (qRT_72A68_1069)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp72a68 (ABC59077. 1)
Forward Primer Sequence: TTGGACGATGGTGTTGTTGAG
Reverse Primer Sequence: TCTAATACTTCCTTCCTTGCACGTT

Primer Set 33 (cyp716a12_3)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp716a12 (ABC59076. 1)
Forward Primer Sequence: ATGGAAGCTTTATTGGAGTGCAA
Reverse Primer Sequence: TCTCTGGCATGGGAAAACATT

Primer Set 34 (cyp716a12_4)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp716a12 (ABC59076. 1)
Forward Primer Sequence: CGGCGAGTTACCTCACATTTATG
Reverse Primer Sequence: GCTGGTTTCGATTTTGCAATTT

Primer Set 37 (cyp72a67_3)
Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp72a67 (ABC59075. 1)
Forward Primer Sequence: ACCAGCATTTGGTGTTACTCGAT
Reverse Primer Sequence: CACTCCAGCAGGTACTTCCATGT

Primer Set 38 (cyp72a67_4)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp72a67 (ABC59075. 1)
Forward Primer Sequence: CACTTTCTCTTTCCCTTTCTGTTTCT
Reverse Primer Sequence: ACCTTTTTACTGGTGTTTTGGAATCT

Primer Set 41 (cyp83g1_3)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp83g1 (ABC59084. 1)
Forward Primer Sequence: TCAGCAAAAATGGCCAAAGAA
Reverse Primer Sequence: CGCGGGTCGGTTACAGAAT

Primer Set 45 (qRT_72A68_1019)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp72a68 (ABC59077. 1)
Forward Primer Sequence: TGCAGGTTATTCCATGTTGCA
Reverse Primer Sequence: TCAACAACACCATCGTCCAAA

Appendix RIII – Primers

Cloning Primers

cyp72a67(NCBI ID: DQ335780. 1)

Used for cloning of gene for yeast expression and genomic DNA locus cloning
Forward Primer Sequence (cyp72a67with kozak and BamH1):
TCCGGATCCGTTATGGAAGCATCATTGGCCATATATTA
Reverse Primer Sequence (cyp72a67with EcoR1 site):
AGGGAATTCTTATGCTTTCACTTTGCGTAGAA

cyp72a68 (NCBI ID: DQ335782. 1)

Used for cloning of gene for yeast expression and genomic DNA locus cloning
Forward Primer Sequence (cyp72a68 with kozak and BamH1):
TCCGGATCCGTTATGGAATTATCTTGGGAAAC
Reverse Primer Sequence (cyp72a68 with EcoR1 site):
AGGGAATTCTTATGTTTTGATTTTGCGTAGAA

cyp83g1 (NCBI ID: DQ335789. 1)

Used for cloning of gene for yeast expression
Forward Primer Sequence (cyp83g1 with kozak and BamH1):
TCCGGATCCGTTATGAACAAAAACATGTCACCCCTTA
Reverse Primer Sequence (cyp83g1 with EcoR1 site):
AGGGAATTCTCACACGCATTCAATTCGCTTCTT

cyp716a12 (NCBI ID: DQ335781. 1)

Used for cloning of gene for yeast expression
Forward Primer Sequence (cyp716a12with kozak and BamH1):
TCCGGATCCGTTATGGAGCCTAATTTCTATCT

Reverse Primer Sequence (cyp716a12with EcoR1 site):
AGGGAATTCTTAAGCTTTGTGTGGATAAAG

cyp88d3 (NCBI ID: AB433176. 1)

Used for cloning of gene for yeast expression
Forward Primer Sequence (cyp716a12with kozak and BamH1):
TCCGGATCCGTTATGGAAATGCAGTGGGTTTA
Reverse Primer Sequence (cyp716a12with EcoR1 site):
AGGGAATTCTTAAGCACGTGAGACTTTAATAACC

Reverse Screen

cyp72a67-f

Forward Primer Sequence:CTTAGCAGAGATCAACCAACTAG

cyp72a68-f

Forward Primer Sequence:GCACGAGGAAAACATTTCACAC

cyp83g1-f

Forward Primer Sequence:CTTAGCAGAGATCAACCAACTAG

cyp83g3-f

Forward Primer Sequence:GAAATGCAGTGGGTTTACATTTG

cyp716a12-f

Forward Primer Sequence:ATGGAGCCTAATTTCTATCT

Tnt1-Fw

Forward Primer Sequence:ACAGTGCTACCTCCTCTGGATG

Tnt1-Re

Reverse Primer Sequence: CAGTGAACGAGCAGAACCTGTG

Appendix RIII – Cloned Sequence

>JHS CYP72A67 YV A10 , 1512 bases

ATGGAAGCATCATTGGCCATATATTATGGCATAATTCTCATCACTGTAAC
ACTTGGTTTAGTATACACATGGAGAGTACTGAATTGGATTTGGTTGAAGC
CAAAGAGGCTAGAGAAGCTCTTACGAGAACAAGGATGTAATGGAAATTCT
TATAGACTTGTGCTTGGGGACTTGAAGGATTCATATAAGATGGGAAAGAA
AGCCAAATCCAAACCCATGGAACTGTCGGATGATATAATCCCTCGTGTCA
TTCCCTACATTCAACAACTTGTTCAAATTTACGGGAAGAATCCTTTCATT
TGGTCTGGAACAACACCAAGGCTGATTCTCACAGAACCAGAGCTAATAAA
AGATGTCTTAAACAGAACTTCTGAATTACAAAAGCCAAAATATGAGATTT
TCAAATTTCTATTTAGTGGTCTTATAATTCACGAGGGAGAAAAGTGGAGA
AAGCATAGAAGGTTAATGAACGCTGCTTTCCAGTTAGAAAAATTGAAGAT
CATGGCACCAAGTTTCCTCACAAGTTGCATTGATATGATTAGCAAATGGG
AGTCAACGTTGTCATCAGATGGATCAGGTGAAATAGACATATGGCCTTCC
CTACAGAATTTGACAAGTGATGTTATTTCTCGAAACGCATTTGGAAGTAG
TTACGAAGAAGGAAAAAGAATATTTGATCTTCAAAGAGAGCAAGGTGAAC

TTGTTATGAAAAATCTAGTGAAATCTTTAATCCCTTTATGGAGGTTTATA
CCTACAGCTACTCAAGGAGGATGCATGAAATTGAAAAAAGATATAGATTC
TTCTCTTAGATATATAATTAACAAAAGAGAGAAAGCAATGAAGGCAGGTG
AAGCAACTGAGAATGACTTGTTAGGTCTTCTTCTAGAGTCAAACCACCAA
GAAATTAGAGATCATGGAAACAACAAGAATATGGGAATGAGTCTTGAAGA
TGTAGTGGGGGAATGCAAGTTATTCTACTTGGCAGGGCAAGAATCTACTT
CAACTATGCTTGTTTGGACAATGATATTGTTGAGTAGGTACCCTGATTGG
CAAGAACGTGCTAGGGAGGAAGTATTACAAATATTTGGCAACAAAAAACC
AGACTATGAAGGACTAAATAAACTTAAGATTCTCCCTATGATTTTGTATG
AGGTTCTAAGGTTGTATCCACCAGCATTTGGTGTTACTCGATATGTTGGC
AAAGACATAAAGTTTGGAAACATGGAAGTACCTGCTGGAGTGGAAGTTTT
CTTACCAATTATTTTGCTTCAACATAACAATGAACTATGGGGTGATGATG
CAAAGATGTTCAATCCTGAGAGATTTGCTGAAGGAATTTCCAAAGCAACA
AATGGTAGATTTATATATTTTCCATTTGGAGGGGGTCCTAGAGTTTGCAT
GGGACAAAACTTTTCCCTATTGGAAGCAAAGATGGCAGTGTCAATGATTT
TACAAAATTTCTATTTTGAACTTTCTCCAACCTATGCTCATACTCCAAAT
TTAGTGATGACT

>jhsCYP72A68, 1563 bases

ATGGAATTATCTTGGGAAACAAAATCAGCCATAATTCTCATCACTGTGAC
ATTTGGTTTGGTATACGCATGGAGGGTATTGAATTGGATGTGGCTGAAGC
CAAAGAAGATAGAGAAGCTTTTAAGAGAACAAGGCCTTCAAGGGAACCCT
TATAGACTTTTGCTTGGAGATGCAAAGGATTATTTTGTGATGCAAAAGAA
AGTTCAATCCAAACCCATGAATCTATCTGATGATATTGCGCCACGTGTCG
CTCCTTACATTCATCATGCTGTTCAAACTCATGGGAAAAAGTCTTTTATT
TGGTTTGGAATGAAACCATGGGTGATTCTCAATGAACCTGAACAAATAAG
AGAAGTATTCAACAAGATGTCTGAGTTCCCAAAGGTTCAATATAAGTTTA
TGAAGTTAATAACTCGCGGTCTTGTTAAACTAGAAGGAGAAAAGTGGAGC
AAGCATAGAAGAATAATCAACCCTGCGTTTCACATGGAAAAATTGAAGAT
TATGACACCAACATTCTTGAAAAGCTGCAATGATTTGATTAGCAATTGGG
AAAAAATGTTGTCTTCAAATGGATCATGTGAAATGGACGTATGGCCTTCC
CTTCAGAGCTTGACAAGTGATGTTATCGCTCGTTCGTCATTTGGAAGTAG
TTATGAAGAAGGAAGAAAAGTATTTCAACTTCAAATAGAGCAAGGTGAAC
TTATAATGAAAAATCTAATGAAATCTTTAATCCCTTTATGGAGGTTTTTA
CCTACCGCTGATCATAGAAAGATAAATGAAAATGAAAAACAAATAGAAAC
TACTCTTAAGAATATAATTAACAAGAGGGAAAAAGCAATTAAGGCAGGTG
AAGCCACTGAGAATGACTTATTAGGTCTCCTCCTAGAGTCGAACCACAGA
GAAATTAAAGAACATGGAAACGTCAAGAATATGGGATTGAGTCTTGAAGA
AGTAGTCGGGGAATGCAGGTTATTCCATGTTGCAGGGCAAGAGACTACTT
CAGATTTGCTTGTTTGGACGATGGTGTTGTTGAGTAGGTACCCTGATTGG
CAAGAACGTGCAAGGAAGGAAGTATTAGAGATATTTGGCAATGAAAAACC
CGACTTTGATGGACTAAATAAACTTAAGATTATGGCCATGATTTTGTATG

AGGTTTTGAGGTTGTACCCTCCTGTAACCGGCGTTGCTCGAAAAGTTGAG
AATGATATAAAACTTGGAGACTTGACATTATATGCTGGAATGGAGGTTTA
CATGCCAATTGTTTTGATTCACCATGATTGTGAACTATGGGGTGATGATG
CTAAGATTTTCAATCCTGAGAGATTTTCTGGTGGAATTTCCAAAGCAACA
AACGGTAGATTTTCATATTTTCCGTTTGGAGCGGGTCCTAGAATCTGCAT
TGGACAAAACTTTTCCCTGTTGGAAGCAAAGATGGCAATGGCATTGATTT
TAAAGAATTTTTCATTTGAACTTTCTCAAACATATGCTCATGCTCCATCT
GTGGTGCTTTCTGTTCAGCCACAACATGGTGCTCATGTTATTCTACGCAA
AATCAAAACATAA

>JHS_CYP83G1, 1521 bases

ATGAACAAAAACATGTCACCCCTTATTCTTTTACCCTTTGCTCTCTTGCT
ATTCTTCTTGTTCAAAAAACACAAAACATCTAAGAAATCAACAACTCTTC
CACCAGGTCCTAAAGGCCTTCCTTTCATTGGAAACTTACACCAACTTGAT
AGTTCAGTTCTTGGTTTAAATTTCTATGAACTCTCTAAGAAATATGGCCC
TATAATCTCCCTTAAACTTGGTTCAAAGCAAACAGTCGTTGTTTCATCAG
CAAAAATGGCCAAAGAAGTAATGAAAACACATGATATCGAATTCTGTAAC
CGACCCGCGTTAATCAGCCATATGAAAATATCATATAATGGATTAGATCA
AATATTTGCACCATATAGAGAATATTGGAGACACACAAAAAAACTTTCCT
TTATTCATTTTCTTAGTGTCAAAAGAGTCTCAATGTTTTACTCAGTTAGA
AAAGATGAGGTGACACGAATGATCAAGAAGATATCAGAAAATGCTTCTTC
CAACAAAGTTATGAACATGCAGGATCTTCTTACTTGTCTTACAAGTACTT
TAGTTTGTAAGACCGCCTTCGGCAGAAGGTATGAAGGGGAAGGAATTGAG
CGTAGCATGTTTCAAGGTCTGCATAAAGAAGTTCAGGATTTGCTAATTTC
GTTCTTTTACGCGGATTATTTGCCCTTTGTTGGAGGGATTGTTGATAAGC
TCACCGGAAAGACGAGTCGCCTTGAGAAAACGTTCAAGGTTTCAGATGAA
CTTTATCAAAGTATTGTTGATGAACATCTTGATCCAGAAAGGAAGAAGTT
GCCTCCACATGAGGATGATGTTATTGATGCCTTGATTGAACTGAAGAATG
ATCCTTACTGCTCAATGGATCTCACTGCAGAACACATCAAGCCCTTGATC
ATGAATATGTCGTTTGCTGTGACAGAAACAATTGCGGCAGCGGTAGTCTG
GGCTATGACTGCGCTAATGAAGAATCCAAGAGCGATGCAGAAAGTACAAG
AAGAGATTCGGAAAGTGTGTGCAGGGAAAGGTTTTATAGAGGAAGAAGAT
GTCGAAAAGCTTCCATATTTCAAGGCCGTTATAAAAGAATCGATGAGATT
GTACCCAATTTTGCCTATACTTTTACCAAGAGAAACAATGACAAATTGCA
ACATTGCAGGGTACGACATTCCAGACAAGACATTGGTGTACGTGAATGCA
TTGGCGATCCATAGAGACCCAGAAGTATGGAAGGATCCAGAAGAGTTTTA
TCCAGAGAGATTCATAGGAAGTGATATAGATTTAAAAGGACAAGATTTTG
AGCTGATTCCGTTTGGTTCTGGGCGAAGAATTTGCCCCGGCTTAAACATG
GCTATTGCTACCATCGACCTTGTACTTTCTAATCTTCTCTATTCATTTGA
CTGGGAAATGCCTGAAGGAGCTAAGAGGGAAGACATTGACACTCATGGTC
AAGCCGGACTTATTCAACACAAGAAAAATCCTCTCTGTCTTGTTGCTAAG
AAGCGAATTGAATGCGTGTGA

>JHS_CYP88D3_d3-3a, 1479 bases
ATGGAAATGCAGTGGGTTTACATTTGTACTGCTGCTTTGTTTGCATGCTA
TGTTTTTGTAAACAAATTTTTGAGGAGGTTTAATGGTTGGTACTATCATC
TCAAATTAAGAAACAAAGAGTACCCTTTGCCTCCAGGTGATATGGGATGG
CCACTTATTGGCAACCTATTATCGTTTAACAAAAACTTCTCATCTGGCCA
ACCTGATTCATTCACCACCAACCTTATTCTCAAATATGGGAGAGATGGTA
TCTACAAAACTCACGTGTGTGGAAATCCAAGTATCATAATTTGTGATCCT
GAGATGTGTAAGCGAGTGCTCTTAGATGATGTAAACTTTAAAATTGGTTA
TCCAAAATCCATCCAAGAATTGACAAAATGTAGACCCATGATTGATGTCT
CGAACGCAAATCACAAGCATTTTCGACGCCTAATCACTGCTCCCATGGTT
GGTCACAAGGTGTTAGACATGTACCTAGAACGTCTCGAGGACATTGCAAT
CAATTCGTTAGAAGAATTGTCTAGCATGAAGCACCCCATCGAGCTCTTGA
AAGAGATGAAGAAGGTTTCCTTTAAATCCATTATCCATGTTTTCATGGGA
ACTTCTAATCAGAACATTGTTAAAAACATTGGAAGTTCATTTACTGATTT
GTCTAAAGGCATGTACTCTATCCCCATCAATGCACCTGGTTTTACTCTCC
ACAAAGCACTCAAGGCACGGAAGAAGATAGCTAAATTATTGCAACCTGTT
GTGGATGAAAGGAGGTTGATGATAAAAAATGGACAACATGTGGGAGAGAA
GAAAGATCTTATGGATATTCTATTGGAAATCAAAGATGAGAATGGTAGAA
AATTGGAGGATCAGGATATCAGTGACCTGTTGATAGGACTTTTATTTGCC
GGACATGAAAGTACAGCAACTGGGATAATGTGGTCAGTTGCACATCTTAC
ACAACATCCACATATCCTACAAAAAGCCAAGGAAGAGCAGGAAGAAATCT
TGAAGATAAGACCAGCCTCCCAAAAACGATTGAGTCTTAATGAAGTCAAG
CAAATGATTTATCTTTCATATGAAATCGATGAAATGTTGCGATTTGCCAA
CATTGCCTTTTCAATTTTTCGAGAGGCTACATCTGATGTTAACATCAACG
GTTATCTCATACCAAAAGGATGGAGAGTGCTAATATGGGCGAGAGCCATT
CATATGGATTCTGAATATTATCCAAATCCTAAAGAATTTAATCCTTCTAG
ATGGAAAGATTATAATGCCAAGGCAGGAACCTTTCTTCCTTTTGGAGCAG
GAAGTAGGCTCTGTCCTGGAGCCGACTTAGCAAAACTTGAAATTTCTATA
TTTCTTCATTATTTCCTCCTTAACTACAGGTTGGAGCGAATAAACCCAGA
TTGCCCTGTTACTACCTTGCCACAATGTAAGCCCACTGATAACTGCCTCG
CTAAGGTTATTAAAGTCTCACGTGCTTAA
>JHS_CYP716A12, 1440 bases
ATGGAGCCTAATTTCTATCTCTCCCTTCTCCTTCTCTTTGTCACTTTCAT
ATCTCTCTCTCTTTTTTTTCATATTCTACAAACAGAAATCTCCATTAAATT
TGCCACCTGGTAAAATGGGTTACCCAATCATAGGTGAAAGCCTTGAGTTC
TTATCAACAGGATGGAAAGGACATCCTGAAAAATTCATTTTCGACCGTAT
GCGTAAATATTCCTCAGAACTCTTTAAAACATCAATCGTAGGAGAATCTA
CGGTGGTTTGTTGCGGAGCAGCAAGTAACAAGTTTTTGTTTTCAAACGAG
AATAAACTTGTGACTGCATGGTGGCCAGATAGTGTAAACAAAATCTTCCC
TACTACTTCTCTTGACTCTAACTTGAAGGAAGAATCCATCAAGATGAGAA

AATTGCTTCCACAATTCTTTAAACCCGAAGCTCTACAACGTTATGTTGGT
GTCATGGATGTTATTGCTCAAAGACATTTTGTTACTCATTGGGATAATAA
AAATGGAATCACCGTCTACCCCTTGGCCAAGAGGTACACCTTTTTGTTAG
CTTGTCGGTTGTTCATGAGCGTTGAAGACGAGAATCATGTAGCAAAATTT
AGTGATCCATTTCAGTTAATTGCGGCCGGAATCATATCTCTACCAATTGA
TTTGCCAGGAACACCATTCAACAAAGCTATAAAGGCCTCAAACTTTATAA
GAAAGGAGTTGATTAAGATCATAAAGCAAAGGAGGGTAGATTTGGCAGAA
GGGACAGCATCACCAACACAAGATATATTGTCTCACATGTTGTTGACAAG
TGATGAAGATGGAAAGAGTATGAATGAACTTAATATTGCTGATAAGATTC
TTGGCCTTTTGATCGGAGGACATGACACTGCTAGCGTCGCATGCACTTTC
CTTGTCAAATATCTCGGCGAGTTACCTCACATTTATGATAAAGTCTATCA
AGAGCAAATGGAAATTGCAAAATCGAAACCAGCAGGAGAATTGTTGAATT
GGGATGACCTGAAGAAAATGAAATACTCTTGGAACGTAGCTTGTGAAGTA
ATGAGACTTTCCCCTCCACTCCAAGGAGGTTTCAGGGAAGCCATCACCGA
CTTTATGTTCAATGGATTCTCAATTCCTAAGGGATGGAAGCTTTATTGGA
GTGCAAATTCAACACATAAGAACGCAGAATGTTTTCCCATGCCAGAGAAA
TTTGACCCAACAAGATTTGAAGGAAATGGACCAGCTCCTTATACTTTTGT
TCCCTTTGGTGGAGGACCAAGGATGTGTCCTGGAAAAGAGTATGCAAGAT
TAGAAATACTTGTTTTCATGCACAATTTGGTGAAAAGGTTTAAGTGGGAA
AAGGTGATTCCAGATGAGAAGATTATTGTTGATCCATTCCCCATCCCTGC
AAAGGATCTTCCAATTCGCCTTTATCCACACAAAGCTTAA

Appendix RIV – Primers

Cloning Primers

cyp88d3 (NCBI ID: BAG68926)

Used for cloning of gene for yeast expression

Forward Primer Sequence (cyp88d3 with kozak and BamH1):

TCCGGATCCGTTATGGAAATGCAGTGGGTTTA

Reverse Primer Sequence (cyp88d3 with EcoR1 site):

AGGGAATTCTTAAGCACGTGAGACTTTAATAACC

cyp88d1 (Removed from NCBI (Seki, 2008))

Used for cloning of gene for yeast expression

Forward Primer Sequence (cyp88d1 with kozak and BamH1):

TCCGGATCCGTTATGGAACTTCAATGGTTTTGGATGTTTGCTGCCACTT

Reverse Primer Sequence (cyp88d3 with EcoR1 site):

AGGGAATTCTTAAGAATCAGAGATCTTTATGACCTTAGCAAGACAA

cyp88d2 (NCBI ID: BAG68925)

Used for cloning of gene for yeast expression

Forward Primer Sequence (cyp88d1 with kozak and BamH1):

TCCGGATCCATGGAATTTCAATGGTTTTGG

Reverse Primer Sequence (cyp88d3 with EcoR1 site):
AGGGAATTCTTAAGCATCAGAGAGCTTTG

Sequencing of Plasmid

gal10 promoter

Used for confirmation of sequence from yeast transformation
Forward Primer Sequence:TCATATGGCATGCATGTGCTCTG

qRT-PCR Primers

Primer Set 8(cyp88d3)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d3 (BAG68926)
Forward Primer Sequence: AAGGAAACCTTCTTCATCTCTTTCAA
Reverse Primer Sequence: AGGACATTGCAATCAATTCGTTAG

Primer Set 4 (cyp88d2)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d2 (BAG68925)
Forward Primer Sequence: ACGGCGACCAGATGAGAAATA
Reverse Primer Sequence: CAATTTCCACTACCTCCTGGTGAT

Primer Set 3 (cyp88d1)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cyp88d1 (BAG68925)
Forward Primer Sequence: TGATATGGCGTATTGTTTCATCAA
Reverse Primer Sequence: GCCAAGGAAGAGCAAGAAGGA

Primer Set 23 (GT2-1 R)

Used for qRT-PCR transcript expression level analysis
Gene Target (Seki, 2008): triterpene udp-glucosyl transferase ugt73k1
Forward Primer Sequence: ACGAAATGAGCAGCCATGTG
Reverse Primer Sequence: TTTCGCTGCTTCCGATAACC

Primer Set 22 (GT2-2 R)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): triterpene udp-glucosyl transferase ugt71g1 (AAW56091)
Forward Primer Sequence: TAGTCCACTCTCAGTCCCAAACC
Reverse Primer Sequence: ATGCAGAACAACAGCTTAATGCTT

Primer Set 19 (CAS-1)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): cycloartenol synthase (CAA75588)
Forward Primer Sequence: GGATTCGGGCTAAATGAAGTTTG
Reverse Primer Sequence: GATAGCGCGTTGGGTTGAAG

Primer Set 18 (BAS1-2)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): beta-aymrin synthase (CAD23247)
Forward Primer Sequence: CCAAGGGAGGCATGAAAAATAG
Reverse Primer Sequence: GCAAACCAGTGATGGCCATT

Primer Set 16 (SE2-2)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): squalene monooxygenase 2 (CAD23248)
Forward Primer Sequence: CCCAAGTGTATGAGCCAAAGC
Reverse Primer Sequence: CGGTGATGCTGATGTTATCATTG

Primer Set 14 (SE1-2)

Used for qRT-PCR transcript expression level analysis
Gene Target (NCBI Protein accession ID): squalene monooxygenase 1 (CAD23249)
Forward Primer Sequence: AAAGGAAATTGTAGAGTGCAGCAA
Reverse Primer Sequence: CGGTTTCGGGTGGATCAC

Primers Used in Reverse Screen

cyp88d2 F

Forward Primer Sequence: ATGGAATTTCAATGGTTTTGGATG

cyp88d2 F2

Forward Primer Sequence: ACGGGCTCTTGCGATCAAGGTAC

Tnt1-Re

Reverse Primer Sequence: CAGTGAACGAGCAGAACCTGTG

Appendix RIV – Targeted Ions

Targeted Analysis
Mass/Charge(Retention Time,Ion ID)
161.023(4.36, IS)
1119.5729(7.845, GlcRha?)
989.4892(8.257, ?)
989.5104(8.928, Glc)
957.507(9.305, ?)
989.4884(9.54, Glc)
1103.517(10.07, Leaf)
1119.564(10.18, Leaf)
1089.5494(10.25, Leaf)
1105.5775(10.337, Glc)
1117.5343(10.661, ?)
987.4865(10.766, GlcGlcGlc)
1383.6057(11.26, O Zhan Stand)
1397.5889(11.29, Leaf)
1397.5887(11.39, Leaf)

1545.6609(11.39, O Zhan Stand)
1677.7001(11.45, O Zhan Stand)
1251.5675(11.5, Leaf)
1265.5474(11.51, Leaf)
1545.6595(11.56, O Zhan Stand)
1119.5602(11.57, ?)
1413.6189(11.609, Leaf)
1235.5402(11.61, Leaf)
1413.6149(11.666, O Zhan Stand)
1515.6458(11.666, O Zhan Stand)
1367.5752(11.666, Leaf)
1103.4962(11.77, Leaf)
1383.6086(11.84, Leaf)
1383.608(11.97, Leaf)
1221.554(12.059, Leaf)
1251.5621(12.138, Leaf)
987.4799(12.206, GlcA-Glc-Glc--Bayogenin)
1089.5117(12.225, Leaf)
1119.567(12.407, GlcRhaGlc?)
811.4469(12.441, GlcGlcBayogenin)
987.4858(12.466, Possible Hex-Hex-Hex-Med Scotts work)
811.4475(12.47, Hex-Hex-Bayogenin ? Higher Mass)
987.4818(12.52, ?)
987.4836(12.536, ?)
1119.5621(12.64, Rha-Hex-Hex-Hex-Bayogenin)
1117.5499(12.83, Rha-Hex-Hex-New Aglycone)
973.5025(12.84, Hex-Hex-Hex-Bayogenin)
985.4704(12.928, ?)
971.4853(12.976, Possibly bayogenin)
987.482(13.13, Possible Hex-Hex-Hex-Med Scotts work)
811.4505(13.2, Hex-Hex-Bayogenin ? Higher Mass)
1103.5669(13.29, Hex-Rha-Hex-Hex-Hed)
971.4885(13.304, Hex-Hex-HexA-Hederagenin)
825.4297(13.485, 3-Glc-28-Glc-Medicagenic Acid (Standard))
1101.5518(13.55, Glc?)
1307.5969(13.562, Leaf)
1367.6135(13.63, 3Glc-Glc-28-Ara-Rha-Xyl-Api-Med)
1057.4866(13.65, ?)
957.507(13.65, Hex-Hex-Hex-Hederagenin)
1087.4955(13.728, 3-Glc-Glc-28-Ara-Rha-Xyl Medicagenic Acid (Stand))
1087.4974(13.81, Leaf)

1145.5433(13.81, Leaf)
1235.5725(13.8147, ?)
809.4337(13.816, Hex-HexA-Hed)
1235.5703(13.827, 3-GlcA-28-Ara-Rha-Xyl Medicagenic Acid)
1235.5771(13.84, Leaf)
957.5068(13.85, Hex-Hex-Hex-Hederagenin)
1161.5363(13.9, Leaf)
1205.5613(13.93, Leaf)
1205.5653(13.94, ?)
957.4758(13.979, Glc-Gal-GlcA-SoyB)
823.4152(14.05, Hex-HexA-New Aglycone)
823.4152(14.06, Hex-HexA-New Aglycone)
911.4344(14.06, 3-Glc-28-Glc-Malonyl-Med)
823.4148(14.09, GlcA-Glc-NewAglycone)
647.3831(14.14, Hex-New Aglycone)
1159.4995(14.143, ?)
1073.5208(14.15, 3-Glc-28-Ara-Rha-Xyl Medicagenic Acid (Stand))
1073.5253(14.152, Leaf)
971.4856(14.16, Not Sure)
955.4971(14.376, ?)
925.5173(14.86, Rha-Hex-Hex-SoyE may be related to 1087)
1087.5739(14.86, Hex-Rha-Hex-Hex-SoyE)
955.495(14.93, Rha-Hex-?)
795.4525(14.99, Hex-Hex-Hederagenin)
927.497(15.1, Hex-Hex-Pent-Hederagenin)
955.4933(15.26, 0)
809.4335(15.29, Hex-HexA-Hederagenin)
955.4926(15.356, GlcA-?)
663.3756(15.49, Hex-Medicagenic Acid)
1129.5471(15.56, Leaf)
749.4464(15.57, Hex-Pent-Soyasapogenol E)
809.4333(15.77, HexA-Hex-Hederagenin)
825.4666(15.77, GlcGlcMed?)
855.4741(15.86, GlcAGlcHed?)
825.4643(15.95, GlcGlc?)
957.5092(16.05, Rha-Hex-Hex-Bayogenin)
971.4877(16.129, HexHex)
1085.5544(16.206, Leaf)
793.4389(16.23, HexA-Hex-Soy E)
1085.5581(16.231, ?)
1027.5154(16.62, ?)

811.4481(16.92, Hex-Hex-Bayogenin)
941.514(16.97, Rha-Hex-Hex-Hederagenin)
825.4304(16.998, 3-Glc-Glc-MedicagenicAcid?)
1073.5581(17.02, ?)
749.4512(17.25, Pent-Hex-SoyE)
1073.5574(17.339, GlcA-?)
957.4825(17.4, Glc-Gal-GlcA-SoyB (Stand))
957.5084(17.414, Soy Mix Stand)
649.394(17.44, HexA-Bayogenin)
1043.5476(17.45, ?)
649.3969(17.47, Hex-Bayogenin)
663.3762(17.48, Hex-Medicagenic Acid)
663.3777(17.53, 3-Glc-MedicagenicAcid (Stand))
941.5107(17.877, Ara-Rha-GlcA-Bayogenin)
809.4349(17.945, Hex-HexA-Hederagenin)
795.4543(17.945, Gal-GlcA-SoyB)
941.5112(17.955, Rha-Gal-GlcA-SoyB (Soy1))
941.5093(17.96, Leaf)
897.4828(18.159, 3-Ara-Glc-Ara-Hederagenin (standard))
705.3849(18.28, 3-Glc-Malonyl-MedicagenicAcid)
911.5005(18.303, Rha-Ara-GlcA-SoyB (Stand))
795.4542(18.357, Hex-Hex-Hederagenin?)
853.4593(18.357, ?)
1113.5524(18.45, Unknown)
765.4431(18.552, Ara-GlcA-SoyB (Stand))
809.4334(18.78, Glc-Glc-hed?)
647.3798(18.78, Hex-New Aglycone)
765.4431(18.994, GlcAHed)
925.5151(19.028, ?)
809.4341(19.05, Unknown - Hed)
809.4313(19.243, Unknown)
1113.5566(19.29, Unknown)
765.4424(19.41, Hex-Hex-Hederagenin)
939.498(19.535, 3-Rha-Xyl-GlcA)
925.4822(19.55, Hex-Hex-Rha-SoyE)
793.4408(19.62, Hex-HexA-455 ?)
795.4526(19.819, Hex-Hex-Hederagenin)
501.3228(19.881, Mediagenic Acid)
1067.5479(19.89, Leaf)
1083.5422(19.89, Leaf)
487.3421(20.637, Bayogenin)

633.4041(20.989, Hex-Hederagenin)
647.4343(20.997, GlcA-Hederagenin)
647.3817(21.57, GlcA-Hederagenin)
647.3814(21.59, GlcA-Hederagenin)
617.4049(22.001, Hex-SoyE?)
631.3854(22.029, ?)
485.3254(22.89, New Aglycone (Aglycone Mix))
1057.5605(23.05, Unknown)
471.3484(23.129, Hederagenin)
515.3385(23.54, Zhanic Acid Aglycone?)
515.3362(25.46, Zhanic Acid Aglycone?)
455.3528(28.886, SoyE)
793.5449(33.93, Rha?)

Bibliography

Chapter I References

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin, (2003) Dissecting plant secondary metabolism; constitutive chemical defences in cereals. *New Phytologist* 159, 101-108.

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

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S. and Sumner, L. W. (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9, 418-425.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X. and Sumner, L. W. (2006a) MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. *Anal. Chem.* 78, 4334-4341.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X. and Sumner, L. W. (2006b) MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. *Analytical Chemistry* 78, 4334-4341.

Bucciarelli, B., Hanan, J., Palmquist, D. and Vance, C. P. (2006) A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development
10. 1104/pp. 106. 082594. *Plant Physiol.* 142, 207-219.

Chan, E. K. F., Rowe, H. C., Hansen, B. G. and Kliebenstein, D. J. (2010) The Complex Genetic Architecture of the Metabolome. *PLoS Genet* 6, e1001198.

Confalonieri, M., Cammareri, M., Biazzi, E., Pecchia, P., Fevereiro, M. P. S., Balestrazzi, A., Tava, A. and Conicella, C. (2009) Enhanced triterpene saponin biosynthesis and root nodulation in transgenic barrel medic (*Medicago truncatula* Gaertn.) expressing a novel β-amyrin synthase (AsOXA1) gene. *Plant Biotechnology Journal* 7, 172-182.

Delis, C., Krokida, A., Georgiou, S., Peña-Rodríguez, L. M., Kavroulakis, N., Ioannou, E., Roussis, V., Osbourn, A. E. and Papadopoulou, K. K. Role of lupeol synthase in Lotus japonicus nodule formation. *New Phytologist* 189, 335-346.

Dixon, R. A. and Sumner, L. W. (2003) Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. *Plant Physiology* 131, 878-885.

Haridas, V., Higuchi, M., Jayatilake, G. S., Bailey, D., Mujoo, K., Blake, M. E., Arntzen, C. J. and Gutterman, J. U. (2001) Avicins: Triterpenoid saponins from Acacia victoriae (Bentham) induce apoptosis by mitochondrial perturbation

10. 1073/pnas. 101619098. *Proceedings of the National Academy of Sciences* 98, 5821-5826.

Harjes, C. E., Rocheford, T. R., Bai, L., Brutnell, T. P., Kandianis, C. B., Sowinski, S. G., Stapleton, A. E., Vallabhaneni, R., Williams, M., Wurtzel, E. T., Yan, J. and Buckler, E. S. (2008) Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification. *Science* 319, 330-333.

Huhman, D. V., Berhow, M. A. and Sumner, L. W. (2005) Quantification of Saponins in Aerial and Subterranean Tissues of *Medicago truncatula*. *Journal of Agricultural and Food Chemistry* 53, 1914-1920.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K. and Osbourn, A. E. (2003) Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. *Plant Molecular Biology* 51, 731-743.

Kirk, D. D., Rempel, R., Pinkhasov, J. and Walmsley, A. M. (2004) Application of Quillaja saponaria extracts as oral adjuvants for plant-made vaccines

doi:10. 1517/14712598. 4. 6. 947. *Expert Opinion on Biological Therapy* 4, 947-958.

Kuljanabhagavad, T., Thongphasuk, P., Chamulitrat, W. and Wink, M. (2008) Triterpene saponins from Chenopodium quinoa Willd. *Phytochemistry* 69, 1919-1926.

Lu, C. D. and Jorgensen, N. A. (1987) Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. *The Journal of Nutrition* 117, 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M. and Jorgensen, N. A. (1987) Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. *Journal of dairy science* 70, 799-805.

Meesapyodsuk, D., Balsevich, J., Reed, D. W. and Covello, P. S. (2007) Saponin Biosynthesis in Saponaria vaccaria. cDNAs Encoding beta-Amyrin Synthase and a Triterpene Carboxylic Acid Glucosyltransferase

10. 1104/pp. 106. 088484. *Plant Physiol.* 143, 959-969.

Nielsen, J., Nagao, T., Okabe, H. and Shinoda, T. (2010) Resistance in the Plant Barbarea vulgaris and Counter-Adaptations in Flea Beetles Mediated by Saponins. *Journal of Chemical Ecology* 36, 277-285.

P. Houghton, N. P., M. Jurzysta, Z. Biely, C. Cheung, (2006) Antidermatophyte activity of medicago extracts and contained saponins and their structure-activity relationships. *Phytotherapy Research* 20, 1061-1066.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. and Osbourn, A. E. (1999) Compromised disease resistance in saponin-deficient plants

Proceedings of the National Academy of Sciences of the United States of America 96, 12923-12928.

Pedersen, M. W., Barnes, D. K., Sorensen, E. L., Griffin, G. D., Nielson, M. W., Hill, R. R., Jr., Frosheiser, F. I., Sonoda, R. M., Hanson, C. H., Hunt, O. J., Peaden, R. N., Elgin, J. H., Jr., Devine, T. E., Anderson, M. J., Goplen, B. P., Elling, L. J. and Howarth, R. E. (1976) Effects of Low and High Saponin Selection in Alfalfa on Agronomic and Pest Resistance Traits and the Interrelationship of these Traits. *Crop Sci* 16, 193-199.

Ronfort, J., Bataillon, T., Santoni, S., Delalande, M., David, J. and 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, 28.

Schilmiller, A., Shi, F., Kim, J., Charbonneau, A. L., Holmes, D., Daniel Jones, A. and Last, R. L. (2010) Mass spectrometry screening reveals widespread diversity in trichome specialized metabolites of tomato chromosomal substitution lines. *The Plant Journal* 62, 391-403.

Sen, S., Makkar, H. P. S. and Becker, K. (1998) Alfalfa Saponins and Their Implication in Animal Nutrition. *Journal of Agricultural and Food Chemistry* 46, 131-140.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T. and Dixon, R. A. (2002) A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. *The Plant Journal* 32, 1033-1048.

Chapter II References

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin,. (2003). Dissecting plant secondary metabolism; constitutive chemical defences in cereals. New Phytologist **159,** 101-108.

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

Ballester, A. -R., Molthoff, J., de Vos, R., Hekkert, B. t. L., Orzaez, D., FernÁ¡ndez-Moreno, J. -P., Tripodi, P., Grandillo, S., Martin, C., Heldens, J., Ykema, M., Granell, A., and Bovy, A. (2011). Biochemical and Molecular Analysis of Pink Tomatoes: Deregulated Expression of the Gene Encoding Transcription Factor SlMYB12 Leads to Pink Tomato Fruit Color. Plant Physiology **152,** 71-84.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. Trends in Plant Science **9,** 418-425.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. Anal. Chem. **78,** 4334-4341.

Bucciarelli, B., Hanan, J., Palmquist, D., and Vance, C. P. (2006). A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development

10. 1104/pp. 106. 082594. Plant Physiol. **142,** 207-219.

Chen, J., Yu, J., Ge, L., Wang, H., Berbel, A., Liu, Y., Chen, Y., Li, G., Tadege, M., Wen, J., Cosson, V., Mysore, K. S., Ratet, P., MadueÃ±o, F., Bai, G., and Chen, R. Control of dissected leaf morphology by a Cys(2)His(2) zinc finger transcription factor in the model legume *Medicago truncatula*. Proceedings of the National Academy of Sciences **107,** 10754-10759.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W. -R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. Plant Physiol. **139,** 5-17.

Dixon, R. A., and Sumner, L. W. (2003). Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. Plant Physiology **131,** 878-885.

Dozmorov, I., and Centola, M. (2003). An associative analysis of gene expression array data. Bioinformatics **19,** 204-211.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol **48,** 155 - 171.

Goodacre, R., Roberts, L., Ellis, D., Thorogood, D., Reader, S., Ougham, H., and King, I. (2007). From phenotype to genotype: whole tissue profiling for plant breeding. Metabolomics **3,** 489-501.

Hannah, M. A., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A. R., and Willmitzer, L. (2010).combined Transcript and Metabolite Profiling of Arabidopsis Grown under Widely Variant Growth Conditions Facilitates the Identification of Novel Metabolite-Mediated Regulation of Gene Expression. Plant Physiology **152,** 2120-2129.

Helliwell, C. A., Chandler, P. M., Poole, A., Dennis, E. S., and Peacock, W. J. (2001). The CYP88A cytochrome P450, ent-kaurenoic acid oxidase, catalyzes three steps of the gibberellin biosynthesis pathway. Proceedings of the National Academy of Sciences **98,** 2065-2070.

Hirai, M., Yano, M., Goodenowe, D., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proc Natl Acad of Sci USA **101,** 10205 - 10210.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research **31,** e15.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. Plant Molecular Biology **51,** 731-743.

Kalo, P., Gleason, C., Edwards, A., Marsh, J., Mitra, R. M., Hirsch, S., Jakab, J. l., Sims, S., Long, S. R., Rogers, J., Kiss, G. r. B., Downie, J. A., and Oldroyd, G. E. D. (2005). Nodulation Signaling in Legumes Requires NSP2, a Member of the GRAS Family of Transcriptional Regulators. Science **308,** 1786-1789.

Krueger, S., Giavalisco, P., Krall, L., Steinhauser, M. -C., BÃ¼ssis, D., Usadel, B., FlÃ¼gge, U. -I., Fernie, A. R., Willmitzer, L., and Steinhauser, D. (2011). A Topological Map of the Compartmentalized <italic>Arabidopsis thaliana</italic> Leaf Metabolome. PLoS ONE **6,** e17806.

Lahoucine Achnine, D. V. H., Mohamed A. Farag, Lloyd W. Sumner, Jack W. Blount, Richard A. Dixon,. (2005). Genomics-based selection and functional

characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. The Plant Journal **41,** 875-887.

Lee, S. M. S., and Lai, P. Y. (2009). Double block bootstrap confidence intervals for dependent data. Biometrika **96,** 427-443.

Li, L., Cheng, H., Gai, J., and Yu, D. (2007). Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*. Planta **226,** 109-123.

Lu, C. D., and Jorgensen, N. A. (1987). Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. The Journal of Nutrition **117,** 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M., and Jorgensen, N. A. (1987). Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. Journal of dairy science **70,** 799-805.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., Sakurai, T., Shimada, Y., and Saito, K. (2010). AtMetExpress Development: A Phytochemical Atlas of Arabidopsis Development. Plant Physiology **152,** 566-578.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., and Dixon, R. A. (2010). Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*. Plant Cell **22,** 850-866.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., and Osbourn, A. E. (1999). Compromised Disease Resistance in Saponin-Deficient Plants. Proceedings of the National Academy of Sciences of the United States of America **96,** 12923-12928.

Parmigiani, G., Garrett, E., Irizarry, R., Zeger, S., Li, C., and Wong, W. (2003). DNA-Chip Analyzer (dChip). In The Analysis of Gene Expression Data, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, and W. Wong, eds (Springer London), pp. 120-141.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RTPCR. Nucleic Acids Research **29,** e45.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M. H., Karlen, Y., Bakker, O., van den Hoff, M. J. B., and Moorman, A. F. M. (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research **37,** e45.

Saito, K., Hirai, M. Y., and Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics - `majority report by precogs'. Trends in Plant Science **13,** 36-43.

Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E. -T., Tschoep, H., Steinhauser, M. C., Guenther, M., Hoehne, M., Rohwer, J. M., Altmann, T., Fernie, A. R., and Stitt, M. (in press). Network Analysis of Enzyme Activities and Metabolite Levels and Their Relationship to Biomass in a Large Panel of Arabidopsis Accessions. The Plant Cell Online.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., and Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. The Plant Journal **32,** 1033-1048.

Tohge, T., and Fernie, A. R. (2010).combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat. Protocols **5,** 1210-1227.

Tohge, T., Nishiyama, Y., Hirai, M., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D., and Kitayama, M. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J **42,** 218 - 235.

Udvardi, M., Kakar, K., Wandrey, M., Montanari, O., Murray, J., Andriankaja, A., Zhang, J., Benedito, V., Hofer, J., Chueng, F., and Town, C. (2007). Legume transcription factors: global regulators of plant development and response to the environment. Plant Physiol **144,** 538 - 549.

Chapter III References

Altschul, S., Gish, W., Miller, W., Meyers, E., and Lipman, D. (1990). Basic Local Alignment Search Tool. J Mol Biol **215,** 403 - 410.

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin,. (2003). Dissecting plant secondary metabolism; constitutive chemical defences in cereals. New Phytologist **159,** 101-108.

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

Benedito, V., Torres-Jerez, I., Murray, J., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P., Tang, Y., and Udvardi, M. (2008). Affymetrix GeneChip Medicago Genome Array

A gene expression atlas of the model legume *Medicago truncatula*. Plant J **55,** 504 - 513.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. Anal. Chem. **78,** 4334-4341.

Broeckling, C. D., Huhman, D. V., Farag, M. A., Smith, J. T., May, G. D., Mendes, P., Dixon, R. A., and Sumner, L. W. (2005). Metabolic profiling of. Journal of Experimental Botany **56,** 323-336.

Chang-Jun Liu, D. H., Lloyd W. Sumner, Richard A. Dixon,. (2003). Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from *Medicago truncatula*. The Plant Journal **36,** 471-484.

Dixon, R. A. (2001). Natural products and plant disease resistance. Nature **411,** 843-847.

Dixon, R. A., and Sumner, L. W. (2003). Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health. Plant Physiology **131,** 878-885.

Firn, R. D., and Jones, C. G. (2003). Natural products - a simple model to explain chemical diversity. Natural Product Reports **20,** 382-391.

Greenhagen, B. T., Griggs, P., Takahashi, S., Ralston, L., and Chappell, J. (2003). Probing sesquiterpene hydroxylase activities in a coupled assay with terpene synthases. Archives of Biochemistry and Biophysics **409,** 385-394.

He, J., Benedito, V., Wang, M., Murray, J., Zhao, P., Tang, Y., and Udvardi, M. (2009). The *Medicago truncatula* gene expression atlas web server. BMC Bioinformatics **10,** 441.

Helliwell, C. A., Poole, A., James Peacock, W., and Dennis, E. S. (1999). Arabidopsis ent-Kaurene Oxidase Catalyzes Three Steps of Gibberellin Biosynthesis. Plant Physiology **119,** 507-510.

Huhman, D. V., and 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,** 347-360.

Iturbe-Ormaetxe, I. a., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus. Plant Molecular Biology **51,** 731-743.

Lahoucine Achnine, D. V. H., Mohamed A. Farag, Lloyd W. Sumner, Jack W. Blount, Richard A. Dixon,. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. The Plant Journal **41,** 875-887.

Lu, C. D., and Jorgensen, N. A. (1987). Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants. The Journal of Nutrition **117,** 919-927.

Lu, C. D., Tsai, L. S., Schaefer, D. M., and Jorgensen, N. A. (1987). Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins. Journal of dairy science **70,** 799-805.

Matsuno, Michiyo, Compagnon, V., Schoch, G. A., Schmitt, M., Debayle, D., Bassard, J. -E., Pollet, B., Hehn, A., Heintz, D., Ullmann, P., Lapierre, C., Bernier, F. o., Ehlting, J. r., and Werck-Reichhart, D. l. (2009). Evolution of a Novel Phenolic Pathway for Pollen Development. Science **325,** 1688-1692.

Million Tadege, J. W., Ji He, Haidi Tu, Younsig Kwak, Alexis Eschstruth, Anne Cayrel, Gabriella Endre, Patrick X. Zhao, Mireille Chabaud, Pascal Ratet, Kirankumar S. Mysore,. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. The Plant Journal **54,** 335-347.

Morant, M., Bak, S., Møller, B. L., and Werck-Reichhart, D. (2003). Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Current Opinion in Biotechnology **14,** 151-162.

Naoumkina, M. A., Modolo, L. V., Huhman, D. V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L. W., and Dixon, R. A. (2010). Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*. Plant Cell **22,** 850-866.

Owen, S. M., and Peñuelas, J. (2005). Opportunistic emissions of volatile isoprenoids. Trends in Plant Science **10,** 420-426.

Pang, Y., Wenger, J. P., Saathoff, K., Peel, G. J., Wen, J., Huhman, D., Allen, S. N., Tang, Y., Cheng, X., Tadege, M., Ratet, P., Mysore, K. S., Sumner, L.

W., Marks, M. D., and Dixon, R. A. (2009). A WD40 Repeat Protein from *Medicago truncatula* Is Necessary for Tissue-Specific Anthocyanin and Proanthocyanidin Biosynthesis But Not for Trichome Development. Plant Physiol. **151,** 1114-1129.

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J., and Osbourn, A. E. (1999). Compromised Disease Resistance in Saponin-Deficient Plants. Proceedings of the National Academy of Sciences of the United States of America **96,** 12923-12928.

Peñuelas, J., and Llusià, J. (2004). Plant VOC emissions: making use of the unavoidable. Trends in ecology & evolution (Personal edition) **19,** 402-404.

Pompon, D., Louerat, B., Bronine, A., Urban, P., Eric, F. J., and Michael, R. W. (1996). [6] Yeast expression of animal and plant P450s in optimized redox environments. In Methods in Enzymology (Academic Press), pp. 51-64.

Ro, D. -K., Arimura, G. -I., Lau, S. Y. W., Piers, E., and Bohlmann, J. r. (2005). Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. Proceedings of the National Academy of Sciences of the United States of America **102,** 8060-8065.

Rozen, S., and Skaletsky, H. (1999). Primer3 on the WWW for General Users and for Biologist Programmers, pp. 365-386.

Sambrook, J., Russell, D. W., and Cold Spring Harbor, L. (2001). Molecular cloning : a laboratory manual / Joseph Sambrook, David W. Russell. (Cold Spring Harbor, N. Y. :: Cold Spring Harbor Laboratory).

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? Trends in Biochemical Sciences **28,** 336-341.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K., and Muranaka, T. (2008). Licorice Î²-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,** 14204-14209.

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

Siminszky, B., Corbin, F. T., Ward, E. R., Fleischmann, T. J., and Dewey, R. E. (1999). Expression of a soybean cytochrome P450 monooxygenase cDNA in yeast and tobacco enhances the metabolism of phenylurea herbicides. Proceedings of the National Academy of Sciences **96,** 1750-1755.

Suzuki, H., Achnine, L., Xu, R., Matsuda, S. P. T., and Dixon, R. A. (2002). A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*

doi:10. 1046/j. 1365-313X. 2002. 01497. x. The Plant Journal **32,** 1033-1048.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., Cayrel, A., Endre, G., Zhao, P. X., Chabaud, M., Ratet, P., and Mysore, K. S. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. The Plant Journal **54,** 335-347.

Urban, P., Mignotte, C., Kazmaier, M., Delorme, F., and Pompon, D. (1997). Cloning, Yeast Expression, and Characterization of the Coupling of Two Distantly Related Arabidopsis thaliana NADPH-Cytochrome P450 Reductases with P450 CYP73A5

10. 1074/jbc. 272. 31. 19176. J. Biol. Chem. **272,** 19176-19186.

Yu, C., Shin, Y. G., Kosmeder, J. W., Pezzuto, J. M., and van Breemen, R. B. (2003). Liquid chromatography/tandem mass spectrometric determination of inhibition of human cytochrome P450 isozymes by resveratrol and resveratrol-3-sulfate. Rapid Communications in Mass Spectrometry **17,** 307-313.

Chapter IV References

Altschul, S., W. Gish, et al. (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology **215**: 403 - 410.

Anne E. Osbourn, X. Q., Belinda Townsend, Bo Qin, (2003). "Dissecting plant secondary metabolism; constitutive chemical defences in cereals." New Phytologist **159**(1): 101-108.

Augustin, J. M., V. Kuzina, et al. (2011). "Molecular activities, biosynthesis and evolution of triterpenoid saponins." Phytochemistry **72**(6): 435-457.

Ballester, A.-R., J. Molthoff, et al. (2011). "Biochemical and Molecular Analysis of Pink Tomatoes: Deregulated Expression of the Gene Encoding Transcription Factor SlMYB12 Leads to Pink Tomato Fruit Color." Plant Physiology **152**(1): 71-84.

Benedito, V., I. Torres-Jerez, et al. (2008). "Affymetrix GeneChip Medicago Genome Array

A gene expression atlas of the model legume *Medicago truncatula*." The Plant Journal **55**: 504 - 513.

Bino, R. J., R. D. Hall, et al. (2004). "Potential of metabolomics as a functional genomics tool." Trends in Plant Science **9**(9): 418-425.

Broeckling, C. D., D. V. Huhman, et al. (2005). "Metabolic profiling of." Journal of Experimental Botany **56**(410): 323-336.

Broeckling, C. D., D. V. Huhman, et al. (2005). "Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism

10.1093/jxb/eri058." Journal of Experimental Botany **56**(410): 323-336.

Broeckling, C. D., I. R. Reddy, et al. (2006). "MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics." Analytical Chemistry **78**(13): 4334-4341.

Bucciarelli, B., J. Hanan, et al. (2006). "A Standardized Method for Analysis of *Medicago truncatula* Phenotypic Development

10.1104/pp.106.082594." Plant Physiology **142**(1): 207-219.

Chan, E. K. F., H. C. Rowe, et al. (2010). "The Complex Genetic Architecture of the Metabolome." PLoS Genetics **6**(11): e1001198.

Chang-Jun Liu, D. H., Lloyd W. Sumner, Richard A. Dixon, (2003). "Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from *Medicago truncatula*." The Plant Journal **36**(4): 471-484.

Chen, J., J. Yu, et al. "Control of dissected leaf morphology by a Cys(2)His(2) zinc finger transcription factor in the model legume *Medicago truncatula*." Proceedings of the National Academy of Sciences **107**(23): 10754-10759.

Chu, H. Y., E. Wegel, et al. (in press). "From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants." The Plant Journal **66**(1): 66-79.

Confalonieri, M., M. Cammareri, et al. (2009). "Enhanced triterpene saponin biosynthesis and root nodulation in transgenic barrel medic (*Medicago truncatula* Gaertn.) expressing a novel β -amyrin synthase (AsOXA1) gene." Plant Biotechnology Journal **7**(2): 172-182.

Czechowski, T., M. Stitt, et al. (2005). "Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis." Plant Physiology **139**(1): 5-17.

Delis, C., A. Krokida, et al. "Role of lupeol synthase in Lotus japonicus nodule formation." New Phytologist **189**(1): 335-346.

Dixon, R. A. (2001). "Natural products and plant disease resistance." Nature **411**(6839): 843-847.

Dixon, R. A. and L. W. Sumner (2003). "Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health." Plant Physiology **131**(3): 878-885.

Dozmorov, I. and M. Centola (2003). "An associative analysis of gene expression array data." Bioinformatics **19**(2): 204-211.

Fiehn, O. (2002). "Metabolomics - the link between genotypes and phenotypes." Plant Molecular Biology **48**: 155 - 171.

Field, B. and A. E. Osbourn (2008). "Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Different Plants." Science **320**(5875): 543-547.

Field, B. and A. E. Osbourn (2008). "Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Plants
10.1126/science.1154990." Science: 1154990.

Firn, R. D. and C. G. Jones (2003). "Natural products - a simple model to explain chemical diversity." Natural Product Reports **20**(4): 382-391.

Frey, M., K. Huber, et al. (2003). "A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-biosynthesis." Phytochemistry **62**(3): 371-376.

Frey, M., K. Schullehner, et al. "Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants." Phytochemistry **70**(15-16): 1645-1651.

Gierl, A. and M. Frey (2001). "Evolution of benzoxazinone biosynthesis and indole production in maize." Planta **213**(4): 493-498.

Goodacre, R., L. Roberts, et al. (2007). "From phenotype to genotype: whole tissue profiling for plant breeding." Metabolomics **3**(4): 489-501.

Greenhagen, B. T., P. Griggs, et al. (2003). "Probing sesquiterpene hydroxylase activities in a coupled assay with terpene synthases." Archives of Biochemistry and Biophysics **409**(2): 385-394.

Hannah, M. A., C. Caldana, et al. (2010). "Combined Transcript and Metabolite Profiling of Arabidopsis Grown under Widely Variant Growth Conditions Facilitates the Identification of Novel Metabolite-Mediated Regulation of Gene Expression." Plant Physiology **152**(4): 2120-2129.

Haridas, V., M. Higuchi, et al. (2001). "Avicins: Triterpenoid saponins from Acacia victoriae (Bentham) induce apoptosis by mitochondrial perturbation 10.1073/pnas.101619098." Proceedings of the National Academy of Sciences **98**(10): 5821-5826.

Harjes, C. E., T. R. Rocheford, et al. (2008). "Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification." Science **319**(5861): 330-333.

He, J., V. Benedito, et al. (2009). "The *Medicago truncatula* gene expression atlas web server." BMC Bioinformatics **10**(1): 441.

Helliwell, C. A., P. M. Chandler, et al. (2001). "The CYP88A cytochrome P450, ent-kaurenoic acid oxidase, catalyzes three steps of the gibberellin biosynthesis pathway." Proceedings of the National Academy of Sciences **98**(4): 2065-2070.

Helliwell, C. A., A. Poole, et al. (1999). "Arabidopsis ent-Kaurene Oxidase Catalyzes Three Steps of Gibberellin Biosynthesis." Plant Physiology **119**(2): 507-510.

Hirai, M., M. Yano, et al. (2004). "Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana." Proceedings of the National Academy of Sciences **101**: 10205 - 10210.

Huhman, D. V., M. A. Berhow, et al. (2005). "Quantification of Saponins in Aerial and Subterranean Tissues of *Medicago truncatula*." Journal of Agricultural and Food Chemistry **53**(6): 1914-1920.

Huhman, D. V. and L. W. Sumner (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.

Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Research **31**(4): e15.

Iturbe-Ormaetxe, I. a., K. Haralampidis, et al. (2003). "Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and Lotus japonicus." Plant Molecular Biology **51**(5): 731-743.

Jonczyk, R., H. Schmidt, et al. (2008). "Elucidation of the Final Reactions of DIMBOA-Glucoside Biosynthesis in Maize: Characterization of Bx6 and Bx7." Plant Physiology **146**(3): 1053-1063.

Kalo, P., C. Gleason, et al. (2005). "Nodulation Signaling in Legumes Requires NSP2, a Member of the GRAS Family of Transcriptional Regulators." Science **308**(5729): 1786-1789.

Kirk, D. D., R. Rempel, et al. (2004). "Application of Quillaja saponaria extracts as oral adjuvants for plant-made vaccines
doi:10.1517/14712598.4.6.947." Expert Opinion on Biological Therapy **4**(6): 947-958.

Krueger, S., P. Giavalisco, et al. (2011). "A Topological Map of the Compartmentalized <italic>Arabidopsis thaliana</italic> Leaf Metabolome." PLoS ONE **6**(3): e17806.

Kuljanabhagavad, T., P. Thongphasuk, et al. (2008). "Triterpene saponins from Chenopodium quinoa Willd." Phytochemistry **69**(9): 1919-1926.

Lahoucine Achnine, D. V. H., Mohamed A. Farag, Lloyd W. Sumner, Jack W. Blount, Richard A. Dixon, (2005). "Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*." The Plant Journal **41**(6): 875-887.

Lee, S. M. S. and P. Y. Lai (2009). "Double block bootstrap confidence intervals for dependent data." Biometrika **96**(2): 427-443.

Li, L., H. Cheng, et al. (2007). "Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*." Planta **226**(1): 109-123.

Lu, C. D. and N. A. Jorgensen (1987). "Alfalfa Saponins Affect Site and Extent of Nutrient Digestion in Ruminants." The Journal of Nutrition **117**(5): 919-927.

Lu, C. D., L. S. Tsai, et al. (1987). "Alteration of Fermentation in Continuous Culture of Mixed Rumen Bacteria by Isolated Alfalfa Saponins." Journal of Dairy Science **70**(4): 799-805.

Matsuda, F., M. Y. Hirai, et al. (2010). "AtMetExpress Development: A Phytochemical Atlas of Arabidopsis Development." Plant Physiology **152**(2): 566-578.

Matsuno, Michiyo, V.compagnon, et al. (2009). "Evolution of a Novel Phenolic Pathway for Pollen Development." Science **325**(5948): 1688-1692.

Meesapyodsuk, D., J. Balsevich, et al. (2007). "Saponin Biosynthesis in Saponaria vaccaria. cDNAs Encoding beta-Amyrin Synthase and a Triterpene Carboxylic Acid Glucosyltransferase
10.1104/pp.106.088484." Plant Physiology **143**(2): 959-969.

Miguel, C. l. and L. Marum (2011). "An epigenetic view of plant cells cultured in vitro: somaclonal variation and beyond." Journal of Experimental Botany.

Million Tadege, J. W., Ji He, Haidi Tu, Younsig Kwak, Alexis Eschstruth, Anne Cayrel, Gabriella Endre, Patrick X. Zhao, Mireille Chabaud, Pascal Ratet, Kirankumar S. Mysore, (2008). "Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*." The Plant Journal **54**(2): 335-347.

Morant, M., S. Bak, et al. (2003). "Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation." Current Opinion in Biotechnology **14**(2): 151-162.

Mugford, S. T., X. Qi, et al. (2009). "A Serine Carboxypeptidase-Like Acyltransferase Is Required for Synthesis of Antimicrobial Compounds and Disease Resistance in Oats." The Plant Cell Online **21**(8): 2473-2484.

Mylona, P., A. Owatworakit, et al. (2008). "Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat
10.1105/tpc.107.056531." Plant Cell **20**(1): 201-212.

Naoumkina, M. A., L. V. Modolo, et al. (2010). "Genomic and Coexpression Analyses Predict Multiple Genes Involved in Triterpene Saponin Biosynthesis in *Medicago truncatula*." Plant Cell **22**(3): 850-866.

Nielsen, J., T. Nagao, et al. (2010). "Resistance in the Plant Barbarea vulgaris and Counter-Adaptations in Flea Beetles Mediated by Saponins." Journal of Chemical Ecology **36**(3): 277-285.

Osbourn, A. (2010). "Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology." Plant Physiology **154**(2): 531-535.

Owen, S. M. and J. Peñuelas (2005). "Opportunistic emissions of volatile isoprenoids." Trends in Plant Science **10**(9): 420-426.

P. Houghton, N. P., M. Jurzysta, Z. Biely, C. Cheung, (2006). "Antidermatophyte activity of medicago extracts and contained saponins and their structure-activity relationships." Phytotherapy Research **20**(12): 1061-1066.

Pang, Y., J. P. Wenger, et al. (2009). "A WD40 Repeat Protein from *Medicago truncatula* Is Necessary for Tissue-Specific Anthocyanin and Proanthocyanidin Biosynthesis But Not for Trichome Development." Plant Physiology **151**(3): 1114-1129.

Papadopoulou, K., R. E. Melton, et al. (1999). "Compromised disease resistance in saponin-deficient plants." Proceedings of the National Academy of Sciences of the United States of America **96**(22): 12923-12928.

Parmigiani, G., E. Garrett, et al. (2003). DNA-Chip Analyzer (dChip). The Analysis of Gene Expression Data. M. Gail, K. Krickeberg, J. Samet, A. Tsiatis and W. Wong, Springer London**:** 120-141.

Pedersen, M. W., D. K. Barnes, et al. (1976). "Effects of Low and High Saponin Selection in Alfalfa on Agronomic and Pest Resistance Traits and the Interrelationship of these Traits." Crop Science **16**(2): 193-199.

Peñuelas, J. and J. Llusià (2004). "Plant VOC emissions: making use of the unavoidable." Trends in Ecology & Evolution (Personal Edition) **19**(8): 402-404.

Pfaffl, M. W. (2001). "A new mathematical model for relative quantification in real-time RTPCR." Nucleic Acids Research **29**(9): e45.

Pompon, D., B. Louerat, et al. (1996). [6] Yeast expression of animal and plant P450s in optimized redox environments. Methods in Enzymology, Academic Press. **Volume 272:** 51-64.

Qi, X., S. Bakht, et al. (2004). "A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants

10.1073/pnas.0401301101." Proceedings of the National Academy of Sciences of the

Qin, B., J. Eagles, et al. "High throughput screening of mutants of oat that are defective in triterpene synthesis." Phytochemistry **71**(11-12): 1245-1252.

Ro, D.-K., G.-I. Arimura, et al. (2005). "Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase." Proceedings of the National Academy of Sciences of the United States of America **102**(22): 8060-8065.

Ronfort, J., T. Bataillon, et al. (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.

Rozen, S. and H. Skaletsky (1999). Primer3 on the WWW for General Users and for Biologist Programmers. **132:** 365-386.

Ruijter, J. M., C. Ramakers, et al. (2009). "Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data." Nucleic Acids Research **37**(6): e45.

Saito, K., M. Y. Hirai, et al. (2008). "Decoding genes with coexpression networks and metabolomics - `majority report by precogs'." Trends in Plant Science **13**(1): 36-43.

Sakamoto, T., K. Miura, et al. (2004). "An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice." Plant Physiology **134**(4): 1642-1653.

Sambrook, J., D. W. Russell, et al. (2001). Molecular cloning : a laboratory manual / Joseph Sambrook, David W. Russell. Cold Spring Harbor, N.Y. :, Cold Spring Harbor Laboratory.

Schenk, R. U. and A. C. Hildebrandt (1972). "Medium and techniques for induction and growth of monocotyledonous and dicotyledonous plant cell cultures." Canadian Journal of Botany **50**(1): 199-204.

Schilmiller, A., F. Shi, et al. (2010). "Mass spectrometry screening reveals widespread diversity in trichome specialized metabolites of tomato chromosomal substitution lines." The Plant Journal **62**(3): 391-403.

Schmidt, S., S. Sunyaev, et al. (2003). "Metabolites: a helping hand for pathway evolution?" Trends in Biochemical Sciences **28**(6): 336-341.

Seki, H., K. Ohyama, et al. (2008). "Licorice beta-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): 14204-14209.

Seki, H., K. Ohyama, et al. (2008). "Licorice Î²-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): 14204-14209.

Sen, S., H. P. S. Makkar, et al. (1998). "Alfalfa Saponins and Their Implication in Animal Nutrition." Journal of Agricultural and Food Chemistry **46**(1): 131-140.

Shibuya, M., M. Hoshino, et al. (2006). "Identification of β -amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay." FEBS Journal **273**(5): 948-959.

Shimura, K., A. Okada, et al. (2007). "Identification of a Biosynthetic Gene Cluster in Rice for Momilactones." Journal of Biological Chemistry **282**(47): 34013-34018.

Siminszky, B., F. T. Corbin, et al. (1999). "Expression of a soybean cytochrome P450 monooxygenase cDNA in yeast and tobacco enhances the metabolism of phenylurea herbicides." Proceedings of the National Academy of Sciences **96**(4): 1750-1755.

Sulpice, R., S. Trenkamp, et al. (in press). "Network Analysis of Enzyme Activities and Metabolite Levels and Their Relationship to Biomass in a Large Panel of Arabidopsis Accessions." The Plant Cell Online.

Suzuki, H., L. Achnine, et al. (2002). "A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*
doi:10.1046/j.1365-313X.2002.01497.x." The Plant Journal **32**(6): 1033-1048.

Swaminathan, S., D. Morrone, et al. (2009). "CYP76M7 Is an ent-Cassadiene C11Î±-Hydroxylase Defining a Second Multifunctional Diterpenoid Biosynthetic Gene Cluster in Rice." The Plant Cell Online **21**(10): 3315-3325.

Tadege, M., J. Wen, et al. (2008). "Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*." The Plant Journal **54**(2): 335-347.

Tohge, T. and A. R. Fernie (2010). "Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function." Nature Protocols **5**(6): 1210-1227.

Tohge, T., Y. Nishiyama, et al. (2005). "Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor." The Plant Journal **42**: 218 - 235.

Udvardi, M., K. Kakar, et al. (2007). "Legume transcription factors: global regulators of plant development and response to the environment." Plant Physiology **144**: 538 - 549.

Urban, P., C. Mignotte, et al. (1997). "Cloning, Yeast Expression, and Characterization of the Coupling of Two Distantly Related Arabidopsis thaliana NADPH-Cytochrome P450 Reductases with P450 CYP73A5
10.1074/jbc.272.31.19176." The Journal of Biological Chemistry **272**(31): 19176-19186.

Wilderman, P. R., M. Xu, et al. (2004). "Identification of Syn-Pimara-7,15-Diene Synthase Reveals Functional Clustering of Terpene Synthases Involved in Rice Phytoalexin/Allelochemical Biosynthesis." Plant Physiology **135**(4): 2098-2105.

Yu, C., Y. G. Shin, et al. (2003). "Liquid chromatography/tandem mass spectrometric determination of inhibition of human cytochrome P450 isozymes by resveratrol

and resveratrol-3-sulfate." Rapid Communications in Mass Spectrometry **17**(4): 307-313.