

The Pennsylvania State University

The Graduate School

Graduate Degree Program in Integrative Biosciences

**MOLECULAR ANALYSIS OF GENES INVOLVED IN THE SYNTHESIS OF  
PROANTHOCYANIDINS IN *THEOBROMA CACAO***

A Dissertation in

Integrative Biosciences

BY

Yi Liu

© 2010 Yi Liu

Submitted in Partial Fulfillment  
of the Requirements  
for the Degree of

Doctor of Philosophy

May 2010

The Dissertation of Yi Liu was reviewed and approved\* by the following:

Mark J. Gultinan  
Professor of Plant Molecular Biology  
Dissertation Advisor  
Chair of Committee

Daniel J. Cosgrove  
Professor of Biology

Ming Tien  
Professor of Biochemistry

Devin Peterson  
Associate Professor of Food Science and Nutrition

Siela Maximova  
Associate Professor of Horticulture

Peter Hudson  
Willaman Professor of Biology  
Director of Huck Institutes of the Life Sciences

\*Signatures are on file in the Graduate School

## ABSTRACT

The flavonoids catechin and epicatechin, and their polymerized oligomers, the proanthocyanidins (PAs, also called condensed tannins), accumulate to levels of up to 15% of the total weight of dry seeds of *Theobroma cacao* L. These compounds have been associated with several health benefits in humans including antioxidant activity, improvement of cardiovascular health and reduction of cholesterol levels. They also play important roles in pest and disease defense throughout the plant. This research focuses on molecularly dissecting the proanthocyanidin biosynthetic pathway of *Theobroma cacao*. To this end, I first isolated candidate genes from *T.cacao* (Tc) encoding key structural enzymes of this pathway which include, anthocyanidin reductase (ANR), leucoanthocyanidin dioxygenase (LDOX, also called anthocyanidin synthase, ANS) and leucoanthocyanidin reductase (LAR). I performed gene expression profiling of candidate *TcANR*, *TcANS* and *TcLAR* in various tissues through different developmental stages and also evaluated PA accumulation levels in those tissues. My results suggested that all PA candidate genes are co-regulated and positively correlated with PA synthesis. To functionally analyze the candidate genes, I used the model plants Arabidopsis and tobacco as expression platforms. Results from Arabidopsis mutant complementation tests and transgenic tobacco plants constitutively overexpressing cacao genes demonstrate that the candidate structural genes isolated from cacao are true *ANS*, *ANR* and *LAR* genes and all actively involved in PA synthesis in cacao. To further explore the transcriptional regulation of the PA synthesis pathway, I then isolated and characterized an R2R3 type MYB transcription factor *TcMYBPA* from cacao. I examined the spatial and temporal gene expression patterns of *TcMYBPA* in cacao and found it to be developmentally expressed in a manner consistent with its involvement in PAs as well as anthocyanin synthesis. Complementation test of *TcMYBPA* in Arabidopsis *tt2* mutant suggested that *TcMYBPA* could functionally substitute Arabidopsis *TT2* gene. Interestingly, except PA accumulation in seeds, I also observed an obvious increase of anthocyanidin accumulation in hypocotyls of transgenic Arabidopsis plants. This is consistent with gene expression analysis which showed that the entire PA pathway could be induced by overexpression of *TcMYBPA* gene, including *DFR*, *LDOX* (*ANS*)

and *BAN (ANR)*. Therefore I concluded that the isolated *TcMYBPA* gene encodes an R2R3 type MYB transcription factor and is involved in the regulation of both anthocyanin and PA synthesis in cacao. This research will not only offer us the knowledge of secondary metabolites production in cacao, but also provides molecular tools for breeding of cacao varieties with improved disease resistance and enhanced flavonoid profiles for nutritional and pharmaceutical applications.



## TABLE OF CONTENTS

<b>LIST OF FIGURES</b> .....	<b>viii</b>
<b>LIST OF TABLES</b> .....	<b>x</b>
<b>CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW</b> .....	<b>1</b>
<b>1.1 Metabolic and Functional Diversity of the Flavonoids</b> .....	<b>1</b>
<b>1.2 The Flavonoid Biosynthetic Pathway</b> .....	<b>2</b>
<b>1.3 Formation and Structures of Proanthocyanidins</b> .....	<b>3</b>
<b>1.4 Genes Encoding Key Enzymes in the Proanthocyanidin Biosynthetic Pathway</b> .....	<b>5</b>
1.4.1 Dihydroflavonol 4-Reductase (DFR) .....	5
1.4.2 Flavonol Synthase (FLS).....	5
1.4.3 Leucoanthocyanidin Reductase (LAR).....	6
1.4.4 Anthocyanidin Reductase (ANR).....	7
1.4.5 Leucoanthocyanidin Dioxygenase (LDOX).....	8
1.4.6 UDP-glucose:Flavonoid 3-O-Glucosyltransferase (UGFT).....	8
<b>1.5 Transcription Factors Regulating PA Synthesis</b> .....	<b>9</b>
<b>1.6 Significance of Flavan-3-ols and PAs in Plant Resistance Mechanisms</b> .....	<b>12</b>
1.6.1 UV Absorption .....	12
1.6.2 Binding of Metals.....	12
1.6.3 Oxidation to Quinones .....	13
1.6.4 Molecular Interaction with Proteins.....	13
<b>1.7 Health Benefits of Flavan-3-ols and Proanthocyanidins</b> .....	<b>14</b>
1.7.1 Inhibition of LDL Oxidation .....	14
1.7.2 Vasodilation.....	15
1.7.3 Inhibition of Platelet Activation.....	15
1.7.4 Anti-Inflammation Activity .....	16
<b>1.8 Cacao Proanthocyanidin Bioavailability and Bioactivity</b> .....	<b>16</b>
<b>1.9 References:</b> .....	<b>18</b>
<b>CHAPTER 2: MOLECULAR BIOLOGY OF PROANTHOCYANIDIN SYNTHESIS IN CACAO: CHARACTERIZATION OF GENES ENCODING ANTHOCYANIDIN REDUCTASE, ANTHOCYANIDIN SYNTHASE AND LEUCOANTHOCYANIDIN REDUCTASE</b> .....	<b>23</b>
<b>2.1 Introduction</b> .....	<b>23</b>
<b>2.2 Materials and Methods</b> .....	<b>26</b>
2.2.1 Plant Material .....	26
2.2.2 Nucleic Acid Purification and cDNA Synthesis .....	28
2.2.3 Isolation of cDNA and Genomic Clones from <i>Theobroma cacao</i> .....	28
2.2.4 Phylogenetic Analysis .....	30
2.2.5 Southern Blot Analysis.....	31
2.2.6 Proanthocyanidin (PA) and Anthocyanin Quantification.....	31
2.2.7 Genetic Transformation of Tobacco and Arabidopsis.....	33
2.2.8 Expression Analysis of <i>TcANS</i> , <i>TcANR</i> and <i>TcLAR</i> .....	33
<b>2.3 Results</b> .....	<b>35</b>
2.3.1 Cloning and Sequence Analysis of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> Genes .....	35
2.3.2 Southern Blot Analysis.....	40

2.3.3	Expression Profiles of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> Genes and PA Accumulation in <i>Theobroma cacao</i> Tissues .....	41
2.3.4	Functional Analysis of the <i>TcANR</i> Gene.....	46
2.3.5	Functional Analysis of the <i>TcANS</i> Gene .....	48
<b>2.4</b>	<b>Discussion .....</b>	<b>54</b>
2.4.1	The Distribution of PA Accumulation Correlates with Expression of the PA Biosynthesis Genes.....	54
2.4.2	Functions of <i>ANS</i> , <i>ANR</i> and <i>LAR</i> in PA Biosynthesis in Cacao .....	56
<b>2.5</b>	<b>Conclusions.....</b>	<b>57</b>
<b>2.6</b>	<b>References .....</b>	<b>59</b>
<b>CHAPTER 3: FUNCTIONAL ANALYSIS OF CACAO LEUCOANTHOCYANIDIN REDUCTASE.....</b>		
<b>3.1</b>	<b>Introduction .....</b>	<b>63</b>
<b>3.2</b>	<b>Materials and Methods .....</b>	<b>64</b>
3.2.1	Extraction and Quantification of PAs .....	64
3.2.2	Protein Sequence Alignment.....	65
3.2.3	Transformation of Arabidopsis with <i>TcLAR</i> .....	66
3.2.4	Gene Expression Analysis.....	66
3.2.5	In Vitro Expression of Recombinant TcLAR protein.....	66
3.2.6	Assay of LAR Activities .....	67
<b>3.3</b>	<b>Results.....</b>	<b>68</b>
3.3.1	Sequence Analysis of the TcLAR protein .....	68
3.3.2	Nature of <i>Theobroma cacao</i> PAs.....	70
3.3.3	Functional Characterization of <i>TcLAR</i> .....	73
3.3.4	In Vitro Enzyme Assay of TcLAR.....	77
<b>3.4</b>	<b>Discussion .....</b>	<b>78</b>
<b>3.5</b>	<b>Conclusions.....</b>	<b>81</b>
<b>3.6</b>	<b>References .....</b>	<b>83</b>
<b>CHAPTER 4: THE <i>THEOBROMA CACAO TCMYBPA</i> GENE ENCODES AN R2R3-MYB TRANSCRIPTION FACTOR INVOLVED IN THE REGULATION OF PROANTHOCYANIDIN SYNTHESIS.....</b>		
<b>4.1</b>	<b>Introduction .....</b>	<b>85</b>
<b>4.2</b>	<b>Materials and Methods .....</b>	<b>87</b>
4.2.1	Isolation of a <i>TcMYBPA</i> cDNA from <i>Theobroma cacao</i> .....	87
4.2.2	Protein Sequence Alignment and Phylogenetic Analysis .....	88
4.2.3	Transformation of Arabidopsis.....	88
4.2.4	Gene Expression Analysis.....	89
4.2.5	DMACA Staining of Tissues.....	90
<b>4.3</b>	<b>Results.....</b>	<b>91</b>
4.3.1	The Cacao <i>TcMybPA</i> Gene Encodes an R2R3-MYB Transcription Factor.....	91
4.3.2	Expression of <i>TcMYBPA</i> Correlates with PA Accumulation in <i>Theobroma cacao</i> .....	94
4.3.3	<i>TcMYBPA</i> Complements the PA-Deficient Phenotype of the Arabidopsis <i>tt2</i> Mutant....	96
<b>4.4</b>	<b>Discussion .....</b>	<b>100</b>
4.4.1	Identification of TcMYBPA as an R2R3-MYB-Type Transcriptional Regulator .....	100
4.4.2	Expression of <i>TcMYBPA</i> Correlates with PA and Anthocyanin Synthesis in Cacao.....	101
4.4.3	<i>TcMYBPA</i> Complements the <i>tt2</i> Seed Phenotype and Induces Anthocyanin Synthesis in Arabidopsis .....	103
<b>4.5</b>	<b>Conclusions.....</b>	<b>104</b>

4.6	References .....	105
<b>CHAPTER 5: GENERAL CONCLUSIONS .....</b>		<b>108</b>
5.1	PAs are Actively Synthesized in Cacao Leaves, Flowers and Pods .....	108
5.2	<i>TcANS</i> , <i>TcANR</i> and <i>TcLAR</i> Encode Key Enzymes Involved in PA Synthesis .....	109
5.3	<i>TcMYBPA</i> Encodes a MYB Transcription Factor that Regulates Both PA and Anthocyanin Synthesis in Cacao .....	111
5.4	References .....	112
<b>Appendix A: Sequences of the Theobroma cacao genes cloned in this thesis .....</b>		<b>121</b>
<b>Appendix B: Protein sequence comparisons of TcANR, TcANS, TcLAR and TcMYBPA with their homologus from other species.....</b>		<b>129</b>
i.	Protein sequence comparison of TcANR and AtBAN (BANYULS).....	129
ii.	Protein sequence comparison of TcANS and AtLDOX (ANS) .....	129
iii.	Protein sequence comparison of TcLAR and DsLAR.....	130
i.	Protein sequence comparison of TcMYBPA and VvMYBPA1.....	130
<b>Appendix C: Multiple sequence alignment of the LAR, ANS and ANR proteins as well as related IFR and DFR proteins of the RED superfamily.....</b>		<b>131</b>

## LIST OF FIGURES

<b>Figure 1-1.</b> Outline of the flavonoid biosynthetic pathway (adapted from Winkel-Shirley, (2002)).....	1
<b>Figure 1-2.</b> Outline of the proanthocyanidin synthesis pathway (adapted from Xie et al., 2003).4	4
<b>Figure 1-3.</b> Working model of anthocyanin and proanthocyanidin synthesis pathway with regulatory genes (adapted from (Dixon et al., 2005)).....	10
<b>Figure 2-1.</b> Developmental stages of cacao leaf (Scavina 6). .....	27
<b>Figure 2-2:</b> Structures of cacao <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> genes with predicted protein domains, and comparison with their orthologues from Arabidopsis ( <i>AtANR</i> and <i>AtANS</i> ), <i>Medicago</i> ( <i>MtLAR</i> ) and <i>Vitis</i> ( <i>VvLAR</i> ). .....	36
<b>Figure 2-3:</b> Phylogenetic analysis of the LAR, ANS and ANR proteins as well as related IFR and DFR proteins of the RED superfamily. The protein sequences were aligned as described in Section 2.2.4. ....	39
<b>Figure 2-4:</b> Cacao genomic Southern blot using full-length cDNA as a probe.....	40
<b>Figure 2-5:</b> Expression of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> genes and accumulation of PAs in <i>Theobroma cacao</i> (Scavina 6; S6) leaves and flowers at various developmental stages. ....	42
<b>Figure 2-6:</b> Expression of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> genes and accumulation of PAs in whole pods of <i>Theobroma cacao</i> (Amelonado) during early stages of pod development (from 2 to 6 weeks after pollination).....	43
<b>Figure 2-7:</b> Expression of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> genes and accumulation of PAs in pod exocarp of <i>Theobroma cacao</i> (Amelonado) during pod development (from 8 to 20 weeks after pollination).....	44
<b>Figure 2-8:</b> Expression of <i>TcANR</i> , <i>TcANS</i> and <i>TcLAR</i> genes and accumulation of PAs in ovules of <i>Theobroma cacao</i> (Amelonado) during pod development (from 8 to 20 weeks after pollination).....	45
<b>Figure 2-9:</b> Complementation of the PA deficient <i>ban</i> mutant phenotype by constitutively expressing <i>TcANR</i> . .....	47
<b>Figure 2-10:</b> Characterization of transgenic tobacco flowers constitutively expressing <i>TcANS</i> . .....	49
<b>Figure 2-11:</b> Complementation of the PA- and anthocyanin-deficient <i>ans</i> ( <i>ldox</i> ) mutant phenotype by constitutively expressing <i>TcANS</i> . Two independent lines showed a similar pattern to Col-0 for PA and anthocyanin accumulation.....	51
<b>Figure 2-12:</b> Characterization of transgenic tobacco flowers constitutively expressing <i>TcLAR</i> . .....	53

<b>Figure 3-1:</b> Multiple sequence alignment of LAR and IFR proteins. ....	69
<b>Figure 3-2:</b> Composition of PAs in leaves and flowers of <i>Theobroma cacao</i> . ....	71
<b>Figure 3-3:</b> Flavan-3-ol accumulation and composition in fruit tissues of <i>Theobroma cacao</i> (Amelonado) at various developmental stages analyzed by HPLC. ....	72
<b>Figure 3-4:</b> Flavan-3-ol accumulation and composition in flowers of wild type (Samsun, ss) tobacco plants and two independent lines of 35S: <i>TcLAR</i> transgenic (LAR7C1, LAR9A2) tobacco plants. ....	73
<b>Figure 3-5:</b> Complementation of the Arabidopsis PA-deficient <i>ldox</i> mutant by constitutively expressing <i>TcLAR</i> . ....	75
<b>Figure 3-6:</b> HPLC analysis of extracts from transgenic Arabidopsis and tobacco samples with and without spiking with known standards to validate the identity of the catechin and epicatechin peaks. ....	76
<b>Figure 3-7:</b> Enzymatic assay of recombinant TcLAR. ....	77
<b>Figure 4-1:</b> Comparison of the amino acid sequences of TcMYBPA and various plant MYB transcription factors. ....	93
<b>Figure 4-2:</b> Transcript levels of <i>TcMYBPA</i> during the development of cacao leaves and flowers (A), young pods (B), exocarp (C) and ovules (D). ....	96
<b>Figure 4-3:</b> Complementation of the PA-deficient <i>tt2</i> mutant phenotype by constitutively expressing <i>TcMYBPA</i> . ....	98
<b>Figure 4-4:</b> Semi-quantitative RT-PCR analysis of expression of flavonoid structural genes in young seedlings of the same <i>Arabidopsis</i> lines analyzed in Figure 4-3. ....	100

## LIST OF TABLES

<b>Table 2-1:</b> Sources of LAR, DFR, ANR and ANS proteins used in Phylogenetic analysis.....	30
<b>Table 2-2:</b> Structures of <i>ANR</i> , <i>ANS</i> and <i>LAR</i> genes.....	37
<b>Table 4-1:</b> Sequences of the primers used in the gene expression study.....	89

## Acknowledgement

This work was not possible without the effort and support of many individuals.

First and foremost, I owe an enormous debt of gratitude to my advisor Mark Gultinan, who has shown me the beauty of science, offered me precious opportunities for collaboration, and supplied sustenance, literally and metaphorically, through long winters at State College. I am very grateful to my committee members, Ming Tian and David Peterson, for offering insightful suggestions and friendly accessibility throughout my graduate study. I would like to thank Daniel Cosgrove for his detailed comments on this dissertation. Special thanks go to Siela Moximova for being there through those hard times. In the Gultinan lab, I would like to thank Joe Verica for teaching me basic molecular biology techniques hand-by-hand, Ann Yuong and Sharon Pishak for countless technical support, and the Marna Yandea-Nelson for teaching me protein biology techniques and numerous suggestions on my experimental design. I would also like to specially thank Yang lab, Ma lab and dePamphilis lab for their advice and help. This dissertation has also received contributions from The Hershey Company, for which I owe my special thanks to Mark Payne, Kenneth Miller and David Stuart who provided equipment and supplies for HPLC analysis and enormous help on data collection and processing. I feel especially thankful to my supervisors and colleagues from Dupont and Genentech. The experience working with you inspired me.

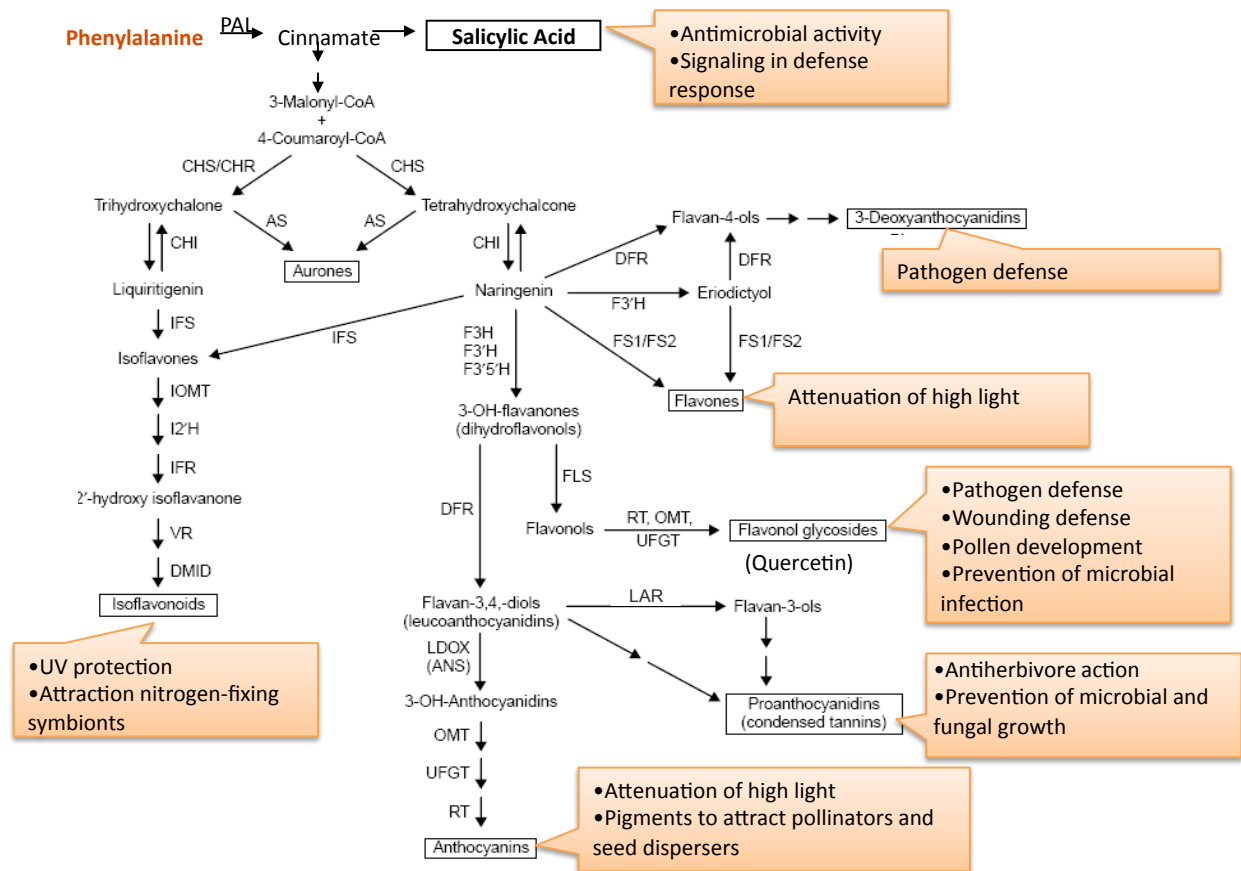
To my “family” in the USA, Ding Zhou, Zi Shi, XiaoYing Meng, Hui Zhang, Huan Xia and Rong Xiao, thank you all for being with me during ups and downs. I love you all.

Finally yet importantly, I am deeply indebted to my parents for bringing me into this world and providing me with a sound education. You have always been my cheerleader. Your endless love supported me through the toughest times.

## CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

### 1.1 Metabolic and Functional Diversity of the Flavonoids

Flavonoids are a diverse group of secondary metabolites that are made in plants and have various biological functions. They are involved in plant defense against insects, pathogens and microbes and in absorption of free radicals and UV light. They also can act as pigments that attract beneficial symbionts and pollinators. Because the flavonoids are important for optimal plant growth and thus maximal agricultural productivity, the biochemistry and molecular biology of flavonoids is an important and very advanced area of research. Much of the knowledge in this field was reviewed by Winkel-Shirley, Xie, Dixon and Lepiniec (Saslowsky and Winkel-Shirley, 2001; Winkel-Shirley, 2002; Dixon et al.,



**Figure 1-1.** Outline of the flavonoid biosynthetic pathway (adapted from Winkel-Shirley, (2002). The names of the major classes of flavonoid endproducts are boxed and the main functions of the products are listed in the rectangular callouts.



2005; Lepiniec et al., 2006). Figure 1-1 shows an outline of the flavonoid biosynthetic pathway and a summary of biological functions of a few key metabolites.

**Flavonols, isoflavonoids and 3-deoxyanthocyanidins** are flavanoids that are synthesized in various plant tissues as a normal part of development, and that further accumulate following stimulation by pathogen attack. The levels of these compounds increase at the sites of infection to concentrations that are toxic to pathogens (Dixon and Paiva, 1995). **Flavones** as well as **anthocyanins** increase in response to high-intensity visible light; these compounds act as photoprotectants that reduce the amount of light reaching the photosynthetic system (Beggs et al., 1987). **Flavonols** such as kaempferol can be induced by wounding and are toxic to potential herbivores (Hahlbrock and Scheel, 1989). They are also involved in pollen development (Ausubel, 1992), and may function to reduce microbial infection (Dixon and Paiva, 1995). **Proanthocyanidins** possess antiherbivore activities owing to their bitter and astringent tastes and negative effects on digestion; they also defend against microbial and fungal growth (Dixon et al., 2005). **Isoflavonoids** can be induced by UV irradiation and protect plant tissues by absorbing UV light (Beggs et al., 1987). Also, isoflavonoids increase in low-nitrogen conditions, and can also act as chemoattractors for nitrogen-fixing symbionts (Graham, 1991). Among all these flavonoids compounds, flavanols and proanthocyanidins have gained special attention because they have shown benefits not only to plants but also to human health, in addition, proanthocyanidins are found at high levels in cacao (Gu et al., 2002; 2006), the subject organism of this thesis research.

## 1.2 The Flavonoid Biosynthetic Pathway

All flavonoids are derived from cinnamic acid, a derivative of the amino acid phenylalanine. Their biosynthetic pathways share some general steps and most start from the condensation of three malonyl-CoA units and p-coumaroyl-CoA catalyzed by chalcone synthase (CHS) to produce tetrahydroxychalcone (Figure 1-1). Yellow-colored tetrahydroxychalcone is then converted into the colorless naringenin through the

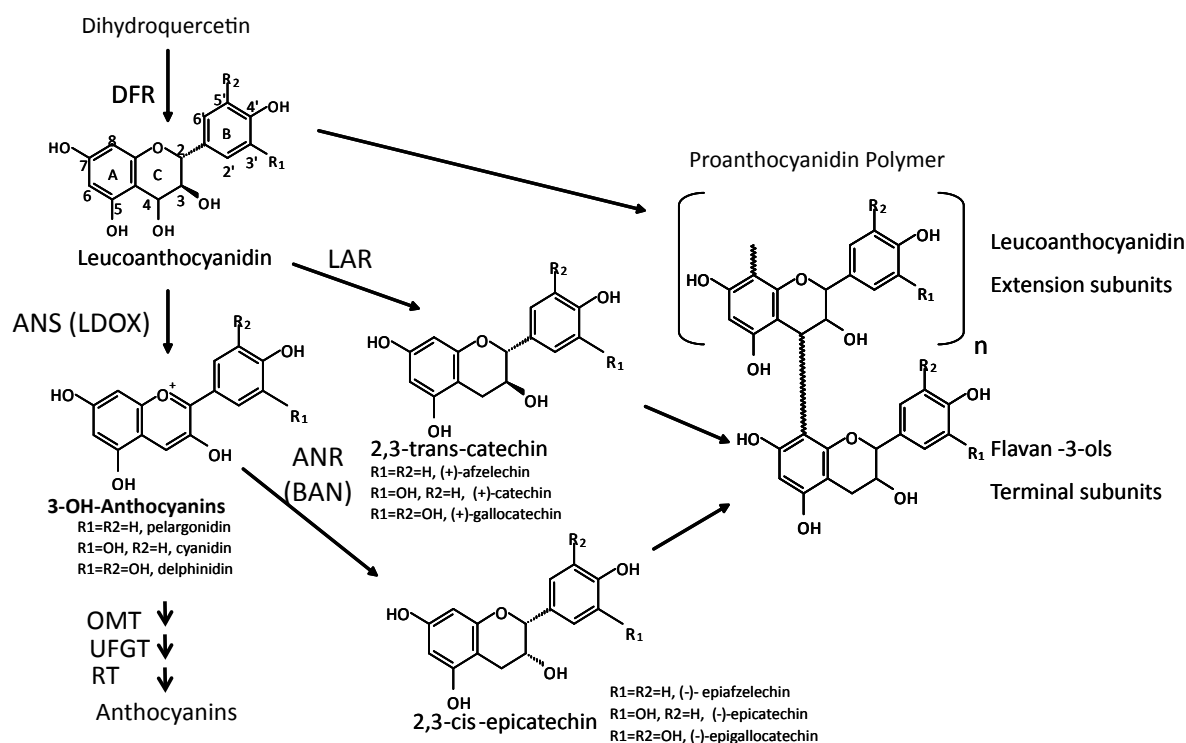
stereospecific isomerization by chalcone isomerase (CHI) (Dixon and Paiva, 1995; Holton and Cornish, 1995). In legume species, tetrahydroxychalcone can also be reduced to trihydroxychalcone by chalcone reductase (CHR), and then converted into liquiritigenin by CHI (Welle and Grisebach, 1989).

Naringenin enters into different pathways as a substrate for the synthesis of six different groups of flavonoids. It can be converted into dihydroflavonols by flavanone 3-hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H) or flavonoid 3', 5'-hydroxylase (F3'5'H). Dihydroflavonols can then be converted into flavonols by flavonol synthase (FLS) and anthocyanins by a series of enzymes including dihydroflavonol reductase (DFR), anthocyanidin synthase (ANS), UDP-glucose flavonol 3-O-glucosyl transferase (UFGT). Alternatively, naringenin can be converted by isoflavone synthase into isoflavones, which are the precursor for the synthesis of isoflavonoids. Naringenin can also be converted by DFR into flavan-4-ols which are the precursors of 3-deoxyanthocyanidins, or it can be converted into flavones by flavone synthase 1 and 2 (FS1/FS2). One set of intermediates in the anthocyanin synthesis pathway, leucoanthocyanidins and 3-OH-anthocyanins are converted into the flavan-3-ols (catechin and epicatechin) which are polymerized into proanthocyanidins (condensed tannins), which is the major topic of study in this thesis.

### **1.3 Formation and Structures of Proanthocyanidins**

The synthesis of proanthocyanidins (PAs) and anthocyanins shares common steps in the flavonoid biosynthesis pathway up to the synthesis of flavan-3,4-diols (such as leucoanthocyanidin), which not only are precursors for anthocyanin and flavan-3-ols synthesis, but also contribute to the extension units of the PA polymers (Figure 1-2) (Dixon et al., 2005). Flavan-3-ols (sometimes referred to as flavanols, such as catechin or epicatechin) derived from leucoanthocyanidins are believed to act as terminal units to initiate PA polymerization, while intermediates derived from leucoanthocyanidins themselves act as extension units to add to flavan-3-ol initiators through C4-C8 linkage

(dominant form of PAs shown in Figure 1-2) or C4-C6 linkage to form branches (not shown).



**Figure 1-2.** Outline of the proanthocyanidin synthesis pathway (adapted from Xie et al., 2003). Enzymes are represented in uppercase letters. DFR, dihydroflavonol 4-reductase, EC 1.1.1.219; ANS, anthocyanidin synthase, EC 1.14.11.19; ANR, anthocyanidin reductase, EC 1.3.1.77; LAR, leucoanthocyanidin reductase, EC 1.17.1.3; OMT, O-methyltransferases, EC 2.1.1.6; UFGT, UDP-glucose: anthocyanidin/flavonol 3-O-glucosyltransferase, EC 2.4.1.115; RT, rhamnosyltransferase, EC 2.4.1.-.

The hydroxylation pattern of the B-ring of the monomeric proanthocyanidins is determined by the presence of the cytochrome P450 monooxygenases flavonoid 3'-hydroxylase (F3'H) and flavonoid 3', 5'-hydroxylase (F3'5'H), enzymes that act early in the flavonoid synthesis pathway after the formation of naringenin (Winkel-Shirley, 2002; Dixon et al., 2005) (Figure 1-1). In the absence of both of these cytochrome P450 enzymes, hydroxylation occurs only at the 4' position of B rings, yielding (epi)afzelechin. In the presence of F3'H, the 3' position will be hydroxylated resulting in the formation of (epi)catechin. In the presence of F3'5'H, the 5' position will also be hydroxylated leading to the formation of (epi)gallocatechin (Figure 1-2). The activity of F3'H and F3'5'H will also

cause similar hydroxylation pattern on the B-ring of anthocyanins, resulting in the formation of pelargonidin with only one hydroxyl group, cyanidin with two hydroxyl groups, and delphinidin with three hydroxyl groups. The pigments derived from each anthocyanin have a characteristic color range since the visible absorption maximum becomes longer with the increase in B-ring hydroxyl groups: pelargonidin derived pigments show orange, pink or red colors, cyanidin-derived pigments show red or magenta colors and delphinidin-derived pigments show purple or blue colors (Zucker et al., 2002).

## **1.4 Genes Encoding Key Enzymes in the Proanthocyanidin Biosynthetic Pathway**

### **1.4.1 Dihydroflavonol 4-Reductase (DFR)**

The enzyme dihydroflavonol 4-reductase (DFR) catalyzes the reduction of dihydroflavonols to leucoanthocyanidins (Dixon et al., 2005). A gene encoding DFR was first isolated from maize (O'Reilly et al., 1985) and snapdragon (Beld et al., 1989) by transposon tagging. The snapdragon clone was used to isolate homologous sequences from petunia, which contains three genes encoding DFR (*DFRa*, *DFRb* and *DFRc*). However, only *DFRa* is expressed in floral tissues. Using the sequences of the snapdragon and maize *DFR* genes, degenerate primers were designed to amplify and isolate the Arabidopsis (*Arabidopsis thaliana*) *DFR* gene (Shirley et al., 1992). Low stringency genomic Southern blot analysis suggested that DFR was encoded by a single copy gene in Arabidopsis.

### **1.4.2 Flavonol Synthase (FLS)**

The enzyme flavonol synthase (FLS) catalyzes the conversion of dihydroflavonols to flavonols. This reaction introduces a double bond between carbons 2 and 3 of the C-ring (Schijlen et al., 2004). The gene encoding FLS was first cloned from petunia, in which flavonol synthesis is controlled by the Fl locus (Holton et al., 1993). Since FLS belongs to the dioxygenase family, dioxygenase homologous gene fragments were amplified by degenerate primers based on conserved dioxygenase sequences from FL/FL wild type and

fl/fl mutant lines of petunia that have greatly reduced flavonol synthesis. PCR products were then used to screening a petunia petal cDNA library for clones that hybridized more strongly to FL/FL wild-type PCR products. In this process, the full-length cDNA clone of FLS was identified. The petunia FLS gene sequence was then used to isolate a homologous gene from Arabidopsis. Low stringency DNA blots showed that, in Arabidopsis, FLS is a multigene family with at least three isoforms (Pelletier et al., 1997). Also, in citrus, FLS is encoded by more than one gene in citrus as demonstrated by DNA hybridization with the *CitFLS* cDNA, which detected two to five fragments in the citrus genomic DNA digests of various restriction enzymes (Moriguchi et al., 2002).

FLS and DFR both use the common substrate dihydroflavonols, which can either enter the side branch of flavonol synthesis through the action of FLS or enter the anthocyanin and proanthocyanidin synthesis pathways after conversion into leucoanthocyanidin by DFR (Figure 1-1). Because of their shared substrate, FLS and DFR exhibit competitive interactions. This was demonstrated by introduction of an antisense petunia FLS cDNA into petunia and tobacco, resulting in the production of darker flower color and increased anthocyanin level plus decreased flavonol level (Holton et al., 1993).

### **1.4.3 Leucoanthocyanidin Reductase (LAR)**

Leucoanthocyanidin reductase (LAR) serves as the first branching point leading to the production of PAs in the anthocyanin pathway (Schijlen et al., 2004). It converts leucoanthocyanidin to catechin (Figure 1-2), which is the terminal unit of the PA polymer. The LAR protein was first purified from *Desmodium uncinatum*, and the gene was subsequently cloned based on the protein sequence (Tanner et al., 2003). But curiously, two studies found that there is no *LAR* ortholog in the Arabidopsis genome although it produces PAs. This suggests that Arabidopsis may use only epicatechin to produce PAs (Abrahams et al., 2003). While *LAR* appears to exist as a single copy gene in *Desmodium*, 2 highly related homologous genes (*VvLAR1* and *VvLAR2*) with distinct expression profiles were discovered in grape (Bogs et al., 2005). Gene expression was correlated to catechin

production in different tissues of grape (*Vitis vinifera*). Similarly, two *LAR* genes were also found in *Lotus corniculatus* (*LcLAR1* and *LcLAR2*), but in the *in vitro* enzyme assays, only *LcLAR1* protein showed enzyme activity and converted leucoanthocyanidin to catechin (Paolucci et al., 2007). Recently, another single copy *LAR* gene was identified from *Medicago truncatula* (Pang et al., 2007). Although this *LAR* protein demonstrated weak activity in the *in vitro* enzyme assays, overexpressing it in tobacco failed to divert the metabolic flow from anthocyanin to catechin synthesis. Surprisingly, the authors also found that the *Medicago* PAs are composed almost entirely of epicatechin, despite the existence of *LAR* gene in the *Medicago* genome and the expression of *LAR* gene in the PA accumulating tissues. Therefore, the precise function of *LAR* protein remains unclear.

#### **1.4.4 Anthocyanidin Reductase (ANR)**

Anthocyanidin reductase (ANR) converts cyanidin to epicatechin, which is also an initiator of PA polymers. Thus, ANR is the second branching point in the anthocyanin pathway leading to PA production. The *ANR* gene (also known as *BANYULS*, *BAN*) was first isolated using Arabidopsis T-DNA mutants (Devic et al., 1999). The function of the *ANR* gene in reducing cyanidin to epicatechin was demonstrated by Xie et al. (2003) by *in vitro* enzyme assays and *in vivo* analysis in transgenic tobacco. Ectopic expression of the Arabidopsis *ANR* gene in tobacco petals resulted in the loss of anthocyanin and accumulation of PAs. *ANR* was also isolated from *Medicago truncatula* and grape, and their functions demonstrated in tobacco (Xie et al., 2003; Bogs et al., 2005). In these three species, *ANR* appears to be present as a single copy in the genome (Devic et al., 1999; Xie et al., 2003; Bogs et al., 2005). The three-dimensional structure of the grape ANR protein was recently described (Gargouri et al., 2009). The study further showed that the grape ANR protein has an epimerase activity in addition to reductase activity, and can convert cyanidin to a 50:50 mixture of both 2,3-*cis* and 2,3-*trans* flavan-3-ols.

#### 1.4.5 Leucoanthocyanidin Dioxygenase (LDOX)

Leucoanthocyanidin dioxygenase (LDOX) converts leucoanthocyanidin to cyanidin, which serves as the precursor of epicatechin and PA synthesis. Since cyanidin also serves as the precursor for anthocyanin, LDOX is also called anthocyanin synthase (ANS). LDOX is a key enzyme that is important to both PA and anthocyanin synthesis. *LDOX* was first cloned in the maize *a2* mutant by transposon tagging (Menssen et al., 1990) and then from snapdragon by differential screening (Martin et al., 1991). *LDOX* was also isolated from Arabidopsis by screening the EST database for *LDOX*-homologous sequences (Pelletier et al., 1997). Complete genome sequence analysis as well as Southern blot analysis showed that *LDOX* is a single copy gene in Arabidopsis. The Arabidopsis *tds4* mutant is defective in the *LDOX* gene and is deficient in both anthocyanins and PAs accumulation (Abrahams et al., 2003). An *LDOX* (*ANS*) gene was also characterized from *Medicago* (Pang et al., 2007). Downregulation of *ANS* gene in *Medicago* leaves resulted in a reduction of both anthocyanin and PAs, consistent with its role in both anthocyanin and PA synthesis.

#### 1.4.6 UDP-glucose:Flavonoid 3-O-Glucosyltransferase (UFGT)

Under normal physiological conditions, flavonoids and anthocyanidins with a free hydroxyl group in the C3 position of the heterocyclic ring are not stable, so they do not accumulate to significant levels. To stabilize anthocyanidins, a glucose moiety can be transferred to the C3 hydroxyl group from UDP-glucose by the enzyme UDP-glucose:flavonoid 3-O-glucosyltransferase (UFGT) (Schijlen et al., 2004). UFGT is required for anthocyanidins and flavonoids stability, water solubility and accumulation in vacuoles.

Using a transposon tagging method, Fedoroff et al. (1984) isolated the first UFGT gene from maize (the *Bz1* gene) . Using the maize gene as a probe, a clone encoding UFGT was then isolated from snapdragon and its function confirmed by expression in *Lisianthus* (Martin et al., 1991; Schwinn et al., 1997). UFGT genes were later isolated from five species including Urad (*Vigna mungo*) (Mato et al., 1998), Green Shiso (*Perilla frutescens*) (Saito et

al., 1999), Baikal Skullcap (*Scutellaria baicalensis*) (Hirotani et al., 2000), grape (Fukuchi-Mizutani et al., 2003) and most recently in Arabidopsis (Kubo et al., 2007).

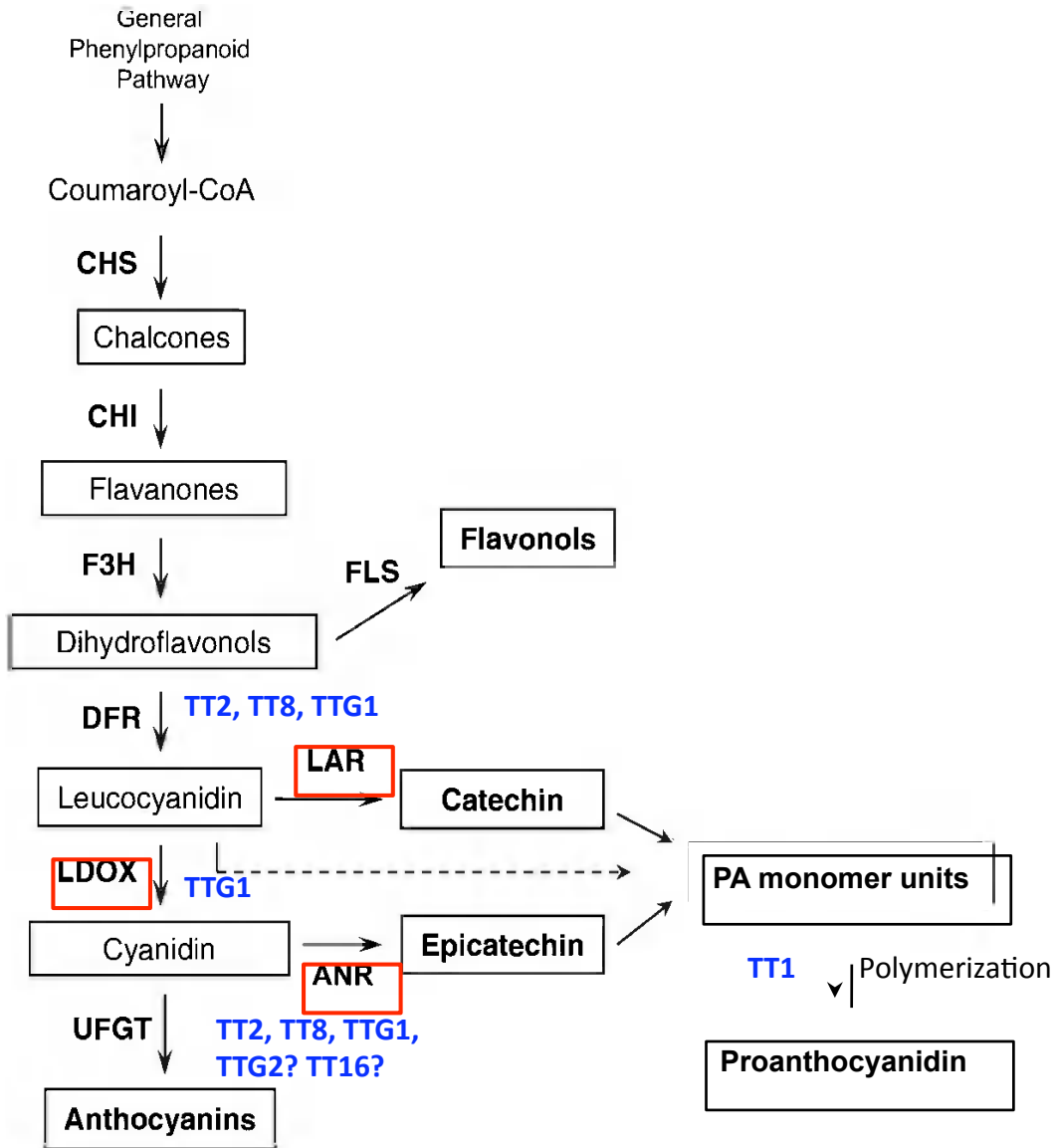
## 1.5 Transcription Factors Regulating PA Synthesis

In the recent years, many genes regulating PA synthesis have been cloned and characterized using mutants having reduced PA or anthocyanin content (Marles et al., 2003). The majority of these genes were cloned from Arabidopsis mainly due to the large collection of *transparent testa* (*tt*) transposon tagging or T-DNA insertion mutants. In Arabidopsis, PAs accumulate specifically in the innermost integumentary layer of the seed coat (endothelium) and will give the mature seed testa a brown color after oxidation (Lepiniec et al., 2006). When genes required for normal PA synthesis are mutated, the mature seed will display a transparent testa phenotype (*TT*).

The *TT* genes isolated to date include a basic helix-loop-helix (bHLH) transcription factor (*TT8*) (Alemanno et al., 1997), a MYB transcription factor (*TT2*) (Nesi et al., 2001), a WD-40 repeat (WDR) protein (*TTG1*) (Walker et al., 1999), a MADS box gene (*TT16*, BSISTER) (Johnson et al., 2002), a WRKY transcription factor (*TTG2*) (Johnson et al., 2002), and a new type of zinc finger protein (WIP) (*TT1*) (Sagasser et al., 2002) (Figure 1-3). The *TRANSPARENT TESTA GLABRA1* (*TTG1*) was the first anthocyanin regulator isolated in Arabidopsis through positional cloning (Walker et al., 1999). *TTG1* is required for the expression of the *DFR* structural gene, and is thus necessary for anthocyanin and PA synthesis. This gene is expressed in almost all major organs in Arabidopsis and has widespread functions other than anthocyanin synthesis, including root hair and trichome formation as well as production of seed mucilage production. *TTG1* encodes a protein with four WR40 repeats, which are crucial for protein-protein interaction, and appears to act upstream of both bHLH and MYB transcription factors. The *TRANSPARENT TESTA 8* (*TT8*) gene was isolated using Arabidopsis T-DNA insertion mutants (Nesi et al., 2000). It encodes a bHLH domain protein required for expression of *DFR* and *BAN* (*ANR*) genes in Arabidopsis siliques. It is expressed only in developing siliques and young seedlings but not in leaves, roots, stems and other tissues. Another transcription factor *TRANSPARENT TESTA 2* (*TT2*), controlling PA and anthocyanin synthesis was isolated from Arabidopsis



also using T-DNA tagging (Nesi et al., 2001). It encodes an R2R3 MYB domain protein that is also required for *DFR* and *BAN* expression, the same as *TT8* and *TTG1*.



**Figure 1-3.** Working model of anthocyanin and proanthocyanidin synthesis pathway with regulatory genes (adapted from (Dixon et al., 2005)). Enzymes are represented in uppercase bold letters; transcription factors regulating the enzymes in Arabidopsis are represented by blue letters beside the names of enzymes. The products in the pathway are given in black boxes. Red boxes indicate the genes that are objects of research in this thesis.

These results suggest that *TT2*, *TT8* and *TTG1* may interact to control the PA and anthocyanin biosynthesis. However, the expression of *TT2* is more restricted than *TT8* and *TTG1*; the gene is expressed only in the seed testa during early embryogenesis. This is consistent with the *BAN* expression and the PA synthesis patterns. Overexpression of *TT2* can induce ectopic transcription of *BAN*, indicating that *TT2* is the limiting factor that determines the spatial and temporal expression of the *BAN* gene. A subsequent study using molecular approaches both in yeast and *in planta* in Arabidopsis showed that *TT2*, *TT8* and *TTG1* can form a stable ternary complex (Baudry et al., 2004). *TT2*, in combination with *TT8*, is responsible for specific recognition of the promoter region of *BAN*. *TTG1* regulates the activity of the complex, potentially by acting on *TT8*. Further study of the *TT8* promoter activity showed that the expression of *TT8* involves the presence of *TT8* itself or homologous bHLH factors as well as *TTG1* and MYB factors. This positive feedback regulation results in a strong restriction of cell-specific expression of *BAN* gene and accumulation of PAs.

*TTG2*, a related gene cloned from Arabidopsis (Johnson et al., 2002), is involved in both trichome development and PA synthesis in the seed coat. It is a WRKY transcription factor that acts downstream of *TTG1*. However, the mutation in *TTG2* does not affect the activity of the *BAN* gene promoter (Debeaujon et al., 2003). The downstream structural gene targets of *TTG2* have yet to be discovered.

*TRANSPARENT TESTA16 (TT16)* was identified through T-DNA tagging (Nesi et al., 2002). It encodes a MADS box protein that belongs to the BSISTER clade (Becker et al., 2002), which consists of genes of unknown function that are mainly expressed in female organs. *TT16* is necessary for the expression of the *BAN* gene as well as the specification of endothelium cells. A mutation in *TT16* has been demonstrated to affect the spatial pattern of *BAN* promoter activity (Debeaujon et al., 2003), but its mechanism and downstream targets still remain unclear.

*TRANSPARENT TESTA 1 (TT1)* was isolated through transposon tagging (Sagasser et al., 2002). The protein contains a novel combination of two TFIIIA-type zinc finger motifs defining a new subfamily of plant zinc finger transcription factors called “WIP” (featuring in

protein domains containing the sequence Trp-Ile-Pro). The mutant has no PAs and a modified spatial pattern of *BAN* gene expression as well as an altered morphology of the seed endothelium where PA is usually deposited (Sagasser et al., 2002; Debeaujon et al., 2003). However, catechin, the monomeric subunit of PAs, was detected in the *tt1* mutant, suggesting that *TT1* may be regulating genes specifically controlling PA polymerization (Sagasser et al., 2002).

## **1.6 Significance of Flavan-3-ols and PAs in Plant Resistance Mechanisms**

Monomeric flavan-3-ols and PAs and their derivatives are present in various plant tissues including fruit, leaf and seed in many plants. Their main functions are thought to provide protection against UV light, microbial pathogens, as well as insect and herbivore invasions (Dixon et al., 2005) and will be reviewed briefly below.

### **1.6.1 UV Absorption**

PAs have UV absorbing properties with a peak absorbance at 280 nm (Revilla, 1991). PAs are often present in epidermal cells of leaves and other tissues that are exposed to high levels of UV light (Winkel-Shirley, 2002). Arabidopsis mutants in *CHS* or *CHI* genes are deficient in PAs and show UV-hypersensitive phenotypes (Li et al., 1993). In poplar, UV irradiation can induce the TT2-like R2R3 MYB transcription factor that is the regulator of the entire PA synthesis pathway. This factor also activates the downstream PA pathway genes and causes a significant increase of PAs accumulation in leaves, suggesting role for PAs in UV stress response (Mellway et al., 2009).

### **1.6.2 Binding of Metals**

Proanthocyanidins bind metals through O-diphenol group complexation (Dixon et al., 2005). This property has been suggested to be one mechanism for antimicrobial activity since iron depletion results in severe limitation of bacterial growth (Scalbert, 1991). The growth of some bacteria such as *Pseudomonas maltophilia* and *Enterobacter cloacae* are inhibited by the proanthocyanidins catechin and epicatechin (Feucht and Treutter, 1999). Similar inhibitory effects were also observed on the growth of some non-phytopathogenic bacteria such as *Staphylococcus aureus*, *Lactobacillus casei* and *Escherichia coli* in the presence of proanthocyanidins and prodelphinidins from grape. The metal binding properties of proanthocyanidins might also function in plant aluminum tolerance, as in the case of the legume *Lotus pedunculatus* where they might complex aluminum in the tannin-rich vacuoles of the root apices (Stoutjesdijk et al., 2001).

### **1.6.3 Oxidation to Quinones**

A recent study of Arabidopsis seeds resulted in the identification of the *TT10* gene, which encodes a laccase-like enzyme that is involved in the oxidative polymerization of PAs (Pourcel et al., 2005). In the presence of the laccase-like oxidase, Arabidopsis epicatechins are oxidized to the corresponding quinones, which would initiate subsequent polymerization resulting in the formation of brown PA polymers. While oxidized, PA polymers will cross-link with the cell wall in seed testa to create a physical barrier to water permeation. Furthermore, since quinones are powerful antibiotics (Pourcel et al., 2007), the PA polymer barrier may also act against pathogens to protect seeds during storage and germination.

### **1.6.4 Molecular Interaction with Proteins**

Proanthocyanidins interact irreversibly with proteins via O-quinone-intermediates as a result of stable covalent bonds that lead to precipitation (Feucht and Treutter, 1999). This

strong interaction is responsible for the astringency of some fruits and other plant tissues, which protects plants from herbivores. Studies in various peanut (*Arachis hypogaea*) genotypes have shown that the fecundity of aphids was inversely proportional to the levels of constitutive proanthocyanidins in the leaf-bud petioles where they feed (Dixon et al., 2005).

## **1.7 Health Benefits of Flavan-3-ols and Proanthocyanidins**

Research of the cardiovascular effects of dietary flavonoids was first stimulated by a study of the Kuna indigenous people conducted by Hollenberg et al. (1997). The study found that Kuna Indians residing on the San Blas Islands north of Panama have very low blood pressure and are free of hypertension and cardiovascular disease. However, when they migrate to urban cities of Panama, they start to suffer age-related blood pressure rise, and a much higher frequency of hypertension. One outstanding feature of the diet differences between island and city-dwelling Kuna includes a very high intake of flavonoid-rich cocoa beverages. Island-dwelling Kuna people use flavonoid-rich raw cocoa powder to flavor their drinking water that is shipped from the mainland. On average, these people consume 40 cups of raw cocoa drinks per week. While city-dwelling Kuna stopped drinking cocoa or drink commercial cocoa that is stripped of flavonoids for a better taste. This study suggested that cocoa flavonoids make an important contribution to a healthy cardiovascular system.

Recent epidemiological and nutritional studies have provided a growing body of evidence that a specific subclass of flavonoids might have a role in improving human cardiovascular health (Norman et al., 2009). This subset of flavonoids includes flavonols such as quercetin and flavan-3-ols that are the monomeric subunits of PA polymers known as catechin and epicatechin. These flavonoids have been conclusively shown to improve cardiovascular health in at least four ways: inhibition of LDL oxidation, vasodilatation, inhibition of platelet activation and prevention of inflammation (Engler and Engler, 2006)

### **1.7.1 Inhibition of LDL Oxidation**

The pathogenesis of atherosclerosis is related to the oxidative modification of low density lipoprotein (LDL), which is the main carrier of cholesterol (Baba et al., 2001). A study conducted in healthy humans indicated that after consuming high polyphenol cacao, the lag time to conjugated diene production was significantly prolonged as a result of LDL oxidation induced by different oxidant agents. It was also observed that a much higher level of epicatechin accumulated in the urine of the cacao consumer. This indicated that the flavan-3-ols in the cacao drink had been absorbed and were associated with resistance against the oxidation of LDL. The antioxidant capacity and diminished production of oxidative products in plasma has also been shown to be related to increased concentrations of cacao epicatechin by another *in vivo* study in healthy human subjects (Wan et al., 2001).

### **1.7.2 Vasodilation**

A study conducted with New Zealand white rabbits demonstrated that the oligomeric proanthocyanidins not only can cause endothelium dependent relaxation but also can activate the synthesis of nitric oxide synthase (NOS) (Karim et al., 2000). Since nitric oxide functions to promote vasodilation, this study indicates that oligomeric proanthocyanidins can induce vasodilatation by modulating endothelium dependent relaxation mediated by activating NOS and increasing the endothelium nitric oxide level. Another study conducted in healthy humans following daily consumption of flavonol rich cacao also reported induced vasodilation as well as improvement in endothelial function via activation of the nitric oxide system (Norman et al., 2009).

### **1.7.3 Inhibition of Platelet Activation**

PAs can also effect human health through interactions with blood components. *In vitro* experiments indicated that after incubating cacao oligomeric proanthocyanidin or dealcoholized red wine (DRW) with whole blood cells, there was a reduction in the

expression of surface protein receptors GPIIb/IIIa that are essential for platelet aggregation and thrombus formation (Rein et al., 2000). Similar results were also observed in blood obtained from healthy human subjects after cacao beverage consumption (Rein et al., 2000). These *in vitro* and *in vivo* studies suggested that cacao flavonoids could suppress platelet activation and platelet microparticle formation and could have an aspirin-like effect on primary vascular homeostasis.

#### **1.7.4 Anti-Inflammation Activity**

Transcription factor nuclear factor- $\kappa$ B (NF $\kappa$ B), which regulates the expression of genes encoding pro-inflammatory proteins such as cytokines, can be activated by reactive oxygen species (ROS) (Christman et al., 2000). Studies in rat indicated that ROS generation was significantly reduced in cultured rat hepatocyte cells treated with quercetin, and that activation of NF $\kappa$ B was also blocked after quercetin treatment (Martinez-Florez et al., 2005). These studies suggested that flavonol might exert its anti-inflammation activity by inhibiting the activation of NF $\kappa$ B and subsequent proinflammatory proteins through diminishing intracellular production of ROS.

### **1.8 Cacao Proanthocyanidin Bioavailability and Bioactivity**

There are a variety of plant-based foods and beverages that serve as a natural sources of flavanols, including cacao, red wine, grapes, apples, and cranberries. Among those, cacao has an extraordinarily high amount of flavanols (Keen et al., 2002), which make up about 15% of the dry weight of unprocessed cacao beans (Hannum and Erdman, 2000). Cacao flavanols typically consist of 20-30% monomers (epicatechin and catechin), over 50% dimers, with the larger proanthocyanidin polymers constitute the remaining 20-30% (Baudry et al., 2004). Differences in total flavanol content exist between cacao

samples differing in cultivars, geographical origin and farming practices (Fisher and Hollenberg, 2005).

The exact steps leading to the polymerization of flavanol monomers is still not known, nor is the precise active form of flavanols (monomers, dimers or polymers) in the human body. To date, in human blood samples after digestion of flavonoid-rich cacao, only flavan-3-ol monomers and dimers have been measured, and no oligomers larger than dimers have yet been detected (Fisher and Hollenberg, 2005). However, *in vitro*, these monomers and dimers have demonstrated very little or no activity in stimulating nitric oxide synthesis. Only tetramers through decamers of catechin showed an effect on endothelium dependent relaxation (Karim et al., 2000). A growing body of research results suggests that flavan-3-ol dimers from cacao are transformed by the microorganisms in the gut into low molecular weight phenolic acids before absorption, and that it may these metabolites that play a major role in health benefits associated with proanthocyanidin consumption (Appeldoorn et al., 2009; Stoupi et al., 2010).

The work presented in this thesis begins to dissect the PA pathway in cacao using a translational genomics strategy. This strategy involved the isolation of cacao genes of interest based on sequence similarities with other species, then the characterization of the structure and expression of each of the genes. To verify the predicted enzymatic activity of the proteins encoded by each of the cacao genes, they were expressed in transgenic *Arabidopsis* plants that had mutations in the cognate structural or regulatory genes (transgenic complementation assay), or were over-expressed in wild-type tobacco (*Nicotiana tabacum*) plants. These results characterized four important cacao genes involved in PA biosynthesis: *TcANR*, *TcANS*, *TcLAR* and *TcMYBPA*. In addition, they also demonstrate that the PA biosynthesis pathway and its regulatory mechanisms are highly conserved between cacao and *Arabidopsis*.



## 1.9 References:

- Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR** (2003) The *Arabidopsis* *TDS4* gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *Plant J* **35**: 624-636
- Abrahams S, Tanner GJ, Larkin PJ, Ashton AR** (2002) Identification and biochemical characterization of mutants in the proanthocyanidin pathway in *Arabidopsis*. *Plant Physiol* **130**: 561-576
- Alemanno L, Berthouly M, MichauxFerriere N** (1997) A comparison between *Theobroma cacao* L. zygotic embryogenesis and somatic embryogenesis from floral explants. *In Vitro Cell. Dev. Biol. Plant* **33**: 163-172
- Appeldoorn MM, Vincken JP, Aura AM, Hollman PC, Gruppen H** (2009) Procyanidin dimers are metabolized by human microbiota with 2-(3,4-dihydroxyphenyl)acetic acid and 5-(3,4-dihydroxyphenyl)-gamma-valerolactone as the major metabolites. *J Agric Food Chem* **57**: 1084-1092
- Ausubel FM** (1992) Short protocols in molecular biology: a compendium of methods from Current protocols in molecular biology, Ed 2nd. Greene Pub. Associates; John Wiley & Sons, New York
- Baba S, Osakabe N, Natsume M, Muto Y, Takizawa T, Terao J** (2001) *In vivo* comparison of the bioavailability of (+)-catechin, (-)-epicatechin and their mixture in orally administered rats. *J. Nutr.* **131**: 2885-2891
- Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L** (2004) TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J* **39**: 366-380
- Becker A, Kaufmann K, Freialdenhoven A, Vincent C, Li MA, Saedler H, Theissen G** (2002) A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. *Mol Genet Genomics* **266**: 942-950
- Beggs CJ, Kuhn K, Böcker R, Wellmann E** (1987) Phytochrome-induced flavonoid biosynthesis in mustard (*Sinapis alba* L.) cotyledons. Enzymic control and differential regulation of anthocyanin and quercetin formation. *Planta* **172**: 121-126
- Beld M, Martin C, Huits H, Stuitje AR, Gerats AG** (1989) Flavonoid synthesis in *Petunia hybrida*: partial characterization of dihydroflavonol-4-reductase genes. *Plant Mol Biol* **13**: 491-502
- Bogs J, Downey MO, Harvey JS, Ashton AR, Tanner GJ, Robinson SP** (2005) Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. *Plant Physiol* **139**: 652-663
- Christman JW, Blackwell TS, Juurlink BH** (2000) Redox regulation of nuclear factor kappa B: therapeutic potential for attenuating inflammatory responses. *Brain Pathol* **10**: 153-162
- Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L** (2003) Proanthocyanidin-accumulating cells in *Arabidopsis* testa: regulation of differentiation and role in seed development. *Plant Cell* **15**: 2514-2531

- Devic M, Guilleminot J, Debeaujon I, Bechtold N, Bensaude E, Koornneef M, Pelletier G, Delseny M** (1999) The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. *Plant J* **19**: 387-398
- Dixon RA, Paiva NL** (1995) Stress-induced phenylpropanoid metabolism. *Plant Cell* **7**: 1085-1097
- Dixon RA, Xie DY, Sharma SB** (2005) Proanthocyanidins - a final frontier in flavonoid research? *New Phytol* **165**: 9-28
- Engler MB, Engler MM** (2006) The emerging role of flavonoid-rich cocoa and chocolate in cardiovascular health and disease. *Nutr Res* **64**: 109-118
- Fedoroff NV, Furtek DB, Nelson OE** (1984) Cloning of the bronze locus in maize by a simple and generalizable procedure using the transposable controlling element Activator (Ac). *Proc Natl Acad Sci U S A* **81**: 3825-3829
- Feucht W, Treutter D** (1999) The role of flavan-3-ols and proanthocyanidins in plant defense. In Inderjit, KMM Dakshi, CL Foy, eds, *Principles and Practices in Plant Ecology: Allelochemical Interactions*. CRC Press, New York, pp 307-338
- Fisher ND, Hollenberg NK** (2005) Flavanols for cardiovascular health: the science behind the sweetness. *J Hypertens* **23**: 1453-1459
- Fukuchi-Mizutani M, Okuhara H, Fukui Y, Nakao M, Katsumoto Y, Yonekura-Sakakibara K, Kusumi T, Hase T, Tanaka Y** (2003) Biochemical and molecular characterization of a novel UDP-glucose:anthocyanin 3'-O-glucosyltransferase, a key enzyme for blue anthocyanin biosynthesis, from gentian. *Plant Physiol* **132**: 1652-1663
- Gargouri M, Manigand C, Mauge C, Granier T, Langlois d'Estaintot B, Cala O, Pianet I, Bathany K, Chaudiere J, Gallois B** (2009) Structure and epimerase activity of anthocyanidin reductase from *Vitis vinifera*. *Acta Crystallogr D* **65**: 989-1000
- Graham TL** (1991) Flavonoid and isoflavonoid distribution in developing soybean seedling tissues and in seed and root exudates. *Plant Physiol* **95**: 594-603
- Gu L, Kelm M, Hammerstone JF, Beecher G, Cunningham D, Vannozzi S, Prior RL** (2002) Fractionation of polymeric procyanidins from lowbush blueberry and quantification of procyanidins in selected foods with an optimized normal-phase HPLC-MS fluorescent detection method. *J Agric Food Chem* **50**: 4852-4860
- Gu LW, House SE, Wu XL, Ou BX, Prior RL** (2006) Procyanidin and catechin contents and antioxidant capacity of cocoa and chocolate products. *J Agric Food Chem* **54**: 4057-4061
- Hahlbrock K, Scheel D** (1989) Physiology and molecular biology of phenylpropanoid metabolism. *Annu Rev Plant Phys* **40**: 347-369
- Hannum SM, Erdman JW** (2000) Emerging health benefits from cocoa and chocolate. *J Med Food* **3**: 73-75
- Hirotsani M, Kuroda R, Suzuki H, Yoshikawa T** (2000) Cloning and expression of UDP-glucose: flavonoid 7-O-glucosyltransferase from hairy root cultures of *Scutellaria baicalensis*. *Planta* **210**: 1006-1013
- Holton TA, Brugliera F, Tanaka Y** (1993) Cloning and expression of flavonol synthase from *Petunia hybrida*. *Plant J* **4**: 1003-1010
- Holton TA, Cornish EC** (1995) Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7**: 1071-1083

- Johnson CS, Kolevski B, Smyth DR** (2002) *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *Plant Cell* **14**: 1359-1375
- Karim M, McCormick K, Kappagoda CT** (2000) Effects of cocoa extracts on endothelium-dependent relaxation. *J Nutr* **130**: 2105S-2108S
- Keen CL, Holt RR, Polagruto JA, Wang JF, Schmitz HH** (2002) Cocoa flavanols and cardiovascular health. *Phytochem Rev* **1**: 231-240
- Kubo H, Nawa N, Lupsea SA** (2007) *Anthocyaninless1* gene of *Arabidopsis thaliana* encodes a UDP-glucose:flavonoid-3-O-glucosyltransferase. *J Plant Res* **120**: 445-449
- Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M** (2006) Genetics and biochemistry of seed flavonoids. *Annu Rev Plant Biol* **57**: 405-430
- Li J, Ou-Lee TM, Raba R, Amundson RG, Last RL** (1993) *Arabidopsis* flavonoid mutants are hypersensitive to UV-B irradiation. *Plant Cell* **5**: 171-179
- Marles MA, Ray H, Gruber MY** (2003) New perspectives on proanthocyanidin biochemistry and molecular regulation. *Phytochemistry* **64**: 367-383
- Martin C, Prescott A, Mackay S, Bartlett J, Vrijlandt E** (1991) Control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. *Plant J* **1**: 37-49
- Martinez-Florez S, Gutierrez-Fernandez B, Sanchez-Campos S, Gonzalez-Gallego J, Tunon MJ** (2005) Quercetin attenuates nuclear factor-kappaB activation and nitric oxide production in interleukin-1beta-activated rat hepatocytes. *J Nutr* **135**: 1359-1365
- Mato M, Ozeki Y, Itoh Y, Higeta D, Yoshitama K, Teramoto S, Aida R, Ishikura N, Shibata M** (1998) Isolation and characterization of a cDNA clone of UDP-galactose: flavonoid 3-O-galactosyltransferase (UF3GaT) expressed in *Vigna mungo* seedlings. *Plant Cell Physiol* **39**: 1145-1155
- Mellway RD, Tran LT, Prouse MB, Campbell MM, Constabel CP** (2009) The wound-, pathogen-, and ultraviolet B-responsive MYB134 gene encodes an R2R3 MYB transcription factor that regulates proanthocyanidin synthesis in poplar. *Plant Physiol.* **150**: 924-941
- Menssen A, Hohmann S, Martin W, Schnable PS, Peterson PA, Saedler H, Gierl A** (1990) The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the *A2* gene. *Embo J* **9**: 3051-3057
- Moriguchi T, Kita M, Ogawa K, Tomono Y, Endo T, Omura M** (2002) Flavonol synthase gene expression during citrus fruit development. *Physiol Plant* **114**: 251-258
- Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L** (2000) The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in *Arabidopsis* siliques. *Plant Cell* **12**: 1863-1878
- Nesi N, Debeaujon I, Jond C, Stewart AJ, Jenkins GI, Caboche M, Lepiniec L** (2002) The *TRANSPARENT TESTA16* locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. *Plant Cell* **14**: 2463-2479
- Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L** (2001) The *Arabidopsis* *TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**: 2099-2114

- Norman KH, Naomi DLF, Marjorie LM** (2009) Flavanols, the Kuna, cocoa consumption, and nitric oxide. *J. Am. Soc. Hypertens* **3**: 105-112
- O'Reilly C, Shepherd NS, Pereira A, Schwarz-Sommer Z, Bertram I, Robertson DS, Peterson PA, Saedler H** (1985) Molecular cloning of the a1 locus of *Zea mays* using the transposable elements En and Mu1. *Embo J* **4**: 877-882
- Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA** (2007) Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiol* **145**: 601-615
- Paolucci F, Robbins MP, Madeo L, Arcioni S, Martens S, Damiani F** (2007) Ectopic expression of a basic helix-loop-helix gene transactivates parallel pathways of proanthocyanidin biosynthesis. structure, expression analysis, and genetic control of leucoanthocyanidin 4-reductase and anthocyanidin reductase genes in *Lotus corniculatus*. *Plant Physiol* **143**: 504-516
- Pelletier MK, Murrell JR, Shirley BW** (1997) Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in Arabidopsis. Further evidence for differential regulation of "early" and "late" genes. *Plant Physiol* **113**: 1437-1445
- Pourcel L, Routaboul JM, Cheynier V, Lepiniec L, Debeaujon I** (2007) Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* **12**: 29-36
- Pourcel L, Routaboul JM, Kerhoas L, Caboche M, Lepiniec L, Debeaujon I** (2005) *TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell* **17**: 2966-2980
- Rein D, Paglieroni TG, Pearson DA, Wun T, Schmitz HH, Gosselin R, Keen CL** (2000) Cocoa and wine polyphenols modulate platelet activation and function. *J Nutr* **130**: 2120S-2126S
- Revilla E** (1991) Analysis of catechins and proanthocyanidins in grape seeds by HPLC with photodiode array detection. *Chromatographia* **31**: 465
- Sagasser M, Lu GH, Hahlbrock K, Weisshaar B** (2002) *A. thaliana* *TRANSPARENT TESTA 1* is involved in seed coat development and defines the WIP subfamily of plant zinc finger proteins. *Genes Dev* **16**: 138-149
- Saito K, Kobayashi M, Gong Z, Tanaka Y, Yamazaki M** (1999) Direct evidence for anthocyanidin synthase as a 2-oxoglutarate-dependent oxygenase: molecular cloning and functional expression of cDNA from a red form of *Perilla frutescens*. *Plant J* **17**: 181-189
- Saslowky D, Winkel-Shirley B** (2001) Localization of flavonoid enzymes in Arabidopsis roots. *Plant J* **27**: 37-48
- Scalbert A** (1991) Antimicrobial properties of tannins. *Phytochemistry* **30**: 3875-3883
- Schijlen EG, Ric de Vos CH, van Tunen AJ, Bovy AG** (2004) Modification of flavonoid biosynthesis in crop plants. *Phytochemistry* **65**: 2631-2648
- Schwinn KE, M. Davies K, Deroles SC, Markham KR, Miller RM, Bradley JM, Manson DG, Given NK** (1997) Expression of an *Antirrhinum majus* UDP-glucose:flavonoid-3-O-glucosyltransferase transgene alters flavonoid glycosylation and acylation in lisianthus (*Eustoma grandiflorum* Grise.). *Plant Sci* **125**: 53-61
- Shirley BW, Hanley S, Goodman HM** (1992) Effects of ionizing radiation on a plant genome: analysis of two Arabidopsis transparent testa mutations. *Plant Cell* **4**: 333-347

- Stoupi S, Williamson G, Viton F, Barron D, King LJ, Brown JE, Clifford MN** (2010) *In vivo* bioavailability, absorption, excretion, and pharmacokinetics of [14C]procyanidin B2 in male rats. *Drug Metab Dispos* **38**: 287-291
- Stoutjesdijk PA, Sale PW, Larkin PJ** (2001) Possible involvement of condensed tannins in aluminium tolerance of *Lotus pedunculatus*. *Aust J Plant Physiol* **28**: 1063-1074
- Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ, Ashton AR** (2003) Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *J Biol Chem* **278**: 31647-31656
- Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N, Blundell TL, Esch JJ, Marks MD, Gray JC** (1999) The *TRANSPARENT TESTA GLABRA1* locus, which regulates trichome differentiation and anthocyanin biosynthesis in *Arabidopsis*, encodes a WD40 repeat protein. *Plant Cell* **11**: 1337-1350
- Wan Y, Vinson JA, Etherton TD, Proch J, Lazarus SA, Kris-Etherton PM** (2001) Effects of cocoa powder and dark chocolate on LDL oxidative susceptibility and prostaglandin concentrations in humans. *Am J Clin Nutr* **74**: 596-602
- Welle R, Grisebach H** (1989) Phytoalexin synthesis in soybean cells: elicitor induction of reductase involved in biosynthesis of 6'-deoxychalcone. *Arch Biochem Biophys* **272**: 97-102
- Winkel-Shirley B** (2002) Biosynthesis of flavonoids and effects of stress. *Curr Opin Plant Biol* **5**: 218-223
- Xie DY, Sharma SB, Paiva NL, Ferreira D, Dixon RA** (2003) Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis. *Science* **299**: 396-399
- Zuker A, Tzfira T, Ben-Meir H, Ovadis M, Shklarman E, Itzhaki H, Forkmann G, Martens S, Neta-Sharir I, Weiss D, Vainstein A** (2002) Modification of flower color and fragrance by antisense suppression of the flavanone 3-hydroxylase gene. *Mol. Breed.* **9**: 33-41

## **CHAPTER 2: MOLECULAR BIOLOGY OF PROANTHOCYANIDIN SYNTHESIS IN CACAO: CHARACTERIZATION OF GENES ENCODING ANTHOCYANIDIN REDUCTASE, ANTHOCYANIDIN SYNTHASE AND LEUCOANTHOCYANIDIN REDUCTASE**

### **2.1 Introduction**

Flavonoids are a diverse group of plant secondary metabolites with various biological functions that play important roles during plant development. They are involved in plant defenses against insects, pathogens and microbes, in absorption of free radicals and UV light, and in attraction of beneficial symbionts and pollinators (Martin et al., 1991; Saslowsky and Winkel-Shirley, 2001; Winkel-Shirley, 2002; Dixon et al., 2005). Proanthocyanidins (PAs, also known as condensed tannins) are components of metabolites synthesized through the general flavonoid pathway. Their main known function is to provide protection against microbial pathogens and invasions of insects and herbivores through the following mechanisms (Dixon et al., 2005): first, PAs have metal depletion activities that result in severe limitation of bacterial growth; second, they can cause irreversible protein precipitation, which is responsible for the astringent taste that repels herbivores; third, PAs can be oxidized to quinones which not only are powerful antibiotics themselves, but also can initiate cross-linking of cell walls to form a physical barrier (Scalbert, 1991; Feucht and Treutter, 1999; Pourcel et al., 2005; Pourcel et al., 2007). Furthermore, PAs, especially cacao PAs, have recently proven to be beneficial to humans by improving cardiovascular health through activation of nitric oxide synthase, by providing cancer chemopreventative effects, and also through neuroprotective activities (Ramassamy, 2006; Aron and Kennedy, 2008; Norman et al., 2009).

The PA biosynthesis pathway that leads to the production of flavan-3-ols ((+)-catechin and (-)-epicatechin), the building blocks of proanthocyanidins, has been well studied in the model plant species maize and Arabidopsis as summarized in Figure 1-2 (Winkel-Shirley, 2001; Lepiniec et al., 2006). Biosynthesis of flavan-3-ols via the anthocyanin pathway involves three principal enzymes: leucoanthocyanidin reductase

(LAR), anthocyanidin synthase (ANS), and anthocyanidin reductase (ANR; in *Arabidopsis*, the product of *BANYULS* gene). The synthesis of PAs and anthocyanins shares common steps leading to flavan-3,4-diols (such as leucoanthocyanidin), which can be converted to catechin (2,3-*trans*-flavan-3-ol) by LAR (Tanner et al., 2003), or to cyanidin by ANS (Saito et al., 1999; Abrahams et al., 2003). Cyanidin then either serves as the substrate for the synthesis of epicatechin (2,3-*cis*-flavan-3-ols) by ANR (Xie et al., 2003), or can otherwise be converted to anthocyanidin by glycosylation (Schijlen et al., 2004). Both catechin and epicatechin act as the initiators for PA polymerization, with intermediates derived from leucoanthocyanidin, catechin or epicatechin added sequentially as extension units (Dixon et al., 2005). However, the details of the polymerization process are unclear and it is not known whether this is a spontaneous or an enzyme-catalyzed reaction. Recent work has identified two new enzymes downstream of flavan-3-ols that are involved in key steps of PA polymer biosynthesis, a laccase-like polyphenol oxidase from *Arabidopsis* (TT10; Pourcel et al., 2005) and an epicatechin 3'-O-glucosyltransferase in *Medicago truncatula* (Pang et al., 2008).

*ANS* and *ANR* genes have been biochemically and genetically characterized in *Arabidopsis* (Pelletier et al., 1997; Devic et al., 1999; Abrahams et al., 2003; Xie et al., 2003), *Medicago* (Xie et al., 2003; Pang et al., 2007) and grape (Bogs et al., 2005). The *Arabidopsis ans* (*ldox*) mutant exhibits a deficiency in both anthocyanin accumulation in hypocotyls and PA deposition in seeds that results in a transparent testa phenotype (Abrahams et al., 2003). Seeds of the *Arabidopsis banyuls* (*anr*) mutant exhibit a lack of PAs and a hyper-accumulation of anthocyanins, resulting in a dark red color reminiscent of the famous Banyuls wine produced in southern France (Devic et al., 1999). Over-expression of *ANR* genes from *Medicago* and grape in tobacco results in a loss of anthocyanin pigments in flower petals and elevated levels of PAs (Xie et al., 2003; Bogs et al., 2005). Antisense down-regulation of *ANS* in *Medicago* results in reduction of both anthocyanins in leaves and PAs in seeds (Pang et al., 2007).

*LAR* genes have been isolated from various plant species including *Desmodium uncinatum* (Tanner et al., 2003), *Vitis vinifera* (Bogs et al., 2005), *Lotus corniculatus* (Paolocci et al., 2007) and *Medicago truncatula* (Pang et al., 2007) and their corresponding

protein functions have been characterized by in vitro recombinant enzyme assays. However, the genetic evidence for LAR function is rather indirect and less convincing than it is for ANR and ANS, as discussed by Pang et al. (2007). It appears that the genomic sequence of *Arabidopsis thaliana* does not contain an intact LAR orthologue, and correspondingly, catechin is not detected in Arabidopsis seed extracts (Abrahams et al., 2003; Tanner et al., 2003; Lepiniec et al., 2006). Intriguingly, grape has 2 LAR genes (*VvLAR1* and *VvLAR2*) with different expression patterns: *VvLAR1* is co-expressed with *VvANR* in flowers and early developmental stage berries; whereas *VvLAR2* is co-regulated with *VvANR* via the *VvMYBPA1* transcription factor in grape skins and seeds, in which the expression level of *VvLAR1* is not appreciable (Bogs et al., 2005; 2007). However, although Bogs et al. (2005; 2007) provided the biochemical functional analysis of *VvLAR1* protein and also promoter activity analysis of only *VvLAR1*, *VvLAR2* is the gene whose expression pattern coincides with the PA synthesis regulator *VvMYBPA1*.

In *Arabidopsis* and *Medicago*, PA accumulation and gene expression is quantitatively and spatially limited to seed coats, making it remarkably difficult for biochemical analysis (Lepiniec et al., 2006; Pang et al., 2007). In contrast, *Theobroma cacao* (*Tc*) produces significant amounts of PAs in various tissues including leaves and beans -- up to 12% of dry weight in leaves (Chaves and Gianfagna, 2007) and approximately 10% in mature beans (Niemenak et al., 2006). Furthermore, large amounts of catechin and epicatechin monomers as well as their related polymers of different length have been detected in cocoa powder (Gu et al., 2002; Chaves and Gianfagna, 2007).

It is also important to understand the stages of development of cacao fruits to provide a biological context for the genetic and biochemical studies that will be presented in this thesis. The development of cacao and anthocyanin synthesis has been described previously. The development of cacao fruits can be divided into three phases. Following pollination and fertilization, the first phase of fruit development is initiated and fruit begins to expand slowly at a rate of about 30-40 cm<sup>3</sup>/ week (Cheesman, 1927). This phase lasts 6-7 weeks until the first division of the fertilized egg, which initiates the second phase of pod development. At the second phase, fruits expand more rapidly at a rate of about 110-130 cm<sup>3</sup>/ week, and embryos enlarge but remain unpigmented till they reach the length of



ovules at about 14-16 weeks after pollination (Lehrian and Keeney, 1980; Alemanno et al., 1997). When the fruits are 14-16 weeks old, the pericarp begins to change color from green to orange (in Scavina 6), denoting onset of the third phase, ripening. Ripe pod color varies from bright red, purple, green, yellow and multi-colored patterns, dependant on genotype. During the third phase, the increase in the fruits external dimensions gradually slows and finally ceases. The seeds begin to solidify and their dry weight increases rapidly at a rate of about 20-40 mg/day. Seed length remains constant as they continue to accumulate anthocyanins and gradually darken until maturity at about 20 weeks after pollination (Cheesman, 1927; Lehrian and Keeney, 1980; Wright et al., 1982; Alemanno et al., 1997).

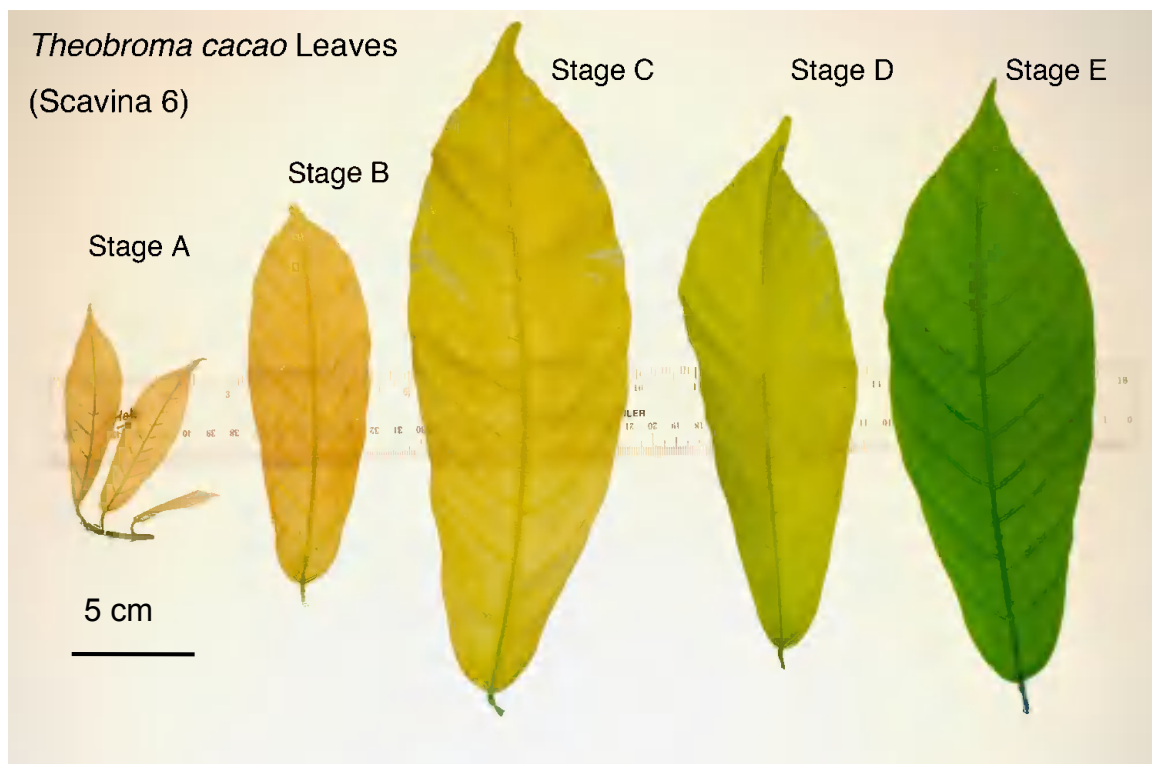
Considering the wide range of health benefits known for PAs and its significance for plant resistance, we targeted this pathway for molecular-genetic analysis. This chapter describes the isolation and expression of the *TcANR*, *TcANS*, and *TcLAR* genes encoding the key enzymes in PA biosynthesis. We measured PA content in different cacao tissues and at different developmental stages and performed functional characterization of the *TcANR*, *TcANS*, and *TcLAR* gene products through *in vivo* tests. To our knowledge, this work represents the first published analysis of the function of an LAR enzyme *in vivo*, providing direct genetic evidence of *LAR* gene function. The results presented here provide background and genetic tools that will be useful in the development of new cacao varieties with altered PA profiles.

## **2.2 Materials and Methods**

### **2.2.1 Plant Material**

Two *Theobroma cacao* varieties, Scavina 6 and Amelonado, were used for this study. Cacao plants were grown in a greenhouse as previous described (Maximova et al., 2003). Leaf tissues were collected from Scavina 6 plants at five stages of development (A-E) as shown in Figure 2-1. Stage A leaves are newly emerged and are 5 to 10 cm long; stage B leaves are larger, soft, red and translucent, 10-15cm long; Stage C leaves are green and remain soft; Stage D leaves are at an early stage of lignification; Stage E leaves are fully

lignified and mature. Stage A and B leaves were pooled together because of the limited amount of Stage A leaves. Flowers were also collected from Scavina 6 plants at two developmental stages: opened flowers and unopened flowers. Unopened flowers samples were a mixture of all stages of flower buds had not yet bloomed into full-size flowers. Cacao pods were obtained by hand pollinating Amelonado (a self-compatible variety). Developing fruits were collected at weekly intervals from 2 weeks after pollination (WAP) to 6 WAP, and then biweekly until 20 WAP when pods were mature (Cheesman, 1927). Whole pods were sampled until 8 WAP, after which pods were bisected, and fertilized ovules and pod exocarps collected separately. Exocarp samples represent the outer 1-2 mm layer of the fruit and were obtained using a fruit peeler. All samples were frozen in liquid nitrogen upon collection and stored at -80°C until biochemical analysis.



**Figure 2-1.** Developmental stages of cacao leaf (Scavina 6). Photographed with transillumination to demonstrate the translucent nature of stage A-C where the ruler can be seen through the leaves, but not stages D and E. Top scale on ruler inches, bottom mm.

Transgenic and wild-type tobacco plants (*Nicotiana tabacum* var. Samsun provided by Wayne Curtis, Department of Chemical Engineering, The Pennsylvania State University) were grown in a greenhouse under the same condition as cacao plants. Arabidopsis plants (*Arabidopsis thaliana*) were grown in soil at 22°C, 50% humidity and a 16 h/8 h light/dark photoperiod in a growth chamber (Conviron, Pembina, ND, USA). Plants grown aseptically were plated on MS medium (Murashige and Skoog, 1962) with 2% (w/v) sucrose solidified with 0.6% (w/v) agar. Arabidopsis ecotype Columbia (Col-0) plants were used as the wild type. T-DNA insertion mutants *ban* (SALK\_040250) and *ldox* (SALK\_028793) were obtained from The Arabidopsis Biological Resource Center (Columbus, OH, USA).

### 2.2.2 Nucleic Acid Purification and cDNA Synthesis

Total RNA from leaves of *Theobroma cacao* (Scavina 6) was isolated using a modified cetyl trimethyl ammonium bromide (CTAB) extraction method (Zeng, 2002). RNA was further purified and concentrated using RNeasy columns (Qiagen, Valencia, CA, USA). The quality of RNA was verified by observing absorbance ratios of A260/A280 (1.8 to 2.0) and A260/A230 (1.8 to 2.2) and by separating 200 ng RNA samples on 0.8% agarose gels to examine intact ribosomal bands. First strand cDNA was synthesized using the SMART RACE cDNA amplification kit (Clontech, Mountain View, CA, USA).

### 2.2.3 Isolation of cDNA and Genomic Clones from *Theobroma cacao*

The putative expressed sequence tag (EST) sequences of cacao anthocyanidin reductase (*TcANR*), anthocyanidin synthase (*TcANS*) and leucoanthocyanidin reductase (*TcLAR*) genes were obtained by searching the *Theobroma cacao* EST database (<http://esttik.cirad.fr/>) (Argout et al., 2008) using the tBLASTn program (Altschul et al., 1990). The query sequences used were the protein sequences of BANYULS and LDOX from *Arabidopsis thaliana* and DuLAR from *Desmodium uncinatum* respectively (Accession numbers: NP\_176365, Q96323 and CAD79341). Based on the sequences of the EST contigs from the ESTtik database (EST Treatment and Investigation Kit; <http://esttik.cirad.fr/>), PCR primers were designed to amplify the entire coding sequences of each gene: ANR\_F (5'-

AGCCATGGCCAGCCAGACCGTAGG-3') and ANR\_R (5'-*GCGGCCGCTCA*CTTGAGCAGCCCCTTAGC-3'), ANS\_F (5'-*CCATGGT*GACTTCAATGGCCCCCAG-3') and ANS\_R (5'-*GCGGCCGCTCA*ATTAGACAGGCCATC-3') and LAR\_F (*CCATGGATATG*AAATCAACAAACATGAATGGTTC) AND LAR\_R (*GCGGCCGCTCA*TGTGCATATCGCAGTG). To facilitate the subsequent cloning into binary T-DNA vectors, *NcoI* sites were added to the 5' end of each start codon and *NotI* sites were added to the 3' end of each stop codon (sites are shown in italics and the start or stop codons are underlined). The coding sequences were amplified from cacao cDNA prepared from young leaves (genotype Scavina 6) with the Advantage cDNA PCR Kit (Clontech, Mountain View, CA, USA) using these primers. PCR reactions were carried out in a total volume of 20 µL at 94°C for 5 min; 5 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 1 min; then another 23 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 1 min; followed by a final extension at 72°C for 5 min. PCR products were gel purified and cloned into the pGEM-T easy vector (Promega, Madison, WI, USA). The correct open reading frames (ORFs) of each of the resulting constructs (pGEMT-TcANR, pGEMT-TcANS and pGEMT-TcLAR) were confirmed by DNA sequencing.

The DNA sequences of the *TcANR*, *TcANS* and *TcLAR* genes were obtained by isolation and sequencing of genomic clones. Briefly, 2 high-density filters arrayed with 18,432 colonies of *Theobroma cacao* (genotype LCT-EEN 37) bacterial artificial chromosome (BAC) clones on each constructed by The Clemson University Genomics Institute (CUGI, <https://www.genome.clemson.edu/>) were screened by hybridization to full-length cDNA of each gene labeled with P<sup>32</sup> using the MEGA Labeling Kit (GE Healthcare, Piscataway NJ). DNA hybridizations were carried out at 65°C in 1 mM ethylenediaminetetraacetic acid (EDTA), 7% sodium dodecyl sulfate (SDS), 0.5M sodium phosphate (pH7.2) for 16-18 h. Filters were washed twice at 65°C in 1 mM EDTA, 1% SDS, 40 mM Na<sub>2</sub>HPO<sub>4</sub> for 20 min, twice at 65°C in 1.5x sodium chloride/sodium citrate (SSC), 0.1% SDS for 20 min and twice at 65°C in 0.5x SSC, 0.1% SDS for 20 min. Two or more BAC clones were identified for each gene and confirmed by PCR using plasmid DNA from individual colonies and gene specific primers. High purify plasmid DNA from individual BAC clones was then isolated using a NucleoBond BAC 100 kit (Macherey-Nagel Inc., Bethlehem, PA, USA) and both strands of

DNA were sequenced using primers designed from the cDNAs and genomic DNA. DNA sequencing results were analyzed and assembled using Vector NTI software (Invitrogen, San Diego, CA), application Contig Assembly. The DNA sequence of each gene was then compared to the corresponding coding sequence by using the BLAST2 online tool ([www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi](http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi)) to obtain exon and intron locations for the gene organization analyses.

## 2.2.4 Phylogenetic Analysis

Deduced protein sequences of all Arabidopsis isoflavone reductase (*IFR*)-like genes were retrieved from The Arabidopsis Information Resource (TAIR) database (<http://www.arabidopsis.org/>) by querying the TAIR protein database with the *Desmodium* LAR protein sequence using the WU-BLAST2 (BLASTP) program. Sources of protein sequences from other species were summarized in Table 2-1.

**Table 2-1:** Sources of LAR, DFR, ANR and ANS proteins used in Phylogenetic analysis

Enzyme	Species	Protein	Accession number	Reference
LAR	<i>Vitis vinifera</i>	VvLAR	CAI26309	(Bogs et al., 2005)
	<i>Desmodium uncinatum</i>	DuLAR	CAD79341	(Tanner et al., 2003)
	<i>Medicago truncatula</i>	MtLAR	CAI56327	(Pang et al., 2007)
	<i>Lotus corniculatus</i>	LcLAR1-1, LcLAR1-2	ABC71324, ABC71325	(Paolocci et al., 2007)
DFR	<i>Zea mays</i>	ZmA1	CAA28734	(Schwarz-Sommer et al., 1987)
	<i>Arabidopsis thaliana</i>	AtDFR	NP_199094	(Shirley et al., 1992)
	<i>Medicago truncatula</i>	MtDFR	AAR27014	(Xie et al., 2004)
ANR	<i>Arabidopsis thaliana</i>	BAN	NP_176365	(Xie et al., 2003)
	<i>Medicago truncatula</i>	MtANR	AAN77735	(Xie et al., 2003)
	<i>Vitis vinifera</i>	VvANR	CAD91911	(Bogs et al., 2005)
	<i>Camellia sinensis</i>	CsANR	AAT68773	(Singh et al., 2009)
	<i>Lotus corniculatus</i>	LcANR,	ABC71337	(Paolocci et al., 2007)
ANS	<i>Perilla frutescens</i>	PfANS,	O04274	(Saito et al., 1999)
	<i>Arabidopsis thaliana</i>	AtANS	Q96323	(Abrahams et al., 2003)
	<i>Zea mays</i>	ZmA2	CAA39022	(Menssen et al., 1990)
	<i>Medicago truncatula</i>	MtANS	ABU40983	(Pang et al., 2007)

Multiple sequence alignment of proteins was performed using the ClustalX algorithm (Thompson et al., 1997) with default parameter settings (gap opening penalty: 10, gap extension penalty: 0.2, delay divergent cutoff: 30%, protein weight matrix: Gonnet series ) and this alignment was used to construct the phylogenetic tree using the neighbor-joining method in the MEGA package (Kumar et al., 2004). Five hundred bootstrapped datasets were used to estimate the confidence of each tree clade.

### **2.2.5 Southern Blot Analysis**

Cacao genomic DNA was isolated from mature cacao leaves of Scavina 6 using a modified CTAB method as previous described (Michiels et al., 2003). Isolated DNA was purified by CsCl density gradient purification (Ausubel et al., 2007). Five µg of DNA was digested overnight independently with the following endonucleases: *EcoRI*, *HindIII*, *NcoI* and *SmaI* (New England Biolabs, Ipswich, MA, USA). DNA was size fractionated on 1% agarose gels and transferred to nylon membranes (Hybond-N+, Amersham Biosciences, Piscataway, NJ, USA) which were dried overnight and then fixed via UV crosslinking (120 mJ) in a GS Gene Linker UV Chamber (Bio-Rad, Hercules, CA, USA). Membranes were hybridized and washed using the standard protocol as described by Ausubel et al. (2007) in buffer containing 0.5 M sodium phosphate (pH 7.2) at either 55°C or 65°C for high and low stringency conditions respectively. Probes containing the ORF regions of *TcANR*, *TcANS* and *TcLAR* were labeled with <sup>32</sup>P-dCTP using Megaprime DNA labeling kit (Amersham Bioscience, Piscataway, NJ, USA) according to the manufacturer's instructions. Hybridized membranes were imaged using a phosphorimager with overnight exposure (Molecular Dynamics, Sunnyvale, CA, USA).

### **2.2.6 Proanthocyanidin (PA) and Anthocyanin Quantification**

For quantitative analysis of anthocyanin levels in transgenic tobacco flowers, fresh petals (0.3-0.5 g fresh weight) from three flowers were immersed in 5 mL ethanol: 6 M HCl (1:1) and incubated at 4°C overnight. The extract was transferred to a new tube and the petals were re-extracted. Absorbance of the pooled extract solution was then measured at 526 nm and total anthocyanin levels were calculated using a standard molar absorbance

curve prepared using cyanidin-3-glucoside (Sigma-Aldrich, MO, USA).

To extract PAs from cacao and tobacco tissues, 0.3 to 0.5 g of frozen tissues were ground into a fine powder in liquid nitrogen and extracted with 5 mL of extraction solution (70% acetone: 29.5% water: 0.5% acetic acid) by vortexing for 5 seconds followed by water bath sonication for 15 min using a bench top ultrasonic cleaner (Model 2510, Branson, Danbury, CT, USA). After sonication, samples were vortexed again and centrifuged at 2500 g for 10 min. The supernatant was transferred to a new tube and the pellet was re-extracted twice as above. Pooled supernatants were extracted twice with hexane to remove fat and chlorophyll and then filtered through a 0.45 µm polytetrafluoroethylene (PTFE) syringe filter (Millipore, Billerica, MA, USA). To quantify PA levels, 50 µL aliquots of samples were mixed with 200 µL of dimethylaminocinnamaldehyde (DMACA; Sigma-Aldrich, MO, USA) reagent (0.1% DMACA, 90% reagent-grade ethanol, 10% HCl) in 96-well microtiter plates. Absorption was measured at 640 nm at one-minute intervals for 20 min to get the highest readings. Triple technical replicates were performed to obtain mean values. The total PA levels were calculated using the standard molar absorbance curve prepared using procyanidin B2 (Indofine, NJ, USA). For each experiment, at least 3 biological replicates were performed.

For quantitative analysis of insoluble PAs from cacao tissues, the residues from soluble PA extractions were air dried in an exhaust hood for two days, weighed, and 5 mL butanol-HCl reagent (95% butan-1-ol: 5% concentrated HCl) was added and the mixture was sonicated for one hour followed by centrifugation at 2500 g for 10 min. An aliquot of clear supernatant was diluted 40-fold in butanol-HCl reagent and absorbance was measured at 550 nm to determine the amount of background absorption. The samples were then boiled for 1 hour with vortexing every 20 min, cooled to room temperature and centrifuged again at 2500 g for 10 min. The supernatant from boiled sample was diluted 40-fold in butanol-HCl reagent and absorbance was measured at 550 nm. The values were normalized by subtraction of the background absorbance and the PA levels were calculated as cyanidin equivalents using cyanidin-3-glucoside (Sigma-Aldrich, MO, USA) as standards.

To visualize the presence of PAs in Arabidopsis seeds, dry seeds were immersed for 2 days in the 0.1% DMACA reagent described above and then washed 3 times with 70% ethanol as described previously (Abrahams et al., 2002).

### **2.2.7 Genetic Transformation of Tobacco and Arabidopsis**

The coding sequences of *TcANS*, *TcANR* and *TcLAR* were excised from the cloning vector (pGEM-T easy) (Promega, Madison, WI, USA) with *NcoI* and *NotI* restriction enzymes and cloned into the pE2113-EGFP (Maximova et al., 2003) intermediate vector to replace the original EGFP coding sequence. As a result, the cacao gene coding sequences are located immediately downstream of the very strong E12- $\Omega$  modified CaMV35S promoter and upstream of the CaMV35S-terminator. The over-expression cassettes were then introduced into the pCAMBIA-1300 binary vector (CAMBIA, Canberra, Australia). All binary transformation constructs were introduced into *Agrobacterium tumefaciens* strain AGL1 (Lazo et al., 1991) by electroporation as described previously (Lin, 1994).

Tobacco leaf disc transformation was performed as previously described (Fisher and Gultinan, 1995) and transgenic shoots were regenerated on MSs (MS shooting) media supplemented with 25 mg/L hygromycin. Only one shoot was selected from each explant to ensure independent transformants. After rooting for 2 weeks in MSr (MS rooting) media supplemented with 25 mg/L hygromycin, hygromycin-resistant plantlets were transferred to soil and grown in a greenhouse as described in section 2.2.1.

Arabidopsis transformation was carried out using the floral dip method (Clough and Bent, 1998), and T<sub>1</sub> transgenic plants were selected on MS media supplemented with 2% sucrose, 0.65% agar and 25 mg/L hygromycin. Hygromycin-resistant T<sub>1</sub> seedlings were transferred to soil 7 days after germination and grown in a growth chamber as described in section 2.2.1.

### **2.2.8 Expression Analysis of *TcANS*, *TcANR* and *TcLAR***

Total RNA from leaves, flowers, whole pods, pod exocarp and ovules of *Theobroma cacao* (Scavina 6 and Amelonado) was isolated as described in section 2.2.2. Total RNA from young leaves of transgenic and wild-type tobacco plants as well as Arabidopsis plants was isolated using the RNeasy Plant mini kit (Qiagen, Valencia, CA, USA). cDNA was synthesized from 1  $\mu$ g of total RNA in a total volume of 20  $\mu$ L using M-MuLV Reverse



Transcriptase (NEB, Ipswich, MA, USA) according to the supplier's protocols, and 2  $\mu$ L were used in the subsequent reverse transcription-PCR (RT-PCR) reactions.

Semi-quantitative RT PCR was performed to measure gene expression levels. The primers for RT-PCR were designed to amplify across at least one intron giving products of approximately 500 bp from cDNA and 700 bp to 1500 bp from genomic DNA. These primer sets were used to check all cDNAs for genomic DNA contamination. The primers used for *TcANS* were TcANSRT\_F (5'-ACCTTGTTAACCATGGGATCTCGG-3') and TcANSRT\_R (5'-GACGGTGTCAACCAATGTGCATGAT-3'); the primers used for *TcANR* were TcANR\_F (5'-TGCTTGAGAAGGGCTACGCTGTTA-3') and TcANR\_R (5'-AAAGATGTGGCAAGGCCAATGCTG-3'); the primers used for *TcLAR* were TcLAR\_F (5'-AATTCCATTGCAGCTTGGCCCTAC-3') and TcLAR\_R (5'-GGCTTGCTCACTGCTTTGGCATT-3'). *TcActin* was used as an internal standard for cacao gene expression using primer set Tc46RT\_F (5'-AGCTGAGAGATTCCGTTGTCCAGA-3') and Tc46RT\_R (5'-CCCACATCAACCAGACTTTGAGTTC-3'). *AtUbi* and *NtrRNA* was chosen as constitutive expression controls for Arabidopsis (ubiquitin) and tobacco (rRNA) respectively with primer pairs AtUbi\_F (5'-ACCGCAAGACCATCACTCT-3') and AtUbi\_R (5'-AGGCCTCAACTGGTTGCTGT-3') (Zhang et al., 2006), and NtrRNA\_F (5'-AGGAATTGACGGAAGGGCA-3') and NtrRNA\_R (5'-GTGCGGCCAGAACATCTAAG-3') (Levy et al., 2004).

The number of PCR cycles was optimized between 20 and 32 to select a cycle number such that amplification is in the linear range; 28 cycles were chosen for all the RT-PCR reactions. The PCR reaction was carried out in a total volume of 20  $\mu$ L at 94°C for 5 min; 28 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 45 sec; followed by a final extension at 72°C for 5 min. The PCR products were visualized on 1% agarose gels stained with ethidium bromide (EtBr) and documented using Molecular Imager Gel Doc XR+ System equipped with a 16-bit CCD camera (Bio-Rad Laboratories, Hercules, CA) and bands were quantified using Quantity One 1-D Analysis Software (Bio-Rad Laboratories, Hercules, CA).

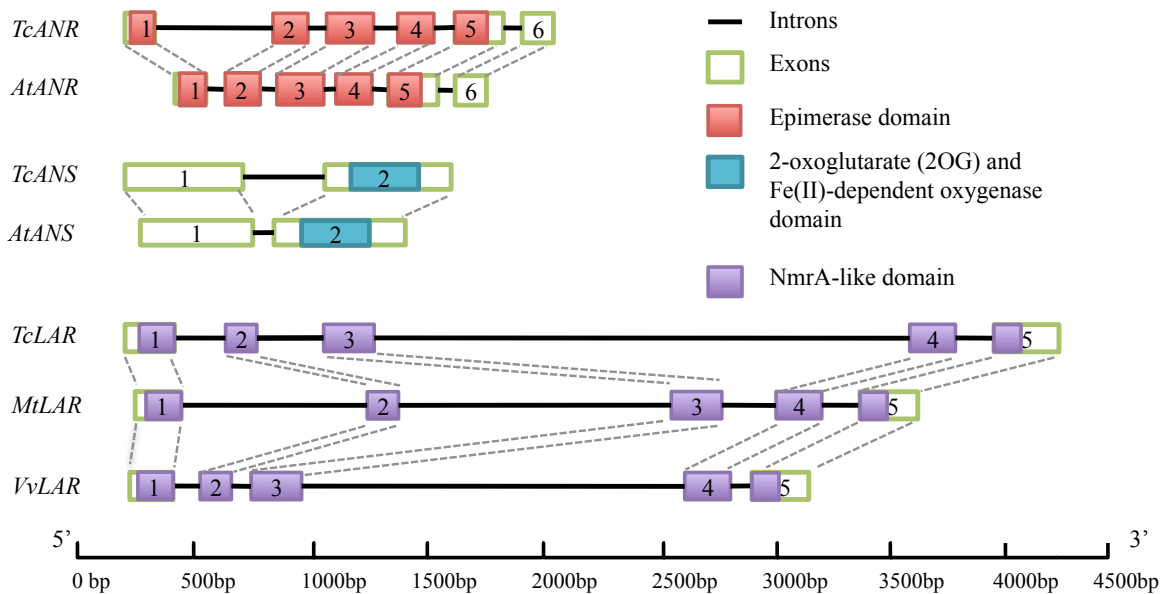
## 2.3 Results

### 2.3.1 Cloning and Sequence Analysis of *TcANR*, *TcANS* and *TcLAR* Genes

Putative *TcANR*, *TcANS* and *TcLAR* cDNA sequences were identified in a collection of *Theobroma cacao* expressed sequence tags (ESTs) (Argout et al., 2008) by querying the cacao ESTtik database (<http://esttik.cirad.fr/>) with protein sequences of Arabidopsis *BANYULS* (accession no. NP\_176365), Arabidopsis *LDOX* (accession no. Q96323) and *Desmodium* LAR (accession no. CAD79341). ESTs similar to each gene were assembled into contigs to determine sequences of full-length open reading frames (ORF) by alignment with cDNAs of homologous genes from other species and predictions from the ORF Finder program ([www.ncbi.nlm.nih.gov/projects/gorf/](http://www.ncbi.nlm.nih.gov/projects/gorf/)). Full-length cDNAs of each gene were amplified by RT-PCR using RNA isolated from young leaves of cacao (Scavina 6) in which PAs are actively synthesized and accumulated (Chaves and Gianfagna, 2007). The *TcANR* cDNA contained a 1,008-bp open reading frame (ORF) encoding a protein of 336 amino acids that showed a 63% identity with the Arabidopsis *BANYULS* gene at the amino acid level. The *TcANS* cDNA contains an ORF of 1,062-bp, which encodes a protein of 354 amino acids. Sequence comparison with the Arabidopsis *LDOX* gene showed an 82% identity at the amino acid level. The *TcLAR* cDNA revealed a 1,083-bp ORF that encodes a protein of 361 amino acids. *TcLAR* showed 61% amino acid identity with the *Desmodium* LAR protein (see sequence comparison in Appendix B, Figure S-1, 2, 3).

To isolate genomic fragments containing the *ANS*, *ANR* and *LAR* genes, the cDNAs were used to screen a cacao BAC library by hybridization. A portion of each hybridizing BAC clone was sequenced using primers designed from the corresponding cDNA. The genomic structure of each gene was established by alignment with its cDNA sequence (Figure 2-2). The genomic organization of the *ANS* and *ANR* genes from Arabidopsis (AT4G22880 and AT1G61720) and *LAR* from *Medicago* and grape (BN000703 and NC\_012007, region 2619277 to 2622652) were also retrieved and compared with the corresponding cacao genes. The position and length of exons, introns and flanking regions are summarized in Table 2-2. The transcribed region of *TcANR* consists of 6 exons and 5

introns distributed over 2,005-bp. The transcribed region of *TcANS* is shorter, having only 1,418-bp and consisting of 2 exons and 1 intron. Their genomic organizations are nearly identical to the corresponding Arabidopsis genes, which have the same exon and intron numbers of similar lengths. Since *LAR* does not have an orthologue in Arabidopsis, we compared the genomic organization of *TcLAR* with those of *MtLAR* and *VvLAR*. The genomic organization of the transcribed region of *TcLAR* shares similarity with those of both *MtLAR* and *VvLAR*, consisting of 5 exons and 4 introns. As observed with the *ANR* gene, the middle three exons (exon 2, 3 and 4) of *LAR* from all three species have identical lengths. *TcLAR* exhibits an extremely long third intron of 2,338 bp; similarly, *VvLAR* also features a long third intron of 1,661 bp, while *MtLAR* contained two long introns (intron 1 and intron 2) that are 812-bp and 1,178-bp respectively. The genomic sequences and complete coding sequences of *TcANR*, *TcANS* and *TcLAR* are listed in Appendix A. The conservation of basic gene structure and sequence between these divergent species suggests their ancient origins and strong evolutionary pressures constraining the evolution of these genes.



**Figure 2-2:** Structures of cacao *TcANR*, *TcANS* and *TcLAR* genes with predicted protein domains, and comparison with their orthologues from Arabidopsis (*AtANR* and *AtANS*), *Medicago* (*MtLAR*) and *Vitis* (*VvLAR*). Green boxes represent exons and lines introns. The scale at the bottom indicates lengths in base pairs.

**Table 2-2:** Structures of *ANR*, *ANS* and *LAR* genes.**(A) Structures of *TcANR* and *AtANR* genes**

	<i>TcANR</i>		<i>AtANR</i>	
	Position	Length	Position	Length
5' flanking	1-2162	2161	1-43	43
Exon1	2163- 2289	127	44-176	133
Intron1	2290 - 2850	561	177-257	80
Exon2	2851 - 3017	<u>167</u>	258-424	<u>167</u>
Intron2	3018 - 3115	98	425-510	85
Exon3	3116 - 3316	<u>201</u>	511-711	<u>201</u>
Intron3	3317 - 3439	123	712-785	73
Exon4	3440 - 3599	<u>160</u>	786-945	<u>160</u>
Intron4	3600 - 3714	115	946-1019	73
Exon5	3715 - 3928	<u>214</u>	1020-1233	<u>214</u>
Intron5	3929 - 4025	97	1234-1327	93
Exon6	4026 - 4167	142	1328-1475	148
3' flanking	4168 - 4940	772	1476-1633	158

**(B) Structures of *TcANS* and *AtANS* genes**

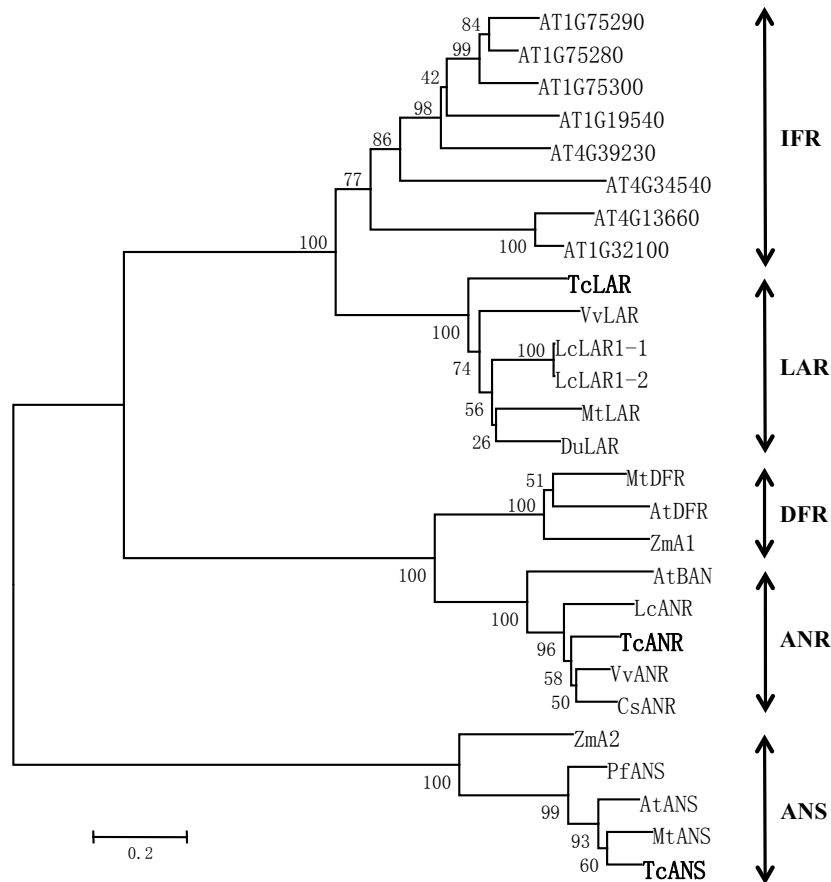
	<i>TcANS</i>		<i>AtANS</i>	
	Position	Length	Position	Length
5' flanking	1-1932	1932	1-150	150
Exon1	1933-2435	503	151-647	497
Intron1	2436-2788	353	648-733	86
Exon2	2789-3350	562	732-1305	574
3' flanking	3351-3940	590	1306-1442	137

**(C) Structures of *TcLAR*, *MtLAR* and *VvLAR* genes**

	<i>TcLAR</i>		<i>MtLAR</i>		<i>VvLAR</i>	
	Position	Length	Position	Length	Position	Length
5' flanking	1-2183	2183	1-620	620	1-96	96
Exon1	2184-2404	221	621-827	207	97-294	198
Intron1	2405-2623	219	828-1639	812	295-404	110
Exon2	2624-2764	<u>141</u>	1640-1780	<u>141</u>	405-545	<u>141</u>
Intron2	2765-3047	283	1781-2958	1178	546-632	87
Exon3	3048-3285	<u>238</u>	2959-3196	<u>238</u>	633-870	<u>238</u>
Intron3	3286-5623	2338	3197-3420	224	871-2531	1661
Exon4	5624-5827	<u>204</u>	3421-3624	<u>204</u>	2532-2735	<u>204</u>
Intron4	5828-5995	168	3625-3793	169	2736-2816	81
Exon5	5996-6273	278	3794-4053	260	2817-3076	260
3' flanking	6274-6499	226	4053-4447	395	3077-3367	291

<sup>a</sup> Identical exon lengths are underlined

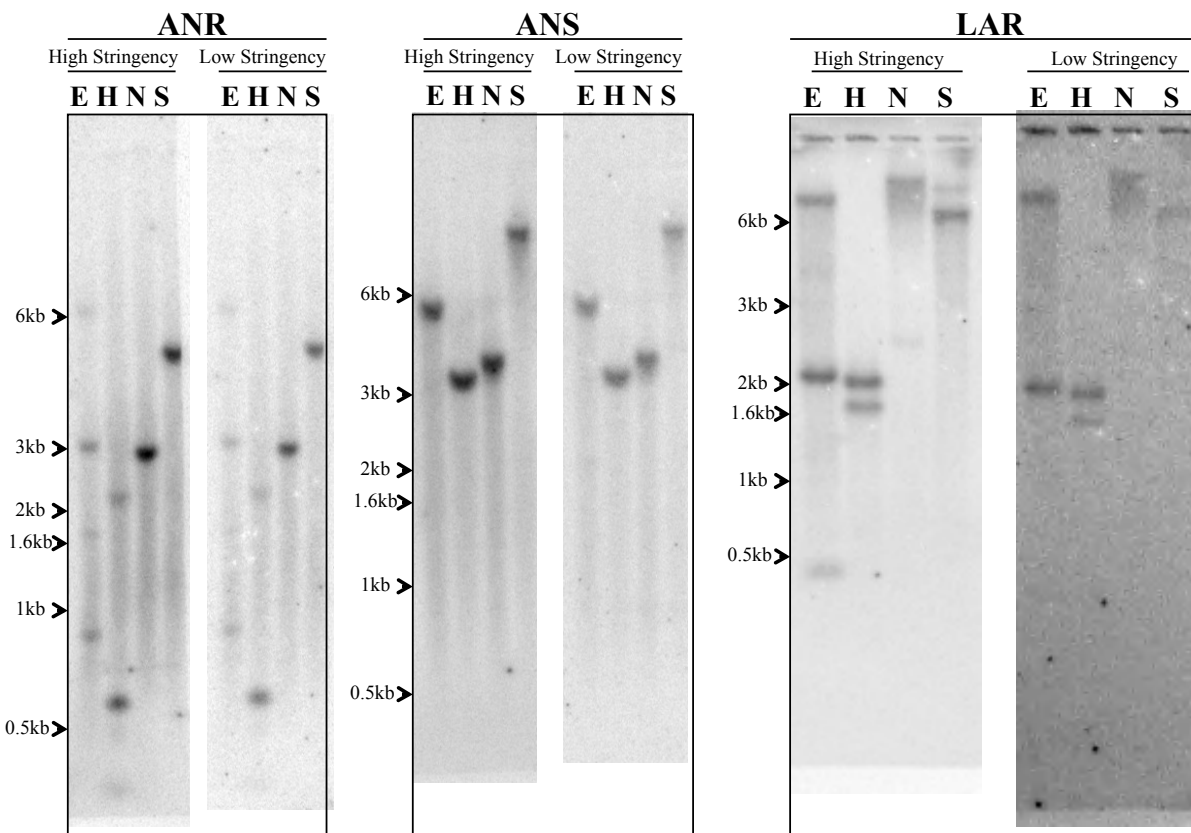
LAR and ANR proteins belong to the reductase-epimerase-dehydrogenase (RED) protein superfamily, although their relationships are rather distant. ANS belongs to a different protein superfamily, the 2-oxoglutarate-dependent dioxygenase (2-ODD) superfamily, although it shares the same substrate with LAR (Figure 2-3). A phylogenetic tree was constructed using the neighbor-joining method with the sequences of functionally-tested proteins of LAR, ANR, ANS and DFR from various plant species. The tree construction also included all the IFR-like proteins from Arabidopsis that are most closely related to LAR proteins (Figure 2-3). The tree was noticeably divided into two clades: all RED proteins (including ANR, LAR and IFR) constituting one clade, and all ANS proteins constituting the other. Within the RED superfamily clade, the IFR, LAR, ANR and DFR proteins clearly are divided into four distinct groups with IFR and LAR forming a subgroup that is distantly related to the subgroup formed by DFR and ANR groups. The cacao ANR, ANS and LAR proteins used in the current research all cluster within their own groups.



**Figure 2-3:** Phylogenetic analysis of the LAR, ANS and ANR proteins as well as related IFR and DFR proteins of the RED superfamily. The protein sequences were aligned as described in Section 2.2.4. The scale bar represents 0.2 substitutions per site. Numbers indicate bootstrap values. The tree includes LAR, ANR and ANS proteins from cacao, which are shown in bold font, as well as from other species whose enzymatic activities have been shown in previous publications. Also included in the tree are 8 IFR-like proteins from Arabidopsis (labeled with their locus numbers). All the other proteins are labeled by the species they come from followed by their catalytic activities. The species represented and their GenBank accession numbers are summarized in Table 2-1.

### 2.3.2 Southern Blot Analysis

Southern blot analysis was performed using full-length cDNAs of *TcANR*, *TcANS* and *TcLAR* as probes to hybridize to cacao (Scavina 6) genomic DNA digested separately with one of four restriction enzymes, *EcoRI*, *HindIII*, *NcoI* and *SmaI* (Figure 2-4). Both high-stringency and low-stringency analysis confirmed the presence of a single hybridizing locus of each cacao PA synthesis genes studied. Multiple bands appeared in the *EcoRI*- and *HindIII*-digested genomic DNA of the *TcANR* hybridization, and *EcoRI*-, *HindIII*- and *SmaI*-digested genomic DNA of the *TcLAR* hybridization. As evidenced by sequence analysis, these multiple bands resulted from the presence of multiple sites of the restriction enzymes in the *TcANS* and *TcLAR* genomic regions.



**Figure 2-4:** Cacao genomic Southern blot using full-length cDNA as a probe. Cacao genomic DNA was digested using four different restriction enzymes: *EcoRI*(E), *HindIII*(H), *NcoI*(N) and *SmaI*(S). Blots were first hybridized and washed at 65°C for high stringency and then stripped and rehybridized and washed at 55°C for low stringency.

### 2.3.3 Expression Profiles of *TcANR*, *TcANS* and *TcLAR* Genes and PA Accumulation in *Theobroma cacao* Tissues

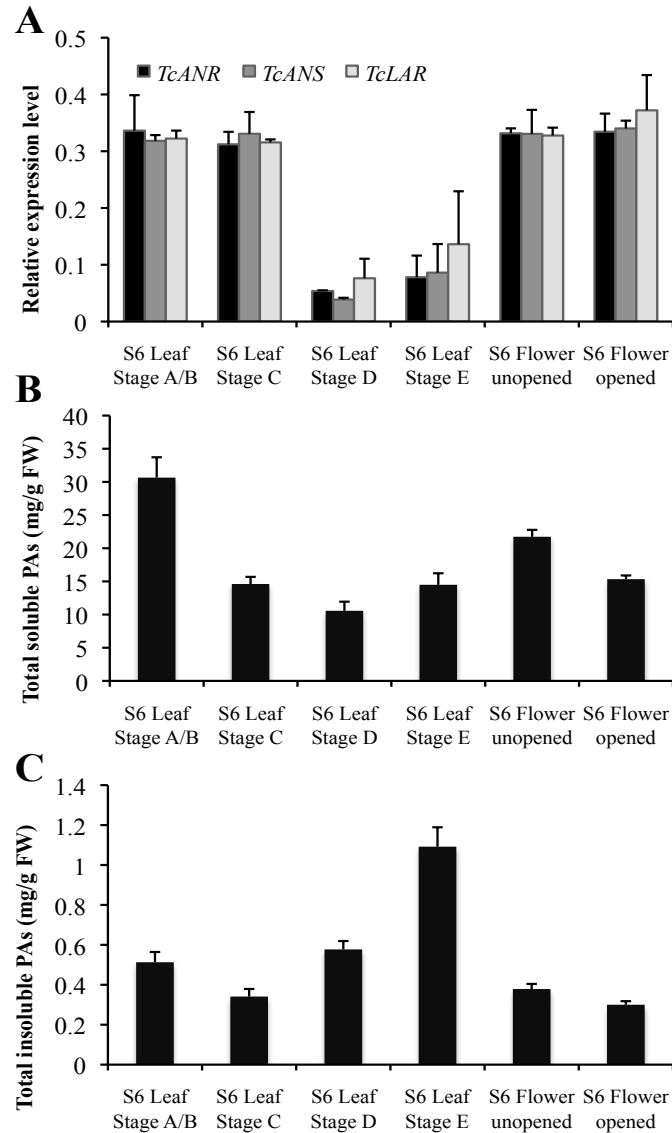
To assess the involvement of *TcANR*, *TcANS* and *TcLAR* in PA biosynthesis, gene expression profiles and PA accumulation patterns were investigated in different developmental stages of leaves, flowers and pods (Figure 2-5). The leaf and flower tissues were from cacao genotype Scavina 6, a self-incompatible genotype and a self-compatible genotype, Amelonado, was used for pod tissue collection, in order to reduce the genomic variation of seeds. Gene transcripts levels were assessed by semi-quantitative RT-PCR. *TcActin* was chosen as a reference gene to normalize gene expression because both cacao microarray analyses and RT-PCR data from this study suggested a relatively constant spatial and temporal expression of this gene for all cacao tissues examined (Z. Shi and S. Maximova, unpublished data) .

Cacao tissues were first treated with 70% acetone containing 0.5% acetic acid to extract soluble PAs. Extracts were then quantified by reaction with DMACA, which specifically interacts with PA monomers and polymers to form a blue complex that has a peak absorption of 526 nm (Thies and Fischer, 1971). The residues left after soluble PA extraction were assayed using butanol-HCl (Yu-Guang Li et al., 1996) to measure the amount of insoluble PAs represented as larger polymers. Because insoluble PA polymers will crystallize and bind to proteins and cell wall components, this interference may reduce the extraction efficiency of the insoluble PAs (Treutter, 2005; Pang et al., 2007). As a result, comparing the relative amount of these two fractions within the same tissue is difficult. However, the accumulation pattern of each fraction is comparable among different tissues.

The steady state levels of *TcANR*, *TcANS* and *TcLAR* transcripts were highest in young leaves and decreased in older leaves (Figure 2-5A). Relatively high levels were present in flower tissues. Moreover, the expression levels of the three PA synthesis genes examined were relatively similar within any of the sampled tissues/developmental stages. Both cacao leaves and flowers contained significant levels of PAs. The highest total soluble PAs were detected in youngest leaves (about 30 mg procyanidin B2 equivalent/g fresh weight (FW), Figure 2-5B). Much lower amounts were detected in older leaves. Total



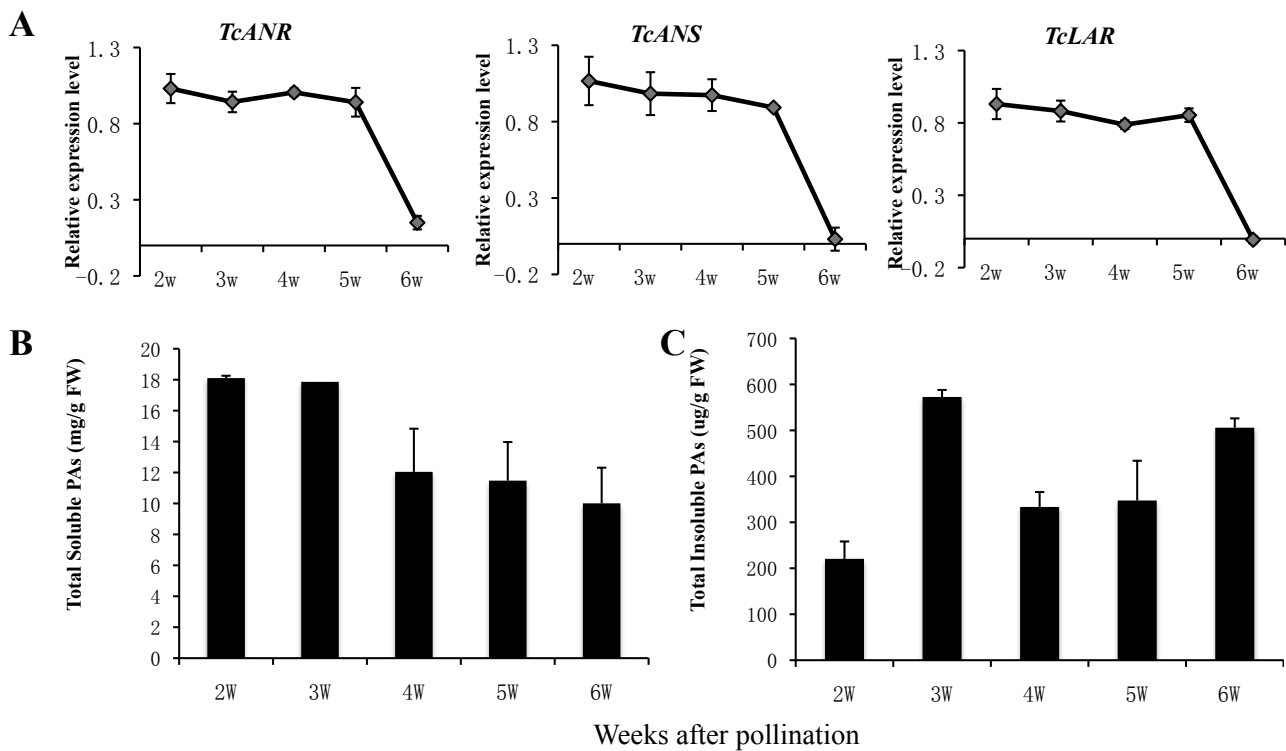
insoluble PAs were relatively lower in young leaves and continued to increase as the leaves aged and became harder. Insoluble PAs reached their maximum level in lignified stage E leaves (about 1.2 mg cyanidin equivalent/g FW, Figure 2-5C). PA levels were also considerable in flowers, with higher soluble PAs levels observed in unopened flowers than in opened flowers, and the levels of the insoluble fraction relatively the same in the two



**Figure 2-5:** Expression of *TcANR*, *TcANS* and *TcLAR* genes and accumulation of PAs in *Theobroma cacao* (Scavina 6; S6) leaves and flowers at various developmental stages. A, Transcript levels of *TcANR*, *TcANS* and *TcLAR*. Expression was determined by semi-quantitative RT-PCR and was calculated relative to the expression of *TcActin* in each sample. B, Levels of soluble PAs expressed as mg PA per g of fresh weight. C, Levels of insoluble PAs expressed as mg PA per g of fresh weight. All data are presented as means  $\pm$  SE, for gene expression data,  $n \geq 3$ , for PA level data,  $n \geq 5$ . FW, Fresh weight.

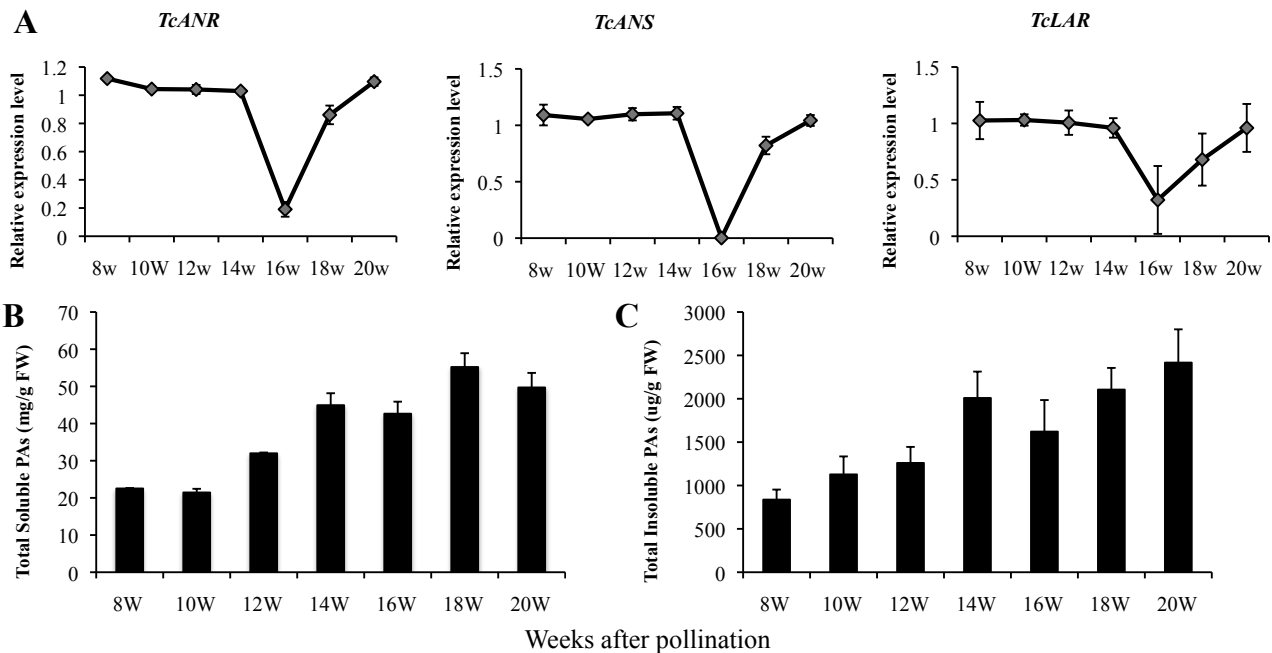
stages of flower development (Figure 2-5B, C).

Figure 2-6 shows both the expression patterns of *TcANR*, *TcANS* and *TcLAR* (Figure 2-6A) and PA levels in whole cacao pods early in their development when the pods are too small to separate ovules and exocarp (Figure 2-6B). The expression of *TcANR*, *TcANS* and *TcLAR* genes was relatively high at two weeks after pollination (WAP) and remained high at 5 WAP, followed by a significant decrease at 6 WAP (Figure 2-5A). Levels of soluble PAs were already close to maximum (approximately 18 mg procyanidin B2 equivalent/g FW) at the earliest sampling time point (Figure 2-6B), whereas insoluble PAs reached maximum levels at 3 WAP (Figure 2-6C).



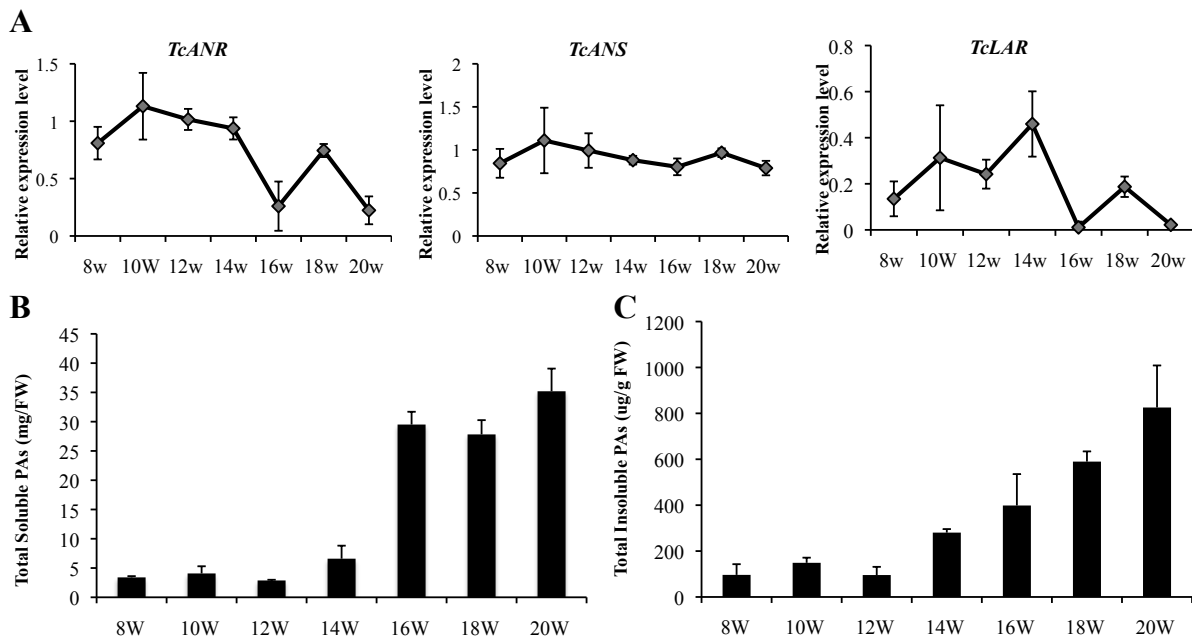
**Figure 2-6:** Expression of *TcANR*, *TcANS* and *TcLAR* genes and accumulation of PAs in whole pods of *Theobroma cacao* (Amelonado) during early stages of pod development (from 2 to 6 weeks after pollination). A, Transcript levels of *TcANR*, *TcANS* and *TcLAR*. B, Levels of total soluble PAs expressed as mg PAs per g of fresh weight. C, Levels of total insoluble PAs expressed as µg PAs per g of fresh weight. All data are presented as means  $\pm$  SE. For gene expression data,  $n \geq 3$ , for PA accumulation data,  $n \geq 5$ . FW, Fresh weight.

At 8 WAP, the pods were large enough to allow dissection into exocarp and ovule samples for separate analysis. Expression patterns of *TcANR*, *TcANS* and *TcLAR* genes and PA levels in cacao pod exocarp tissues were examined at two-week intervals, from 8 WAP to 20 WAP, when pods fully ripened. The expression of all three genes examined was similar (Figure 2-7A). They were all relatively high from 8 WAP to 14 WAP but decreased significantly at 16 WAP, increasing again at 18 WAP and reaching a maximum at 20 WAP. In accordance with gene expression patterns, the deposition of both soluble and insoluble PAs continued to increase during the development of the pods, reaching a maximum (soluble PA at approximately 50 mg procyanidin B2 equivalent/g FW; insoluble PA at approximately 2.5 mg cyanidin equivalent/g FW) around the time of ripening (Figure 2-7B,C), while a pause of the PA accumulation occurred at 16 WAP, at which time point, soluble PAs were about the same level as 14 WAP and insoluble PAs slightly decreased.



**Figure 2-7:** Expression of *TcANR*, *TcANS* and *TcLAR* genes and accumulation of PAs in pod exocarp of *Theobroma cacao* (Amelonado) during pod development (from 8 to 20 weeks after pollination). A, Transcript levels of *TcANR*, *TcANS* and *TcLAR*. B, Levels of total soluble PAs expressed as mg PAs per g of fresh weight. C, Levels of total insoluble PAs expressed as  $\mu$ g PAs per g of fresh weight. All data are presented as means  $\pm$  SE, for gene expression data,  $n \geq 3$ , for PA level data,  $n \geq 5$ . FW, Fresh weight.

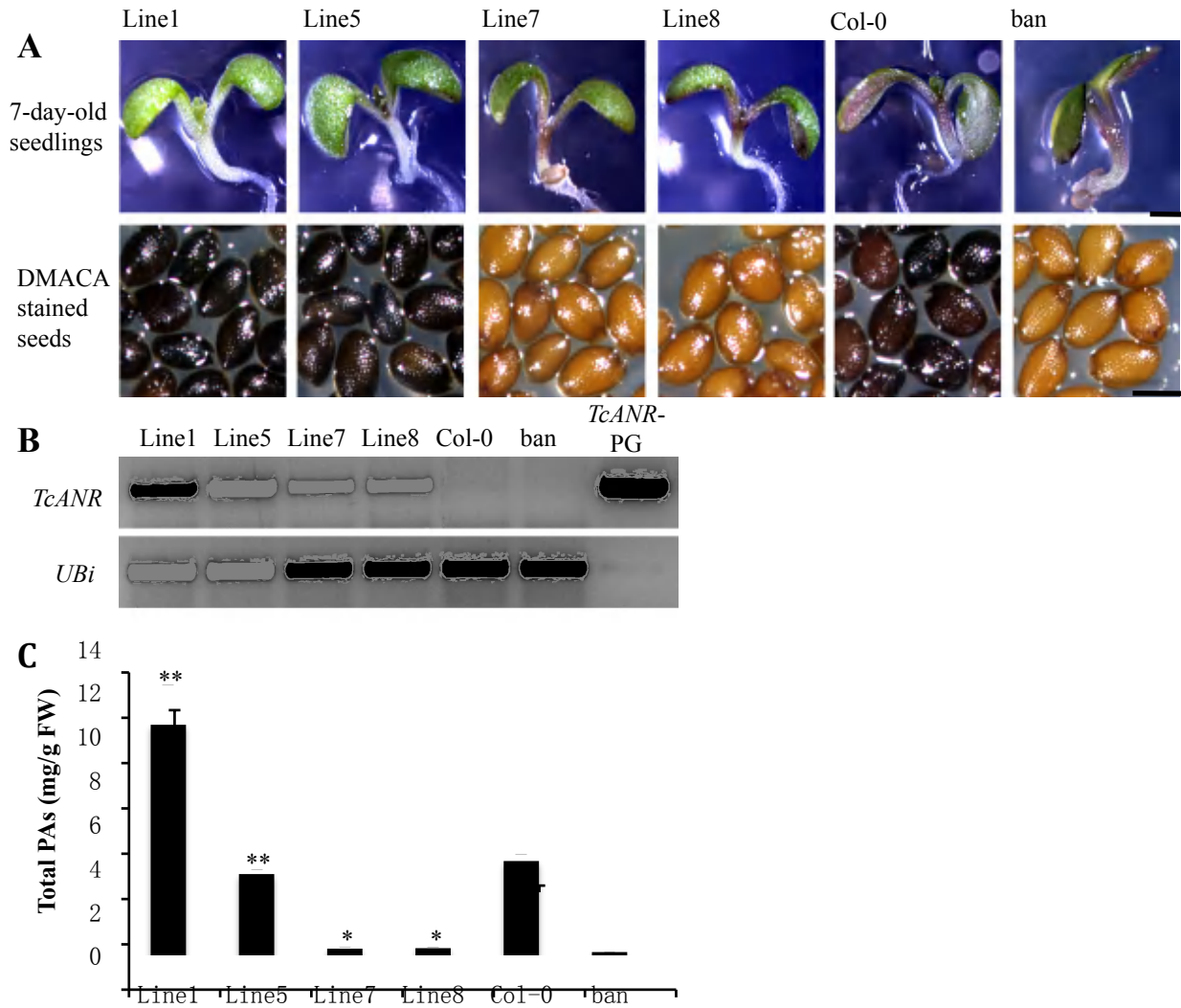
Unlike the co-regulated pattern of gene expression in exocarp, the expression pattern of *TcANS* differed quite significantly from that of *TcANR* and *TcLAR* in ovules (Figure 2-8A). The expression of *TcANR* and *TcLAR* in ovules was quite similar, maintaining relatively high levels before 14 WAP but significantly decreasing at 16 WAP, then increasing at 18 WAP and dropping again at 20 WAP. The overall expression level of *TcLAR* was lower than that of *TcANR*. *TcANS* expression did not significantly decrease at 16 WAP but remained relatively stable (0.7 to 1.2 relative to *TcActin*) throughout pod development, from 8 WAP to 20 WAP, although a slight increase did occur after 16 WAP followed by slight decrease at 20 WAP. The PA concentrations of both soluble and insoluble fractions in cacao ovules were lower than in exocarp (Figure 2-8B,C). The ovule soluble PA accumulation was relatively low before 16 WAP and significantly increased at 16 WAP, reaching a maximum at 20 WAP (about 35 mg procyanidin B2 equivalent/g FW). However, throughout the development of ovules, the insoluble PA increased at a relatively constant rate from 14 WAP.



**Figure 2-8:** Expression of *TcANR*, *TcANS* and *TcLAR* genes and accumulation of PAs in ovules of *Theobroma cacao* (Amelonado) during pod development (from 8 to 20 weeks after pollination). A, Transcript levels of *TcANR*, *TcANS* and *TcLAR*. B, Levels of total soluble PAs expressed as mg PAs per g of fresh weight. C, Levels of total insoluble PAs expressed as ug PAs per g of fresh weight. All data are presented as means  $\pm$  SE, for gene expression data,  $n \geq 3$ , for PA level data,  $n \geq 5$ . FW, Fresh weight.

### 2.3.4 Functional Analysis of the *TcANR* Gene

To investigate the *in vivo* function of *TcANR*, a genetic complementation experiment was performed. The *TcANR* coding sequence was placed under the control of an enhanced tobacco mosaic virus promoter (E12Ω) and introduced into the *Arabidopsis banyuls (ban)* mutant, which is defective in the gene encoding ANR. *ban* mutant seeds are pale yellow due to a lack of PAs, and there is more purple anthocyanin pigment deposited in hypocotyls compared to *Arabidopsis* ecotype Columbia control plants. From eight independent transgenic *TcANR* over-expressing lines tested, 5 lines showed white hypocotyls, and 3 lines showed reduced pigmentation in hypocotyls as compared to *ban* mutants. The lines with white hypocotyls also produced seeds exhibiting the wild-type phenotype that stained blue with DMACA reagent, suggesting the deposition of PAs in the seed coat (Figure 2-9A). After PA extraction and quantification, all lines showed significantly increased levels of PAs (Figure 2-9B). RT-PCR analysis confirmed expression of *TcANR*, and the expression levels positively correlated with PA accumulation (Figure 2-8C). Therefore, *TcANR* functionally complemented the PA deficient phenotype of *Arabidopsis ban* mutant.



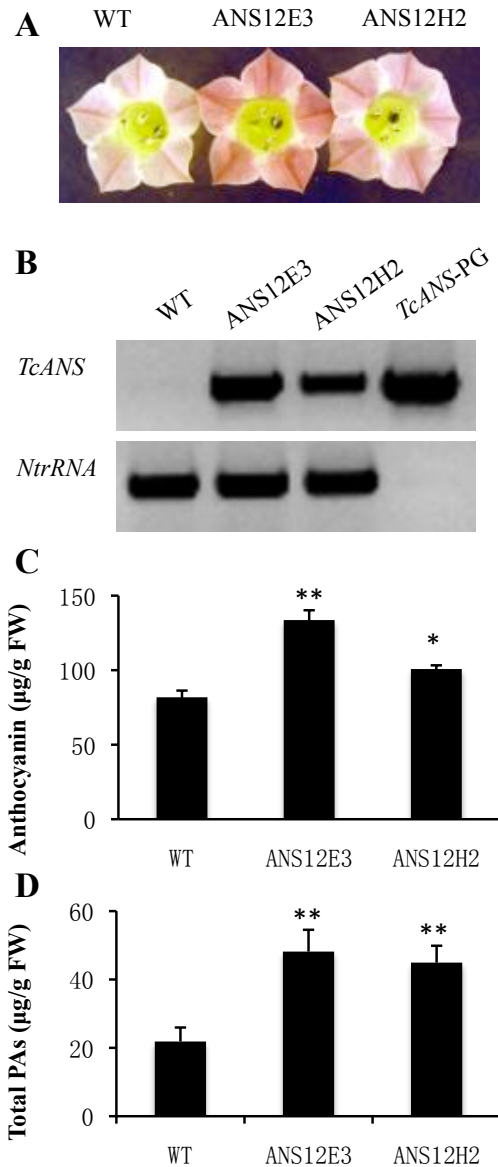
**Figure 2-9:** Complementation of the PA deficient *ban* mutant phenotype by constitutively expressing *TcANR*. A, 7-day-old seedlings and DMACA-stained seeds from Col-0 the *ban* mutant (SALK\_040250) and four independent transgenic lines (*ban* 35S:*TcANR*). The bar represents 1 mm. B, Analysis of *TcANR* and *AtUbi* transcripts in total RNA from leaves of plants shown in (A) by semi-quantitative RT-PCR. PCR products from *TcANR*-pGEM plasmid were loaded on the last lane acting as a positive control for the *TcANR* primer set and a negative control for the *AtUbi* primer set. C, PA levels in mature seeds of plants shown in (A). PA levels were determined by extraction and DMACA reaction with procyanidin B2 as standards. All data are presented as mean values  $\pm$  SE, n=3. \*P < 0.05 versus *ban*; \*\*P < 0.01 versus *ban*.

### 2.3.5 Functional Analysis of the *TcANS* Gene

To investigate the *in vivo* function of *TcANS*, two model plants, tobacco and *Arabidopsis* were utilized. We used an *Arabidopsis ans (ldox)* mutant to perform tests of transgenic complementation and the metabolic flux between PA and anthocyanin synthesis resulting in flower petal color changes in tobacco, which can result from elevated levels of ANS activity.

#### 2.3.5.1 Expression of *TcANS* in Tobacco

ANS is involved in both PA synthesis and anthocyanin synthesis, as cyanidin, an anthocyanin precursor, can also be reduced to epicatechin by ANR (Figure 1-2). This was demonstrated in a recent study describing that down regulation of *MtANS* in *Medicago* resulted in decreased levels of both PAs and anthocyanins (Pang et al., 2007). Based on these results, we reasoned that if the putative cacao ANS protein is truly a functional ANS, over-expression of *TcANS* in tobacco should result in increased accumulation of both anthocyanin and PAs. The ORF of *TcANS*, driven by the E12 $\Omega$  promoter, was introduced into wild-type tobacco (cv. Samsun) for constitutive ectopic expression. Twelve independent hygromycin-resistant lines were generated, of which nine showed a visible increase in pink color intensity in flower petals. The two lines displaying the greatest increase in petal color were chosen for further analysis. RT-PCR analysis confirmed high *TcANS* transcript levels in these two tobacco transgenic lines, which positively correlated with the color of the petals (Figure 2-10A and B). Amplification of the tobacco ribosomal RNA gene *NtrRNA*, which served as an internal control, showed a relatively similar expression level in both wild-type control and transgenic tobacco plants. As predicted, anthocyanin levels increased in the two transgenic lines (Figure 2-10C). The levels of PAs in the petals of transgenic lines, quantified by DMACA assays, were also significantly higher as compared to untransformed Samsun plants (Figure 2-10D). On average, a two-fold increase of PA and anthocyanin was observed in the two lines compared to wild type.

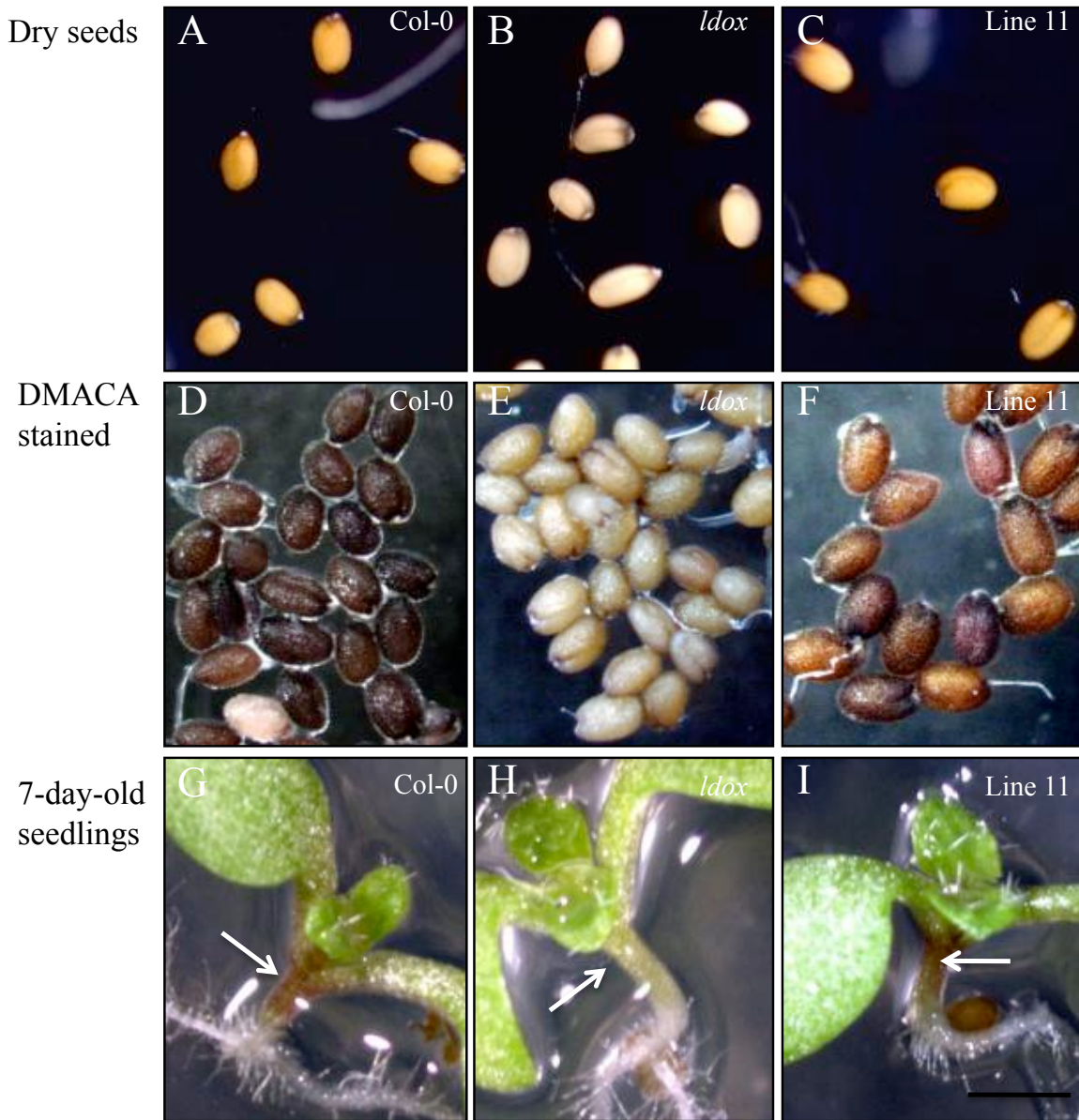


**Figure 2-10:** Characterization of transgenic tobacco flowers constitutively expressing *TcANS*. A, Pigmentation of flower petals from wild type (WT) and two independent lines of *TcANS* transgenic (ANS12E3, ANS12H2) tobacco plants. B, Analysis of *TcANS* and *NtrRNA* transcripts in total RNA from leaves of plants shown in (A) by RT-PCR. PCR products from *TcANS*-pGEM plasmid were loaded on the last lane acting as a positive control for *TcANS* primer set and a negative control for the *NtrRNA* primer set. C, Anthocyanin levels in flower petals of plants shown in (A). Anthocyanin levels were determined by extraction and absorption with cyanidin 3-glucoside as standards. D, Total soluble PA levels in flower petals of plants shown in (A). PA levels are determined by extraction and DMACA reaction with procyanidin B2 as standards. All data are presented as mean values  $\pm$  SE,  $n \geq 7$ . \* $P < 0.05$  versus WT; \*\* $P < 0.01$  versus WT.



### 2.3.5.2 Transgenic Complementation in Arabidopsis

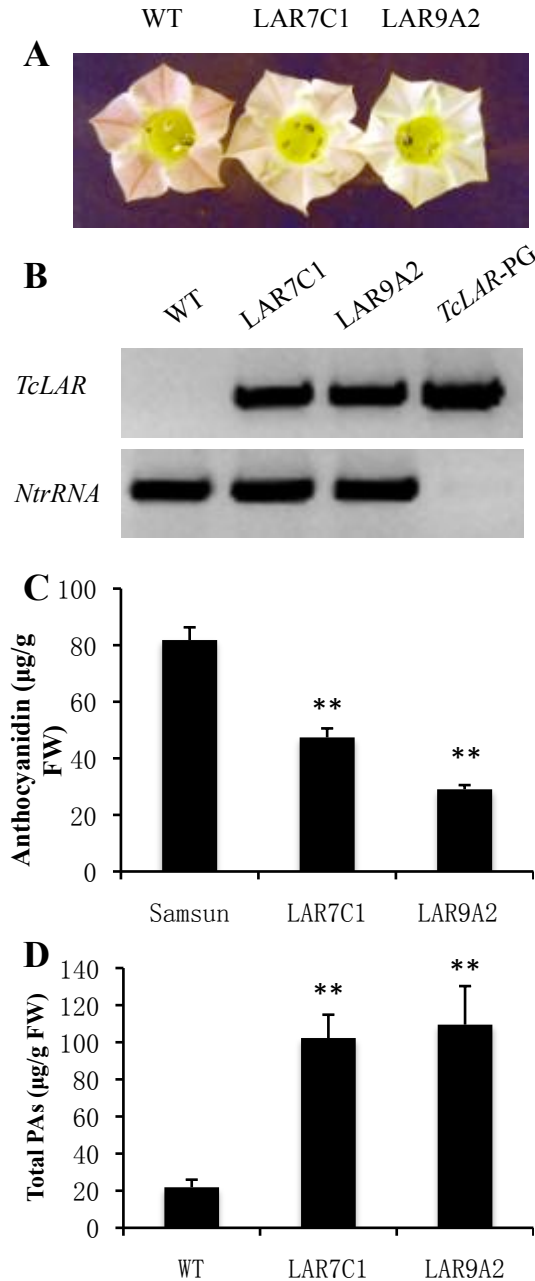
The 35S-*TcANS* transgene was also introduced into the *Arabidopsis ans (ldox)* T-DNA mutant, which produces hypocotyls that appear white to light green due to lack of anthocyanins, and seeds that appear light yellow due to lack of PAs. Eighteen independent hygromycin resistant transgenic T1 seedlings were selected. From these, 2 lines developed wild-type purple colored hypocotyls (Figure 2-11) and produced wild-type brown-colored seeds that stained blue after reacting with the DMACA reagent. The color suggests deposition of PAs in the seed coat, however the color intensity was lower than in wild type. RT-PCR using RNA extracted from T2 seedlings confirmed expression of *TcANS* genes (data not shown).



**Figure 2-11:** Complementation of the PA- and anthocyanin-deficient *ans* (*ldox*) mutant phenotype by constitutively expressing *TcANS*. Two independent lines showed a similar pattern to Col-0 for PA and anthocyanin accumulation. A, Brown Col-0 wild type seeds. B, Pale yellow *ldox* mutant seeds (SALK\_028793). C, Brown *ldox* 35S:*TcANS* seeds. D to F, PAs were localized by staining the seeds with DMACA, which specifically stains PAs blue. D, Col-0 wild type seeds. E, *ldox* mutant seeds. F, *ldox* 35S:*TcANS* seeds. G and H, Anthocyanin accumulation in the upper hypocotyls of 7-day-old seedlings. G, Col-0 wild type seedling. H, *ldox* mutant seedling. I, *ldox* 35S:*TcANS* seedling. Arrows in G, H and I point to the upper hypocotyl region. The bar represents 2 mm (A to I).

### 2.3.6 Functional Analysis of the *TcLAR* Gene

Because the *LAR* gene is not present in *Arabidopsis* (Abrahams et al., 2003), we cannot obtain direct evidence of the *in vivo* function of *TcLAR* through genetic complementation analysis. Therefore, to functionally characterize *TcLAR*, we used tobacco as a model system. As over-expression of the *ANR* gene can divert the metabolic flow from anthocyanin synthesis to PA synthesis (Xie et al., 2003), we predicted that over-expression of the cacao *LAR* gene in tobacco would result in a decrease in anthocyanin pigment and an increase in PA accumulation in flower petals. Transgenic tobacco plants were generated that constitutively expressed the ORF of *TcLAR* under the control of the E12 $\Omega$  promoter. Twenty-two independent transgenic lines that are resistant to hygromycin were generated. Nine of these exhibited a decrease in intensity of the visible pink color of petals. Two lines (Lines LAR7C1 and LAR9A2) exhibited virtually white petals (Figure 7A). RT-PCR showed that both lines expressed high levels of *TcLAR* transcripts (Figure 7B). Quantification of PA and anthocyanin levels indicated that anthocyanin levels in these two lines were about half those of WT controls, and that total PAs accumulated to an approximately 5-fold higher level than in controls (Figure 7C and D). Moreover, the levels of PAs in transgenic tobacco petals were inversely proportional to the concentrations of anthocyanin, indicating diversion of metabolic flow from anthocyanin to PA synthesis.



**Figure 2-12:** Characterization of transgenic tobacco flowers constitutively expressing *TcLAR*. A, Flowers from wild type (WT) and two independent lines of *TcLAR* transgenic (LAR7C1, LAR9A2) tobacco plants. B, Analysis of *TcLAR* and *NtrRNA* transcripts in total RNA from young leaves of plants shown in (A) by RT-PCR. PCR products from the *TcLAR*-PGEM plasmid alone were loaded on the last lanes to act as a positive control for the *TcLAR* primer set and a negative control for the *NtrRNA* primer set. C, Anthocyanin levels in flower petals of WT and transgenic plants. Anthocyanin levels were determined by extraction and absorption with Cyanidin-3-glucoside as standards. D, Total soluble PAs levels in flower petals of WT and transgenic plants. PA levels were determined by extraction and DMACA reaction with procyanidin B2 as standards. All data are presented as mean values  $\pm$  SE,  $n \geq 7$ . \*\* $P < 0.01$  versus WT.

## 2.4 Discussion

The research described in this chapter resulted in the isolation of the cacao cDNAs and genes coding for the three key enzymes involved in PA biosynthesis: *ANR*, *ANS*, and *LAR*. The proteins encoded by these genes showed 63% identity to Arabidopsis BAN, 82% identity to Arabidopsis LDOX and 61% identity to the *Desmodium* LAR, respectively, at the amino acid level. Isolation of genomic fragments containing each of the genes allowed the determination of the genomic organization of each gene. Generally, the number and length of exons are highly evolutionarily conserved in all of these genes. The phylogenetic analysis also suggests that the PA pathway genes evolved before the divergence of monocots (maize) and dicots (cacao, Arabidopsis, *Medicago*, *Desmodium* and grape).

### 2.4.1 The Distribution of PA Accumulation Correlates with Expression of the PA Biosynthesis Genes

The localization of PAs in plant tissues has been well studied in the model plants Arabidopsis and the legume *Medicago truncatula*, and in the fruit species grape (*Vitis vinifera*). In Arabidopsis, PA accumulation is limited to the seed coat (Abrahams et al., 2002); similarly in *Medicago* the major localization of PAs is also in the seed coat with very small amounts present in flowers, leaves, roots and stems (Pang et al., 2007). In contrast, in grape, large amounts of PAs accumulate in leaves, fruit skins and seeds (Bogs et al., 2005). Both Arabidopsis and *Medicago* PAs consist almost entirely of epicatechin units, although in *Medicago*, the *LAR* gene is expressed in the tissues that accumulate PAs. In grape, both ANR (encoded by *VvANR*) and LARs (encoded by *VvLAR1* and *VvLAR2*) contribute to PA synthesis in the fruits, in which PAs consist of both catechin and epicatechin. However, in grape leaves, although catechin is still present, *LAR* genes are expressed at very low levels.

Although *Theobroma cacao* and Arabidopsis are phylogenetically closely related, the PA accumulation profiles are quite different. In Arabidopsis, PAs do not accumulate in vegetative tissues but accumulate only in the seed coat (Devic et al., 1999). Our results demonstrated that in cacao leaves, flowers, and fruits, in which PAs accumulate at high

levels, the expression of *ANS*, *ANR*, and *LAR* were co-regulated in most of the tissues, and they all significantly contribute to PA biosynthesis.

Cacao leaves accumulate significant amounts of PAs that appear to be important for the plant defense response against pathogens (Chaves and Gianfagna, 2007). Only the newly-emerging leaves (Stage A/B) are susceptible to pathogen infection (Purdy and Schmidt, 1996). The data indicated that young leaves (Stage A/B) that are usually susceptible to pathogens accumulate the highest levels of extractable PAs (Figure 2-5). This is consistent with the observation of Chaves and Gianfagna (2007), who showed that cacao (genotype Scavina 6) flush leaves less than 6 cm long have the highest level of total procyanidins (98.7 mg per g of dry weight), while mature leaves accumulate only 20 to 40 mg per gram of dry weight. The data also showed that as soluble PAs decreased when leaves grew older and lignified, the insoluble PAs increased, suggesting that synthesized PAs do not simply degrade but become insolubilized by polymerization and binding to proteins. Previous studies have shown that this structures could provide a physical barrier against pathogen attack (Treutter, 2005). The patterns of transcript abundance of *TcANR*, *TcANS* and *TcLAR* genes are nearly identical and are consistent with their roles in the synthesis of PAs—gene transcripts were highest in early stage leaves where PAs are most actively synthesized and decreased in later stage leaves where soluble PA levels are much lower.

Cacao flowers contained significant levels of PAs (Figure 2-5), which paralleled the high-level expression of *TcANR*, *TcANS*, and *TcLAR*, indicating that PA synthesis occurred in developing flowers prior to pollination. Following pollination, even higher expression of PA synthesis genes was detected in young pods (Figure 2-6). The levels of extractable PAs decreased slightly, possibly due to the dilution effect of fruit expansion. The PAs in young pods (before 8 WAP) are mainly synthesized in fruit skin (exocarp) because embryos do not begin to develop until 6-7 WAP. My results indicated that PA synthesis in the pod exocarp continued during the second and third phases of fruit development, with all three PA related genes (*TcANR*, *TcANS*, and *TcLAR*) co-expressed (Figure 2-7). Cacao seeds also contained significant amounts of PAs, but the expression of PA-specific genes (*TcANR* and *TcLAR*) in relation to anthocyanin PA-common gene *TcANS* is different than in exocarps

(Figure 2-8). *TcANR* and *TcLAR* gene expression levels were co-regulated and were higher in the second phase of fruit development, but decreased at the beginning of the third phase. In contrast, *TcANS* expression remained steady, which related to the onset of anthocyanin synthesis.

Similar gene expression patterns were observed in grape and apple skins, in which PA synthesis genes and anthocyanin synthesis genes were separately regulated, and the occurrence of anthocyanin synthesis diverted the metabolic flow from PA synthesis (Bogs et al., 2005; Takos et al., 2006). In grape skins and seeds, as well as apple skins, PA synthesis completely stops after anthocyanin synthesis begins and PA levels decrease in both grape skins and seeds and apple skins, and expression of both *ANR* and *LAR* genes are no longer detected (Bogs et al., 2005). Unlike grape seeds, cacao seeds still actively synthesized PAs after anthocyanin synthesis began (from 16WAP to 20WAP), with significant increases of both soluble and insoluble PAs concentrations until maturation (Figure 2-7 B, C). Correspondingly, *TcANR* and *TcLAR* were expressed in the middle of the third phase. This suggests that the regulatory mechanism of anthocyanin and PA synthesis in cacao is different from that in apple and grape. In grape and apple, PAs synthesis is regulated separately from anthocyanin synthesis. In contrast, in cacao, the temporal synthesis of PAs and anthocyanins significantly overlapped, as did the expression patterns of their biosynthesis genes, suggesting that the regulation of PA and anthocyanin synthesis may share some common mechanisms. However, further characterization of the anthocyanidin-specific structural genes as well as the transcription factors controlling these structural genes are needed to test this hypothesis.

#### **2.4.2 Functions of *ANS*, *ANR* and *LAR* in PA Biosynthesis in Cacao**

The Southern blot analysis results presented in this work showed that *ANS*, *ANR* and *LAR* genes are all single copy in the cacao genome, thus providing a simple model to investigate enzyme functions. But we cannot exclude the possibility of existence of additional gene copies due to the sensitivity of the Southern blot analysis. To draw an

exclusive conclusion, future studies may involve searching the cacao whole genome sequence when it is available.

*In vivo* genetic analysis of *TcANS* and *TcANR* suggested their roles in PA biosynthesis. Over-expression of *TcANS* in tobacco resulted in elevated levels of both anthocyanin and PAs in flower petals. It also restored anthocyanin synthesis in hypocotyls as well as PA accumulation in seeds of the Arabidopsis *ans (ldox)* mutant. Similarly, over-expression of *TcANR* in the Arabidopsis *banyuls (anr)* mutant restored PA synthesis in seeds. Moreover, ectopic over-expression of *TcANR* in hypocotyls diverted metabolites from the anthocyanin synthesis branch and resulted in decreased anthocyanin pigments related to gene expression levels.

The results presented in this work show that cacao has only one copy of *LAR* that is co-regulated with *TcANR* in all tissues examined, thus providing a simple model to investigate *LAR* enzyme functions. Over-expression of the *TcLAR* gene in tobacco plants resulted in a significant reduction of anthocyanin synthesis in flower petals and an increase of PA synthesis indicating diverted metabolic flow from anthocyanin to PAs. This data provides, for the first time, direct genetic evidence for a clear role of *LAR* in PA biosynthesis.

## **2.5 Conclusions**

The PA accumulation analysis results presented in this chapter demonstrate that cacao synthesizes significant amount of PAs in various tissues including leaves, flowers and fruits. PAs are more actively synthesized in newly-emerging leaves than in mature leaves, consistent with their role in pathogen defense. PA synthesis in cacao pods largely correlates with anthocyanin synthesis, suggesting their coordinated regulatory mechanisms.

The molecular-genetic results presented in this chapter demonstrate successful cloning of genes encoding three important enzymes involved in the synthesis of PAs: *TcANR*, *TcANS* and *TcLAR*. Phylogenetic analysis demonstrated the high degree of conservation of each of the candidate cacao gene sequences to well-established PA biosynthesis genes from model organisms. The mRNA expression levels of all three genes



significantly correlated with PA synthesis levels in various tissues, suggesting their involvement in PA synthesis. *TcANR* gene can substitute for the function of the Arabidopsis *ANR* (*BANYULS*) gene in transgenic complementation assays. The *TcANS* gene can substitute for the function of the Arabidopsis *ANS* (*LDOX*) gene in transgenic complementation assays; when overexpressed in tobacco plants, it can also divert the metabolic flow, resulting in an increase of both PA and anthocyanin levels in flower petals. Overexpression of the *TcLAR* gene in tobacco plants can divert the metabolic flow from anthocyanin synthesis to PA synthesis, and results in an increase of PA levels together with a decrease of anthocyanin levels in flower petals. These *in vivo* analysis results strongly suggest the *TcANR*, *TcANS* and *TcLAR* genes encode proteins performing the predicted enzyme activities.

Future studies could use HPLC to study the composition of cacao PAs and study the contributions of *TcANR*, *TcANS* and *TcLAR* genes to each PA building block. Such an investigation would provide tools for furthering the bioengineering of specific PA profiles with designed content and composition.

## 2.6 References

- Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR** (2003) The *Arabidopsis* TDS4 gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *Plant J* **35**: 624-636
- Abrahams S, Tanner GJ, Larkin PJ, Ashton AR** (2002) Identification and biochemical characterization of mutants in the proanthocyanidin pathway in *Arabidopsis*. *Plant Physiol* **130**: 561-576
- Alemanno L, Berthouly M, MichauxFerriere N** (1997) A comparison between *Theobroma cacao* L. zygotic embryogenesis and somatic embryogenesis from floral explants. *In Vitro Cell. Dev. Biol. Plant* **33**: 163-172
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Looor G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C** (2008) Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. *BMC Genomics* **9**: 512
- Aron PM, Kennedy JA** (2008) Flavan-3-ols: nature, occurrence and biological activity. *Mol Nutr Food Res* **52**: 79-104
- Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K** (2007) Current protocols in molecular biology. John Wiley and Sons, Inc, New York
- Bogs J, Downey MO, Harvey JS, Ashton AR, Tanner GJ, Robinson SP** (2005) Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. *Plant Physiol* **139**: 652-663
- Bogs J, Jaffe FW, Takos AM, Walker AR, Robinson SP** (2007) The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiol* **143**: 1347-1361
- Chaves FC, Gianfagna TJ** (2007) Cacao leaf procyanidins increase locally and systemically in response to infection by *Moniliophthora perniciosa* basidiospores. *Physiol. Mol. Plant Pathol.* **70**: 174-179
- Cheesman EE** (1927) Fertilization and embryogeny in *Theobroma cacao*, L. *Ann. Bot.* **41**: 107-126
- Clough SJ, Bent AF** (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* **16**: 735-743
- Devic M, Guillemot J, Debeaujon I, Bechtold N, Bensaude E, Koornneef M, Pelletier G, Delseny M** (1999) The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. *Plant J* **19**: 387-398
- Dixon RA, Xie DY, Sharma SB** (2005) Proanthocyanidins - a final frontier in flavonoid research? *New Phytol* **165**: 9-28

- Feucht W, Treutter D** (1999) The role of flavan-3-ols and proanthocyanidins in plant defense. In Inderjit, KMM Dakshi, CL Foy, eds, Principles and Practices in Plant Ecology: Allelochemical Interactions. CRC Press, New York, pp 307-338
- Fisher DK, Gultinan MJ** (1995) Rapid, efficient production of homozygous transgenic tobacco plants with *Agrobacterium tumefaciens* : A seed-to-seed protocol. Plant Mol Biol Reprtr **13**: 278-289
- Gu L, Kelm M, Hammerstone JF, Beecher G, Cunningham D, Vannozzi S, Prior RL** (2002) Fractionation of polymeric procyanidins from lowbush blueberry and quantification of procyanidins in selected foods with an optimized normal-phase HPLC-MS fluorescent detection method. J Agric Food Chem **50**: 4852-4860
- Kumar S, Tamura K, Nei M** (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform **5**: 150-163
- Lazo GR, Stein PA, Ludwig RA** (1991) A DNA transformation-competent *Arabidopsis* genomic library in *Agrobacterium*. Biotechnology (N Y) **9**: 963-967
- Lehrian DW, Keeney PG** (1980) Changes in lipid components of seeds during growth and ripening of cacao fruit. J. Am. Oil Chem. Soc. **57**: 61-65
- Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M** (2006) Genetics and biochemistry of seed flavonoids. Annu Rev Plant Biol **57**: 405-430
- Levy M, Edelbaum O, Sela I** (2004) Tobacco mosaic virus regulates the expression of its own resistance gene N. Plant Physiol **135**: 2392-2397
- Lin JJ** (1994) Optimization of the transformation efficiency of *Agrobacterium tumefaciens* cells using electroporation. Plant Sci **101**: 11-15
- Martin C, Prescott A, Mackay S, Bartlett J, Vrijlandt E** (1991) Control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. Plant J **1**: 37-49
- Maximova S, Miller C, Antunez de Mayolo G, Pishak S, Young A, Gultinan MJ** (2003) Stable transformation of *Theobroma cacao L.* and influence of matrix attachment regions on GFP expression. Plant Cell Rep **21**: 872-883
- Menssen A, Hohmann S, Martin W, Schnable PS, Peterson PA, Saedler H, Gierl A** (1990) The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the *A2* gene. Embo J **9**: 3051-3057
- Michiels A, Van den Ende W, Tucker M, Van Riet L, Van Laere A** (2003) Extraction of high-quality genomic DNA from latex-containing plants. Anal Biochem **315**: 85-89
- Murashige T, Skoog F** (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol Plant **15**: 473-497
- Niemenak N, Rohsius C, Elwers S, Ndoumou DO, Lieberei R** (2006) Comparative study of different cocoa (*Theobroma cacao L.*) clones in terms of their phenolics and anthocyanins contents. J. Food Comp Anal. **19**: 612-619
- Norman KH, Naomi DLF, Marjorie LM** (2009) Flavanols, the Kuna, cocoa consumption, and nitric oxide. J. Am. Soc. Hypertens **3**: 105-112
- Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA** (2007) Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. Plant Physiol **145**: 601-615
- Paolucci F, Robbins MP, Madeo L, Arcioni S, Martens S, Damiani F** (2007) Ectopic expression of a basic helix-loop-helix gene transactivates parallel pathways of proanthocyanidin biosynthesis. structure, expression analysis, and genetic control

- of leucoanthocyanidin 4-reductase and anthocyanidin reductase genes in *Lotus corniculatus*. *Plant Physiol* **143**: 504-516
- Pelletier MK, Murrell JR, Shirley BW** (1997) Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in *Arabidopsis*. Further evidence for differential regulation of "early" and "late" genes. *Plant Physiol* **113**: 1437-1445
- Pourcel L, Routaboul JM, Cheynier V, Lepiniec L, Debeaujon I** (2007) Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* **12**: 29-36
- Pourcel L, Routaboul JM, Kerhoas L, Caboche M, Lepiniec L, Debeaujon I** (2005) *TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell* **17**: 2966-2980
- Purdy LH, Schmidt RA** (1996) Status of cacao witches' broom: biology, epidemiology, and management. *Annu Rev Phytopathol* **34**: 573-594
- Ramassamy C** (2006) Emerging role of polyphenolic compounds in the treatment of neurodegenerative diseases: A review of their intracellular targets. *Eur J Pharmacol* **545**: 51-64
- Saito K, Kobayashi M, Gong Z, Tanaka Y, Yamazaki M** (1999) Direct evidence for anthocyanidin synthase as a 2-oxoglutarate-dependent oxygenase: molecular cloning and functional expression of cDNA from a red form of *Perilla frutescens*. *Plant J* **17**: 181-189
- Saslowsky D, Winkel-Shirley B** (2001) Localization of flavonoid enzymes in *Arabidopsis* roots. *Plant J* **27**: 37-48
- Scalbert A** (1991) Antimicrobial properties of tannins. *Phytochemistry* **30**: 3875-3883
- Schwarz-Sommer Z, Shepherd N, Tacke E, Gierl A, Rohde W, Leclercq L, Mattes M, Berndtgen R, Peterson PA, Saedler H** (1987) Influence of transposable elements on the structure and function of the *A1* gene of *Zea mays*. *Embo J* **6**: 287-294
- Shirley BW, Hanley S, Goodman HM** (1992) Effects of ionizing radiation on a plant genome: analysis of two *Arabidopsis* transparent testa mutations. *Plant Cell* **4**: 333-347
- Singh K, Rani A, Paul A, Dutt S, Joshi R, Gulati A, Ahuja PS, Kumar S** (2009) Differential display mediated cloning of anthocyanidin reductase gene from tea (*Camellia sinensis*) and its relationship with the concentration of epicatechins. *Tree Physiol* **29**: 837-846
- Takos AM, Ubi BE, Robinson SP, Walker AR** (2006) Condensed tannin biosynthesis genes are regulated separately from other flavonoid biosynthesis genes in apple fruit skin. *Plant Sci* **170**: 487-499
- Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ, Ashton AR** (2003) Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *J Biol Chem* **278**: 31647-31656
- Thies M, Fischer R** (1971) New color reaction for microchemical detection and quantitative determination of catechins. *Mikrochim Acta*: 9-13
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG** (1997) The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882
- Treutter D** (2005) Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol.* **7**: 581-591

- Winkel-Shirley B** (2001) It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. *Plant Physiol* **127**: 1399-1404
- Winkel-Shirley B** (2002) Biosynthesis of flavonoids and effects of stress. *Curr Opin Plant Biol* **5**: 218-223
- Wright DC, Park WD, Leopold NR, Hasegawa PM, Janick J** (1982) Accumulation of lipids, proteins, alkaloids and anthocyanins during embryo development *in vivo* of *Theobroma cacao* L. *J. Am. Oil Chem. Soc.* **59**: 475-479
- Xie DY, Jackson LA, Cooper JD, Ferreira D, Paiva NL** (2004) Molecular and biochemical analysis of two cDNA clones encoding dihydroflavonol-4-reductase from *Medicago truncatula*. *Plant Physiol* **134**: 979-994
- Xie DY, Sharma SB, Paiva NL, Ferreira D, Dixon RA** (2003) Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis. *Science* **299**: 396-399
- Yu-Guang Li, Greg Tanner, Phil Larkin** (1996) The DMACA-HCl protocol and the threshold proanthocyanidin content for bloat safety in forage legumes. *J. Sci. Food Agric.* **70**: 89-101
- Zeng Y** (2002) RNA isolation from highly viscous samples rich in polyphenols and polysaccharides. *Plant Mol Biol Reprtr* **20**: 417
- Zhang W, Sun Y, Timofejeva L, Chen C, Grossniklaus U, Ma H** (2006) Regulation of Arabidopsis tapetum development and function by *DYSFUNCTIONAL TAPETUM1* (*DYT1*) encoding a putative bHLH transcription factor. *Development* **133**: 3085-3095

## CHAPTER 3: FUNCTIONAL ANALYSIS OF CACAO LEUCOANTHOCYANIDIN REDUCTASE

### 3.1 Introduction

Leucoanthocyanidin reductase (LAR), which catalyzes the NADPH-dependent reduction of (+) 3,4-*cis*-leucocyanidin to catechin by removing the C4 hydroxyl group (Figure 1-2), represents the first branching point from the anthocyanin synthesis pathway to PA synthesis (Tanner et al., 2003). The activity of LAR has been demonstrated *in vitro* using extracts of various plant species including barley (Kristiansen, 1986) and various legumes (Tanner and Kristiansen, 1993; Singh et al., 1997; Skadhauge et al., 1997). The protein was first purified and the corresponding gene cloned from the legume *Desmodium uncinatum* (Tanner et al., 2003).

*LAR* genes have subsequently been characterized from a number of species including *Vitis vinifera* (grape), *Lotus corniculatus* and *Medicago truncatula*, and the corresponding protein function has been characterized by *in vitro* recombinant enzyme assays (Bogs et al., 2005; Pang et al., 2007; Paolucci et al., 2007). However, none of these studies has provided genetic evidence for *LAR* function. The only genetic evidence of *LAR* function is the observation that the genomic sequence of *Arabidopsis thaliana* appears not to contain an intact *LAR* orthologue, and correspondingly, catechin is not detected in *Arabidopsis* seed extracts (Abrahams et al., 2003; Tanner et al., 2003; Lepiniec et al., 2006). *LAR* genes are expressed in other plant species that accumulate not only epicatechin but also catechin (Tanner et al., 2003; Bogs et al., 2005; Paolucci et al., 2007). For example, grape and *Lotus* express both *LAR* and anthocyanidin reductase (*ANR*) genes and synthesize PAs consisting of both catechin and epicatechin (Foo et al., 1996; Bogs et al., 2005; Paolucci et al., 2007). In *Medicago*, although both *LAR* and *ANR* are expressed, the PAs are composed almost entirely of epicatechin (Pang et al., 2007). These intriguing results put the precise function of *LAR* gene in question.

LAR and ANR both belong to the reductase-epimerase-dehydrogenase (RED) superfamily. All members in this superfamily have the N-terminal Rossmann-fold NAD(P)(+)-binding domain that is found in numerous dehydrogenases of various metabolic pathways (Marchler-Bauer et al., 2009). A recent study demonstrated that an ANR protein from grape can catalyze the conversion of cyanidin to a 50:50 mixture of 2,3-*cis* and 2,3-*trans* flavan-3-ols (Gargouri et al., 2009) and suggested the existence of a reductase/epimerase dual-functional enzyme in the RED superfamily.

The research in Chapter 2 described the isolation of PA synthetic structural genes from cacao, including *TcLAR*. It also demonstrated that *TcLAR* is co-regulated with other PA synthetic genes such as *TcANR* and *TcANS*, and correlated *TcLAR* expression with PA synthesis in leaves, flowers and fruits. Furthermore, the research in Chapter 2 provided direct genetic evidence of the involvement of cacao *LAR* in PA biosynthesis using transgenic tobacco plants with a constitutively expressed *TcLAR* gene. This chapter presents research exploring the precise function of *TcLAR* and its contribution to PA synthesis. PAs extracted from various tissues of cacao were analyzed by HPLC to examine their composition. Transgenic tobacco and *Arabidopsis ldox* mutants over-expressing the *TcLAR* gene were generated to analyze the *in vivo* function of *TcLAR* in more detail. Surprisingly, the results suggest that cacao *LAR* might have a dual function in producing both catechin and epicatechin.

## **3.2 Materials and Methods**

### **3.2.1 Extraction and Quantification of PAs**

PAs from cacao and tobacco tissues were extracted as described in Chapter 2. To extract *Arabidopsis* PAs, the methods of (Bogs et al., 2007) were adapted as follows. Eight to twelve green young siliques from single plants were pooled and ground into a fine powder in liquid nitrogen-filled microtubes using matching pestles (A. Daigger & Company, Inc., Vernon Hills, IL, USA). The ground samples were extracted using 500  $\mu$ L of extraction

solution (70% acetone, 29.5% water, 0.5% acetic acid) by vortexing for 5 seconds followed by sonication in a bench top ultrasonic cleaner (Model 2510, Branson, Danbury, CT, USA) for 15 min. After sonication, samples were vortexed again and centrifuged at 16,000g for 2 min. The supernatants were transferred to a new tube and pellets were re-extracted as above. Pooled supernatants were extracted twice with 500  $\mu$ L of hexane to remove fat and chlorophyll and then filtered through 0.45  $\mu$ m polytetrafluoroethylene (PTFE) syringe filters (Millipore, Billerica, MA, USA).

Catechin and epicatechin content was determined by reverse-phase HPLC using an Alliance separations module (Model 2695; Waters, Milford, MA, USA) equipped with a multi  $\lambda$  fluorescence detector (Model 2475; Waters, Milford, MA, USA). Samples (10  $\mu$ L) were separated on a 250 mm  $\times$  4.6 mm Luna 5- $\mu$ m Phenyl Hexyl column (Phenomenex, Torrance, CA, USA) and then assayed by fluorescence emission at 315 nm following excitation at 280 nm. The HPLC separation utilized a binary mobile phase gradient mixture of A+B where mobile phase A was 0.5% trifluoroacetic acid (TFA) (v/v with water) and mobile phase B was 0.5% TFA (v/v with methanol). The gradient conditions were: 0 min, 16% mobile phase B; 4 min, 16% mobile phase B; 14 min, 50% mobile phase B; 18 min, 50% mobile phase B; 22min, 100% mobile phase B; 26 min, 100% mobile phase B; 30 min, 16% mobile phase B. The column was maintained at 30°C and the flow rate was 1 mL/min. Catechin and epicatechin standards were purchased from Sigma-Aldrich (St. Louis, MO, USA). This work was performed at the Hershey Technical Center (Hershey, PA, USA) in collaboration with Mark Payne.

### **3.2.2 Protein Sequence Alignment**

Protein sequences of isoflavone reductase (IFR) were retrieved from The Arabidopsis Information Resource (TAIR) database (<http://www.arabidopsis.org/>). The cacao IFR-like gene EST sequence (CL222Contig1) was retrieved by BLAST (Altschul et al., 1990) of the cacao EST database (<http://esttik.cirad.fr/index.html>) and the protein sequence was obtained using the NCBI ORF finder program (<http://www.ncbi.nlm.nih.gov/projects/gorf/>). All other LAR proteins from different



species were retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>), including *Vitis vinifera* (VvLAR1, CAI26309; VvLAR1, CAI26308) (Bogs et al., 2005), *Desmodium uncinatum* (DuLAR, CAD79341) (Tanner et al., 2003), *Medicago truncatula* (MtLAR, CAI56327) (Pang et al., 2007) and *Lotus corniculatus* (LcLAR1-1, ABC71324 ; LcLAR1-2, ABC71325; LcLAR2-1, ABC71330; LcLAR2-2, ABC71331). Multiple protein sequence alignment was performed using ClustalX (version 2.0.11) with default parameters.

### 3.2.3 Transformation of Arabidopsis with TcLAR

*Agrobacterium tumefaciens* strains carrying TcLAR over-expression constructs were prepared and Arabidopsis transformation was performed as described in Chapter 2 (section 2.2.7).

### 3.2.4 Gene Expression Analysis

RNA samples from tobacco flowers and Arabidopsis young seedlings and leaves were isolated using the RNeasy Plant mini kit (Qiagen, Valencia, CA, USA). First strand cDNA synthesis and semi-quantitative RT-PCR were performed as described in Chapter 2 (section 2.2.8).

### 3.2.5 In Vitro Expression of Recombinant TcLAR protein

The open reading frame (ORF) of the TcLAR gene was PCR amplified from pGEMT-TcLAR (Chapter 2, section 2.2.3) using Advantage 2 polymerase mix (Clontech, Mountain View, CA) and the following primers: TcLARCDF1 (5'-*GAGCTC*atggatatgaaatcaacaaacatg-3'; the *SacI* site is in italics and the start codon is underlined) and TcLARCDR2 (5'-*CTCGAG*tgtgcatatcgcatg-3'; the *XhoI* site is in italics and the stop codon was removed to incorporate the C-terminal His-tag sequence of the expression vector at the 3' end of the ORF of TcLAR). It was then subcloned into the *SacI* and *XhoI* sites of the pET-21a expression vector (Novagen, Gibbstown, NJ, USA). After confirmation by sequencing, the resulting vector pET21a-TcLAR was transformed into *Escherichia coli* strain Rosetta (DE3) (Novagen, Gibbstown, NJ, USA). For protein expression, a single bacterial colony was inoculated into Luria-Bertani medium (10g/L

tryptone, 5 g/L yeast extract, 10 g/L NaCl) containing 100 µg/mL ampicillin and grown at 37°C overnight. An overnight culture was then diluted into terrific broth (TB) medium (12 g/L Tryptone, 24g/L Yeast Extract, 0.4% glycerol, 2.31 g/L KH<sub>2</sub>PO<sub>4</sub>, 12.54 g/L K<sub>2</sub>HPO<sub>4</sub>) containing 100 µg/mL ampicillin and grown at 37°C until the OD<sub>600</sub> reached 0.6-0.8, at which time IPTG (isopropyl β-D-1-thiogalactopyranoside), was added to a final concentration of 1 mM to induce protein expression for 18 hours. Recombinant TcLAR protein with a 6×His tag at the C terminus was purified using a Magne-His kit (Promega, Madison, WI, USA) and the protein concentration measured by the Bradford method (Bradford, 1976). This work was performed in collaboration with Yongzhen Pang and Richard A. Dixon at Samuel Roberts Nobel Foundation (Ardmore, Oklahoma, USA).

### 3.2.6 Assay of LAR Activities

<sup>3</sup>H-3,4-*cis*-leucocyanidin was synthesized as described by (Tanner and Kristiansen, 1993). Assay of recombinant TcLAR protein with <sup>3</sup>H-3,4-*cis*-leucocyanidin was carried out in a final volume of 100 µL containing 10% (w/v) glycerol, 100 mM potassium phosphate (pH 7.0), 4 mM dithiothreitol (DTT), 0.5 mM NADPH, 0.4 mM <sup>3</sup>H-leucocyanidin and 30 µg of purified recombinant TcLAR protein. The reaction was initiated by the addition of enzyme and incubated at 30°C for 1 h. The assay was terminated by the addition of 20 µL of methanol followed by centrifugation. Products were analyzed by HPLC, with absorbance monitoring at 280 nm. Products eluting at retention times between 13 to 31 min were collected (1 min/tube) and the fractions containing labeled product were identified by liquid scintillation counting. Boiled pure protein was used as a control.

Reverse-phase HPLC analysis of enzymatic products was performed using an Agilent HP1100 HPLC (Agilent Technologies, Inc., Santa Clara, CA, USA) with the following gradient using solvents A (1% phosphoric acid) and B (acetonitrile) at a 1 mL/min flow rate: 0 to 5min, 6% B; 5 to 10 min, 6% to 10% B; 10 to 20 min, 10% to 11% B; 20 to 25 min, 11% to 12.5% B; 25 to 45 min, 12.5% to 37% B; 45 to 48 min, 37% to 100% B; 48 to 58 min, 100%, 58 to 60 min 100% to 6% B. Absorbance data were collected at 280 nm. Identifications were based on comparison of chromatographic behavior and UV spectra with authentic

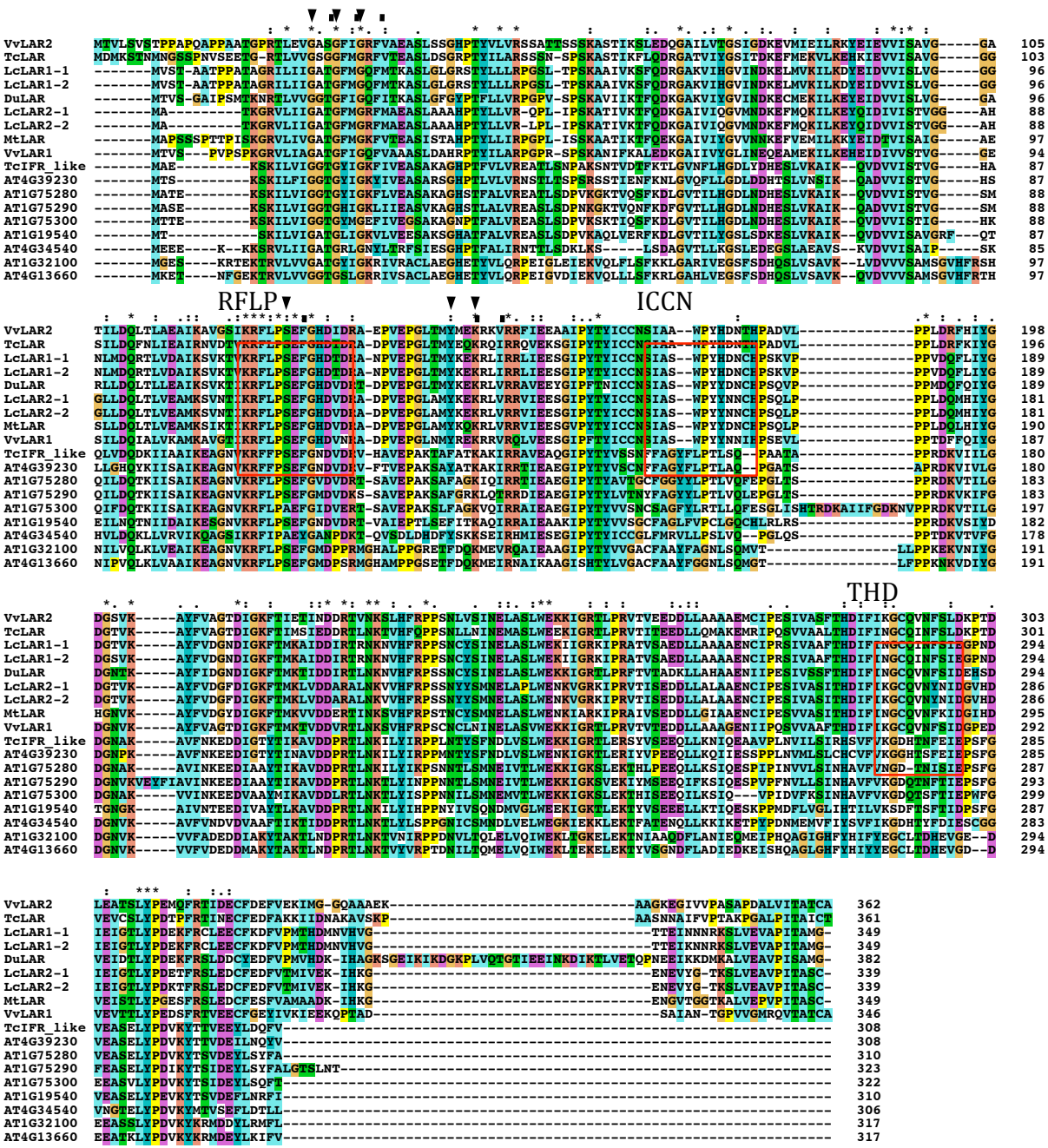
standards. This work was performed in collaboration with Yongzhen Pang and Richard A. Dixon at Samuel Roberts Nobel Foundation (Ardmore, Oklahoma, USA).

### 3.3 Results

#### 3.3.1 Sequence Analysis of the TcLAR protein

A cDNA containing a predicted full-length ORF of TcLAR was isolated as described in Chapter 2. In order to examine the relationship between *LAR* genes from different species and the closely related IFR proteins, a sequence comparison was performed using 7 *LAR* proteins from various species and 8 IFR-like proteins from *Arabidopsis*. Although no *LAR* sequence is thought to exist in *Arabidopsis*, it seems reasonable that one of the related IFR-like sequences might encode a protein with *LAR* activity. Sequences of *LAR* proteins from various species were obtained from GenBank, including two from grape (*VvLAR1* and *VvLAR2*), four from *Lotus* (*LcLAR1-1*, *LcLAR1-2*, *LcLAR2-1* and *LcLAR2-2*), one from *Desmodium* (*DuLAR*) and one from *Medicago* (*MtLAR*). The TcLAR protein sequence was used to perform BLAST (tBLASTn) analysis on the cacao EST collection (Argout et al., 2008). This analysis showed that the highest matching sequence other than TcLAR itself was annotated as an IFR gene in the database (CL222Contig1, E-value=3e-65); the predicted protein sequence was designated as TcIFR-like protein.

Multiple protein sequence alignment of the *LAR* and IFR-like proteins from cacao and *Arabidopsis* revealed several highly conserved amino acid motifs that are present in all *LAR* proteins but absent from all IFR proteins as previously described (a 13-amino acid RFLP motif, an 11-amino acid ICCN motif and a 10-amino acid THD motif) (Tanner et al., 2003; Bogs et al., 2005) (Figure 3-1). The TcLAR protein was highly conserved in all the *LAR*-specific motifs, and the sequences following these motifs were identical in *LcLAR1-1*, *LcLAR1-2* and TcLAR. The NADP<sup>+</sup> binding motif Gly-X-X-X-Gly-X-Gly is conserved in both *LAR* and IFR-like predicted cacao proteins, but the catalytic site Ser-Tyr-Lys is highly conserved only in *LAR* proteins but absent from all of the IFR-like proteins except AT4G39230.

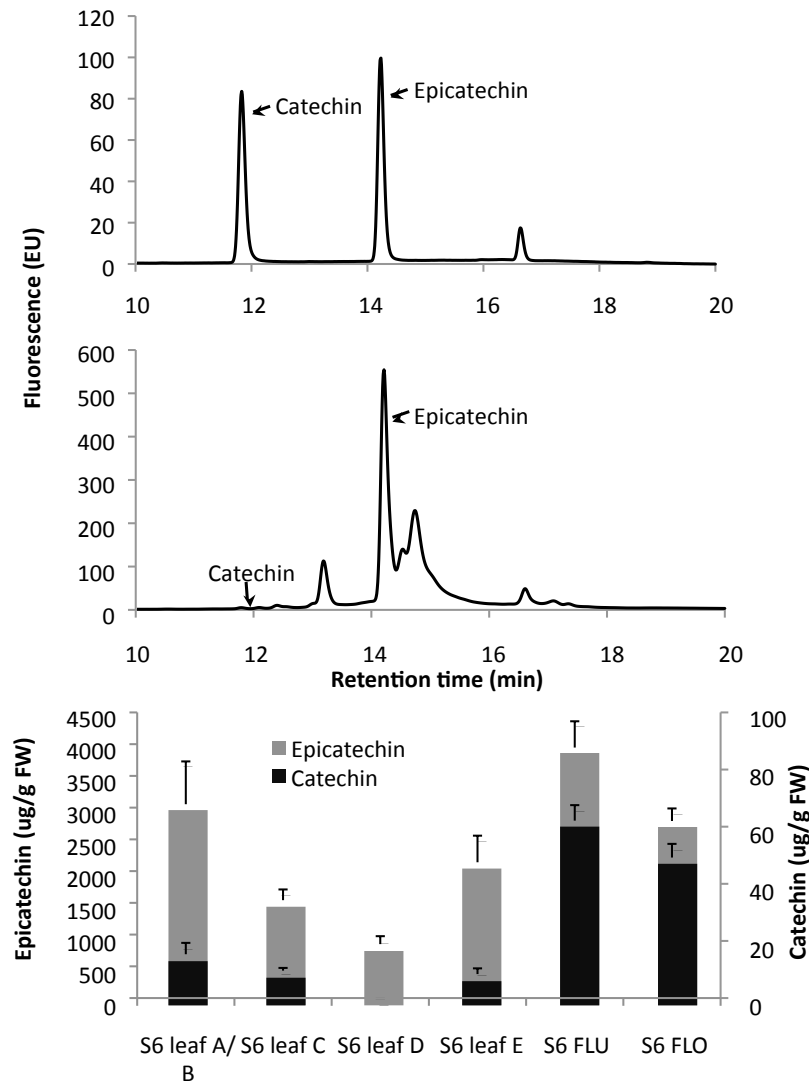


**Figure 3-1:** Multiple sequence alignment of LAR and IFR proteins. The alignment includes the cacao LAR protein as well as a cacao IFR-like protein (CU53395; CU508598). The alignment also includes all the LAR proteins characterized from various species in previous publications. The species represented, the gene names, and their GenBank accession numbers are *Vitis vinifera* (VvLAR1, CAI26309; VvLAR2, CAI26308), *Desmodium uncinatum* (DuLAR, CAD79341), *Medicago truncatula* (MtLAR, CAI56327) and *Lotus corniculatus* (LcLAR1-1, ABC71324; LcLAR1-2, ABC71325; LcLAR2-1, ABC71330; LcLAR2-2, ABC71331). All related IFR-like proteins from the Arabidopsis genome that are most closely related to the LAR protein are also included in the alignment and are designated by their locus number. The multiple protein alignment was performed using ClustalX 2.0.11 with default parameters. The RFLP, ICCN and THD motifs are boxed. The catalytic amino acids are labeled with black arrows.

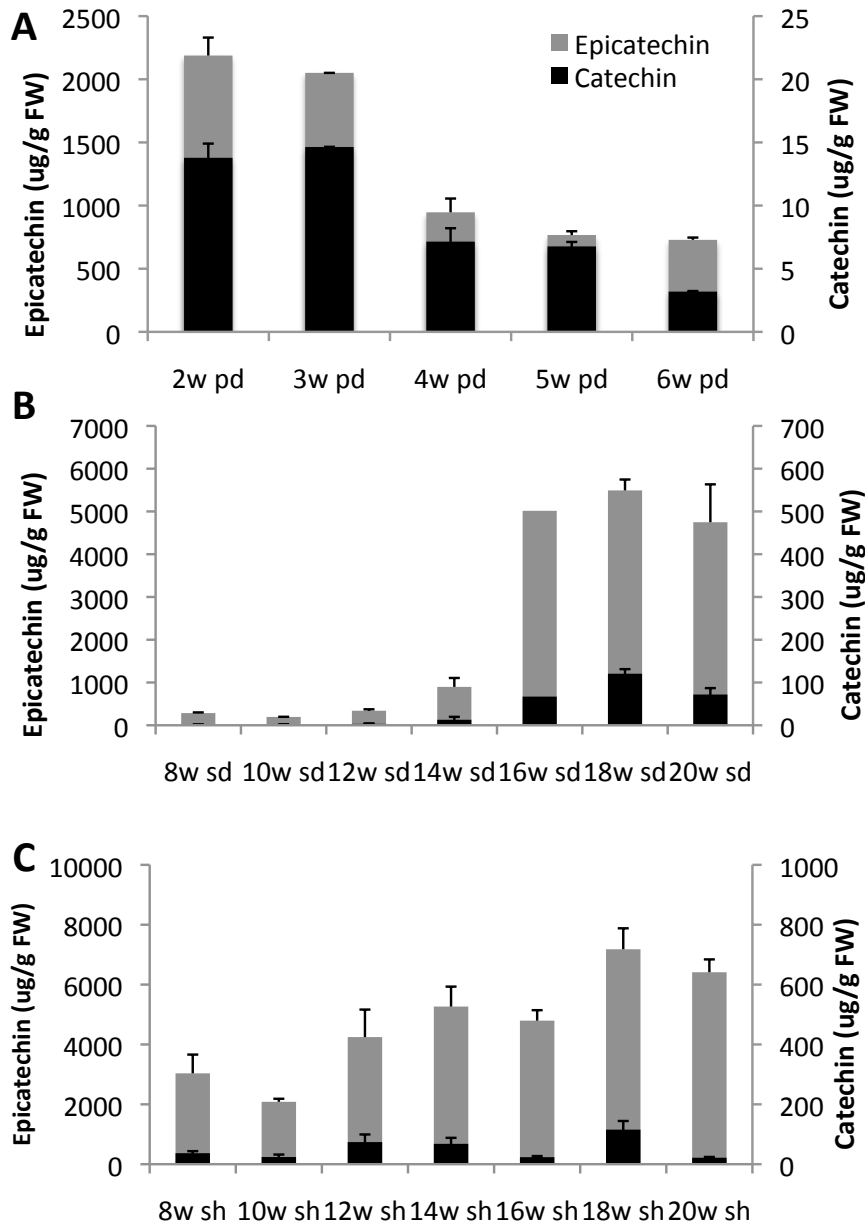
### 3.3.2 Nature of *Theobroma cacao* PAs

To determine the composition of PAs in cacao, soluble PAs extracts prepared from cacao leaves, flowers, young pods, pod exocarps and ovules at different developmental stages were separated and quantified by HPLC. The data are presented for catechin and epicatechin concentrations as  $\mu\text{g}$  of catechin or epicatechin per g of fresh tissues (Figure 3-2, Figure 3-3). In all the tissues examined, the accumulation patterns at different developmental stages of both catechin and epicatechin followed the same trend of total PA accumulation shown in Figures 2-5 to 2-8 in Chapter 2. Briefly, in cacao leaves and flowers, both PA monomers (catechin and epicatechin) accumulated at higher levels in young leaves, maximizing at about 3000  $\mu\text{g}$  of epicatechin and 15  $\mu\text{g}$  of catechin per g of fresh tissues, while mature leaves had lower levels of both catechin and epicatechin. Stage D leaves had the lowest concentration of PA monomers—epicatechin was detected at about 1000  $\mu\text{g}$  per g of fresh tissues and catechins were not detected. There were also significant levels of both PA monomers in flowers, and their concentrations were slightly higher in unopened flowers than in opened flowers. In young pods (from 2 WAP to 6 WAP), both catechin and epicatechin reached maximum concentrations of about 2000  $\mu\text{g}$  and 15  $\mu\text{g}$ , respectively, per g of fresh tissue at the earliest time point sampled (2 WAP). Then the concentrations of both PAs decreased slightly over time to a minimum of about 1000  $\mu\text{g}$  of epicatechin and 5  $\mu\text{g}$  of catechin per g of fresh tissue. Pods at later stages of development (from 8 WAP to 20 WAP) were dissected into exocarps and ovules. In exocarps, both epicatechin and catechin levels remained relatively constant from 8 WAP to 14 WAP at about 700  $\mu\text{g}$  and 50  $\mu\text{g}$ , respectively, per g of fresh tissue. Subsequently, both PA monomers significantly increased from the onset of fruit ripening (16 WAP) and remained high (about 5000  $\mu\text{g}$  of epicatechin and 100  $\mu\text{g}$  of catechin per g of fresh tissue) until maturation at 20 WAP. In ovules, the PA monomer concentrations were even higher than in exocarps, starting from 3000  $\mu\text{g}$  of epicatechin and 30  $\mu\text{g}$  of catechin per g of fresh tissue at 8 WAP and gradually increasing over time, reaching maximum concentrations of about 8000  $\mu\text{g}$  of epicatechin and 120  $\mu\text{g}$  of catechin per g of fresh tissue at 18 WAP. The data also

showed that the relative abundance of these PAs was similar in all cacao tissues examined. PAs consisted almost entirely of epicatechin with only 0.5% to 2% catechin.



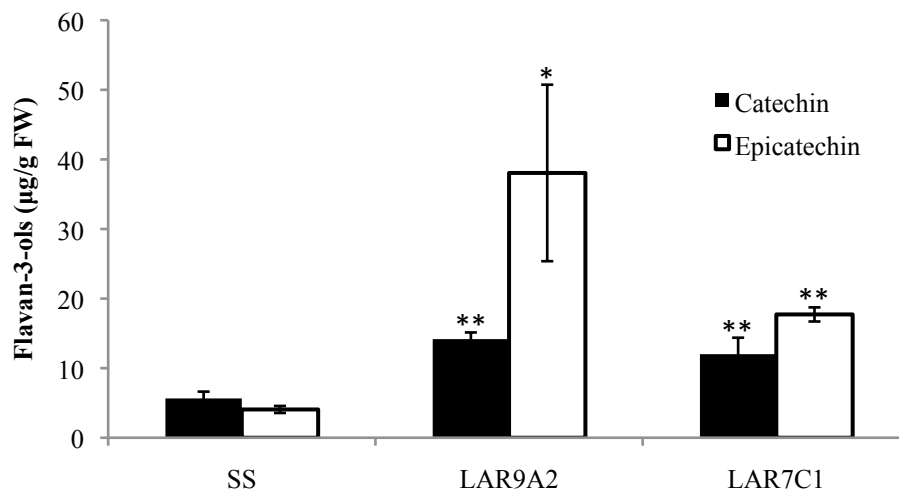
**Figure 3-2:** Composition of PAs in leaves and flowers of *Theobroma cacao*. A, Catechin and epicatechin standards analyzed by HPLC. B, Representative HPLC analysis of flavan-3-ols extracted from cacao stage A/B leaves. C, Flavan-3-ol accumulation and composition in leaves and flowers at various stages of development analyzed by HPLC. Data are presented as means  $\pm$  SE,  $n \geq 5$ . FW, Fresh weight.



**Figure 3-3:** Flavan-3-ol accumulation and composition in fruit tissues of *Theobroma cacao* (Amelonado) at various developmental stages analyzed by HPLC. A, Whole pods (pd) during early stages of pod development (from 2 to 6 WAP). B, Isolated ovules/seeds (sd) at various stages of development (from 8 to 20 WAP). C, Pod exocarp (sh) tissues at various stages of development (from 8 to 20 WAP). Data are presented as means  $\pm$  SE,  $n \geq 3$  except for 3 WAP pod and 16 WAP seeds where only 2 biological replicates and 1 biological replicate respectively were performed due to sample limitations. FW, Fresh weight.

### 3.3.3 Functional Characterization of *TcLAR*

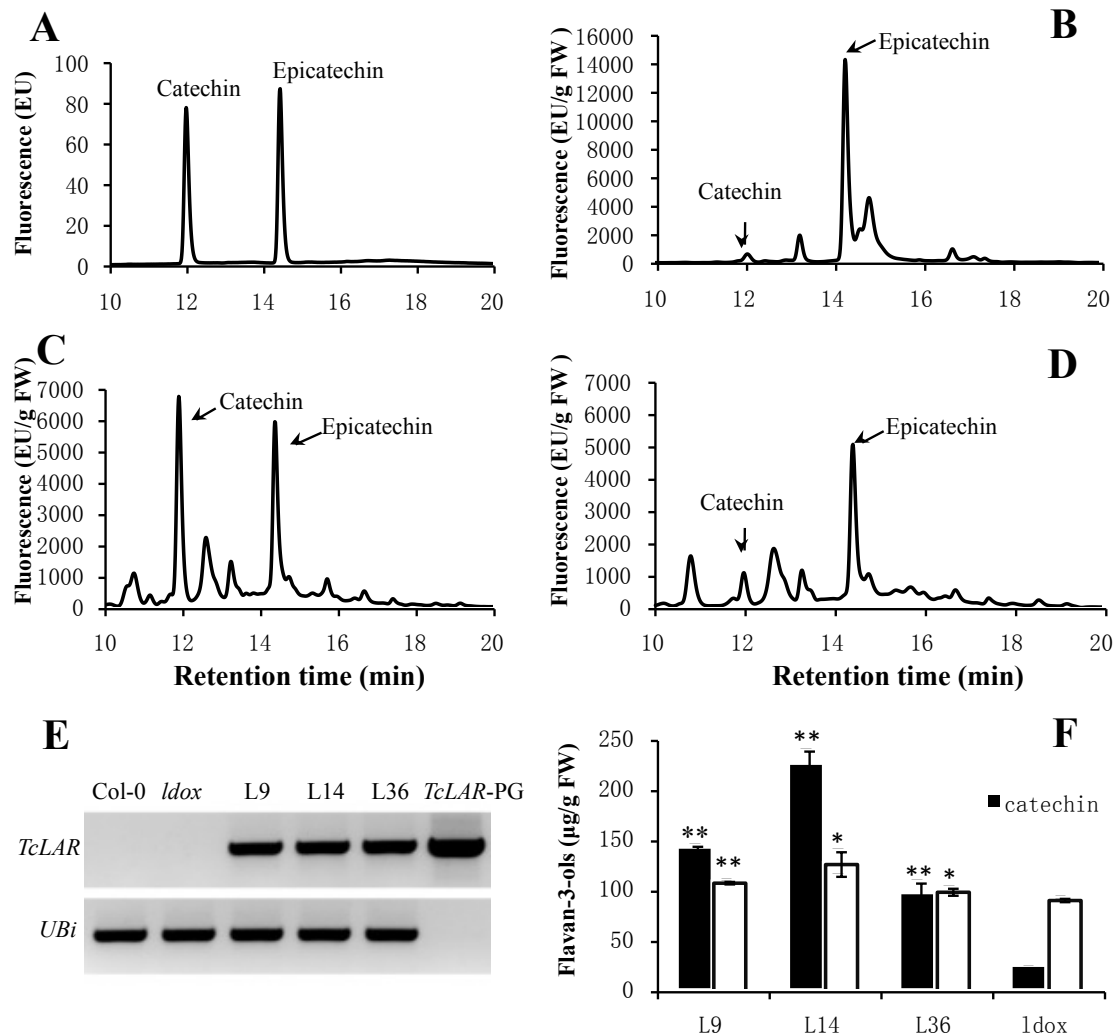
To further investigate the function of *TcLAR*, HPLC was used to analyze the PAs extracted from transgenic tobacco flowers over-expressing *TcLAR* as described in Chapter 2. Since LAR functions in the production of catechin as evidenced by biochemical assays, I predicted that the increased level of PAs in transgenic tobacco resulted from the increased synthesis of catechin. In flower extracts of both transgenic tobacco lines (LAR7C1 and LAR9A2) there was indeed a significant increase of catechin levels. However, there was also a significant increase of epicatechin levels (Figure 3-4).



**Figure 3-4:** Flavan-3-ol accumulation and composition in flowers of wild type (Samsun, ss) tobacco plants and two independent lines of 35S:*TcLAR* transgenic (LAR7C1, LAR9A2) tobacco plants. Flavan-3-ol levels were determined by extraction, HPLC separation, and quantification. All the columns are presented as means  $\pm$  SE,  $n \geq 7$ . \* $P < 0.05$  versus WT, \*\* $P < 0.01$  versus WT. FW, Fresh weight.

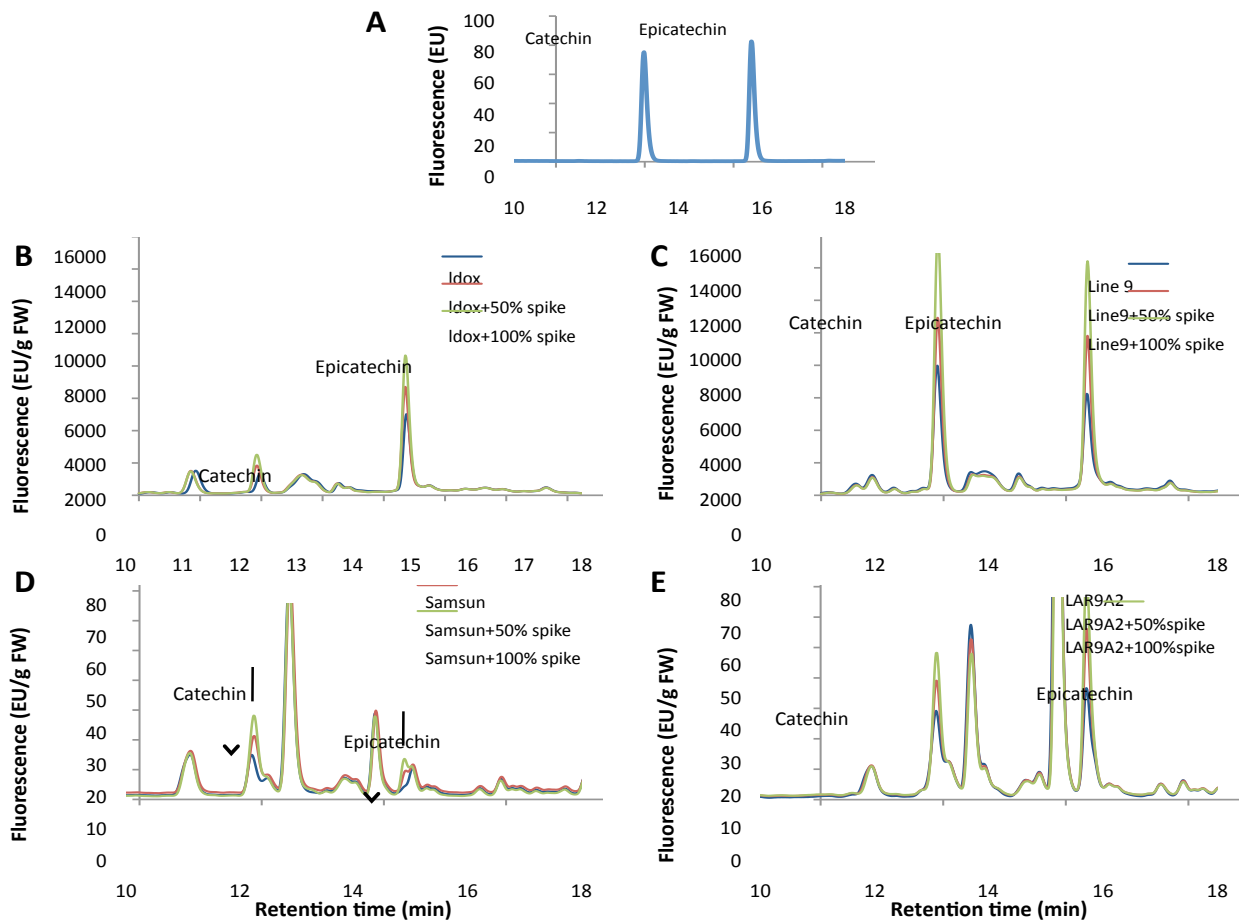


I also took advantage of the Arabidopsis *ldox* T-DNA mutant to examine *LAR* function. Because Arabidopsis lacks an *LAR* gene (Abrahams et al., 2003; Tanner et al., 2003; Lepiniec et al., 2006) and the *ldox* mutant is deficient in cyanidin (the substrate for *ANR*), the *ldox* mutant exhibits a significant decrease of epicatechin and PAs synthesis. I reasoned that since the *ldox* mutant accumulates leucoanthocyanidin which can provide the substrate for a heterologous *LAR* protein, the *ldox* mutant was potentially a good *in vivo* model to analyze *LAR* function. I predicted that over-expressing *TcLAR* should result in the synthesis of catechin in developing siliques of the *ldox* mutant Arabidopsis, even if PAs were not produced due to lack of epicatechin synthesis. HPLC separation and quantification of PA extracts from Arabidopsis transgenic lines over-expressing *TcLAR* gene in the *ldox* mutant background revealed not only a significant increase of catechin, but also surprisingly, a modest yet significant increase of epicatechin (Figure 3-5). Quantification of total PAs extracted from mature seeds also revealed significant PAs increases in all transgenic lines compared to the *ldox* mutant (data not shown).



**Figure 3-5:** Complementation of the Arabidopsis PA-deficient *ldox* mutant by constitutively expressing *TcLAR*. A, Catechin and epicatechin standards analyzed by HPLC. B, HPLC analysis of Col-0 young siliques. C, HPLC analysis of *ldox* 35S:*TcLAR* Line14 young siliques. D, HPLC analysis of *ldox* mutant young siliques. E. *TcLAR* and *AtUbiquitin* transcripts in total RNA from leaves of Col-O, *ldox* mutant and *ldox* 35S:*TcLAR* transgenic plants, (line9, line 14 and line36), by RT-PCR. PCR products from the *TcLAR*-PGEM plasmid were loaded on the last lane as a positive control for the *TcLAR* primer set and as a negative control for the *AtUbiquitin* primer set. F, Catechin and epicatechin levels in young siliques of plants that were the source of the total RNA used in (E). Catechin and epicatechin levels were determined by extraction and HPLC analysis. The data are presented as means  $\pm$  SE, n=3. \*P < 0.05 versus *ldox*; \*\*P < 0.01 versus *ldox*. FW, fresh weight; EU, emission units.

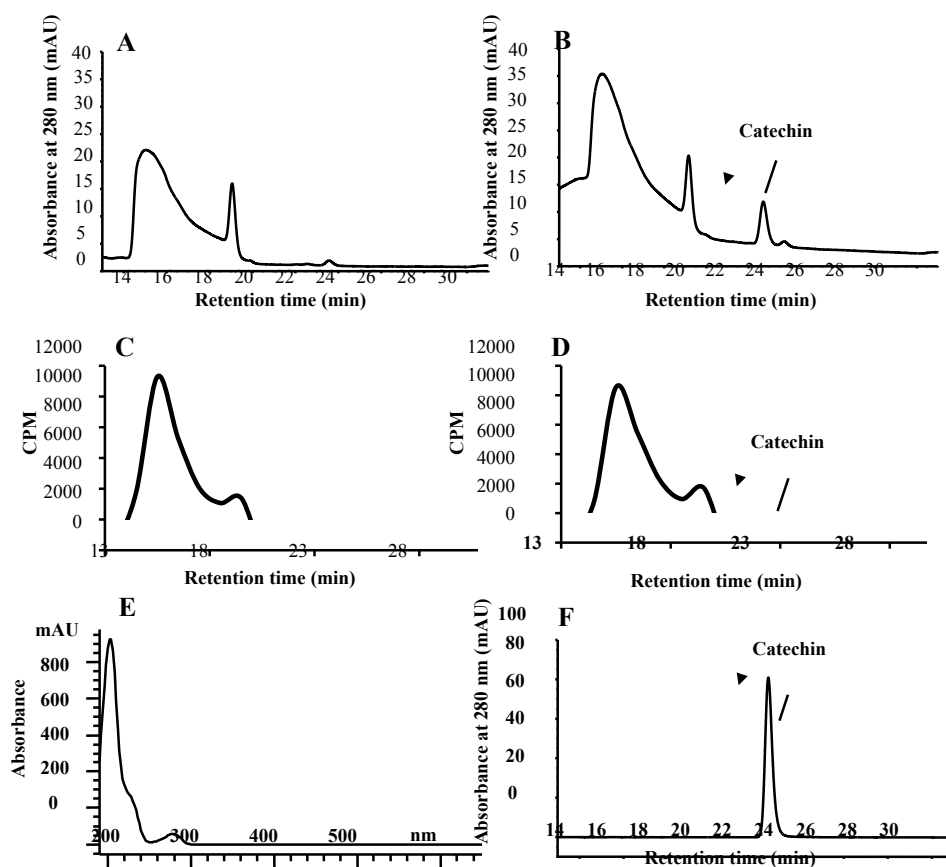
To verify the identity of peaks identified in the HPLC analysis, several key samples were spiked with known concentrations of catechin and epicatechin standards. As shown in Figure 3-6, for samples spiked with 50% and 100% of catechin and epicatechin, there is an obvious increase of the peaks that were identified as catechin and epicatechin in both *Arabidopsis* and tobacco samples. On the other hand, all the other minor peaks observed on the chromatograms, including the shoulder peaks near the catechin and epicatechin peaks, were of similar heights as in unspiked samples.



**Figure 3-6:** HPLC analysis of extracts from transgenic *Arabidopsis* and tobacco samples with and without spiking with known standards to validate the identity of the catechin and epicatechin peaks. A, Catechin and epicatechin standards used in spiking. B, Young siliques extracts from *Arabidopsis ldox* mutant with and without standard spikes. C, Young siliques extracts from transgenic *Arabidopsis* 35S:*TcLAR* Line9 with and without standard spikes. D, Flower petal extracts from wildtype tobacco (Samsun) with and without standard spikes. E, Flower petal extracts from transgenic tobacco 35S:*TcLAR*:Samsun Line LAR9A2 with and without standard spikes. FW, fresh weight; EU, emission units.

### 3.3.4 In Vitro Enzyme Assay of TcLAR

Recombinant cacao LAR protein was expressed and purified from *E.coli* and then assayed using  $^3\text{H}$ -labeled leucocyanidin as substrate in the presence of NADPH, followed by analysis of products by HPLC-UV and radioactivity detection. The negative control reaction (boiled protein) showed no product formation (Figure 3-7 A and C), whereas a peak with the same retention time and UV-spectrum as that of the pure (-)-catechin standard (Figure 3-7 E and F) was detected when TcLAR was incubated with  $^3\text{H}$ -labeled leucocyanidin (Figure 3-7, B and D). However, no epicatechin product was detected (retention time of 26-27 min).



**Figure 3-7:** Enzymatic assay of recombinant TcLAR. A, HPLC analysis of products from incubation of leucocyanidin with boiled protein as control, monitored by UV spectroscopy. B, as above, for incubation of purified recombinant TcLAR with  $^3\text{H}$ -leucocyanidin. C, As in (A), but monitoring radioactivity of fractions; D, As in (B), but monitoring radioactivity of fractions; (E), UV spectrum of catechin product from (B); F, HPLC analysis of an authentic standard of (-)-catechin. mAU, milliabsorbance units; CPM, counts per minute.

### 3.4 Discussion

PA composition in relation to expression of PA synthesis genes has been described in various plant species including Arabidopsis, apple, *Medicago*, and grape (Devic et al., 1999; Abrahams et al., 2003; Bogs et al., 2005; Takos et al., 2006; Pang et al., 2007). Arabidopsis lacks an *LAR* gene but has an *ANR* gene that is expressed spatially and temporally with PA synthesis. Correspondingly, Arabidopsis PAs are composed exclusively of epicatechin (Devic et al., 1999; Abrahams et al., 2003; Tanner et al., 2003; Routaboul et al., 2006). In other plants such as apple, *Medicago* and grape, in which both *ANR* and *LAR* genes are present and expressed, the PA (epicatechin) accumulation patterns are closely correlated with *ANR* expression (Bogs et al., 2005; Takos et al., 2006; Pang et al., 2007). Expression levels of *LAR* genes, on the other hand, show a rather poor correlation with catechin accumulation in apple and grape, in which PAs are composed of both catechin and epicatechin. Moreover, in *Medicago*, although the *LAR* gene is expressed, there is a lack of catechin and the PAs are composed almost entirely of epicatechin. This poor correlation brought into question the role of *LAR* in PA synthesis and its precise function.

In the Chapter 2, it was shown that in cacao, *ANR* and *LAR* genes are co-regulated and that their expression correlated well with PA accumulation in various tissues, suggesting significant roles in PA synthesis for both *ANR* and *LAR*. In this chapter, I examined the composition of PAs and further explored the individual contributions of these proteins to catechin and epicatechin synthesis. HPLC analysis showed that cacao PAs in all tissues examined were composed almost entirely of epicatechin units with catechin units comprising only 0.5% to 2%. This result is consistent with previous work in cacao beans (Gu et al., 2002), in which PA extension units were reported to be composed exclusively of epicatechin, and terminal units were composed mostly of epicatechin with about 1% catechin. The close correlation of *LAR* expression and PA accumulation, especially epicatechin accumulation, suggests the possibility that *LAR* may also contribute to epicatechin synthesis in addition to catechin synthesis.

The catalytic activities of LAR proteins from all plant species discussed above except apple have been verified by *in vitro* recombinant assays (Tanner et al., 2003; Bogs et al., 2005; Pang et al., 2007). *Lotus* LAR showed weak activity in a recombinant DFR protein coupled assay (Paolucci et al., 2007). However, none of those enzyme assays was able to recover sufficient flavan-3-ol products for stereochemical analysis in order to determine whether they produce catechin or epicatechin. Although Tanner et al. (2003) was able to distinguish the product (+)-catechin (2R, 3S) from (-)-catechin (2S, 3R) on a chiral HPLC system, they did not discriminate between catechin and epicatechin (Tanner et al., 2003). In my results, *TcLAR* showed *in vivo* activity in both transgenic tobacco and Arabidopsis plants. But to my surprise, ectopic expression of *TcLAR* resulted in elevated levels of both catechin and epicatechin units in tobacco flowers, in contrast to the prediction that only catechin should be formed. Based on the sequence similarity between LAR and ANR, it is possible that LAR could function redundantly to ANR and convert cyanidin to epicatechin. Alternatively, LAR may perform as a dual functional enzyme and convert leucoanthocyanidin to both catechin and epicatechin. To test the possibility that LAR performs the same activity as ANR and uses cyanidin as a substrate to form epicatechin, I took advantage of the Arabidopsis *ldox* (*ans*) mutant. Arabidopsis does not have an LAR gene and synthesizes only epicatechin through the ANS and ANR pathway. Although Arabidopsis has eight IFR-like genes that are homologous to LAR, LAR-specific protein motifs and the C-terminal extension, revealed by deduced protein sequence comparison, distinguished the LAR protein family from the IFR protein family (Figure 3-1). Arabidopsis *ldox* mutants have a significantly reduced PA (epicatechin) level due to lack of a supply of cyanidin, the epicatechin precursor. Thus, the *ldox* mutant provides a system that has leucoanthocyanidin but not cyanidin. Over-expression of *TcLAR* in the *ldox* mutant resulted in high-level formation of catechin, confirming its enzymatic function in converting leucoanthocyanidin to catechin. Moreover, HPLC quantification showed that there is also a significant elevation of epicatechin levels, suggesting that LAR may also convert leucoanthocyanidin to epicatechin.

LAR proteins belong to the reductase/epimerase/dehydrogenase (RED) superfamily (Tanner et al., 2003). This protein superfamily contains the short-chain dehydrogenase

family that is composed of a set of proteins about 250 amino acids long homologous to 20- $\beta$ -hydroxysterol dehydrogenase (HSD), the NDP-sugar epimerases and some medium-chain dehydrogenases/reductases of about 330-370 amino acids that are homologous to UDP-galactose-4-epimerase (Labesse et al., 1994). Although these proteins have low sequence similarities (about 15% identity), distinct functions and specific substrates, they commonly feature a single domain comprising both the NAD(P)<sup>+</sup>-binding site and the catalytic site (Thoden et al., 1997; Rizzi et al., 1998; Kavanagh et al., 2008). The functional N-terminal domain displays a six- to seven-stranded Rossmann fold (Rao and Rossmann, 1973) that binds NAD(P)<sup>+</sup> at the Gly-X-X-X-Gly-X-Gly motif. The catalytic triad Ser124-Tyr149-Lys153 (numbers refer to *E. coli* UDP-galactose 4-epimerase) is also highly conserved throughout the RED protein family including dehydrogenases, reductases and epimerases. The common catalytic mechanism shared by RED superfamily enzymes is a hydride and proton transfer involving NAD(P)<sup>+</sup> and typically an active tyrosine (Tyr) residue in the N-terminal domain, whereas variable C-terminal domains determine substrate specificity (Kavanagh et al., 2008). Theoretically, proteins possessing active NAD(P)<sup>+</sup>-binding sites and the Ser-Tyr-Lys catalytic triad have the potential to possess reductase activity, epimerase activity or dual activity. For example, two RED family proteins, the anthocyanidin reductase from grape and the GDP-4-keto-6-deoxy-D-mannose epimerase/reductase from *E. coli* have been shown to possess the dual reductase/epimerase activity (Rizzi et al., 1998; Gargouri et al., 2009). Crystal structural analysis of these proteins revealed that they are conserved in both the Gly-rich motif and the Ser-Tyr-Lys catalytic triad. Sequence analysis of TcLAR and other LAR proteins revealed that they are highly conserved in both the Gly-rich motif and the Ser-Tyr-Lys catalytic triad (Figure 3-1). This suggests that LAR also has the potential to possess dual reductase/epimerase activity based on the domain conservation.

However, when recombinant TcLAR was assayed *in vitro*, while we observed the production of catechin from leucoanthocyanidin, we did not detect even trace amounts of epicatechin. Our assay and detection methods are sufficiently sensitive to detect such dual-functionality as demonstrated by the assay of recombinant ANR enzyme, using cyanidin as the substrate in a single enzymatic reaction in which the production of both catechin and epicatechin was detected (data not shown). The failure to detect epicatechin in assays with

the cacao LAR enzyme suggests that it functions solely in converting leucoanthocyanidin to catechin. However we cannot rule out the possibility that the synthesis of epicatechin by TcLAR requires additional co-factors that are not present in the *in vitro* enzyme assay.

An alternative explanation for the increase in epicatechin in *TcLAR* over-expressing Arabidopsis and tobacco plants is the possible existence of a gene(s) encoding a catechin-epicatechin epimerase. However, this is less likely to occur in Arabidopsis because Arabidopsis itself does not synthesize catechin; it would be surprising if a catechin-epicatechin-specific epimerase were still present even in the absence of its substrate. Nevertheless, we cannot exclude the possible existence of such an epimerase that may also serve other functions important in plant productivity. A second alternative hypothesis to explain epicatechin formation would be racemization of catechins by polymerization to proanthocyanidins followed by nonstereo-specific depolymerization. It is also possible that the production of catechin push the metabolic flow to the epicatechin branch as a feedback effect to balance the relative amount of catechin and epicatechin. To test these possibilities, future investigations could determine whether there is a decrease in the concentration of polymeric catechin derivatives corresponding to the increase of free epicatechin. Also, it would be interesting to measure *ANR* gene transcript levels and *ANR* protein activities in *TcLAR* over-expressing transgenic plants to determine whether the increased synthesis of epicatechin is a result of increased activity of the *ANR* pathway.

### **3.5 Conclusions**

In this chapter, I further explored the function of TcLAR by monitoring the production of catechin using both *in vitro* recombinant enzyme assays and *in vivo* functional analysis in transgenic tobacco and Arabidopsis plants over-expressing *TcLAR*. The results clearly suggest that TcLAR is indeed an LAR protein and is active in converting leucoanthocyanidin to catechin.

In addition, the *in vivo* analysis showed that in *TcLAR* over-expressing transgenic plants, there is an increase of not only catechin but also epicatechin. The synthesis of



epicatechin could be the result of an endogenous catechin-to-epicatechin epimerase, the nonstereo-specific depolymerizaion of catechin, or a feedback effect of excessive catechin synthesis that diverts the metabolic flow to epicatechin. Further investigation is needed to fully understand the mechanism of increased epicatechin production.

### 3.6 References

- Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR** (2003) The *Arabidopsis* *TDS4* gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *Plant J* **35**: 624-636
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Loor G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C** (2008) Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. *BMC Genomics* **9**: 512
- Bogs J, Downey MO, Harvey JS, Ashton AR, Tanner GJ, Robinson SP** (2005) Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. *Plant Physiol* **139**: 652-663
- Bogs J, Jaffe FW, Takos AM, Walker AR, Robinson SP** (2007) The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiol* **143**: 1347-1361
- Bradford MM** (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**: 248-254
- Devic M, Guilleminot J, Debeaujon I, Bechtold N, Bensaude E, Koornneef M, Pelletier G, Delseny M** (1999) The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. *Plant J* **19**: 387-398
- Foo LY, Newman R, Waghorn G, McNabb WC, Ulyatt MJ** (1996) Proanthocyanidins from *Lotus corniculatus*. *Phytochemistry* **41**: 617-624
- Gargouri M, Manigand C, Mauge C, Granier T, Langlois d'Estaintot B, Cala O, Pianet I, Bathany K, Chaudiere J, Gallois B** (2009) Structure and epimerase activity of anthocyanidin reductase from *Vitis vinifera*. *Acta Crystallogr D* **65**: 989-1000
- Gu L, Kelm M, Hammerstone JF, Beecher G, Cunningham D, Vannozzi S, Prior RL** (2002) Fractionation of polymeric procyanidins from lowbush blueberry and quantification of procyanidins in selected foods with an optimized normal-phase HPLC-MS fluorescent detection method. *J Agric Food Chem* **50**: 4852-4860
- Kavanagh K, Jornvall H, Persson B, Oppermann U** (2008) The SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell Mol Life Sci* **65**: 3895-3906
- Kristiansen KN** (1986) Conversion of (+)-dihydroquercetin to (+)-2,3-trans-3,4-cis-leucocyanidin and (+)-catechin with an enzyme extract from maturing grains of barley. *Carlsberg Res. Comm* **51**: 51-60
- Labesse G, Vidalcros A, Chomilier J, Gaudry M, Mornon JP** (1994) Structural comparisons lead to the definition of a new superfamily of NAD(P)(H)-accepting

- oxidoreductases: the single-domain reductases/epimerases/dehydrogenases (the RED family). *Biochem J* **304**: 95-99
- Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M** (2006) Genetics and biochemistry of seed flavonoids. *Annu Rev Plant Biol* **57**: 405-430
- Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH** (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205-210
- Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA** (2007) Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiol* **145**: 601-615
- Paolucci F, Robbins MP, Madeo L, Arcioni S, Martens S, Damiani F** (2007) Ectopic expression of a basic helix-loop-helix gene transactivates parallel pathways of proanthocyanidin biosynthesis. structure, expression analysis, and genetic control of leucoanthocyanidin 4-reductase and anthocyanidin reductase genes in *Lotus corniculatus*. *Plant Physiol* **143**: 504-516
- Rao ST, Rossmann MG** (1973) Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**: 241-&
- Rizzi M, Tonetti M, Vigevani P, Sturla L, Bisso A, De Flora A, Bordo D, Bolognesi M** (1998) GDP-4-keto-6-deoxy-D-mannose epimerase/reductase from *Escherichia coli*, a key enzyme in the biosynthesis of GDP-L-fucose, displays the structural characteristics of the RED protein homology superfamily. *Structure* **6**: 1453-1465
- Routaboul JM, Kerhoas L, Debeaujon I, Pourcel L, Caboche M, Einhorn J, Lepiniec L** (2006) Flavonoid diversity and biosynthesis in seed of *Arabidopsis thaliana*. *Planta* **224**: 96-107
- Singh S, McCallum J, Gruber MY, Neil Towers GH, Muir AD, Bohm BA, Koupai-Abyazani MR, Glass ADM** (1997) Biosynthesis of flavan-3-ols by leaf extracts of *Onobrychis viciifolia*. *Phytochemistry* **44**: 425-432
- Skadhauge B, Gruber MY, Thomsen KK, Diter Von W** (1997) Leucocyanidin reductase activity and accumulation of proanthocyanidins in developing legume tissues. *Am J Bot* **84**: 494-503
- Takos AM, Ubi BE, Robinson SP, Walker AR** (2006) Condensed tannin biosynthesis genes are regulated separately from other flavonoid biosynthesis genes in apple fruit skin. *Plant Sci* **170**: 487-499
- Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ, Ashton AR** (2003) Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *J Biol Chem* **278**: 31647-31656
- Tanner GJ, Kristiansen KN** (1993) Synthesis of 3,4-cis-[<sup>3</sup>H]leucocyanidin and enzymatic reduction to catechin. *Anal Biochem* **209**: 274-277
- Thoden JB, Hegeman AD, Wesenberg G, Chapeau MC, Frey PA, Holden HM** (1997) Structural analysis of UDP-sugar binding to UDP-galactose 4-epimerase from *Escherichia coli*. *Biochemistry* **36**: 6294-6304

## CHAPTER 4: THE *THEOBROMA CACAO* *TCMYBPA* GENE ENCODES AN R2R3-MYB TRANSCRIPTION FACTOR INVOLVED IN THE REGULATION OF PROANTHOCYANIDIN SYNTHESIS

### 4.1 Introduction

The mechanisms regulating the transcription of the flavonoid biosynthetic pathway genes are well studied in several model systems like Arabidopsis (*Arabidopsis thaliana*) and maize (*Zea mays*) (Lepiniec et al., 2006). Transcriptional regulation of these genes is mediated by members of three protein families: the R2R3-MYB transcription factors, the MYC-like basic helix-loop-helix (bHLH) proteins and the WD40 repeat proteins (Ramsay and Glover, 2005; Grotewold, 2006; Lepiniec et al., 2006). The combinatorial interactions of different members from these three protein families determine their specificity of target gene activation (Winkel-Shirley, 2001). This interaction has been shown for several flavonoids synthesis regulators isolated from Arabidopsis (Nesi et al., 2001; Baudry et al., 2004; Lepiniec et al., 2006). This has also been shown for several anthocyanin synthesis regulators isolated from *Zea mays* (Grotewold et al., 2000; Allan et al., 2008), and *Petunia hybrida* (de Vetten et al., 1997; Quattrocchio et al., 1999; Spelt et al., 2000).

The flavonoid pathway and its genetic regulation have been subject to intense research using the model plant Arabidopsis (Lepiniec et al., 2006). The regulation of proanthocyanidin (PA) synthesis has been well characterized by the analysis of transparent testa (TT) mutants that fail to accumulate PAs in the seed coat (Abrahams et al., 2002; Lepiniec et al., 2006). Three TT loci, *TT2*, *TT8* and *TTG1*, which encode R2R3-MYB, bHLH and WD40 repeat proteins respectively, are necessary for proper temporal and spatial accumulation of PAs (Lepiniec et al., 2006). The three proteins interact and form a ternary transcriptional protein complex to activate “late” PA-specific genes including *DFR* (dihydroflavanol reductase), *LDOX* (leucoanthocyanidin dioxygenase) and *BAN* (*ANR*, anthocyanidin reductase) (Pelletier et al., 1997; Nesi et al., 2000; Nesi et al., 2001; Baudry et al., 2004). Another three TT loci, *TT16*, *TT1* and *TTG2* that encode a MADS box protein, a zinc-finger protein and a WRKY transcription factor, respectively, are also important for PA

synthesis (Lepiniec et al., 2006). These proteins have been shown to regulate the expression of BAN protein through a posttranscriptional mechanism and thus are involved in the differentiation of PA-accumulating cells (Lepiniec et al., 2006).

The *TT2* gene product (TT2) is a key regulator of PA synthesis and confers target gene specificity to the MYB-bHLH-WD40 complex. It is specifically expressed in PA-accumulating cells in Arabidopsis but can induce ectopic expression of the *BAN* gene when constitutively expressed in the presence of a functional TT8 protein (Nesi et al., 2001). TT2 belongs to the large R2R3-MYB protein family that has 133 members in Arabidopsis. These proteins are typically involved in many aspects of plant secondary metabolism, plant cell identity and cell fate determination (Stracke et al., 2001; Stracke et al., 2007). Members of the R2R3-MYB protein family are characterized by the presence of two highly conserved head-to-tail MYB motifs in the N-terminal region, the R2 and R3 repeats, although their C-terminal regions are very divergent. Each of the R2R3 repeats consists of three  $\alpha$ -helices (Grotewold et al., 2000); helix 3 of each motif is involved in interaction with DNA and helix 1 of the R3 repeat is important for corresponding bHLH recognition.

In addition to Arabidopsis, the TT2-like PA-specific R2R3-MYB transcription factors (TFs) have been characterized in grape (*Vitis vinifera*), *Lotus* (*Lotus japonicus*) and poplar (*Populus tremuloides*) (Bogs et al., 2007; Yoshida et al., 2008; Mellway et al., 2009; Terrier et al., 2009). In grape, two TT2-like MYB TFs (VvMYBPA1 and VvMYBPA2) have been identified. These TFs exhibit tissue-specific functions in inducing PA structural gene expression and synthesis: VvMYBPA1 is mainly expressed in grape seeds; and VvMYBPA2 is mainly expressed in exocarp of young berries and in the leaves. Similar observations were reported in *Lotus*, in which three copies of TT2-like R2R3-MYB TFs were identified that differ in organ-specific expression and responsiveness to stress. Each of the TFs mentioned above is capable of activating the *ANR* promoter in transient reporter assays. In poplar, a *MYB134* gene encoding a TT2-like TF was recently shown to be responsive to wounding, pathogen presence and UV-B irradiation, consistent with the biological roles of PAs in anti-herbivore, anti-pathogen and UV damage protection. Overexpression of MYB134 in poplar resulted in transcriptional activation of the genes encoding enzymes of

the full PA biosynthesis pathway from PAL1 to ANR and LAR , but not FLS, which is specific to flavonol synthesis.

In Chapter 2, I described the isolation of three structural genes (*TcANR*, *TcANS* and *TcLAR*) from cacao and the functional verification of their involvement in the PA synthesis pathway. In this chapter, I will describe the isolation and characterization of a cacao gene, *TcMYBPA*, that encodes an R2R3-MYB transcription factor involved in regulating the biosynthesis of cacao PAs. Constitutive expression of *TcMYBPA* in the *Arabidopsis tt2* mutant not only successfully complimented its primary phenotype (a PA-deficient seed coat) but also resulted in increased anthocyanin accumulation in young seedlings, suggesting that *TcMYBPA* regulates both the anthocyanin and PA pathways in cacao.

## 4.2 Materials and Methods

### 4.2.1 Isolation of a *TcMYBPA* cDNA from *Theobroma cacao*

Total RNA from stage A/B leaves of *Theobroma cacao* (Scavina 6) was isolated as described in Chapter 2. First strand cDNA was synthesized using the SMART RACE cDNA amplification kit (Clontech, Mountain View, CA, USA). The putative EST sequence of *TcMYBPA* was obtained by searching the *Theobroma cacao* EST database (<http://esttik.cirad.fr/>) (Argout et al., 2008) using BLAST (program: tBLASTn) (Altschul et al., 1990) with the protein sequence of *TT2* (AT5G35550) from *Arabidopsis thaliana* as the query sequence. The ORF of putative *TcMYBPA* was amplified with the Advantage cDNA PCR Kit (Clontech, Mountain View, CA, USA) using cDNA from stage A/B leaves as template with the following primer pairs: *TcMYBPA\_F* (5'- *GTCCATGGGAAAGGGCTCCTTGTGTTC* -3') and *TcMYBPA\_R* (5'- *AGCGGCCGCTCAGATCAAATAATGATTCAGC* -3'). To facilitate the subsequent cloning into binary vectors, an *NcoI* site (CCATGG) was added at the start codon (ATG) and a *NotI* site (GCGGCCGC) was added immediately 3' to the stop codon (TCA) respectively (sites are shown in italics and the start or stop codons are underlined). The PCR reaction was carried out in a total volume of 20 µL at 94°C for 5 min; 5 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min; another 23 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 1 min; followed by a final extension at 72°C for 5 min. The PCR products were

gel purified and cloned into the pGEM-T Easy plasmid (Promega, Madison, WI, USA) and replicated in *E. coli* strain DH5 $\alpha$ . DNA sequencing was performed using twelve of the resulting DNA clones (pGEMT-TcMYBPA), and two clones had the precise sequence of the consensus sequences. One clone (pGEMT-TcMYBPA-3) was chosen for cloning into the binary vector for plant transformation and subsequent experiments.

#### 4.2.2 Protein Sequence Alignment and Phylogenetic Analysis

PA-specific R2R3-MYB protein sequences were retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>), including AtTT2 from *Arabidopsis* (CAC40021) (Nesi et al., 2001), VvMYBPA1 and VvMYBPA2 from grape (AM259485, ACK56131) (Bogs et al., 2007; Terrier et al., 2009), LjTT2a, LjTT2b and LjTT2c from *Lotus japonicus* (AB300033, AB300034, AB300035) (Yoshida et al., 2008) and MYB134 from *Populus tremuloides*. (FJ573151)(Mellway et al., 2009). A protein sequence alignment performed with the ClustalW algorithm was used to construct a phylogenetic tree using the neighbor-joining method in the MEGA package (Kumar et al., 2004). One thousand bootstrap datasets were used to estimate the confidence of each tree clade. Protein sequence alignment of anthocyanin- and proanthocyanin-specific MYB proteins was performed using the same method as was for the phylogenetic tree but was edited and displayed using GENEDOC software (Version 2.6.02, <http://www.nrbsc.org/gfx/genedoc/>).

#### 4.2.3 Transformation of *Arabidopsis*

The coding sequence of *TcMYBPA* was excised from the intermediate cloning vector (pGEMT-TcMYBPA-3) with *NcoI* and *NotI* restriction enzymes and introduced into the pE2113-EGFP (Maximova et al., 2003) intermediate vector to substitute the coding sequence of *TcMYBPA* for the original EGFP coding sequence. As a result, the cacao gene coding sequence is located immediately downstream of the very strong E12- $\Omega$  promoter (a modified CaMV35S promoter) and upstream of the CaMV35S terminator. The over-

expression cassette was then excised out from pE2113 vector with *EcorI* and *PvuII* restriction enzymes and introduced into the pCAMBIA-1300 binary vector (CAMBIA, Canberra, Australia).

This binary transformation construct were introduced into *Agrobacterium tumefaciens* strain AGL1 (Lazo et al., 1991) by electroporation as previously described in Chapter 2 (Lin, 1994).

#### 4.2.4 Gene Expression Analysis

Total RNA from leaves, flowers, pods, pod exocarp and ovules of *Theobroma cacao* (Scavina 6 and Amelonado) was isolated as described above. Total RNA from young Arabidopsis seedlings was isolated using the RNeasy Plant mini kit (Qiagen, Valencia, CA, USA). cDNA was synthesized from 1 µg of total RNA in a total volume of 20 µL using M-MuLV Reverse Transcriptase (NEB, Ipswich, MA, USA) according to the supplier's protocols, and 2 µL of this reaction were used in the subsequent RT-PCR reactions.

Semi-quantitative RT-PCR was performed to measure gene expression levels as previously described (Ahn, 2009) with the following modifications: The primers for Arabidopsis cDNA span two exons, giving products of about 500 bp, and thus are mRNA specific, avoiding potential amplification from genomic DNA contamination. The primers sets used are listed in Table 4-1 below.

**Table 4-1:** Sequences of the primers used in the gene expression study

Primer	Sequence (5' to 3')
AtDFR_RTF	CCGTGTGTGTAACCGGCGCT
AtDFR_RTR	TGCGCCTCGTTCCGAGTGAT
AtLDOX_RTF	CACCAAGTGATTACATAGAAGCAACG
AtLDOX_RTR	TCACCATCTCCGGCAACGGC
AtBAN_RTF	TGGTGGCACGGGAAACTTAGCC
AtBAN_RTR	CGGTCACATGCATTTCTTTCCCGGT



AtUFGT_RTf	ACCATCGGTGTCAAAGAAGTAGGTG
AtUFGT_RTR	GCGGTGCCCATGGAACCACT
AtCHS_RTf	TCAAGCGCATGTGCGACAAGT
AtCHS_RTR	CGGCGGCGCCATCACTGAAA
AtCHI_RTf	TCATCCAACGCCTGCGCCTC
AtCHI_RTR	ACGCAACCGTAAGAGAGCCGGT
AtF3H_RTf	ACCTCCAGGGAGAGGCTGTGC
AtF3H_RTR	AAATGGCCGTGGTCGCCGAG
AtFLS_RTf	ATCTGGCCACCGTCATGCGTC
AtFLS_RTR	CCTCCATTACTCAACCTCAGAATC

---

To ensure the linearity of the semi-quantitative RT-PCR measurements, each primer set was tested in time course PCR reactions to measure amplification kinetics and to determine the optimal PCR cycle in which the reaction is well within the linear range. The number of PCR cycles was optimized from 20 to 32, and 28 cycles was chosen as optimal for all the RT-PCR reactions. The PCR reaction was carried out in a total volume of 20  $\mu$ L at 94°C for 5 min; 28 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 45 s; followed by a final extension at 72°C for 5 min. The PCR products were visualized on 1% agarose gels stained with ethidium bromide and photographed using a Molecular Imager Gel Doc XR+ System equipped with a 16-bit CCD camera (Bio-Rad Laboratories, Hercules, CA). Relative fluorescent intensity of the separated PCR products was quantified using Quantity One 1-D Analysis Software (Bio-Rad Laboratories, Hercules, CA).

#### 4.2.5 DMACA Staining of Tissues

To visualize the presence of PAs in Arabidopsis young seedlings and dry seeds, tissues were immersed in 4-dimethylaminocinnamaldehyde (DMACA) reagent (2% (w/v)DMACA, 90% ethanol, 10% HCl) for 2 days as described previously (Abrahams et al., 2002) and then washed 3 times with 70% ethanol.

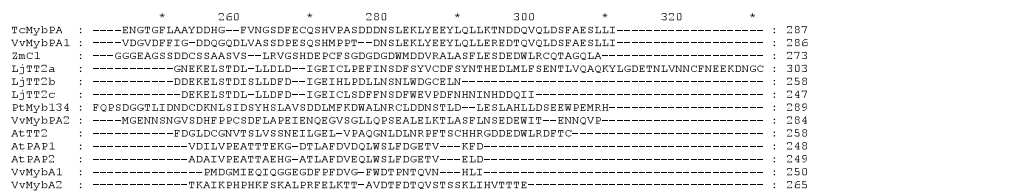
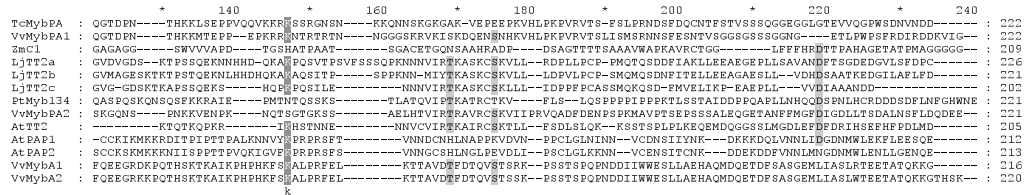
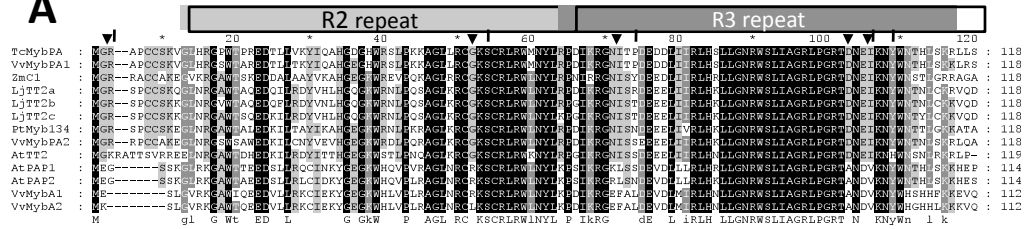
## 4.3 Results

### 4.3.1 The Cacao *TcMybPA* Gene Encodes an R2R3-MYB Transcription Factor

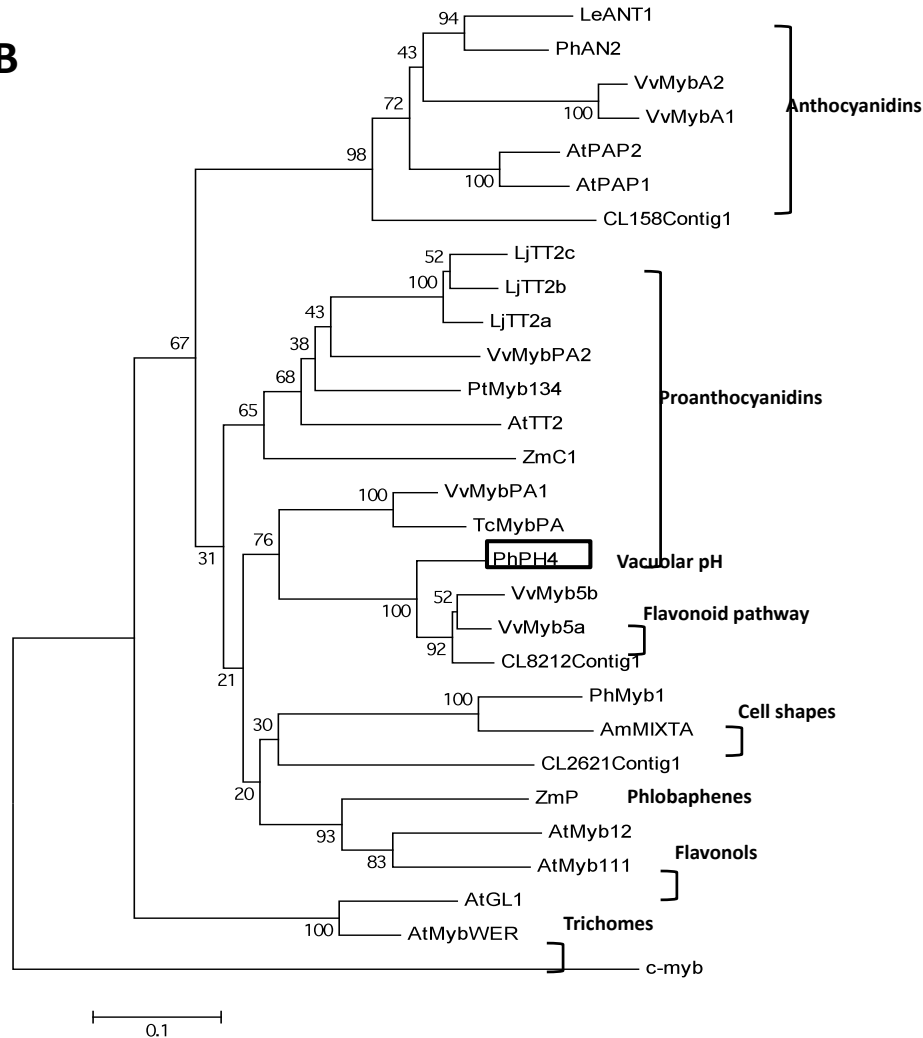
Four putative *TcMYBPA* cDNA sequences were identified in a collection of *Theobroma cacao* expressed sequence tags (ESTs) (Argout et al., 2008) by querying the cacao ESTtik database (<http://esttik.cirad.fr/>) with the protein sequence of Arabidopsis TT2 (accession no. Q2FJA2). Cacao ESTs showing sequence similarity to the *TT2* gene were assembled into a contig to recover full-length open reading frames (ORFs) by alignment with cDNAs of homologous genes from other species and predictions from the ORF Finder program ([www.ncbi.nlm.nih.gov/projects/gorf/](http://www.ncbi.nlm.nih.gov/projects/gorf/)). The full-length coding sequence of *TcMYBPA* was amplified by RT-PCR using cDNA isolated from young leaves of cacao (Scavina 6), in which PAs are actively synthesized and accumulated (Chaves and Gianfagna, 2007). The isolated ORF was named *TcMYBPA* (accession no. GU324346, see sequence in Appendix A).

The 864-bp ORF of *TcMYBPA* encodes a protein of 287 amino acids that shares 68 % identity with grape *VvMYBPA1* (see protein sequence comparison in Appendix B). A protein sequence alignment of *TcMYBPA* with other PA- and anthocyanin-regulating MYB proteins revealed that *TcMYBPA* contains an N-terminal R2R3 repeat that corresponds to the DNA-binding domain of plant MYB-type proteins (Fig. 4-1A). Like the high sequence similarity observed between the R2R3 repeat regions shared by 126 members of Arabidopsis (Stracke et al., 2001; Czemmél et al., 2009), the *TcMYBPA* R2R3 repeat region is highly conserved when compared to other plant R2R3 MYBs. The *TcMYBPA* N-terminal region also contains the [D/E]LX<sub>2</sub>[R/K]X<sub>3</sub>LX<sub>6</sub>LX<sub>3</sub>R motif for interaction with bHLH partners in the R3 repeat region (Grotewold et al., 2000), whereas the C-terminal region shows little homology to other plant MYBs.

**A**



**B**



**Figure 4-1:** Comparison of the amino acid sequences of TcMYBPA and various plant MYB transcription factors. A, Alignment of the deduced amino acid sequences of the R2R3-MYB proteins functioning in anthocyanin and PA synthesis, including TcMYBPA (cacao), ZmC1 (maize), VvMybPA1, VvMybPA2, VvMybA1, VvMybA2 (grape), PtMyb134 (poplar), LjTT2a, LjTT2b, LjTT2c (*Lotus*), and Arabidopsis regulators AtTT2, AtPAP1 and AtPAP2. The R2 and R3 repeats of the MYB domain are indicated above the alignment. Identical amino acids are indicated in black, similar amino acids in gray. Arrowheads indicate amino acids that are conserved in all PA-regulating MYBs but absent in anthocyanin-regulating MYBs. Sequences were aligned using the ClustalW program and were displayed using the GeneDoc program. B, Phylogenetic tree showing selected plant MYB transcription factors from GenBank. Human c-myb is included as a outgroup. Functions of the MYB proteins are given on the right side in bold. The alignment was performed using the ClustalW program and the tree was constructed using the neighbor-joining algorithm of the MEGA package (Version 3.1) (Kumar et al., 2004). The scale bar represents 0.1 substitutions per site and the numbers next to each node are bootstrap values from 1,000 replicates. The GenBank accession numbers of the MYB proteins are as follows: AtGL1 (P27900), ZmP (P27898), ZmC1 (AAA33482), VvMybA1 (BAD18977), VvMybA2 (BAD18978), AtPAP1 (AAG42001), PhAN2 (AAF66727), LeANT1 (AAQ55181), OsMyb4 (T02988), AmMixta (CAA55725), AtMyb111 (AAK97396), AtMyb12 (NM\_130314), PmMybF1 (AAA82943), PhPH4 (AAY51377), AtPAP2 (AAG42002), AtMybWER (CAC01874), VvMyb5a (AAS68190), VvMYB5b (Q58QD0), VvMYBPA1 (AM259485), VvMybPA2 (ACK56131), c-myb (AAB49039), PtMyb134 (FJ573151), PhMyb1 (Z13996), LjTT2a (AB300033), LjTT2b (AB300034), LjTT2c (AB300035), AtTT2 (Q2FJA2). Also included in the tree are one putative cacao PA specific MYB (TcMYBPA), and three MYB-like proteins found in the cacao EST collections (CL158Contig1, CL8212Contig1 and CL2621Contig1)

To investigate these relationships more closely, a phylogenetic tree was constructed using the full-length amino acid sequences of TcMYBPA and sequences of all functionally tested MYBs involved in regulating proanthocyanidin and anthocyanin biosynthesis, as well as MYBs associated with several other biological processes (Fig 4-1B). By searching the cacao EST database using tBLASTn with the protein sequence of putative cacao MYB TcMYBPA as the query, three EST contigs (CL8212Contig1, CL2621Contig1 and CL158Contig1) containing MYB-like proteins were also identified as the next best cacao matches to TcMYBPA. The results show that the putative cacao proanthocyanidin regulatory protein TcMYBPA is most closely related to the grape PA regulatory MYB protein VvMYBPA1 and clusters in the same clade with all the anthocyanidin and proanthocyanidin regulatory MYB proteins.

Within the same clade are also VvMYB5a and VvMYB5b from grape, which are involved in regulating the entire flavonoid pathway, and PhPH4 from petunia, which is involved in regulating vacuolar pH. R2R3-type MYB proteins that regulate other biochemical and physiological processes such as phlobaphene and flavonol synthesis, cell shape determination and trichome development clustered into separate subgroups. The other three cacao MYB-like proteins cluster together with MYBs that have functions other than proanthocyanidin regulation, such as flavonoid pathway regulation (CL8212Contig1), cell shape determination (CL2621Contig1) and anthocyanidin synthesis regulation (CL158Contig1). ZmC1, the maize anthocyanin synthesis regulator that was shown to activate the Arabidopsis *ANR* promoter (Baudry et al., 2004), clustered together in the same subgroup with AtTT2 and VvMYPPA2, which are functionally verified PA regulators. This was consistent with the protein alignment analysis in which ZmC1 was more similar to PA regulatory MYBs than to anthocyanin regulatory MYBs. Protein alignment also revealed that some conserved amino acids present in the N-terminal region of TcMYBPA as well as all PA regulatory MYB proteins and ZmC1 were absent in all the other anthocyanin MYB factors (Fig. 4-1A); this could indicate similarity in function. These included, according to position on TcMYBPA, Gly2, Arg3, Gly50, Asn69, Ile70, Asp101, Glu103, and Ile104.

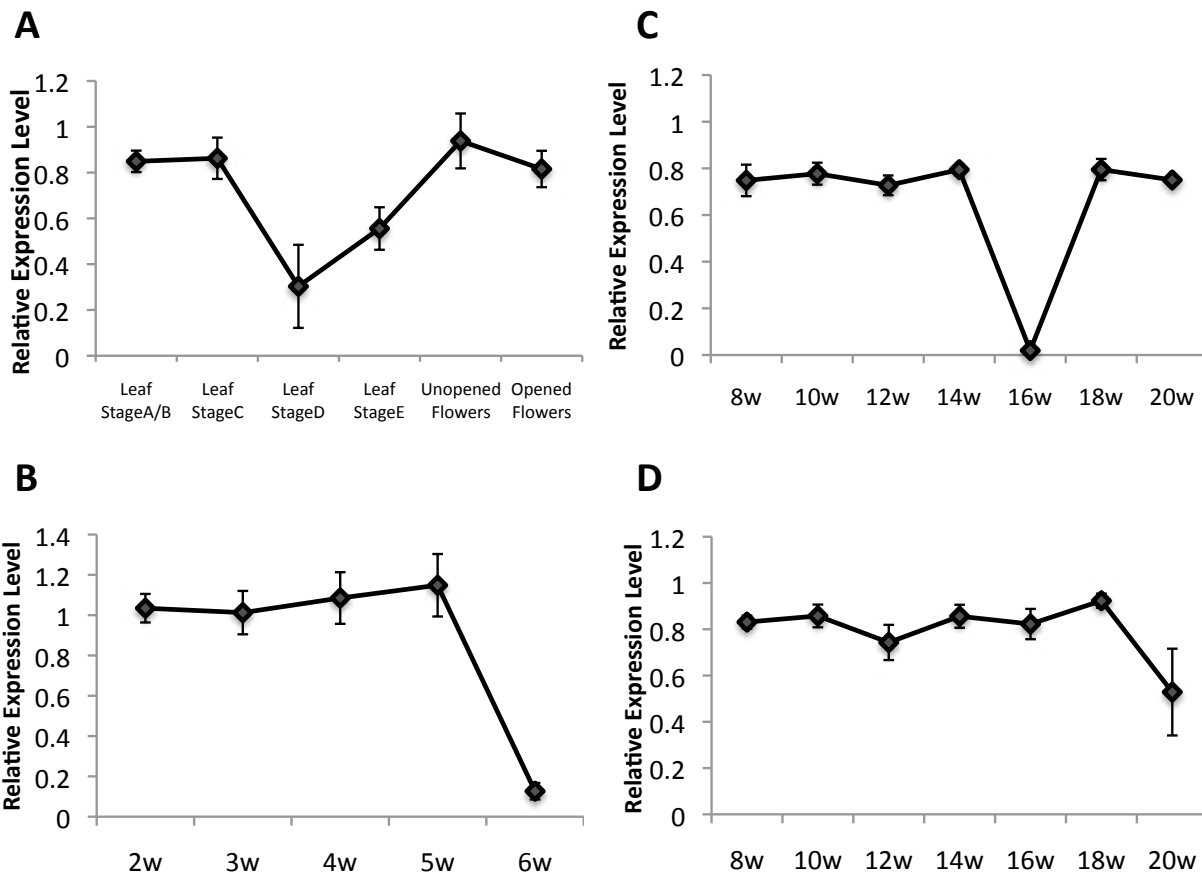
In summary, the TcMYBPA protein sequence includes conserved R2R3 regions typical of plant MYB transcription factors. Moreover, in TcMYBPA, I was able to detect conserved amino acid homologies shared with all the TT2-like MYB regulators but absent in anthocyanin regulators. These conserved amino acids can be used to identify candidate PA-specific MYB regulators from other plant species.

#### **4.3.2 Expression of *TcMYBPA* Correlates with PA Accumulation in *Theobroma cacao***

To assess the involvement of *TcMYBPA* in PA biosynthesis, the expression of the putative TcMYBPA gene was examined in tissue samples from different developmental stages of leaves, flowers and pods in which PAs accumulate. The leaf and flower tissues were from cacao genotype Scavina 6. Because Scavina 6 is a self-incompatible genotype, to reduce the genetic variability in pods, an alternative genotype, Amelonado, was used for pod tissue collection. Gene transcript levels were assessed by semi-quantitative RT-PCR.

*TcActin* was chosen as a reference gene to normalize gene expression because both cacao microarray analyses and RT-PCR data from our laboratory suggested a relatively constant spatial and temporal expression pattern for all cacao tissues examined (Z. Shi and S. Maximova, unpublished data) .

Figure 4-2A and B show that *TcMYBPA* was expressed in leaves, flowers and whole pods in early developmental stages when the pods were too small for collection of separate seed and exocarp samples. Figure 4-2C and D show that *TcMYBPA* was also expressed in both exocarp and ovules of cacao pods from 8 weeks after pollination (WAP) to 20 WAP at the time of ripening. The expression of *TcMYBPA* in developing leaves, flowers, young pods and exocarps correlated with the accumulation of PAs and the expression of structural genes *TcANS*, *TcANR* and *TcLAR*, which are specifically involved in PA synthesis (Chapter 2). In ovules, unlike the co-regulated pattern of gene expression in leaves, flowers, young pods and exocarps, the expression pattern of *TcANS* differed quite significantly from that of *TcANR* and *TcLAR*. *TcANS* showed a relatively stable level of expression throughout pod development (including at 16 WAP when anthocyanin synthesis begins), in contrast to both *TcANR* and *TcLAR*, which showed a significant decrease at 16 WAP (Chapter 2). In ovules, *TcMYBPA* expression was detected during PA accumulation and was similar to that of *TcANS*; neither gene exhibited a reduction in expression at 16 WAP, the onset of anthocyanin synthesis. This suggests that *TcMYBPA* may contribute to the regulation of anthocyanin synthesis as well as PA synthesis. Nevertheless, the regulation of the PA-specific genes *TcANR* and *TcLAR* may also involve other transcription factors such as bHLH and WD40 repeat proteins whose interactions with *TcMYBPA* determine their specific expression patterns, which are slightly different from *TcANS*. To gain a better understanding of their regulation, further characterization and expression analysis of *bHLH* and *WD40* genes will be helpful.



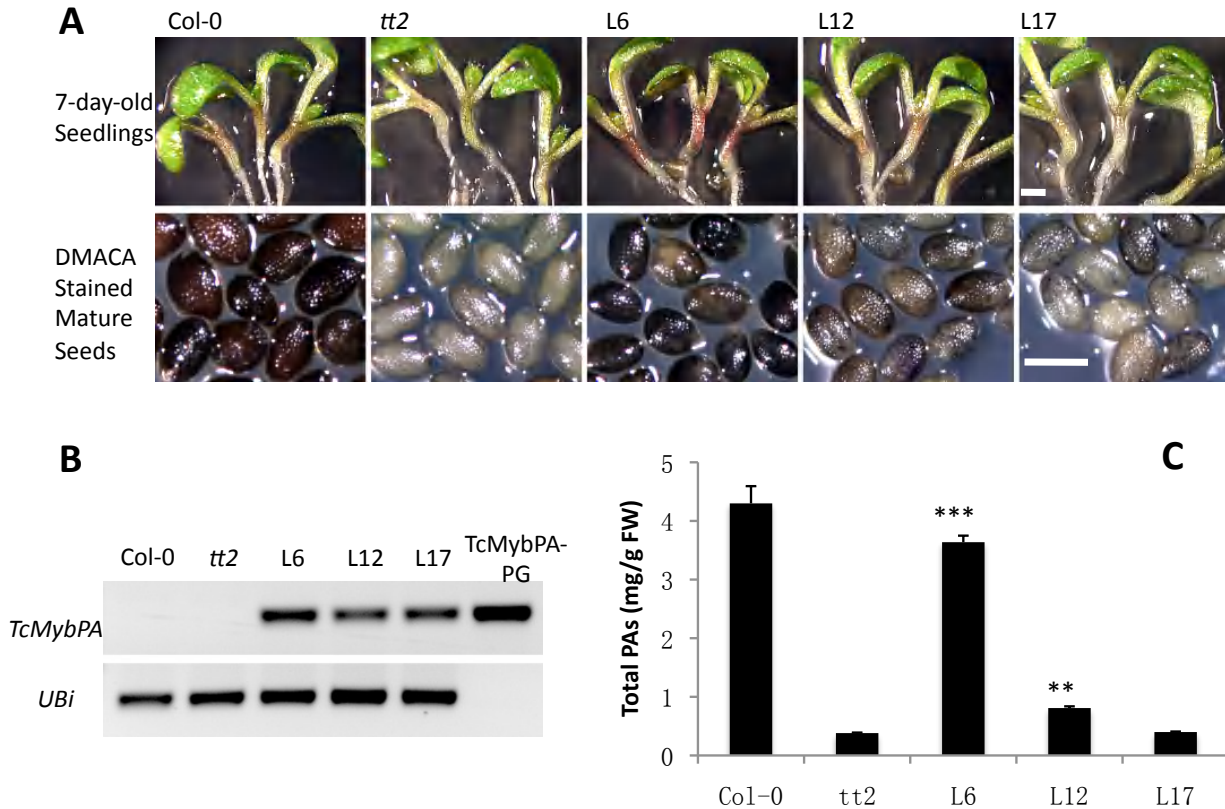
**Figure 4-2:** Transcript levels of *TcMYBPA* during the development of cacao leaves and flowers (A), young pods (B), exocarp (C) and ovules (D) Gene expression was determined by semi-quantitative RT-PCR and is shown relative to expression of *TcActin* in each sample. All data are presented as the mean of three replicates with error bars indicating  $\pm$ SE.

#### 4.3.3 *TcMYBPA* Complements the PA-Deficient Phenotype of the Arabidopsis *tt2* Mutant

Based on the very high degree of sequence conservation with Arabidopsis TT2 (see above) I hypothesized that the candidate gene *TcMYBPA* encodes a protein transcription factor that participates in the regulation of the PA biosynthesis genes *LAR*, *ANR* and *LDOX*. To test this hypothesis, a genetic complementation test was performed by transforming a constitutively expressed *TcMYCPA* coding sequence into the Arabidopsis *tt2* mutant (Nesi et al., 2001), creating *TcMYBPA-tt2* transgenic plants. Twenty one hygromycin-resistant transgenic T1 plants were generated and all of them developed a normal phenotype

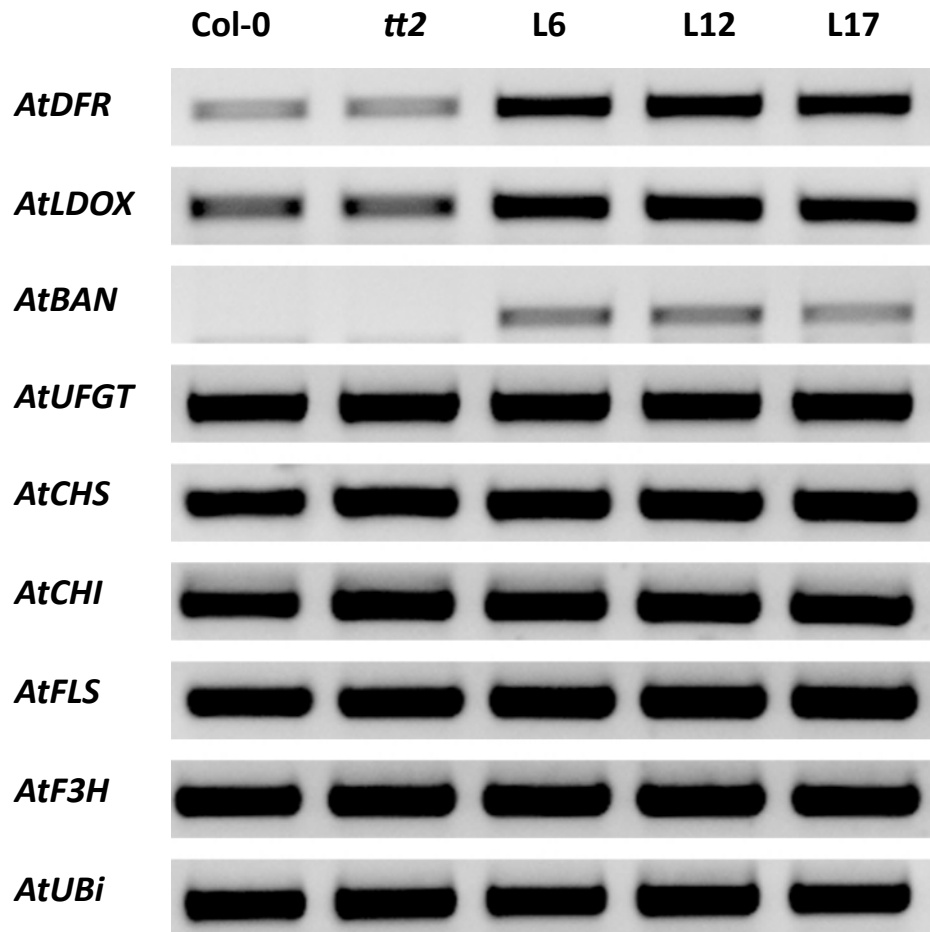
regarding general plant health, vigor, size and height. Three independent hygromycin-resistant transgenic T1 plants of *TcMYBPA-tt2* were selected because of their increased seed coat color by visual observation. After DMACA staining, 2 lines (Line 6 and Line 12) stained blue with DMACA (Fig 4-3A), suggesting deposition of PAs in the seed coat. In Line 6, the DMACA staining resulted in nearly the same intense color as in Col-0; while in line 12, the blue color was less intense than in Col-0, suggesting decreased PA levels compared to wild-type. RT-PCR using RNA extracted from T2 seedlings confirmed expression of the *TcMYBPA* gene in these transgenic lines and indicated that Line 6 had the highest expression level, which correlated with the highest PA levels as suggested by DMACA staining (Fig 4-3B). PA levels in the two *TcMYBPA-tt2* lines were 2- to 8-fold higher than in the *tt2* background. *TcMYBPA-tt2* line 6, which had the highest *TcMYBPA* expression, had nearly the same PA concentration as in the Col-0 seeds. In the young seedlings, two transgenic lines (Line 6 and Line 12) accumulated elevated levels of anthocyanins in the hypocotyls compared to *tt2* mutant plants. Line 6, which has the highest *TcMYBPA* gene expression level, accumulated the most red/purple anthocyanin pigments.





**Figure 4-3:** Complementation of the PA-deficient *tt2* mutant phenotype by constitutively expressing *TcMYBPA*. A, 7-day old seedlings and DMACA stained seeds from Col-0, the *tt2* mutant (SALK\_005260) and three independent transgenic lines of *tt-35S:TcMYBPA*. The bar represents 1 mm. B, RT-PCR analysis of *TcMYBPA* and *AtUbiquitin* transcripts in total RNA from the young seedlings shown in (A). PCR products from the *TcMYBPA*-pGEM plasmid were loaded on the last lane as a positive control for the *TcMYBPA* primer set and as a negative control for the *AtUbiquitin* primer set. C, PA levels in mature seeds of plants shown in (A). PA levels were determined by extraction and DMACA reaction using procyanidin B2 as a standard. All the data are presented as means  $\pm$  SE, n=3. \*\*P < 0.01 versus *tt2*; \*\*\*P < 0.001 versus *tt2*. FW, fresh weight.

In order to confirm that TcMYBPA activates PA synthesis genes, we used semi-quantitative RT-PCR to examine the expression of relevant genes in young seedlings of transgenic *tt2-35S:TcMYBPA* lines, untransformed *tt2* mutant and wild-type plants (Figure 4-4). Expression levels were measured for the PA-related structural genes (*DFR*, *LDOX* and *BAN*) as well as the general flavonoid pathway genes (chalcone synthase, *CHS*; chalcone isomerase, *CHI*; and flavonoid 3'-hydroxylase, *F3H*), a flavonol-specific gene (flavonol synthase; *FLS*) and an anthocyanin-specific gene (UDP-Glc-flavonoid glucosyltransferase, *UFGT*). Gene expression of *DFR* and *LDOX* was at about the same level as in the wild-type (Col-0) control and the *tt2* mutant, a result consistent with their contribution to anthocyanidin synthesis. In all transgenic lines, overexpression of *TcMYBPA* was found to activate the flavonoid late biosynthesis genes (Nesi et al., 2001) related to PA synthesis (*DFR*, *LDOX* and *BAN*). There was a 2-fold increase of *DFR* gene expression in all transgenic lines, and an approximate 1.5- to 1.7-fold increase of *LDOX* gene expression. *BAN* was not expressed in either *tt2* or Col-0 seedlings but it was significantly activated in the transgenic lines, suggesting that *TcMYBPA* controls its activation. However, no significant gene activation was detected for all the other flavonoid genes including *CHS*, *CHI*, *F3H* representing the general flavonoid pathway, *FLS* representing the flavonol-specific pathway and *UFGT* representing the anthocyanin-specific pathway. Due to limited transgenic plant material, only one biological replicate was performed on each transgenic line, so this experiment should be repeated on T<sub>2</sub> transgenic plants to improve confidence in these conclusions.



**Figure 4-4:** Semi-quantitative RT-PCR analysis of expression of flavonoid structural genes in young seedlings of the same *Arabidopsis* lines analyzed in Figure 4-3. *DFR*, dihydroflavonol reductase; *LDOX*, leucoanthocyanidin dioxygenase; *BAN*, banyuls (anthocyanidin reductase); *UGT*, UDP-Glc flavonoid glucosyltransferase; *CHS*, chalcone synthase; *CHI*, chalcone isomerase; *F3H*, flavonoid 3'-hydroxylase; *FLS*, flavonol synthase, *UBi*, Ubiquitin.

## 4.4 Discussion

### 4.4.1 Identification of TcMYBPA as an R2R3-MYB-Type Transcriptional Regulator

In this study, amino acid sequence motifs specific to the PA-regulating clade of MYB transcription factors from other species were used to identify a candidate cacao ortholog. We compared five genes from four species including *Arabidopsis* and *Lotus* TT2 (Nesi et al., 2001; Stracke et al., 2007), grape VvMYBPA1 and VVMYBPA2 (Bogs et al., 2007; Terrier et

al., 2009) and poplar MYB134 (Mellway et al., 2009). Each of these has been experimentally demonstrated to play a key role in regulation of transcription of PA biosynthesis genes. Arabidopsis and *Lotus* TT2, poplar MYB134 and grape VvMYBPA2 formed a phylogenetic cluster with ZmC1 from maize, which has been shown to activate the Arabidopsis *ANR* promoter (Nesi et al., 2001). However, cacao TcMYBPA and grape VvMYBPA1 are not in the clade that contains most of the PA-regulating MYBs; they formed another cluster that is significantly closer to the TT2/C1 clade than to other functionally unrelated MYB regulators. By contrast, the multiple protein sequence alignment including all the known PA and anthocyanin-regulatory MYB proteins revealed some PA specific motifs in the N-terminal domain. Five sites (1 or 2 amino acids) were conserved in all PA-specific MYBs, including ZmC1, but were absent from all other anthocyanin-specific MYBs. The discrepancy between the phylogenetic analysis, which showed a separate clade of TcMYBPA and VvMYBPA1 distinct from all other PA-regulatory MYBs, and the protein alignment, which clearly showed highly conserved PA-specific protein motifs in all PA MYBs, may result from the low homology C-terminal domain of those R2R3 MYB proteins. Similar to the results of Bogs et al. (2007), none of the conserved motifs in the C-terminal domain described by Stracke et al. (2001) were found. By contrast, phylogenetic analysis seems to be a strong predictor of the anthocyanin regulatory MYB proteins, with all the functionally proven anthocyanin specific MYB transcription factors falling into the same subgroup (Quattrocchio et al., 1999; Mathews et al., 2003; Deluc et al., 2006; Deluc et al., 2008).

#### **4.4.2 Expression of *TcMYBPA* Correlates with PA and Anthocyanin Synthesis in Cacao**

The data presented in Chapter 2 revealed that PA synthesis in cacao leaves occurs at higher levels in young leaves than in older leaves. This correlates with the amount of synthesis of anthocyanins, which are present at a much higher concentrations in younger stage leaves than in mature leaves (Lee et al., 1987). Anthocyanin and PA synthesis share common structural enzymes in the PA synthesis pathway, including anthocyanin synthase (ANS), which produces cyanidins for the ANR reaction leading to epicatechins or for the

UFGT reaction leading to anthocyanidins. Consistent with the PA and anthocyanin accumulation pattern, the cacao PA-specific structural genes ANR and LAR and the anthocyanin PA-common gene ANS were all co-regulated in developing leaves and more highly expressed in younger leaves but less in older leaves as described in Chapter 2. It was also demonstrated in this chapter that the *TcMYBPA* gene is expressed during cacao leaf development (Fig 4-2A). Its expression pattern correlates with PA accumulation and expression of the PA biosynthetic genes *TcANR*, *TcANS* and *TcANR*. Similar results were observed from *TcMYBPA* transcript profiling in young pods and exocarp tissues, in which *TcMYBPA* exhibits the exactly same pattern with the co-regulated PA synthesis genes *TcANR*, *TcANS* and *TcANR*, suggesting that the TcMYBPA protein is involved in regulation of PA biosynthesis in leaves, young pods and exocarp.

In cacao reproductive tissues, PA synthesis began in developing flowers prior to pollination and continued in fruits until maturation, while anthocyanin synthesis began at the onset of fruit ripening and paralleled PA synthesis until maturation. Distinct from co-regulated expression of *TcANS*, *TcANR* and *TcLAR* genes in fruit exocarp, the *TcANS* gene had a different expression pattern from that of *TcANR* and *TcLAR* in ovules. *TcANR* and *TcLAR* were still co-regulated in ovules throughout developmental stages and both dropped at 16 WAP when fruit ripening commences and anthocyanin synthesis begins, while *TcANS* expression remained relatively high at 16 WAP, likely contributing to anthocyanin synthesis (Chapter 2). Surprisingly, *TcMYBPA* shared the same expression pattern with *TcANS* rather than with the PA-specific genes *TcANR* and *TcLAR*, and the expression level remained stable, showing no decrease at 16 WAP. Similar observations were observed regarding the expression pattern of *VvMYBPA1* in grape skins, in which *VvMYBPA1* retained a relatively high transcript level two weeks after the onset of ripening and PA synthesis completely stopped when anthocyanin synthesis began (Bogs et al., 2007). One interpretation is that the high levels of *VvMYBPA1* could also contribute to anthocyanin synthesis, as it could activate the promoter of the *VvANS* (*VvLDOX*) gene. Overall, the expression pattern of *TcMYBPA* suggests that the encoded protein is involved in regulation of PA biosynthesis; moreover, it may also be involved in regulation of anthocyanin biosynthesis.

#### 4.4.3 *TcMYBPA* Complements the *tt2* Seed Phenotype and Induces Anthocyanin Synthesis in Arabidopsis

Overexpression of *TcMYBPA* in the Arabidopsis *tt2* mutant complemented the PA-deficient phenotype in Arabidopsis mature seeds (Figure 4-3). This indicated that this R2R3-type MYB transcription factor was able to substitute for the function of the key Arabidopsis PA regulator TT2. In contrast to grape *VvMYBPA1* (the MYB protein most similar to *TcMYBPA1*), which can induce ectopic PA accumulation when overexpressed in Arabidopsis, *tt2-35S:TcMYBPA* transgenic plants accumulated PAs only in the seed coat. This was similar to Arabidopsis *TT2*, which also failed to induce PA accumulation in tissues other than seed coat when ectopically expressed. Gene expression analysis of *tt2-35S:TcMYBPA* transgenic plants showed that overexpression of *TcMYBPA* induced only late flavonoid biosynthetic genes, *DFR*, *LDOX* and *BAN*, similar to Arabidopsis *TT2*, which also controls only the late flavonoid biosynthetic genes *DFR* and *BAN* (Nesi et al., 2001). By contrast, *VvMYBPA1* regulates the entire flavonoid pathway branch leading to PA synthesis, including both early and late flavonoid biosynthetic genes (Bogs et al., 2007).

In transgenic Arabidopsis, an increased accumulation of anthocyanins was also observed in hypocotyls of young seedlings; especially in Line 6, which showed an obvious visual color difference compared to untransformed controls. This could be explained by the ability of *TcMYBPA* to induce the expression of *LDOX (ANS)*, which is a structural gene contributing to both the anthocyanin and the proanthocyanin pathway. This is different from the Arabidopsis *TT2* MYB transcription factor, which has been shown to be involved specifically in the genetic control of flavonoid late biosynthesis genes (LBGs) including *DFR*, *LDOX* and *BAN* only in seeds (Nesi et al., 2001). However, both *BAN* and *TT2* are not expressed in seedlings, while both *DFR* and *LDOX* are expressed in seedlings, contributing to anthocyanin synthesis. Their expression is controlled by another MYB transcription factor, *AtPAP1* (Borevitz et al., 2000; Cominelli et al., 2008; Dare et al., 2008), whereas overexpression of *AtTT2* did not increase the expression levels of LBGs in seedlings, with the exception of *BAN*, suggesting its specific involvement in PA synthesis (Nesi et al., 2001).

The activity of TcMYBPA was in contrast to grape VvMYBPA1. Although VvMYBPA1 could activate the *VvLDOX* gene promoter in transient reporter gene assays, it failed to induce anthocyanin synthesis when overexpressed in Arabidopsis (Bogs et al., 2007). Bogs et al. also showed that anthocyanin synthesis in grape was regulated by another MYB transcription factor VvMYBA2 (Walker et al., 2007). However, the data from this research in transgenic Arabidopsis demonstrated that activation of anthocyanin synthesis was consistent with the *TcMYBPA* gene expression pattern in cacao, which was co-regulated with the *TcANS* gene and coincided with anthocyanin synthesis. Taken together, in cacao, *TcMYBPA* appeared to be capable of regulating both the PA and anthocyanin pathway by activating late PA biosynthetic genes. Potentially, this could provide a means to manipulate the amount and composition of PAs and anthocyanin together in cacao and possibly in other fruits.

#### 4.5 Conclusions

My results support the conclusion that TcMYBPA from cacao is involved in regulation of transcription of several PA biosynthesis genes. This is based on several lines of evidence. First, protein sequence comparison showed that TcMYBPA was most similar to the grape PA transcriptional regulator VvMYBPA1 and shared the conserved motifs of all the other functionally characterized R2R3-MYB PA synthesis regulators. Second, transcript profiling showed that *TcMYBPA* was expressed in all tissues accumulating PAs and consistently co-regulated with PA biosynthesis structural genes including *TcANR*, *TcANS* and *TcLAR*. Third, over-expression of *TcMYBPA* in Arabidopsis was able to functionally complement the PA-deficient phenotype in the seeds of the *tt2* mutant and resulted in a significant increase of PA accumulation compared to the *tt2* mutant. This was the result of activation of the PA biosynthetic genes including *DFR*, *LDOX* and *ANR* as shown by gene expression analysis of transgenic plants relative to untransformed *tt2* and Col-0 plants.

## 4.6 References

- Abrahams S, Tanner GJ, Larkin PJ, Ashton AR** (2002) Identification and biochemical characterization of mutants in the proanthocyanidin pathway in Arabidopsis. *Plant Physiol* **130**: 561-576
- Ahn JH** (2009) Semiquantitative analysis of Arabidopsis RNA by reverse transcription followed by noncompetitive PCR. *Cold Spring Harbor Protocols* **2009**: pdb.prot5296
- Allan AC, Hellens RP, Laing WA** (2008) MYB transcription factors that colour our fruit. *Trends Plant Sci.* **13**: 99-102
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Loor G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C** (2008) Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao L.* generated from various tissues and under various conditions. *BMC Genomics* **9**: 512
- Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L** (2004) TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J* **39**: 366-380
- Bogs J, Jaffe FW, Takos AM, Walker AR, Robinson SP** (2007) The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiol* **143**: 1347-1361
- Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C** (2000) Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **12**: 2383-2394
- Chaves FC, Gianfagna TJ** (2007) Cacao leaf procyanidins increase locally and systemically in response to infection by *Moniliophthora perniciosa* basidiospores. *Physiol. Mol. Plant Pathol.* **70**: 174-179
- Cominelli E, Gusmaroli G, Allegra D, Galbiati M, Wade HK, Jenkins GI, Tonelli C** (2008) Expression analysis of anthocyanin regulatory genes in response to different light qualities in *Arabidopsis thaliana*. *J Plant Physiol* **165**: 886-894
- Czemmel S, Stracke R, Weisshaar B, Cordon N, Harris NN, Walker AR, Robinson SP, Bogs J** (2009) The grapevine R2R3-MYB transcription factor VvMYBF1 regulates flavonol synthesis in developing grape berries. *Plant Physiol* **151**: 1513-1530
- Dare AP, Schaffer RJ, Lin-Wang K, Allan AC, Hellens RP** (2008) Identification of a cis-regulatory element by transient analysis of co-ordinately regulated genes. *Plant Methods* **4**: 17
- de Vetten N, Quattrocchio F, Mol J, Koes R** (1997) The an11 locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes Dev* **11**: 1422-1434
- Deluc L, Barrieu F, Marchive C, Lauvergeat V, Decendit A, Richard T, Carde JP, Merillon JM, Hamdi S** (2006) Characterization of a grapevine R2R3-MYB



- transcription factor that regulates the phenylpropanoid pathway. *Plant Physiol* **140**: 499-511
- Deluc L, Bogs J, Walker AR, Ferrier T, Decendit A, Merillon JM, Robinson SP, Barrieu F** (2008) The transcription factor VvMYB5b contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiol* **147**: 2041-2053
- Grotewold E** (2006) The genetics and biochemistry of floral pigments. *Annu Rev Plant Biol* **57**: 761-780
- Grotewold E, Sainz MB, Tagliani L, Hernandez JM, Bowen B, Chandler VL** (2000) Identification of the residues in the Myb domain of maize C1 that specify the interaction with the bHLH cofactor R. *Proc. Natl. Acad. Sci. USA* **97**: 13579-13584
- Kumar S, Tamura K, Nei M** (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**: 150-163
- Lazo GR, Stein PA, Ludwig RA** (1991) A DNA transformation-competent *Arabidopsis* genomic library in *Agrobacterium*. *Biotechnology (N Y)* **9**: 963-967
- Lee DW, Brammeier S, Smith AP** (1987) The selective advantages of anthocyanins in developing leaves of mango and cacao. *Biotropica* **19**: 40-49
- Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M** (2006) Genetics and biochemistry of seed flavonoids. *Annu Rev Plant Biol* **57**: 405-430
- Lin JJ** (1994) Optimization of the transformation efficiency of *Agrobacterium tumefaciens* cells using electroporation. *Plant Sci* **101**: 11-15
- Mathews H, Clendennen SK, Caldwell CG, Liu XL, Connors K, Matheis N, Schuster DK, Menasco DJ, Wagoner W, Lightner J, Wagner DR** (2003) Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. *Plant Cell* **15**: 1689-1703
- Maximova S, Miller C, Antunez de Mayolo G, Pishak S, Young A, Gultinan MJ** (2003) Stable transformation of *Theobroma cacao* L. and influence of matrix attachment regions on GFP expression. *Plant Cell Rep* **21**: 872-883
- Mellway RD, Tran LT, Prouse MB, Campbell MM, Constabel CP** (2009) The wound-, pathogen-, and ultraviolet B-responsive MYB134 gene encodes an R2R3 MYB transcription factor that regulates proanthocyanidin synthesis in poplar. *Plant Physiol* **150**: 924-941
- Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L** (2000) The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in *Arabidopsis* siliques. *Plant Cell* **12**: 1863-1878
- Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L** (2001) The *Arabidopsis* *TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**: 2099-2114
- Pelletier MK, Murrell JR, Shirley BW** (1997) Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in *Arabidopsis*. Further evidence for differential regulation of "early" and "late" genes. *Plant Physiol* **113**: 1437-1445
- Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, Koes R** (1999) Molecular analysis of the *ANTHOCYANIN2* gene of petunia and its role in the evolution of flower color. *Plant Cell* **11**: 1433-1444
- Ramsay NA, Glover BJ** (2005) MYB bHLH WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci* **10**: 63-70

- Spelt C, Quattrocchio F, Mol JN, Koes R** (2000) *ANTHOCYANIN1* of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural anthocyanin genes. *Plant Cell* **12**: 1619-1632
- Stracke R, Ishihara H, Huel G, Barsch A, Mehrrens F, Niehaus K, Weisshaar B** (2007) Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J* **50**: 660-677
- Stracke R, Werber M, Weisshaar B** (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr Opin Plant Biol* **4**: 447-456
- Terrier N, Torregrosa L, Ageorges A, Violet S, Verries C, Cheynier V, Romieu C** (2009) Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggests additional targets in the pathway. *Plant Physiol* **149**: 1028-1041
- Walker AR, Lee E, Bogs J, McDavid DA, Thomas MR, Robinson SP** (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J* **49**: 772-785
- Winkel-Shirley B** (2001) It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. *Plant Physiol* **127**: 1399-1404
- Yoshida K, Iwasaka R, Kaneko T, Sato S, Tabata S, Sakuta M** (2008) Functional differentiation of *Lotus japonicus* TT2s, R2R3-MYB transcription factors comprising a multigene family. *Plant Cell Physiol* **49**: 157-169

## CHAPTER 5: GENERAL CONCLUSIONS

Proanthocyanidins (PAs) are a subgroup of plant secondary metabolites known as flavonoids. Due to their importance for plant defense and their beneficial role in human health, our understanding of PAs and the PA biosynthetic pathway has greatly improved in the past decades. The key structural and regulatory genes in this pathway have been cloned and functionally analyzed in several plant species. However, little is known about the PA synthesis pathway in *Theobroma cacao*, a plant species that, among all food plants, produces one of the highest amounts of PAs. The overarching objective of this thesis research was to characterize the PA biosynthesis pathway in *Theobroma cacao*, focusing on genes encoding the principal enzymes and the regulation of these genes. This knowledge will provide us with genetic tools and rational approaches for further attempts to develop new cacao varieties with tailored PA profiles.

Specific objectives of this thesis research were to address three aspects of PA biosynthesis: first, the PA accumulation pattern in various tissues and developmental stages of cacao was examined; second, genes encoding key structural enzymes were isolated and their expression pattern in relation to PA accumulation and function were characterized; third, a key transcription factor was characterized that specifically regulated the PA synthesis branch of the general flavonoid pathway.

### 5.1 PAs are Actively Synthesized in Cacao Leaves, Flowers and Pods

The accumulation pattern and composition of PAs in various cacao tissues were presented in Chapters 2 and 3. Analysis of cacao leaves, flowers, young pods, exocarps and ovules/seeds at various developmental stages showed that PAs accumulated in all these tissues at very high levels. Some tissues had even higher concentrations of PAs than did seeds, which were previously considered to be the major source of PAs in chocolate. In vegetative tissues, soluble PAs accumulated at higher levels in young leaves and at lower levels in older leaves; this was consistent with the role of PAs in UV protection and

pathogen defense because young leaves are more susceptible to these challenges (Winkel-Shirley, 2002; Chaves and Gianfagna, 2007; Mellway et al., 2009). It was also found that, whereas the soluble PAs levels decreased when leaves aged, the insoluble levels increased, indicating that PAs that were previously soluble were incorporated into longer polymers and then became bound to cell wall components (Treutter, 2005; Pang et al., 2007) and became insoluble with the aging of leaves. Unlike *Arabidopsis*, which can only synthesize PAs in the seed coat after fertilization (Devic et al., 1999), cacao accumulated high levels of PAs in reproductive tissues as early as in unopened flowers, and continued to synthesize PAs in pods until maturation. In the later stages of cacao seed development after pod maturation commenced, anthocyanin synthesis began while PA synthesis continued. This is different from grape, in which anthocyanin and PA synthesis are separately regulated and temporally divided and PA synthesis is completed when anthocyanin synthesis begins (Bogs et al., 2005). This suggests that the PA and anthocyanin synthesis pathway is differently regulated in cacao as compared to grape.

It is also interesting to note that cacao PAs in leaves, flowers and fruits are composed almost entirely of epicatechin units with only 0.5 to 2% catechin units. These results showed the same trend observed by Cu et al. (2002) who analyzed the PA composition in cacao seeds and found that these PAs consisted of an epicatechin to catechin ratio of 99:1.

## **5.2 *TcANS*, *TcANR* and *TcLAR* Encode Key Enzymes Involved in PA Synthesis**

The isolation and characterization of structural genes encoding key enzymes in the PA biosynthesis pathway was described in Chapters 2 and 3. I used bioinformatic analysis to identify candidate *ANTHOCYANIDIN REDUCTASE (ANR)*, *ANTHOCYANIDIN SYNTHASE (ANS)* and *LEUCOANTHOCYANIDIN REDUCTASE (LAR)* gene homologs from the cacao EST database. I then used RT-PCR to clone the coding sequences of each gene from cacao cDNA that was prepared using RNA from young leaves in which PAs were being actively synthesized. Sequencing results suggested that all three candidate genes showed more than 60% identity to homologous genes at the protein level. Gene expression profiling of

candidate *TcANR*, *TcANS* and *TcLAR* genes was performed in various tissues from different developmental stages. These genes are in part, co-regulated and their expression positively correlated with the level of PA synthesis. Functional analysis of the candidate genes was performed by generating transgenic plants over-expressing each of the cacao genes. In each case, the cacao genes were able to complement the corresponding Arabidopsis mutant, and each functioned as expected in the tobacco flower flavonoid pathway. Thus, I can confidently conclude that the structural genes isolated are true *ANS*, *ANR* and *LAR* genes and that all are actively involved in PA synthesis in cacao. To my knowledge, these results also provide the first genetic evidence for *LAR* gene function.

A further examination of flower extracts of *TcLAR* over-expressing transgenic tobacco plants revealed that there was a significant increase of both catechin and epicatechin. This was different from our prediction that only catechins would be synthesized (based on the role of the TcLAR protein in converting leucoanthocyanidin to catechin). This also suggested that TcLAR may have a reductase/epimerase dual function and the ability to convert leucoanthocyanidin to both catechin and epicatechin. Other hypotheses are that TcLAR acts redundantly to TcANR and converts cyanidin to epicatechin, or that an epimerase exists that can convert catechin, the product of TcLAR, to epicatechin. To differentiate between two hypothesis, the *TcLAR* gene was over-expressed in the Arabidopsis *ldox* mutant that does not synthesize cyanidin due to absence of the *LDOX* gene and is less likely to have a catechin specific epimerase due to lack of catechin. In young siliques of transgenic plants, a significant increase of both catechin and epicatechin was detected, suggesting that both products were likely to be synthesized directly from leucoanthocyanidin by TcLAR. However, *in vitro* enzyme assay of recombinant TcLAR protein with leucoanthocyanidin as the substrate showed only the production of catechin but not epicatechin, suggesting that TcLAR has only reductase activity. Thus, the possibility that *TcLAR* encodes a reductase/epimerase dual function protein was not supported. Further investigation is needed to determine the mechanism of increased epicatechin synthesis in transgenic *TcLAR* over-expressing plants.

### 5.3 *TcMYBPA* Encodes a MYB Transcription Factor that Regulates Both PA and Anthocyanin Synthesis in Cacao

Chapter 4 described the isolation and functional analysis of *TcMYBPA*, an R2R3-type MYB transcription factor in cacao. Using as a query the protein sequence of *TT2*, a key regulator of PA biosynthesis in *Arabidopsis*, bioinformatic analysis revealed one candidate *TT2*-like MYB homolog from the cacao EST database. Sequence analysis showed that the *TT2*-like *TcMYBPA* protein has common R2R3 repeat domains shared by all R2R3-type MYB transcription factors, and also contained several short amino acid motifs that are specific to PA-regulatory MYBs. I examined the spatial and temporal gene expression patterns of *TcMYBPA* and found that it was developmentally expressed in a manner consistent with its involvement in PAs biosynthesis. *TcMYBPA* has the same expression pattern as all PA-related structural genes in leaves, flowers, young pods and exocarps. However in ovules, *TcMYBPA* expression coincided with expression of *TcANS* rather than with expression of *TcANR* and *TcLAR*. Overexpression of *TcMYBPA* in the *Arabidopsis tt2* mutant complemented the PA-deficient phenotype in seeds, demonstrating that *TcMYBPA* can functionally substitute for the *Arabidopsis TT2* gene. Interestingly, in addition to PA accumulation in seeds, I also observed an obvious increase of anthocyanidin accumulation in hypocotyls of transgenic *Arabidopsis* plants. Gene expression analysis of *TcMYBPA* overexpressing lines relative to untransformed *tt2* mutant and wild type Col-0 plants showed that the entire PA-specific pathway was induced by overexpression of the *TcMYBPA* gene, including *DFR*, *LDOX (ANS)* and *BAN (ANR)*. Because the *ANS* gene is specific for both anthocyanidin and PA synthesis, this explains the increase of anthocyanin synthesis in hypocotyls. Therefore, I concluded that the cacao *TcMYBPA* gene encodes an R2R3-type MYB transcription factor and is involved in the regulation of both anthocyanin and PA synthesis in cacao.

## 5.4 References

- Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR (2003)** The *Arabidopsis* *TDS4* gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. *Plant J* **35**: 624-636
- Abrahams S, Tanner GJ, Larkin PJ, Ashton AR (2002)** Identification and biochemical characterization of mutants in the proanthocyanidin pathway in *Arabidopsis*. *Plant Physiol* **130**: 561-576
- Ahn JH (2009)** Semiquantitative analysis of *Arabidopsis* RNA by reverse transcription followed by noncompetitive PCR. *Cold Spring Harbor Protocols* **2009**: pdb.prot5296
- Alemanno L, Berthouly M, Michaux-Ferriere N (1997)** A comparison between *Theobroma cacao* L. zygotic embryogenesis and somatic embryogenesis from floral explants. *In Vitro Cell. Dev. Biol. Plant* **33**: 163-172
- Allan AC, Hellens RP, Laing WA (2008)** MYB transcription factors that colour our fruit. *Trends Plant Sci.* **13**: 99-102
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990)** Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Llor G, Babin R, Sounigo O, Ducamp M, Gultinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C (2008)** Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. *BMC Genomics* **9**: 512
- Aron PM, Kennedy JA (2008)** Flavan-3-ols: nature, occurrence and biological activity. *Mol Nutr Food Res* **52**: 79-104
- Ausubel FM (1992)** Short protocols in molecular biology: a compendium of methods from Current protocols in molecular biology, Ed 2nd. Greene Pub. Associates; John Wiley & Sons, New York
- Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (2007)** Current protocols in molecular biology. John Wiley and Sons, Inc, New York
- Baba S, Osakabe N, Natsume M, Muto Y, Takizawa T, Terao J (2001)** *In vivo* comparison of the bioavailability of (+)-catechin, (-)-epicatechin and their mixture in orally administered rats. *J. Nutr.* **131**: 2885-2891
- Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L (2004)** TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J* **39**: 366-380
- Becker A, Kaufmann K, Freialdenhoven A, Vincent C, Li MA, Saedler H, Theissen G (2002)** A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. *Mol Genet Genomics* **266**: 942-950
- Beggs CJ, Kuhn K, Böcker R, Wellmann E (1987)** Phytochrome-induced flavonoid biosynthesis in mustard (*Sinapis alba* L.) cotyledons. Enzymic control and differential regulation of anthocyanin and quercetin formation. *Planta* **172**: 121-126

- Beld M, Martin C, Huits H, Stuitje AR, Gerats AG** (1989) Flavonoid synthesis in *Petunia hybrida*: partial characterization of dihydroflavonol-4-reductase genes. *Plant Mol Biol* **13**: 491-502
- Bogs J, Downey MO, Harvey JS, Ashton AR, Tanner GJ, Robinson SP** (2005) Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. *Plant Physiol* **139**: 652-663
- Bogs J, Jaffe FW, Takos AM, Walker AR, Robinson SP** (2007) The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiol* **143**: 1347-1361
- Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C** (2000) Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **12**: 2383-2394
- Bradford MM** (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**: 248-254
- Chaves FC, Gianfagna TJ** (2007) Cacao leaf procyanidins increase locally and systemically in response to infection by *Moniliophthora perniciosa* basidiospores. *Physiol. Mol. Plant Pathol.* **70**: 174-179
- Cheesman EE** (1927) Fertilization and embryogeny in *Theobroma cacao*, L. *Ann. Bot.* **41**: 107-126
- Christman JW, Blackwell TS, Juurlink BH** (2000) Redox regulation of nuclear factor kappa B: therapeutic potential for attenuating inflammatory responses. *Brain Pathol* **10**: 153-162
- Clough SJ, Bent AF** (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* **16**: 735-743
- Cominelli E, Gusmaroli G, Allegra D, Galbiati M, Wade HK, Jenkins GI, Tonelli C** (2008) Expression analysis of anthocyanin regulatory genes in response to different light qualities in *Arabidopsis thaliana*. *J Plant Physiol* **165**: 886-894
- Czemmel S, Stracke R, Weisshaar B, Cordon N, Harris NN, Walker AR, Robinson SP, Bogs J** (2009) The grapevine R2R3-MYB transcription factor VvMYBF1 regulates flavonol synthesis in developing grape berries. *Plant Physiol* **151**: 1513-1530
- Dare AP, Schaffer RJ, Lin-Wang K, Allan AC, Hellens RP** (2008) Identification of a cis-regulatory element by transient analysis of co-ordinately regulated genes. *Plant Methods* **4**: 17
- de Vetten N, Quattrocchio F, Mol J, Koes R** (1997) The an11 locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes Dev* **11**: 1422-1434
- Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L** (2003) Proanthocyanidin-accumulating cells in *Arabidopsis* testa: regulation of differentiation and role in seed development. *Plant Cell* **15**: 2514-2531
- Deluc L, Barrieu F, Marchive C, Lauvergeat V, Decendit A, Richard T, Carde JP, Merillon JM, Hamdi S** (2006) Characterization of a grapevine R2R3-MYB transcription factor that regulates the phenylpropanoid pathway. *Plant Physiol* **140**: 499-511



- Deluc L, Bogs J, Walker AR, Ferrier T, Decendit A, Merillon JM, Robinson SP, Barrieu F** (2008) The transcription factor VvMYB5b contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiol* **147**: 2041-2053
- Devic M, Guillemot J, Debeaujon I, Bechtold N, Bensaude E, Koornneef M, Pelletier G, Delseny M** (1999) The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. *Plant J* **19**: 387-398
- Dixon RA, Paiva NL** (1995) Stress-induced phenylpropanoid metabolism. *Plant Cell* **7**: 1085-1097
- Dixon RA, Xie DY, Sharma SB** (2005) Proanthocyanidins - a final frontier in flavonoid research? *New Phytol* **165**: 9-28
- Engler MB, Engler MM** (2006) The emerging role of flavonoid-rich cocoa and chocolate in cardiovascular health and disease. *Nutr Res* **64**: 109-118
- Fedoroff NV, Furtek DB, Nelson OE** (1984) Cloning of the bronze locus in maize by a simple and generalizable procedure using the transposable controlling element Activator (Ac). *Proc Natl Acad Sci U S A* **81**: 3825-3829
- Feucht W, Treutter D** (1999) The role of flavan-3-ols and proanthocyanidins in plant defense. *In* Inderjit, KMM Dakshi, CL Foy, eds, *Principles and Practices in Plant Ecology: Allelochemical Interactions*. CRC Press, New York, pp 307-338
- Fisher DK, Guiltinan MJ** (1995) Rapid, efficient production of homozygous transgenic tobacco plants with *agrobacterium tumefaciens* : A seed-to-seed protocol. *Plant Mol Biol Repr* **13**: 278-289
- Fisher ND, Hollenberg NK** (2005) Flavanols for cardiovascular health: the science behind the sweetness. *J Hypertens* **23**: 1453-1459
- Foo LY, Newman R, Waghorn G, McNabb WC, Ulyatt MJ** (1996) Proanthocyanidins from *Lotus corniculatus*. *Phytochemistry* **41**: 617-624
- Fukuchi-Mizutani M, Okuhara H, Fukui Y, Nakao M, Katsumoto Y, Yonekura-Sakakibara K, Kusumi T, Hase T, Tanaka Y** (2003) Biochemical and molecular characterization of a novel UDP-glucose:anthocyanin 3'-O-glucosyltransferase, a key enzyme for blue anthocyanin biosynthesis, from gentian. *Plant Physiol* **132**: 1652-1663
- Gargouri M, Manigand C, Mauge C, Granier T, Langlois d'Estaintot B, Cala O, Pianet I, Bathany K, Chaudiere J, Gallois B** (2009) Structure and epimerase activity of anthocyanidin reductase from *Vitis vinifera*. *Acta Crystallogr D* **65**: 989-1000
- Graham TL** (1991) Flavonoid and isoflavonoid distribution in developing soybean seedling tissues and in seed and root exudates. *Plant Physiol* **95**: 594-603
- Grotewold E** (2006) The genetics and biochemistry of floral pigments. *Annu Rev Plant Biol* **57**: 761-780
- Grotewold E, Sainz MB, Tagliani L, Hernandez JM, Bowen B, Chandler VL** (2000) Identification of the residues in the Myb domain of maize C1 that specify the interaction with the bHLH cofactor R. *Proc. Natl. Acad. Sci. USA* **97**: 13579-13584
- Gu L, Kelm M, Hammerstone JF, Beecher G, Cunningham D, Vannozzi S, Prior RL** (2002) Fractionation of polymeric procyanidins from lowbush blueberry and quantification of procyanidins in selected foods with an optimized normal-phase HPLC-MS fluorescent detection method. *J Agric Food Chem* **50**: 4852-4860

- Gu LW, House SE, Wu XL, Ou BX, Prior RL** (2006) Procyanidin and catechin contents and antioxidant capacity of cocoa and chocolate products. *J Agric Food Chem* **54**: 4057-4061
- Hahlbrock K, Scheel D** (1989) Physiology and molecular biology of phenylpropanoid metabolism. *Annu Rev Plant Phys* **40**: 347-369
- Hannum SM, Erdman JW** (2000) Emerging health benefits from cocoa and chocolate. *J Med Food* **3**: 73-75
- Hirotsani M, Kuroda R, Suzuki H, Yoshikawa T** (2000) Cloning and expression of UDP-glucose: flavonoid 7-O-glucosyltransferase from hairy root cultures of *Scutellaria baicalensis*. *Planta* **210**: 1006-1013
- Holton TA, Brugliera F, Tanaka Y** (1993) Cloning and expression of flavonol synthase from *Petunia hybrida*. *Plant J* **4**: 1003-1010
- Holton TA, Cornish EC** (1995) Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7**: 1071-1083
- Johnson CS, Kolevski B, Smyth DR** (2002) *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *Plant Cell* **14**: 1359-1375
- Karim M, McCormick K, Kappagoda CT** (2000) Effects of cocoa extracts on endothelium-dependent relaxation. *J Nutr* **130**: 2105S-2108S
- Kavanagh K, Jornvall H, Persson B, Oppermann U** (2008) The SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell Mol Life Sci* **65**: 3895-3906
- Keen CL, Holt RR, Polagruto JA, Wang JF, Schmitz HH** (2002) Cocoa flavanols and cardiovascular health. *Phytochem Rev* **1**: 231-240
- Kristiansen KN** (1986) Conversion of (+)-dihydroquercetin to (+)-2,3-trans-3,4-cis-leucocyanidin and (+)-catechin with an enzyme extract from maturing grains of barley. *Carlsberg Res. Comm* **51**: 51-60
- Kubo H, Nawa N, Lupsea SA** (2007) *Anthocyaninless1* gene of *Arabidopsis thaliana* encodes a UDP-glucose:flavonoid-3-O-glucosyltransferase. *J Plant Res* **120**: 445-449
- Kumar S, Tamura K, Nei M** (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**: 150-163
- Labesse G, Vidalcros A, Chomilier J, Gaudry M, Mornon JP** (1994) Structural comparisons lead to the definition of a new superfamily of NAD(P)(H)-accepting oxidoreductases: the single-domain reductases/epimerases/dehydrogenases (the RED family). *Biochem J* **304**: 95-99
- Lazo GR, Stein PA, Ludwig RA** (1991) A DNA transformation-competent *Arabidopsis* genomic library in *Agrobacterium*. *Biotechnology (N Y)* **9**: 963-967
- Lee DW, Brammeier S, Smith AP** (1987) The selective advantages of anthocyanins in developing leaves of mango and cacao. *Biotropica* **19**: 40-49
- Lehrian DW, Keeney PG** (1980) Changes in lipid components of seeds during growth and ripening of cacao fruit. *J. Am. Oil Chem. Soc.* **57**: 61-65
- Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M** (2006) Genetics and biochemistry of seed flavonoids. *Annu Rev Plant Biol* **57**: 405-430
- Levy M, Edelbaum O, Sela I** (2004) Tobacco mosaic virus regulates the expression of its own resistance gene N. *Plant Physiol* **135**: 2392-2397

- Li J, Ou-Lee TM, Raba R, Amundson RG, Last RL** (1993) Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation. *Plant Cell* **5**: 171-179
- Lin JJ** (1994) Optimization of the transformation efficiency of *Agrobacterium tumefaciens* cells using electroporation. *Plant Sci* **101**: 11-15
- Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH** (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205-210
- Marles MA, Ray H, Gruber MY** (2003) New perspectives on proanthocyanidin biochemistry and molecular regulation. *Phytochemistry* **64**: 367-383
- Martin C, Prescott A, Mackay S, Bartlett J, Vrijlandt E** (1991) Control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. *Plant J* **1**: 37-49
- Martinez-Florez S, Gutierrez-Fernandez B, Sanchez-Campos S, Gonzalez-Gallego J, Tunon MJ** (2005) Quercetin attenuates nuclear factor-kappaB activation and nitric oxide production in interleukin-1beta-activated rat hepatocytes. *J Nutr* **135**: 1359-1365
- Mathews H, Clendennen SK, Caldwell CG, Liu XL, Connors K, Matheis N, Schuster DK, Menasco DJ, Wagoner W, Lightner J, Wagner DR** (2003) Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. *Plant Cell* **15**: 1689-1703
- Mato M, Ozeki Y, Itoh Y, Higeta D, Yoshitama K, Teramoto S, Aida R, Ishikura N, Shibata M** (1998) Isolation and characterization of a cDNA clone of UDP-galactose: flavonoid 3-O-galactosyltransferase (UF3GaT) expressed in *Vigna mungo* seedlings. *Plant Cell Physiol* **39**: 1145-1155
- Maximova S, Miller C, Antunez de Mayolo G, Pishak S, Young A, Guiltinan MJ** (2003) Stable transformation of *Theobroma cacao L.* and influence of matrix attachment regions on GFP expression. *Plant Cell Rep* **21**: 872-883
- Mellway RD, Tran LT, Prouse MB, Campbell MM, Constabel CP** (2009) The wound-, pathogen-, and ultraviolet B-responsive MYB134 gene encodes an R2R3 MYB transcription factor that regulates proanthocyanidin synthesis in poplar. *Plant Physiol* **150**: 924-941
- Menssen A, Hohmann S, Martin W, Schnable PS, Peterson PA, Saedler H, Gierl A** (1990) The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the *A2* gene. *Embo J* **9**: 3051-3057
- Michiels A, Van den Ende W, Tucker M, Van Riet L, Van Laere A** (2003) Extraction of high-quality genomic DNA from latex-containing plants. *Anal Biochem* **315**: 85-89
- Moriguchi T, Kita M, Ogawa K, Tomono Y, Endo T, Omura M** (2002) Flavonol synthase gene expression during citrus fruit development. *Physiol Plant* **114**: 251-258
- Murashige T, Skoog F** (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol Plant* **15**: 473-497
- Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L** (2000) The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in Arabidopsis siliques. *Plant Cell* **12**: 1863-1878

- Nesi N, Debeaujon I, Jond C, Stewart AJ, Jenkins GI, Caboche M, Lepiniec L** (2002) The *TRANSPARENT TESTA16* locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. *Plant Cell* **14**: 2463-2479
- Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L** (2001) The Arabidopsis *TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**: 2099-2114
- Niemenak N, Rohsius C, Elwers S, Ndoumou DO, Lieberei R** (2006) Comparative study of different cocoa (*Theobroma cacao* L.) clones in terms of their phenolics and anthocyanins contents. *J. Food Comp Anal.* **19**: 612-619
- Norman KH, Naomi DLF, Marjorie LM** (2009) Flavanols, the Kuna, cocoa consumption, and nitric oxide. *J. Am. Soc. Hypertens* **3**: 105-112
- O'Reilly C, Shepherd NS, Pereira A, Schwarz-Sommer Z, Bertram I, Robertson DS, Peterson PA, Saedler H** (1985) Molecular cloning of the *a1* locus of *Zea mays* using the transposable elements *En* and *Mu1*. *Embo J* **4**: 877-882
- Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA** (2007) Early steps in proanthocyanidin biosynthesis in the model legume *Medicago truncatula*. *Plant Physiol* **145**: 601-615
- Paolocci F, Robbins MP, Madeo L, Arcioni S, Martens S, Damiani F** (2007) Ectopic expression of a basic helix-loop-helix gene transactivates parallel pathways of proanthocyanidin biosynthesis. structure, expression analysis, and genetic control of leucoanthocyanidin 4-reductase and anthocyanidin reductase genes in *Lotus corniculatus*. *Plant Physiol* **143**: 504-516
- Pelletier MK, Murrell JR, Shirley BW** (1997) Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in Arabidopsis. Further evidence for differential regulation of "early" and "late" genes. *Plant Physiol* **113**: 1437-1445
- Pourcel L, Routaboul JM, Cheynier V, Lepiniec L, Debeaujon I** (2007) Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* **12**: 29-36
- Pourcel L, Routaboul JM, Kerhoas L, Caboche M, Lepiniec L, Debeaujon I** (2005) *TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell* **17**: 2966-2980
- Purdy LH, Schmidt RA** (1996) Status of cacao witches' broom: biology, epidemiology, and management. *Annu Rev Phytopathol* **34**: 573-594
- Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, Koes R** (1999) Molecular analysis of the *ANTHOCYANIN2* gene of petunia and its role in the evolution of flower color. *Plant Cell* **11**: 1433-1444
- Ramassamy C** (2006) Emerging role of polyphenolic compounds in the treatment of neurodegenerative diseases: A review of their intracellular targets. *Eur J Pharmacol* **545**: 51-64
- Ramsay NA, Glover BJ** (2005) MYB bHLH WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci* **10**: 63-70
- Rao ST, Rossmann MG** (1973) Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**: 241-&
- Rein D, Paglieroni TG, Pearson DA, Wun T, Schmitz HH, Gosselin R, Keen CL** (2000) Cocoa and wine polyphenols modulate platelet activation and function. *J Nutr* **130**: 2120S-2126S

- Revilla E** (1991) Analysis of catechins and proanthocyanidins in grape seeds by HPLC with photodiode array detection. *Chromatographia* **31**: 465
- Rizzi M, Tonetti M, Vigevani P, Sturla L, Bisso A, De Flora A, Bordo D, Bolognesi M** (1998) GDP-4-keto-6-deoxy-D-mannose epimerase/reductase from *Escherichia coli*, a key enzyme in the biosynthesis of GDP-L-fucose, displays the structural characteristics of the RED protein homology superfamily. *Structure* **6**: 1453-1465
- Routaboul JM, Kerhoas L, Debeaujon I, Pourcel L, Caboche M, Einhorn J, Lepiniec L** (2006) Flavonoid diversity and biosynthesis in seed of *Arabidopsis thaliana*. *Planta* **224**: 96-107
- Sagasser M, Lu GH, Hahlbrock K, Weisshaar B** (2002) *A. thaliana* *TRANSPARENT TESTA 1* is involved in seed coat development and defines the WIP subfamily of plant zinc finger proteins. *Genes Dev* **16**: 138-149
- Saito K, Kobayashi M, Gong Z, Tanaka Y, Yamazaki M** (1999) Direct evidence for anthocyanidin synthase as a 2-oxoglutarate-dependent oxygenase: molecular cloning and functional expression of cDNA from a red form of *Perilla frutescens*. *Plant J* **17**: 181-189
- Saslowsky D, Winkel-Shirley B** (2001) Localization of flavonoid enzymes in *Arabidopsis* roots. *Plant J* **27**: 37-48
- Scalbert A** (1991) Antimicrobial properties of tannins. *Phytochemistry* **30**: 3875-3883
- Schijlen EG, Ric de Vos CH, van Tunen AJ, Bovy AG** (2004) Modification of flavonoid biosynthesis in crop plants. *Phytochemistry* **65**: 2631-2648
- Schwarz-Sommer Z, Shepherd N, Tacke E, Gierl A, Rohde W, Leclercq L, Mattes M, Berndtgen R, Peterson PA, Saedler H** (1987) Influence of transposable elements on the structure and function of the *A1* gene of *Zea mays*. *Embo J* **6**: 287-294
- Schwinn KE, M. Davies K, Deroles SC, Markham KR, Miller RM, Bradley JM, Manson DG, Given NK** (1997) Expression of an *Antirrhinum majus* UDP-glucose:flavonoid-3-O-glucosyltransferase transgene alters flavonoid glycosylation and acylation in *lisianthus* (*Eustoma grandiflorum* Grise.). *Plant Sci* **125**: 53-61
- Shirley BW, Hanley S, Goodman HM** (1992) Effects of ionizing radiation on a plant genome: analysis of two *Arabidopsis* transparent testa mutations. *Plant Cell* **4**: 333-347
- Singh K, Rani A, Paul A, Dutt S, Joshi R, Gulati A, Ahuja PS, Kumar S** (2009) Differential display mediated cloning of anthocyanidin reductase gene from tea (*Camellia sinensis*) and its relationship with the concentration of epicatechins. *Tree Physiol* **29**: 837-846
- Singh S, McCallum J, Gruber MY, Neil Towers GH, Muir AD, Bohm BA, Koupai-Abyazani MR, Glass ADM** (1997) Biosynthesis of flavan-3-ols by leaf extracts of *Onobrychis viciifolia*. *Phytochemistry* **44**: 425-432
- Skadhauge B, Gruber MY, Thomsen KK, Diter Von W** (1997) Leucocyanidin reductase activity and accumulation of proanthocyanidins in developing legume tissues. *Am J Bot* **84**: 494-503
- Spelt C, Quattrocchio F, Mol JN, Koes R** (2000) *ANTHOCYANIN1* of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural anthocyanin genes. *Plant Cell* **12**: 1619-1632

- Stoupi S, Williamson G, Viton F, Barron D, King LJ, Brown JE, Clifford MN** (2010) *In vivo* bioavailability, absorption, excretion, and pharmacokinetics of [14C]procyanidin B2 in male rats. *Drug Metab Dispos* **38**: 287-291
- Stoutjesdijk PA, Sale PW, Larkin PJ** (2001) Possible involvement of condensed tannins in aluminium tolerance of *Lotus pedunculatus*. *Aust J Plant Physiol* **28**: 1063-1074
- Stracke R, Ishihara H, Huep G, Barsch A, Mehrstens F, Niehaus K, Weisshaar B** (2007) Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J* **50**: 660-677
- Stracke R, Werber M, Weisshaar B** (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr Opin Plant Biol* **4**: 447-456
- Takos AM, Ubi BE, Robinson SP, Walker AR** (2006) Condensed tannin biosynthesis genes are regulated separately from other flavonoid biosynthesis genes in apple fruit skin. *Plant Sci* **170**: 487-499
- Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ, Ashton AR** (2003) Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *J Biol Chem* **278**: 31647-31656
- Tanner GJ, Kristiansen KN** (1993) Synthesis of 3,4-cis-[3H]leucocyanidin and enzymatic reduction to catechin. *Anal Biochem* **209**: 274-277
- Terrier N, Torregrosa L, Ageorges A, Vialet S, Verries C, Cheynier V, Romieu C** (2009) Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggests additional targets in the pathway. *Plant Physiol* **149**: 1028-1041
- Thies M, Fischer R** (1971) New color reaction for microchemical detection and quantitative determination of catechins. *Mikrochim Acta*: 9-13
- Thoden JB, Hegeman AD, Wesenberg G, Chapeau MC, Frey PA, Holden HM** (1997) Structural analysis of UDP-sugar binding to UDP-galactose 4-epimerase from *Escherichia coli*. *Biochemistry* **36**: 6294-6304
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG** (1997) The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882
- Treutter D** (2005) Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol*. **7**: 581-591
- Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N, Blundell TL, Esch JJ, Marks MD, Gray JC** (1999) The *TRANSPARENT TESTA GLABRA1* locus, which regulates trichome differentiation and anthocyanin biosynthesis in *Arabidopsis*, encodes a WD40 repeat protein. *Plant Cell* **11**: 1337-1350
- Walker AR, Lee E, Bogs J, McDavid DA, Thomas MR, Robinson SP** (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J* **49**: 772-785
- Wan Y, Vinson JA, Etherton TD, Proch J, Lazarus SA, Kris-Etherton PM** (2001) Effects of cocoa powder and dark chocolate on LDL oxidative susceptibility and prostaglandin concentrations in humans. *Am J Clin Nutr* **74**: 596-602
- Welle R, Grisebach H** (1989) Phytoalexin synthesis in soybean cells: elicitor induction of reductase involved in biosynthesis of 6'-deoxychalcone. *Arch Biochem Biophys* **272**: 97-102

- Winkel-Shirley B** (2001) It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. *Plant Physiol* **127**: 1399-1404
- Winkel-Shirley B** (2002) Biosynthesis of flavonoids and effects of stress. *Curr Opin Plant Biol* **5**: 218-223
- Wright DC, Park WD, Leopold NR, Hasegawa PM, Janick J** (1982) Accumulation of lipids, proteins, alkaloids and anthocyanins during embryo development *in vivo* of *Theobroma cacao* L. *J. Am. Oil Chem. Soc.* **59**: 475-479
- Xie DY, Jackson LA, Cooper JD, Ferreira D, Paiva NL** (2004) Molecular and biochemical analysis of two cDNA clones encoding dihydroflavonol-4-reductase from *Medicago truncatula*. *Plant Physiol* **134**: 979-994
- Xie DY, Sharma SB, Paiva NL, Ferreira D, Dixon RA** (2003) Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis. *Science* **299**: 396-399
- Yoshida K, Iwasaka R, Kaneko T, Sato S, Tabata S, Sakuta M** (2008) Functional differentiation of *Lotus japonicus* TT2s, R2R3-MYB transcription factors comprising a multigene family. *Plant Cell Physiol* **49**: 157-169
- Yu-Guang Li, Greg Tanner, Phil Larkin** (1996) The DMACA-HCl protocol and the threshold proanthocyanidin content for bloat safety in forage legumes. *J. Sci. Food Agric.* **70**: 89-101
- Zeng Y** (2002) RNA isolation from highly viscous samples rich in polyphenols and polysaccharides. *Plant Mol Biol Repr* **20**: 417
- Zhang W, Sun Y, Timofejeva L, Chen C, Grossniklaus U, Ma H** (2006) Regulation of *Arabidopsis* tapetum development and function by *DYSFUNCTIONAL TAPETUM1* (*DYT1*) encoding a putative bHLH transcription factor. *Development* **133**: 3085-3095
- Zuker A, Tzfira T, Ben-Meir H, Ovadis M, Shklarman E, Itzhaki H, Forkmann G, Martens S, Neta-Sharir I, Weiss D, Vainstein A** (2002) Modification of flower color and fragrance by antisense suppression of the flavanone 3-hydroxylase gene. *Mol. Breed.* **9**: 33-41

## APPENDIX A: SEQUENCES OF THE THEOBROMA CACAO GENES CLONED IN THIS THESIS

>gi|290579516|gb|GU324348.1| Theobroma cacao cultivar Scavina 6 anthocyanidin complete cds

```
ATGGCCAGCCAGACCGTAGGCCAAAAAGACCGCTTGTGTCGTAGGTGGCACCGGATACGTTGCATC
TTTGTGGTCAAGCTGTTGCTTGAGAAGGGCTACGCTGTTAACTACTGTCAGGGACCCAGACA
ACCAGAAAAAGATCCCTCACCTCGTAACACTACAAAAGCTAGGAGACTTGAAAATCTTTCGAGCA
GATTTGACTGATGAAGGCAGCCTTGATGTCCCATAGCTGGTTGTGACCTTGTCTTCCATGTTGC
AACACCCGTCAATTTTGGCTTCTCAAGATCCTGAGAATGACATGATCAAACCAGCAATCCAGGGAG
TGCTGAACGTTTTTGAAGCTTGTGCCAAAGCAAAAACAGTCAAACGGGTCTGCTTACTTCTTCA
GCCGCAGCTGTGTCTATCAACACACTCAAGGGGACAGATCTGGTCCTGACTGAGAAAGACTGGAC
CGACGTTGAGTTCTTATCGTCGGCAAAGCCACCAACTTGGGGGTACCCTGCATCCAAGACATTGG
CTGAAAAGGCAGCATGGAAATTTGCTCAAGAAAACAACATCGATCTCATCACGGTCATCCCTTCT
CTCATGACCGGTCTTCTCTCACCCAGACGTGCCCAGCAGCATTGGCCTTGCACATCTTTGCTT
TCAGGCAACGAATTCCTTGTAATGCTTTGAAAGGTATGCAAATGTTGTCAGGTTCAATCTCTA
TCACTCATGTGGAGGACGTCTGTGCGGGCCATGTTTTTCTGGCAGAAAAAGAATCTGCATCCGGC
CGATATATATGCTGTGCTGTCAATTCAGTGTTCCCTGAGCTTGCTAAGTTCCTCAACCAAAGATA
CCCTGAGTTCAAAGTCCCTACTGATTTTGGAGATTTCCCTCTAAAGCCAAGTTGATCATTTCCT
CGGATAAGCTTATTAATGAAGGATTCAGCTTTAAGTTTGGGATTGAGGAAATCTACGACCAAAC
TGTAGAATACATGAACGCTAAGGGGCTGCTCAAGTGA
```

>gi|290579514|gb|GU324347.1| Theobroma cacao cultivar LCT-EEN 37 anthocyanidin reductase (ANR) gene

```
cdsATGCTCGATTGGCATGAAAAACACATCAAGCCCATGCACATTAAGAAATGCGGAATATGTC
AAATTCTAAGACATGGATTCTCCACCATGGAGAAATCCAATGGTTTCCGGGGTGTTTTACATCT
TCCATAGGTAAAAGAGCTTTAGAGTGATTGAGCTTGATGAAGAGAACCGATGTCTAAGAAAGG
CTCTGGTAGTGTGCAGGGTAATTGCTGGAAGGGTTCAAAACCCTTTAGAAAATGCGCAAGAAAT
GGCAAGTCAATCAAGCTTTGATTCACTGGCTGGAACCTTTGATAGCCACTCGAATATTGAGGAA
CTCTATTCACTAAATCCTAGAGCTCTCTTCCCTTGTTTTGTGGTCATCTGCAAACCCTCAAAGCA
AAGTGCTCAAAAATTATAACTTAGGACCACGTTCCCTTGTCATTGTGATTGTGAGGTCTAATTTCT
TTATCCAACCTTCTCTGTAATTATTTGTTCTTATAGTAAATATATCTTCTCTTTCTGCAAGGTAT
TTGTTCTCTCAAACTTCATCTTGAACCGCTATGAGATTGCATCAATCTACATAGAGCTGTAGCT
AGTCTAGCTGCTTGAAGTTTCTGACCAGCCTTCTCTTGACCACCGCACCCTCTCTCTCTCCCCTCA
TTTTTCTAGTGAAACTGCCCGAGGATTGAATTCTGTTCTGCAGATAAGTCAAATTTGACCATCA
ATCAAGCTCCTATCCATTAATTCTGCTTGTATATCCCTTTTAGACAAATGTAATTAATAACAC
AAACCCTAGTACAGAATCTATTAAGTCTAATTTGAATTCATCCTGGAAGTGCCTGATCCTGCT
ATGCGTGGTCTAATTCGTCAACAACATTGTCCCTCTTCTTCTGTCTGCACTGTCTTGTCACTA
CTATGGCTCTTCTTGGGCTCAGATCTAATCCTGATACTGTGTTTTGATAATAAGAGTTGCTTACA
GATACGAGTTAGGGTATTTAATTACAGACCTATTTCGGGAAGGGGAAGGGAAACTTGTTACAATG
TACTTAACAAAGATCGTCAGGATATCTCCGGGTGTTCTTTGATCCTCCTATGGGGTTAATCTTAT
GTTTCCCTCGATTAAACAAATGAAAAATAAATAAATAAAGGATTCCTGGCAACAGTTCTCTGTT
GCTTGCTATAATTGAAATTGTGAATCATAAGAAGTAAAACCACCTAACCACAGGCCACGTGA
```



GGCATCATTTGGAGCAACTTGAGGTTTGGCCAGCTACCCCTCTTTTGCCTTCCAAAGATTTTAACT  
TTGACATCAGTTGAGCCTTCACTTCCAACATTTTCAGCAAACCTTCATCTACTGTTTTTCTGCACCA  
CTGCATGTCCTATATGATATCCAAAAACAACCTCCCCTTTCTAACATACAATAAATTTTGTGCT  
CGAAATCTGATTATCGCTTATGATCATTACTGGAAGATGCCAAGCTCCAAAACCTTCACCTAAATG  
CTTTTGTGTGCCTCTTCTTAAGTCCATACTTTCTTAGTAAAAGAATGTGGATGCAAAAACCCATT  
TTGGGAATACTTTTCCACAGTAAAAAATAATTATTAGATTATGTCTTGATTCAATCAAATTT  
CCTCAATGAACCATAGAACAATTTTAGTGACTAAAGCAGTTGATAAATACTCAACCCATTTACT  
ATTATTTAATGGTTTTCTCATAGAACAATTTTTTAAGAAGTACACAAGCAATAAACTCTACATT  
TACTCTACATTTATCAATCGAATACAAAATATTTTTGGTAGGGGTCACTGTAACTCAGTCA  
TTATCAAAAAGAAGCGATGAGAGGAAATGAGTTTCTAGCTAAAAGAACTCGTGGGTAAGTACT  
GGCTACCCCTTGTAGTAGCAGTTTGGGGAGTGCAGTACACCACCGATGGTTTTGAAAGACTTTTT  
GAGTCGTTGGTATGCACAAGGGCACGTGCTCACCTTCTCCATCTAAAATCTACTCAAGCCCTGG  
GTAAGTGCCCATCGTCTATAAAAACAATAATGCAATAAGTTTATTCCACCTATGCATCTTTGTCTG  
AACGGTTGAAGGGTTCAAAAACAAGCCCAAAAATCGAAACGGAAAAGCAAAAAGTAAGGTACCCG  
GTCAAGAAAAGGAATATAGTCATTGAAGCCATGGCCAGCCAGACCGTAGGCAAAAAGACCGCTT  
GTGTCGTAGGTGGCACCGGATACGTTGCATCTTTGTTGGTCAAGCTGTTGCTTGAGAAGGGCTAC  
GCTGTTAACTACTGTCAGGGACCCAGGTTGATCTTCTCTTCTTCTCATCTTCTTCTGTTTTT  
CTTGTTCAATTTGTTTCTACTGCTTTGCTTTGGTGGGTCAATCCAGTATTTTACTTTCTTCCCCTTC  
CTTGGTTTTCTTGTTTTTATATATAATATATTGGTATGGCTGCTGCTGCAATTTAGGAATTTCTA  
CGATTTATGCCCCATTGTAGCATTAGTTCTTTGTTCTTTGTTTTTCACTTTAAGCTTAAACTA  
TAAATTCCTACTACTCTGTATCGAGCATGTTGGAAGTTAATAAGCGAGAACAACCGAGGAACA  
TACCGCCTTGTCTTGTGTCAGTTGGTGTTTTTTAGGGGGTACCCACGATATCCGTTGCCTGAGCAGG  
AGAGAATACTATCAATTCCTTGGGTTTGGGTTTCAACCCCTCTCGAAGAGTTTCTTACCAAATTA  
TCACATTTTTTGCAGATAAAGTGTAATGAGTAGAATCTTTTTTCTTTTTTCTTTTTGGGGGTTT  
ATTTTCATTTTTCTGGCAACACCCAGAGTAAGTAAACATGAATGGGTGTAATGCTGTGTCTTTTCT  
GCAGACAACCAGAAAAAGATCCCTCACCTCGTAACACTACAAAAGCTAGGAGACTTGAAAATCT  
TTCGAGCAGATTTGACTGATGAAGGCAGCTTTGATGTCCCATAGCTGGTTGTGACCTTGTCTTC  
CATGTTGCAACACCCGTCAATTTTGCTTCTCAAGATCCTGAGGTATGTAAAACCATTAAACTGCT  
TTTCCAGTGATGATCAAATTCCTTCTGGTTTTGAGGAATGATGACAAGGTTTACTTTATTGGAT  
TTTGATTATAGAATGACATGATCAAACCAGCAATCCAGGGAGTGCTGAACGTTTTGAAAGCTTG  
TGCCAAAGCAAAAACAGTCAAACGGGTGCTCTTGACTTCTTCAAGCCGAGCTGTGTCTATCAACA  
CACTCGAGGGGACAGATCTGGTCTGACTGAGAAAGACTGGACCGACGTTGAGTTCTTATCGTCG  
GCAAAGCCACCAACTTGGGTAACAATTTTCATGCTAATCCATTCCTCTTCTCTTATCTTCGGGG  
GAATTGCAGAAGAGGGCAAGGTAACAAAATAAATTGGTGTGCATAATCTGAAGTAAGCTTTTTAT  
CCATGAATGCAGGGGTACCCTGCATCCAAGACATTGGCTGAAAAGGCAGCATGGAAATTTGCTCA  
AGAAAACAACATCGATCTCATCACGGTCATCCCTTCTCTCATGACCGGTCTTCTCTCACCCCAG  
ACGTGCCCAGCAGCATTGGCCTTGCCACATCTTTGCTTTCAGGTATTAAGTTAGAACCTCGTGTC  
CTGGCCTTGTCTTAGATGTAAAACCTGATGCATAAAGAAGTAGCCTGGAGCACCATGAACTGTA  
ACTGATGGGAATTTTAAACATTTTTGCAGGCAACGAATTCCTTGTAATGCTTTGAAAGGTATGC  
AAATGTTGTCAGGTTCAATCTCTATCACTCATGTGGAGGACGTCTGTCTGGGCCCATGTCTTTCTG  
GCAGAAAAAGAATCTGGATCCGGCCGATATATATGCCTGTGCTGTCAATTCAGTGTTCTGAGC  
TTGCTAAGTTCTCAACCAAGATACCCTGAGTTCAAAGTCCCTACTGAGTAAGCCAACCTGCAT  
TCAATATCACAATCTAACTTCTCTTCTTCTGCTAGAATTGTGGTTAATCTTAGTTTTGTTTGC  
TTTTGTTACAATTGCAGTTTTGGAGATTTCCCCTCTAAAGCCAAGTTGATCATTTCCCTCGGATAAG  
CTTATTAATGAAGGATTCAGCTTTAAGTTTGGGATTGAGGAAATCTACGACCAAACTGTAGAAT  
ACATGAACGCTAAGGGGCTGCTCAAGTGAAGAGTCCGCCTAACATTGTCCTAATGACTGTGATG

TTTGGTTGCTTAAGATGTATGCTGTCTTTTGTATATTATCCTAATAACTTGATGTTCTGCAAAT  
CAAGCAAATACCATATGGCGAATATCATTGCTTTCCCAAAGAAAAAGAAAAAAAAAAAAAGAA  
ATCCAAAGTATCCTATTTAGTATTGGAAGACCAAAAATCAAATCACCAACTGAATCATGGAATG  
GGTTCTTGTGTACTIONTATCAAATGACTATCATACTTTCTTCTGCGTCCAATTCTTCAACGTTCAA  
TTAAAGAAGGATCAACAGTCCCTTGTAGATCCAGTTTAAAAGTTGATCTTCAAAAAAAAAAAAA  
TCCAGTTTAAAAGTAGTTTCTCTGCTTTAGCAGGTGGTTTTTGGCCATGTTGCACCCTTGAGTTT  
CTTGGCTTGTGGGTCTCGGACAATGTCAGCACGATGTCCCTCATTGAGGGCCTTTTCTGGAGAC  
AGGATTGATACAAGTATAAGCAAGGGCAGCCATATGATTTCAGTTGTTGAACATCAAATGTTCT  
TGAAGTCGAGGATCTGCAATCTCTTCCACCCTACATCGTTTTCCACATCAATAGCCGCCTATTC  
ATGATCACAGAAAACAAGAAAAATGATATTGAATCATTTCATGTTTTGTGCATGCATAGGCCTCA  
AAATCAGGTGAGCAAAGAGCAGA

>gi|290579520|gb|GU324350.1| Theobroma cacao cultivar Scavina 6 anthocyanidin  
synthase (ANS) complete cds

ATGGTGACTIONTCAATGGCCCCCAGAGTAGAGAGCTTGGCAAGCAGTGGGATTCAGTCCATCCCGAA  
GGAGTACATTAGACCTCAGGAAGAGCTTACAAGCATTGGTAATGTGTTTTGAAGAAGAGAAAAA  
GAGGAAGGGCCTCAGGTTCCAACCATTTGATTTAAAGGAAATTGACTCAGAGGACAGAGAGGTAC  
GGGAGAGATGTCGCCAGGAGTTGAAGAGAGCTGCCACGGAGTGGGGTGTGATGCACCTTGTTAA  
CCATGGGATCTCGGACGAGCTCATGGAACGTGTCAAGAAAGCTGGACAGAAGTTCTTTGAACTT  
TCTGTCGAGGAGAAAAGAGAAGTATGCCAACGACCAGACTTTGGGGAAGATTCAGGGGTATGGCA  
GCAAGCTAGCTAACAATGCTAGTGGTCAGCTTGAGTGGGAGGACTACTTCTTCCATCTTGTGTAT  
CCCGAGGACAAGAGAGACTTGTCCATCTGGCCTCAAACACCAAGCGACTACACTGAAGTCACAAG  
TGAGTACGCAAGGCAACTCCGAGTCCTTGGCAGCAAATTTCTTTCGGCACTATCACTTTGCTTAG  
GATTGGAAGAAGGAAGGCTAGAGAAGGAAGTTGGTGGATTGGAAGAGCTCCTTCTTCAAATGAA  
AATCAATTACTATCCCAAATGCCCTCAACCAGAACTCGCTCTCGGTGTGGAAGCTCACACAGATG  
TAAGTGCACCTTACCTTCACTTCTCCACAACATGGTCCCTGGCCTGCAACTTTTCTACGAAGGCAAG  
TGGATCACCGCAAATGTGTTCCAACTCCATCATCATGCACATTGGTGACACCGTCGAGATCCT  
CAGCAATGGTAAGTACAAGAGCATTCTTACAGGGGTCTGGTTAACAAGGAGAAGGTTAGGATC  
TCATGGGCAGTTTTCTGTGAGCCGCCAAAGGAGAAGATCATTCTCAAGCCACTGCCAGAGACTGT  
GTCCGAGACGGAGCCTCCGTTGTTCCCTCCTCGCACCTTTGCTCAGCATATTCACCACAAGCTGTT  
TAGGAAGACCCAGGATGGCCTGTCTAATTGA

>gi|290579518|gb|GU324349.1| Theobroma cacao cultivar LCT-EEN 37 anthocyanidin  
synthase (ANS) gene

TATTGTGTGGAAAA  
ACTAAAAGGCTTTTACTCTTCTGGTCAAGAACAAGAAGAATGGGCAAGCACCAAAAAAGCAAAA  
AAGTCTTTGTTCCCTTCTTTTGTTCGACCTGTTTATCCCATAGTTCATATAAAATCACATTTTT  
GGCCAATTTTTTAGGGACGAAGAAACAAGGAACGGAGGCCAAGGCAAGATAGGGGCCGGTGGTT  
TTGCTAGCAAATACATACGAAAATTAATTGAACTAGGTAGCAGCAGGCATATATCCTGCTGACT  
GAAAGCTCGTAGAGATGAAGCACGAGCAACCAACTACCTCATTGTTCTTCCAGAAGCAACTCCTA  
GTTTTCGATCCCATGCAAGATCTTCAATCATATAACGTCTAGAACTTTCTTCTTTTCGTACATAATA  
AGTAATGTTCAAATCAACCATTTGTTAAAAGCAAACCATATGGAATAAATTTGAATTAACC  
TATTTAAATTTCAATTGAAGCTTTTTTTCGGAATGAATGGTCCATAAACTAGACTTTCTAATGCT  
TAGACTAACAAGGTGTATATATATATATATATGTTAAGAGTATATAATTTTTGATCATTTTTATTTCG  
TTATTAAAAATATATTTTACAAGTTTTATTACTTTATAATATATAATAAAACGAAAGGAGTA  
TTATTATTCATAAAAAAAGAAATCCAATTCTCATCTCATGCAATTGTTGAGTCAAGGCCTT

AATGTTTTTTGAGTTCAATCAAACCTTTAATGTTTCCAAAAAGAGGGCAGGGGAGGGATTCAATT  
AACTCCGCTAATGATGATTAGCTGTTGAAATCATTTGAGTCTCTCTGCCATTTGGGGTTAAATG  
AATCCAAATTAAGATGGGTTAGATGAAACGTGCAGTCTGGCTTGGTAGTTGGACTTTCCAAGT  
AGAAATTTTGGTCGTTATTTATCCGCGCTCTGCTTAATTAATTAGTCAACTCTCCTGTAAAGCAA  
ATCAGCTAATTTGCTTAAACTACCCATTACTACTATGTACATTAGCTCAAGAAATGTGCACCTTA  
GGCATTGCTCCATTGCCTGGTGTAAATTAAGTTAAAGTACAAAGTGACTTAATAGAAAGAGTGT  
TTATTATGACATTATTAATACTTTTAATTTCTCTCAAATAACATTATTTAAGACATGGATAATT  
AACTTTTATGTATGTATAATCTTTTCATTTTATAAAAAGTTAAACCATTGATAAACAGGTTATCT  
AGCATGGTTCAAAAAACAGTAAGTAATTTAGAATAGTACAATTTAATATTTAAATTAAGAGAT  
ATTGAAAACCTATATAGTATTAGATGTATATATTGAACATTAATTACTTGTGTAATGGATGTTT  
TCATTTTTACATATAATTTTAGGTTACAAGAATTATATGATGAATGGAAAAGAAACAAAAGCA  
AAACAAGTTCTACCTCACAGGCGCGTTTGGTTGAGATAGATTAGCAAATTAGAGCAGAGGGTGT  
TAGGTCCAAGCTTCCAGTCAACTCACCTTGGGACAACCAAAGTTGTATGACCACTGCTCTAACTC  
AGACCTTGGTGGAGCTCATCACGTGTATGACTTACCAGTTACATCTATTTTTCTTCAGTATTTCT  
TTCTCTTGATTTGGTAGCTCTACCCATTTGCATGTTCACTAAGGCAACTGATTTTTTTTTTTTT  
TAATATCATAGCTTTCTTGATCTTCCGTGTTAAAATTTTCTCGAACCAGATCATTATAAAAAGGC  
CACTAAAGATCAGCACTACGGTATATTCCTGAGAGTGAGGTTACCACAAAAGCAAAAAA  
AAAAAGGGTTGTTGTTACAGAGTGGAAACAAGGAACCTTCTAAAACAAGTTTAGAAGATCGCAAG  
AATGGTGACTTCAATGGCCCCAGAGTAGAGAGCTTGGCAAGCAGTGGGATTCAGTCCATCCCGA  
AGGAGTACATTAGACCTCAGGAAGAGCTTACAAGCATTGGTAATGTGTTTGAAGAAGAGAAAA  
AGAGGAAGGGCTCAGGTTCCAACCATTGATTTAAAGGAAATTGACTCAGAGGACAGAGAGGTA  
CGGGAGAGATGTCGCCAGGAGTTGAAGAAAGCTGCCACGGAGTGGGGTGTGATGCACCTTGTTA  
ACCATGGGATCTCGGACGAGCTCATGGAACGTGTCAAGAAAGCTGGACAGAAGTTCTTTGAACT  
TTCTGTGAGGAGAAAGAGAAGTATGCCAACGACCAGGCTTTGGGGAAGATTCAGGGGTATGGC  
AGCAAGCTAGCTAACAATGCTAGTGGTCAGCTTGAAGTGGGAGGACTACTTCTTCCATCTTGTGTA  
TCCCGAGGACAAGAGAGACTTGTCCATCTGGCCTCAAACACCAAGCGACTACACGTGAGTTTATG  
GCTTTTGGTTTATTTTACATACTGCTTTTTGCAATTACTAGATTCTTTGATCGATTAATGTTAAT  
GTTTCTTGAGCATCATATCAAACAAGCTGTATATGTCCACCGGTTTCATTGAACACTATCACAAT  
TTTTTTTTTAAAAGTAAAACCTTTCACATTTAATAAAAAGATCTACAAGGTTGGCAATTATCTG  
TCTGCCTGATTAGATAGAAAATTTTCTAATATTCAGGATACTTATTACAGTAAGAACAATATT  
TCTGTGATATGAAATATTAAGTTAAACGTAAACTATCCGTATGGATTTTAAACAATTCACCACT  
GTTCAATTGGTTACTATGCAGTGAAGTACAAGTGAGTACGCAAGGCAACTCCGAGTCTTGGCAG  
CAAAATCTTTTGGCACTATCACTTTGCTTAGGATTGGAAGAAGGAAGGCTAGAGAAGGAAGTT  
GGTGGATTGGAAGAGCTCCTTCTCAAATGAAAATCAATTAATCTATCCCAAATGCCCTCAACCAGA  
ACTCGCTCTCGGTGTGGAAGCTCACACAGATGTAAGTGCACTTACCTTCACTTCCACAACATGG  
TCCCTGGCCTGCAACTTTTCTACGAAGGCAAGTGGATCACCGCAAATGTGTTCCAAACTCCATC  
ATCATGCACATTGGTGACACCATCGAGATCCTCAGCAATGGTAAGTACAAGAGCATTCTTCACAG  
GGTCTGGTTAACAAGGAGAAGGTTAGGATCTCATGGGCAGTTTTCTGTGAGCCGCCAAAGGAG  
AAGATCATTCTCAAGCCACTGCCAGAGACTGTGTCCGAGACGGAGCCTCCGTTGTTCCCTCCTCG  
CACTTTGTCTCAGCATATTCACCACAAGCTGTTTAGGAAGACCCAGGATGGCCTGTCTAATTGAG  
GCTAGTCATTAGTTAAATTAATAATATCTTCTTGTTTTTAACGTCTTTATAAGCTGTTTACGGG  
TCTGGTGATGCTATATTATCTTGGGTTAAACCTTTGGTTGTGGTAGGCTGATGCCGGGGTGGTGT  
CTGTCTTTCACTCCTTGGCTTCTCTTTACCTGCTTTATTGAATAATGGCAGACTGATTTGCTTCC  
TTGTGTTAAGCAGTTTGTGAATTAATGGTCTTTGTTTACATTTCTTCCCTCAATTTACGTCTCCA  
TAAACAGAACTCTTTCCCTCCACACTTTCCTTCTTAGTCTAAATTTTTTAATACAATAGCAATC  
TTTTTCTTAAACAAATCAAGTGAAGTACCTGTAATTATCTAAGTAGTGAATCAAACCCTAAACA

GGCAAGTTTTTGCCTCCCTTCCGTTCTTTTTATTCCGAACACCCAGGAAATTAACAAAAGGTAAA  
TTGTCCCCAGTGGCACTCGGTTAATTGTCGTTTAGATTTTGATATGTATAACTTTGTGTTGGGGG  
CATTTTCTGCCCTGTGAAGGTCAAGCAGCCATGCTAACAGTATAACTATTAAGTAGTCTCAATAA  
TGAAG

>gi|290579524|gb|GU324352.1| Theobroma cacao cultivar Scavina 6 leucoanthocyanidin reductase (LAR) complete cds

ATGGATATGAAATCAACAAACATGAATGGTTCCTCTCCTAATGTCTCGGAAGAACTGGTCGGA  
CCTTAGTCGTTGGTTCGGGTGGGTTTATGGGCCGGTTCGTCACCGAAGCCAGCCTAGACTCCGGC  
CGTCCTACGTATATTTTGGCTCGGTCTAGTTCGAACTCTCCTTCCAAAGCCTCCACCATCAAGTT  
TCTTCAAGACAGAGGAGCCACTGTTATTTACGGCTCTATCACAGACAAAGAATTCATGGAGAAG  
GTTCTGAAAGAACATAAGATAGAAGTTGTAATATCTGCAGTGGGAGGGGAAGCATCTTAGACC  
AGTTCAATCTGATAGAGGCTATCAGGAATGTTGACACTGTCAAGAGGTTCTTACCGTCTGAATTC  
GGCACGACACAGACAGGGCTGACCCGGTGGAGCCAGGGCTGACCATGTATGAACAAAAGAGGC  
AGATTAGGAGGCAGGTAGAGAAATCTGGGATTCCTTACACTTACATATGTTGCAATTCCATTGC  
AGCTTGGCCCTACCACGACAACACTCACCTGCAGATGTTCTGCCACCCCTTGATAGGTTCAAAA  
TATACGGTGATGGCACTGTCAAAGCATACTTTGTGGCGGGTACCGATATTGGGAAGTTCACTAT  
AATGTCGATAGAAGATGATCGAACACTGAACAAAACACTGTCCATTTTCAACCTCCAAGCAACCTAC  
TAAACATAAACGAGATGGCCTCACTATGGGAGGAGAAGATTGGACGTACACTTCTAGGGTCAC  
CATCACAGAAGAAGATCTGCTGCAGATGGCCAAAGAGATGCGGATCCCACAGAGTGTGGTTGCA  
GCATTAACTCATGATATTTTCATAAATGGCTGCCAAATAAACTTTAGCTTGGACAAGCCAACCTG  
ATGTTGAAGTCTGCTCCCTCTACCCAGACACTCCTTTTTCGAACCATCAACGAGTGCTTCGAGGAC  
TTTGCCAAGAAGATAATTGATAATGCCAAAGCAGTGAGCAAGCCAGCGGCAAGCAACAATGCAA  
TATTTGTGCCAACTGCTAAGCCAGGAGCATTGCCTATCACTGCGATATGCACATGA

>gi|290579522|gb|GU324351.1| Theobroma cacao cultivar LCT-EEN 37 leucoanthocyanidin reductase (LAR) gene

TGAGCAGCACTGATGTAAATTAATAAAGTTTAGGACTAATAAAAAAATTCATTCAAAAAATTT  
GGTAAATAACAAAAAATTTACCATATGACTTGAAGACGAAAAAATTTGTTTTCAAAAAATCAA  
ACTCCATGTGACCCAAGAATCATAATAAAATCCCTATTAATAATGATTCTCACTTTTAATTTAAA  
AAAAAAAACCAACTCTATATGTAAGATAGATGATAAAATTTGATTAATTCAATCTCACATG  
TTGGGACTAGTAAATTCATAAGATAATATGATTCTTTCTGACAACCAATCAGGAAGAATTC  
AATCAATTTTTGCTTAATAAAAAAAGATTATCCTGTCTGCCATTGTTGTCAGAGGGTTTTTTTT  
TTTTCACAGTTTAAGATCTATGTTTTATATAATTAATGAGGCCATTCTTCTATATATATATATA  
TATATATGGAGTACTTGCCTAATTTAATCTCATTTAAATTTAATTTACTAAAAAAGTATATA  
TTAATTTTGGTCCATGCCTGCACAACAAAAAATATTATTTCTAAATTATATCGATAATTACAAT  
TTACAAAGCTAAAATAAAATAAAATAAAATTAATTAAGCCTATTTGGTTTGATTTTTTTTA  
AACTTAAAAATTAATATAAACTCTTATGAAAAATAATAGCTTCAAAATTAAGTTAAATCTG  
TTTGGTAAATTTACTTTTATAAGCTCTATTTTATAGATAAGTTGTTTTGCAAAACAATTTATAT  
AAGTTTTAATTTTAGTTAAAATTACAATCAAGGGTGTGTGTAATAAATAACTTTTTATTAATCA  
TAAATTTTTTTTAGAACATAAAAAAAGTTAATTTTATTTCTTGTTCAATTTGAAATAAAAAAT  
ATAAAATTTATAATTAATAAATAAATCATAAAAAAAGAAAAACAGATAATATTAATAATTTTT  
TAATAGATAAGAACATTTTGAACAAAAAGAAAAGTTTTTCCGGAAAAAATATATGTTTTTAAA  
AATAGAGAAGAGAAATCTCTTCCAAAGGTTATTTTAAAGCTCTTTTTTTTTTTGTTTTGAAAT  
TTTTGTTTTGTTAAAAGAAATTTTTTTTTAAAAATTTCTCAAAATATCTATTTGACTTGATTTTT  
ATTTTTAAGAGACAGATAAGTTGAAAAAAAATCAAAATTAATCGGCACTAAATGTTGAAAAA

ATTTAAAATTAATTACTAGAAGAAATATTTATGGGGGAAGAAATTTTAATTTTCAGAAGAAAAAT  
AGAAAAAATTATATGTTTGGAGTAGCTAGGCACCTAGCTGTTTTAGTAGAGTTGACTTGGGCCA  
CAATGGGAATGTGCTGTCTGACAGTCAAGGACGTGGCTATCCAACCACACTGTCAATCAAAAACC  
CACCAATCAGTCTTCCTCCTCCCCCTCCCCAGCCTCTTTTACAAAAGTACATACGTATATTTTCTTT  
TCTGTATTTTTGTTAGTTATAGTACAAATCAAATCGGGTTTTAAGGGAATTTTTAATCTACAAA  
CTATTTCTATACACAAAATTAGCCATACTAAGGAAGAAAAAAAATTAATAATATAGCACATTA  
GTAATATAAATTAATTACTCGATACAATGATAAAATTGCTTCAGTTTTACATCAAAACTTGATT  
AAAAAAATTATCGATTGATTTTTCTTTTTAATCTTTAAATTTAATAATTTTTTTCCCTTAAATT  
TAACACGATAAGAATGTTCTGAATGAGAGTCCGGATGCTAGCTGAAGTTACTAATTAAGAAAAAAG  
TATACAACCTTTTTAGCAAAAATGAATAATAGGATATTTGGTTTTATTATCCTATTATTCACTAT  
TTTGCTAAAAAGTTGTATAATTTTTATAAAAAAAATAAATAAAGGGGGAAAAAGAGGATAAAA  
AAATACTCAATGCTACCTAATAAAATGGCTACATACGGGTAGACAACAACCTCATGCTACGAAAA  
TTGCAATCCATGTTCCCTGTTGCTAATTTGCGCCATTGCTTTTTGCTTACCTGCCCTTAATTGC  
TAACCTCTATATAAGCACAAAGTCCATATTGCTTTTTGGTCACCGCCACATTCCTCACTCTCTCGT  
CACTCTTTTTATTTTTTTTTCTGGTTTCCTTTGTGCGCCAAAACCTTAAGCTTAAGTAAAAGCAAA  
CAATATGAAATCAACAAACATGAATGGTTCCTCTCCTAATGTCTCGGAAGAACTGGTCGGACCT  
TAGTCGTTGGTTCGGGTGGGTTTATGGGCCGGTTCGTCACCGAAGCCAGCCTAGACTCCGGCCGT  
CCTACGTATATTTTGGCTCGGTCTAGTTCTGAACCTCTCCTTCCAAAGCCTCCACCATCAAGTTTCT  
TCAAGACAGAGGAGCCACTGTTATTTACGTATGTACAATTCTCCCTCGACACCTCTTCCATTTTC  
TGTTACATTTCCACACGTATACAAATACATATACATTTCTAATGTGTAATTATTTGTGTATAT  
TTATATATATGTAATGTATAATGTGTAATTATTTATGTATGTATATATGTATGTATGTATGTAT  
GTATGTATATATGTATGTACGGACGTTATACAATCTTCGGAATTGTTGTAACAGGGCTCTATCA  
CAGACAAAGAATTCATGGAGAAGGTTCTGAAAGAACATAAGATAGAAGTTGTAATATCTGCAGT  
GGGAGGGGAAGCATCTTAGACCAGTTCAATCTGATAGAGGCTATCAGGAATGTTGACACTGTC  
AAGGTATATGCTCAAAACAACAACCTAACATTCATAGGGGAAGAACTTAGATCTTGTATATGGT  
CAATGTAGTGACTTGTTTGGTATGTTTGGCTTCTAGTTAGAATAAAACACTTATTGCATGCCT  
AGCTAAAAGTTAGGAACTTCTTTTGAACCTAGTTTAGCTAGAGCTAAGCTATTCTAAGCAAGA  
AGACATTAATAGTACCTAAAGCTATGTTTTTCTATTTAATTCAACAAGCATTGCACAAAATGG  
GTTAATGAGTCAAAGGTGAATCTGTTGCAGAGGTTCTTACCGTCTGAATTCGGGCACGACACAG  
ACAGGGCTGACCCGGTGGAGCCAGGGCTGACCATGTATGAACAAAAGAGGCAGATTAGGAGGCA  
GATAGAGAAATCTGGGATTCCTTACACTTACATATGTTGCAATTCCATTGCAGCTTGGCCCTACC  
ACGACAACACTCACCTGCAGATGTTCTGCCACCCCTAGATAGGTTCAAATCTACGGTGATGGC  
ACTGTCAAAGGTACCTCATCTTTCTTTTTCTTTCATTGGTTTTGTTTTTGTGTATCTTGACTTT  
AGTGTGGTTGGTAGATGGAACCGATAGCATAGAAATTTGGGAAGAGAAGGATGGATGAGAGA  
TTTTGGTTTTCCAAGAAAATCAATGTCCAGGATCTCTCATTAAATAACCACCTCCCACATGATA  
TTCCATCTTATCATATCAAAGTAAATAGATTACTAGCATTTGTTTGGTTGAGTTTTAAATGTCTT  
CTATTAGAAGCTAACATTGGGAAATTAATTAGAGGTCAGATCATAGGAAAAAATTTAATGGTT  
GAGGTAAATGTTGCCTGCAATTTAATTCTGACATTGGTTGAGTTTGGTGAAACAAGAAAAATTT  
GAGTTAATTAGTGGTTTGTCTCTGTTGTTTACAAAGATAGATAATACTGGATCTCTTAGTTTGG  
CTGGCAAATATTCAGCAGTCTTACAGTTTTAAGAACAAATGCTGCCTACCTACTGCCATCCCATT  
TTCTAGTAGAAGAAGGAAGAGACAAGGGACTGAATCTTTTGAATGAAAAACCAAAACACATG  
TAAATGATTGGATAAGAAAAAATATTAGGGAAAAAAGAAATAGCAGTCATAGTCAAAGTGCT  
GATCAGGTGTCTAGATATATACAGGTATAGCATGTTATATTCTAGACGAAGGCACTGTATGGCA  
GCAGGTATAGGCTTCAAATAAACTTTTTATCTTATATGGCCTGCTGCTTTGACGAAATTGAAAAT  
TTATCATCAACTAGGCCAATTATCGTTTAAATTCAAAATAGCTTTCAAACCTAAATGGCCAT  
TTCTTCAAAGCCCCGAGTAAACCTTTTTGTCCCATCTTTTTGGTAGTCGATAGTATCCACTTCAT

TTGGTTAATGCAATTATTTTCATTCTAAACATTTCTGCTTCAATAATGACTTCCTCCAAATCTGGA  
TAGCCAGAAGGGATATTTTCTACCAAATGGGAAGCTTTGAACTCACAGGCGAAAAGGGGTAATT  
TTTTTTCTACCATGCCGTAACCAGCATAATATCATCACAAATCCATGATCATATTTTTACTAAAT  
AGATCTCATATTTTCATAGTATTTCTTACAGCTCTAAACCTACTCATTTCAGCATTGTAAAATTGAG  
CAGCCAAGAGAGCTAGTCCAGCTTGTCTTTTGATAGAGGACAAAGGAGAAAGAGTCTTAGTCA  
AAGTACAATGCACCCTGCTTCCTCTTTTTCTTTAGTAAAATAGTAGAAGCGACACCAGTTCTAAA  
TAGGTTTTGCATCTTGGCTAGTTCCAAGAAATTTGCTAGTCATTAAGGCAATGCCCCATTAGGAA  
GGAAAAGTCGTAAGAATCGGTGGACCTCCCTTAGATTCCCAACGGATAGAATGATGTTTTGCTTC  
TTTCTTCTTGTGTTGTTGAGCAGCACTGAGAGGCACGTGACTGCTATGATGGGTTAGGTAGCAGATT  
ATCACGTGACCCATCCTCTCCTAACACTTTCCCATCCCTTTCTTCTTTTACCCCCAACAAACA  
CACAAAAGGGTTATTTATTAGCTACCGGGAACATTACATTAAGCATCAAGTTAATAATAGTT  
TGGGAATTGAATTTTTAAACCTTTGACTCCGTCAGTTTAGATCTTTCATTTTCAAATTGAGTTAT  
TAATAAATTATTACAAAATAATTGACATAATATAGGACAACCCAATTTACTTGGGTAATCACTA  
TATTTTAAAGGCTATGTTGTTAGGCAGCCTATTCTAGGAGGAGTCAATGTCGACAACCTGGACAT  
TTGGGGTAAAAAGAAGTCCAAGATTTGATCATTTCATAGGCTGTCCTAAGCTAATCGAAAAGGAG  
GAAGTCCCAACTAATTAAGTGTTTGGTAAACAAGTTTATTCTCAAGAGAGGACCACCGAATTC  
ATGTCAGAGATTTGCTTATTAATTCAAAGATTTGGACTTTTGGATGTTGCCCGTGAGTTTCTGAC  
GTTGGCTCAGAGCAAGTCTTTGATCTTCTTGTCAAAGAACTGCTCTACCCTTCTAATAATGAGT  
AAACCAGTCACAGAAAGCAACCCCCAGGTCCAACCTACTCTTCAACAGTTTGCTAGTTTAAAAAA  
AGAAAAGAAAAATTGGGGTTTGCATGTACCTAAGTTACCATCTCTGCATTGAATTTAAGTTGT  
AGTAGTAGTTGTAAAATAAAGTCTAGAACCAAAGCTTTTAATGAAAGACAACCTAGGTCCATA  
ACTCACCATCATTATGAGAATTTTGAATATCAATCACCTTACTGTATATTATTTATGATGTTGAC  
TCTATGATTTTTGTGGTGAATGCCCAGCATACTTTGTGGCGGTACCGATATTGGGAAGTTCACTA  
TAATGTGCATAGAAGATGATCGAACACTGAACAAAACCTGTCCATTTTCAACCTCCAAGCAACCTA  
CTAAACATAAACGAGATGGCCTCACTATGGGAGGAGAAGATTGGACGTACACTTCCTAGGGTCA  
CCATCACAGAAGAAGATCTGCTGCAGATGGCCAAAGGTTTGTCTAATTATTTTTCAGTTTTCTTT  
AAGTTTTGGTTCAAGCAACTTAACCTTCTCCAAGGAACCTATATGCCACTCGGTTGGCTCCATT  
AAGCATTAAATCCATGAAGCAGTAAGTTCTTGCCTAACAAAATGGATGCTAACCCAACCTTCTGAT  
ATAAATGCAGAGATGCGGATCCCACAGAGTGTGGTTGCAGCATTAACTCATGATATTTTCATAA  
ATGGCTGCCAAATAAACTTTAGCTTGGACAAGCCAACCTGATGTTGAAGTCTGCTCCCTTACCCA  
GACTCCTTTTGAACCATCAACGAGTGCTTCGAGGACTTTGCCAAGAAGATAATTGATAATGC  
CAAAGCAGTGAGCAAGCCAGCGCAAGCAACAATGCAATATTTGTGCCAACTGCTAAGCCAGGA  
GCATTGCCTATCACTGCGATATGCACATGAGAAATATCTCACTCTATCCATTTCCACATCAATAA  
TTCTTTTACAAGTTCTTTTAATCGTACAATGGTAAGAGACTTATCTGTTGCCAGTGTTCGGCA  
AAAATAATCANATGTATCTCTTGAATAAATATC

>gi|290579512|gb|GU324346.1| Theobroma cacao cultivar Scavina 6 TT2 like MYB  
transcriptoin factor (TcMYBPA) complete cds  
ATGGGAAGGGCTCCTTGTGTTCTAAAGTTGGGTTGCATAGAGGTCCCTGGACTCCTAGAGAAG  
ACACATTGCTTGTCAAGTACATTCAAGCTCATGGTGACGGTCACTGGAGATCACTTCCTAAGAAA  
GCCGGGCTTCTTAGGTGTGAAAGAGTTGCAGGCTCAGATGGATGAACTATTTAAGACCAGATA  
TAAAGAGAGGGAATATAACTCCCGATGAGGATGATCTTATCATCAGATTACATTCCCTCCTCGGC  
AATCGGTGGTCACTCATTGCCGGAAGGCTTCTGGTGAACCGATAACGAGATTAAAAATTACTG  
GAACACCCATCTGAGTAAAAGACTTCTAAGCCAAGGGACTGACCCTAACACCCACAAGAACTAT  
CAGAGCCCCAGTTCAACAAGTGAAGAAGAGAAAAAGCAGCAGAGGCAACAGCAACAAGAAGCA  
GAACAATAGCAAGGGCAAAGGCGCAAAGGTTGAGCCAGAAGAGCCCAAAGTCCATCTCCCTAAG

CCCGTTAGAGTAACTTCTTTCTCTTTACCAAGAAACGACAGCTTTGACCAATGTAATACGTTTAG  
CACGGTGTCTTCAAGCCAAGGAGGAGAGGGAGGATTGGGTACAGAGGTTGTACAGGGACCTTGG  
TCAGATAATGTCAACGATGATGAAAATGGGACCGGATTTCTTGCTGCTTATGATGATCATGGTT  
TTGTTAACGGTTCAGATTTTCGAGTGCCAGTCTCATGTACCAGCAAGTGATGACGATAATTCTCTC  
GAGAAGCTTTACGAAGAGTATCTCCAGCTTCTGAAGACAAACGATGATCAAGTGCAGTTGGATT  
CTTTCGCTGAATCATTATTGATCTGA

## APPENDIX B: PROTEIN SEQUENCE COMPARISONS OF TCANR, TCANS, TCLAR AND TCMYBPA WITH THEIR HOMOLOGUS FROM OTHER SPECIES

### i. Protein sequence comparison of TcANR and AtBAN (BANYULS)

Score = 437 bits (1125), Expect = 7e-121, Method: Compositional matrix adjust.  
Identities = 211/330 (63%), Positives = 264/330 (80%), Gaps = 0/330 (0%)

Query	7	GKKTACVVGGTGYVASLLVKKLLEKGYAVNTTVRDPDQKKIPHLVTLQKLGDLKIFRAD	66
		G K ACV+GGTG +AS+L+K LL+ GY VNTTVRDP+N+KKI HL LQ+LGDLKIF+AD	
Sbjct	9	GSKKACVIGGTGNLASILIKHLLQSGYKVNNTTVRDPENEKKAHLRKLQELGDLKIFKAD	68
Query	67	LTDEGSLDVPIAGCDLVFHVATPVNFASQDPENDMIKPAIQVNLVNLKACAKAKTVKRVV	126
		LTDE S + +GC+ +FHVATP+NF S+DPE DMIKPAIQGV+NVLK+C K+K+VKRV+	
Sbjct	69	LTDEDSFESSFSGCEYIFHVATPINFKSEDPEKDMIKPAIQGVINVLKSCLSKSVKRVV	128
Query	127	LTSSAAAVSINTLKGTDLVLTEKDWTDVEFLSSAKPPTWGYPAKTLAEKAAWKFAQENN	186
		TSSAAAVSIN L GT +V+ E++WTDVEFL+ KP WGYP SK LAEK AW+FA+EN	
Sbjct	129	YTSSAAAVSINNLSTGTGIVMNEENWTDVEFLTEEKPFNWGYPISKVLAEKTAWEFAKENK	188
Query	187	IDLITVIPSLMTGPSLTPDVPSSIGLATSLLSGNEFLVNLKGMQMLSGSISITHVEDVC	246
		I+L+TVIP+L+ G SL D PSS+ L+ S ++G E V LK MQ LSGSIS HV+D+	
Sbjct	189	INLVTVIPALIAGNSLLSDPPSSLSLSMSFITGKEMHVTGLKEMQKLSGSISFVHVDLLA	248
Query	247	RAHVFLAEKESASGRYICCAVNSSVPELAKFLNQRYPEFKVPTDFGDFPSKAKLISSDK	306
		RAH+FLAEKE+ASGRYICCA N+SVPE+A FL QRYP++ V ++F + S KL +SS K	
Sbjct	249	RAHLFLAEKETASGRYICCAVNTSVPEIADFLIQRYPKYNVLESEFEGLSIPKLTLSQK	308
Query	307	LINEGFSFKFGIEEYDQTVVEYMNAKGLLK 336	
		LINEGF F++GI E+YDQ +EY +KGL+K	
Sbjct	309	LINEGFRFEYGINEMYDQMIIEYFESKGLIK 338	

### ii. Protein sequence comparison of TcANS and AtLDOX (ANS)

Score = 591 bits (1524), Expect = 3e-167, Method: Compositional matrix adjust.  
Identities = 282/345 (81%), Positives = 311/345 (90%), Gaps = 0/345 (0%)

Query	8	RVESLASSGIQSIPKEYIRPQEELTSIGNVFEEKKEEGPQVPTIDLKEIDSEDRVRRER	67
		RVESLA SGI SIPKEYIRP+EEL SI +VF EEKKE+GPQVPTIDLK I+S+D ++RE	
Sbjct	6	RVESLAKSGIISIPKEYIRPKEELESINDVFLEEKKEDGPQVPTIDLKNIESDDEKIREN	65
Query	68	CRQELKRAATEWGMHLVNHGISDELMEVKKAGQKFFELSVEEKEKYANDQTLGKIQGY	127
		C +ELK+A+ +WGMHL+NHGI +LMERVKKAG++FF LSVEEKEKYANDQ GKIQGY	
Sbjct	66	CIEELKKASLDWGMHLINHGIPADLMERVKKAGEEFFSLSVEEKEKYANDQATGKIQGY	125
Query	128	GSKLANNASGQLEWEDYFFHLVYPEDKRDLSIWPQTPSDYTEVTSEYARQLRVLASKILS	187
		GSKLANNASGQLEWEDYFFHL YPE+KRDLSIWP+TPSDY E TSEYA+ LR+LA+K+	
Sbjct	126	GSKLANNASGQLEWEDYFFHLAYPEEKRDLSIWPKTPSDYIEATSEYAKCLRLLLATKVFK	185
Query	188	ALSLCLGLEEGRLEKEVGGLEELLQMKINYYPKCPQPELALGVEAHTDVSALTFILHNM	247
		ALS+ LGLE RLEKEVGGLEELLQMKINYYPKCPQPELALGVEAHTDVSALTFILHNM	
Sbjct	186	ALSVGLGLEPDRLEKEVGGLEELLQMKINYYPKCPQPELALGVEAHTDVSALTFILHNM	245
Query	248	VPGLQLFYEGKWITAKCVPNSIIMHIGDTVEILSNGKYKSIHRGLVNKEKVRISWAVFC	307
		VPGLQLFYEGKW+TAKCVP+SI+MHIGDT+EILSNGKYKSIHRGLVNKEKVRISWAVFC	
Sbjct	246	VPGLQLFYEGKWVTAKCVPDSIVMHIGDTLEILSNGKYKSIHRGLVNKEKVRISWAVFC	305
Query	308	EPPKEKIILKPLPETVSETEPPLFPRTFAQHIHKKLFRKTQDGL 352	
		EPPK+KI+LKPLPE VS P FPPRTFAQHI HKLF K Q+ L	
Sbjct	306	EPPKDKIVLKLPEMVSVESPAKFPRTFAQHIEHKLFGKEQEEL 350	



### iii. Protein sequence comparison of TcLAR and DsLAR

Score = 434 bits (1116), Expect = 8e-120, Method: Compositional matrix adjust.  
 Identities = 199/326 (61%), Positives = 259/326 (79%), Gaps = 2/326 (0%)

```

Query 8  MNGSSPNVSEETGRTLTVVGGSGGFMGRFVTEASLDSGRPTYILARSSNSPSKASTIKFLQ 67
      ++G+ P++++  RTLTVV  GF+G+F+T+ASL  G PT++L R  SPSKA  IK  Q
Sbjct 3  VSGAIPSMTK--NRTLTVVGGTGFIGQFITKASLGFGYPTFLLVVRPGVPVSPSKAVIIKTFQ 60

Query 68  DRGATVIYGSITDKFMEKVLKEHKIEVVISAVGGGSILDQFNLEAIRNVDTVKRFLPS 127
      D+GA VIYG I  DKE MEK+LKE++I+VVIS VGG  +LDQ  L+EAI++V T+KRFLPS
Sbjct 61  DKGAKVIYGVINDKECKMEKILKEYEIDVVISLVGGARLLDQLTLEAIKSVKTIKRFLPS 120

Query 128  EFGHDTDRADPVEPGLTMYEQKRQIRROVEKSGIPYTYICCNSIAAWPYHDNTHPADVLP 187
      EFGHD DR DPVEPGLTMY++KR +RR VE+ GIP+T ICCNSIA+WPY+DN HP+ V P
Sbjct 121  EFGHDVDRDTPVEPGLTMYKEKRLVRRAVEEYGIPTNICNSIASWPYYDNCHPSQVPP 180

Query 188  PLDRFKIYGDGTVKAYFVAGTDIGKFTIMSIEDDRTLNKTVHFQPPSNLLNINEMASLWE 247
      P+D+F+IYGDG  KAYF+ G DIGKFT+ +I+D RTLNK VHF+P SN  +INE+ASLWE
Sbjct 181  PMDQFQIYGDGNTKAYFIDGNDIGKFTMKTIDDRTLKNVHFRPSSNCYSINELASLWE 240

Query 248  EKIGRTLPRVTITEEDLLQMAKEMRIPQSVVAALTHDIFINGCQINFSLDKPTDVEVCSL 307
      +KIGRTLPR T+T + LL  A E  IP+S+V++ THDIFINGCQ+NFS+D+ +DVE+ +L
Sbjct 241  KKIGRTLPRFTVTADKLLAAHAENIIPESIVSSFTHDIFINGCQVNFSDIHSDEVIDTL 300

Query 308  YPDTPFRTINECFEDFAKKIIDNAKA 333
      YPD  FR++++C+EDF  + D  A
Sbjct 301  YPDEKFRSLDDCYEDFVPMVHDKIHA 326
  
```

### i. Protein sequence comparison of TcMYBPA and VvMYBPA1

Score = 364 bits (935), Expect = 5e-99, Method: Compositional matrix adjust.  
 Identities = 201/295 (68%), Positives = 225/295 (76%), Gaps = 17/295 (5%)

```

Query 1  MGRAPCCSKVGLHRGPWTPREDTLLVKYIQAHGDGHWRSPLPKAGLLRCGKSCRLRWMNY 60
      MGRAPCCSKVGLHRG WT REDTLL KYIQAHG+GHWRSPLPKAGLLRCGKSCRLRWMNY
Sbjct 1  MGRAPCCSKVGLHRGSWTARED TLLTKYIQAHGEGHWRSPLPKAGLLRCGKSCRLRWMNY 60

Query 61  LRPDIKRGNITPDEDDLIIRLHSLGNRWSLIAGRLPGRTDNEIKNYWNTHLSKRLLSQG 120
      LRPDIKRGNITPDEDDLIIRLHSLGNRWSLIAGRLPGRTDNEIKNYWNTHLSK+L SQG
Sbjct 61  LRPDIKRGNITPDEDDLIIRLHSLGNRWSLIAGRLPGRTDNEIKNYWNTHLSKLLRSQG 120

Query 121  TDPNTHKKLSEPPVQVQVKKRKSRRGNSNKKQNNKSGKGA-VEPEEPKVHLPKPVVTSF 179
      TDPNTHKK++EPP + K+RK++R +N  + + K +K  E  KVHLPKPVVTS
Sbjct 121  TDPNTHKKMTEPP--EPKRRKNTRTRTNNGGSKRVKISKDQENSNHKVHLPKPVVTSL 178

Query 180  -SLPRNDSFDQCNTFSTVSSSQGEGGLGTEVVQGPWSDNVNDE----NGTGFLAAYDD 234
      S+ RN+SF+  +TVS  G  G  E +  P  ++ DD+  +G  F  DD
Sbjct 179  ISMSRNSFES----NTVSGGSGSSSGGNETLPWPSFRDIRDDKVIGVDGVDFFIG-DD 233

Query 235  HG--FVNGSDFECQSHVPASDDNSLEKLYEEYLQLLKTNDQVQLDSFAESLLI 287
      G  V  SD  E  QSH+P +  DNSLEKLYEEYLQLL+  D  QVQLDSFAESLLI
Sbjct 234  QQQDLVASSDPESQSHMPPT--DNSLEKLYEEYLQLLEREDTQVQLDSFAESLLI 286
  
```

**APPENDIX C: MULTIPLE SEQUENCE ALIGNMENT OF THE LAR, ANS AND ANR PROTEINS AS WELL AS RELATED IFR AND DFR PROTEINS OF THE RED SUPERFAMILY.**

```

                *           20           *           40           *           60
AT1G75290 : -----MASE-----KSKILVIGCTGHIGKLTIEASVKAGHSTLALVREASLSDPNK : 46
AT1G75280 : -----MATE-----KSKILVIGCTGYIGKFLVEASAKAGHSTFALVREATLSDPVK : 46
AT1G75300 : -----MTTE-----KSKILVIGCTGYMGEFIVEGSAKAGNPTFALVREASLSDPVK : 46
AT1G19540 : -----MT-----SKILVIGATGLIGKVLVEESAKSGHATFALVREASLSDPVK : 43
AT4G39230 : -----MTS-----KSKILFIGCTGYIGKYIVEASARSGHPTLVLVNSTLTSPSR : 45
AT4G34540 : -----MEEEEK--KKSRLVITIGATGRIGNYLTRFSTIESGHPTFALLINTTLLSDKLL : 48
AT4G13660 : -----MKETNFGKTRVLVVGCTGSLGRRIVSACLAEGHETYVLOEPEIGVDIEK : 50
AT1G32100 : -----MGESKRTEKTRVLVVGATGYIGKRIVRACLAEGHETYVLOEPEIGLEIEK : 50
TcLAR : MDMKSTNMNGSSPNVSEETGRTLIVVSGGFMGRFVTEASLDGRPTYILLASSSNS-PSK : 59
VvLAR : -----MTVSP--VSPKGRVLIAGATGFIIGQFVAAAASLDAHRPTYILLAPGPRS-PSK : 50
LcLAR1-1 : -----MVSTAATPPATAGRILIIIGATGFVQGFMTKASLGLGRSTYLLLPGLS-PSK : 52
LcLAR1-2 : -----MVSTAATPPATAGRILIIIGATGFVQGFMTKASLGLGRSTYLLLPGLS-PSK : 52
MtLAR : -----MAPSSPTTPIKGRVLIIVGATGFVQGFVTEASISTAHPTYLLLPGLI-SSK : 53
DuLAR : -----MTVSGAIPSMTKNRTLVVCGTGEIIGQFITKASLGFYPTFLLVPPGVS-PSK : 52
MtDFR : -----MGSMAETVVCVTGASGFVGSWLVMLRMERGMVRATVDPENLKKVS : 46
AtDFR : -----MVSQKETVVCVTGASGFVGSWLVMLRLLERGYFVRATVDPGNLKKVQ : 46
ZmA1 : -----MERGAGASEKGTVLVTGASGFVGSWLVMLKLLQAGYTVRATVDPANVGKTK : 51
AtBAN : -----MDQTLTHPTGSKKACVIGCTGNIASILIKHLLQSGYKVNTTVDPENEKKIA : 51
LcANR : -----MASIKPIYEKKKACVIGCTGFVASLLIKOLLEKGYAVNTTVDPDNHKKIS : 51
TcANR : -----MASQTVG-KKTACVVGCTGFVASLLVKLLEKGYAVNTTVDPDNQKKIP : 49
VvANR : -----MATQHPIG-KKTACVVGCTGFVASLLVKLLEKGYAVNTTVDPDNQKKS : 50
CsANR : -----MEAQPTA-PKAACVVGCTGFVAATLIKLLLEKGYAVNTTVDPGNQKKS : 49
ZmA2 : -----MESSPLLQLPAARVEALSIISLSAISAIPEEYVRPADERAG-LGDAFDLARTHANDHT : 54
PfANS : -----MVTSAMGPSPRVEELARSGLDTPKDYVRPEELKSIIGNILAEKSSE---- : 49
AtANS : -----MVAVERVESLAKSCLISIPKEYIRPKEELES-INDVLEEKKED---- : 43
MtANS : -----MGTVAQRVESLALSIISIPKEYVRPKEELAN-IGNIFDEEKKE---- : 43
TCANS : -----MVTSMAPRVESLASSCLISIPKEYIRPQEELTS-IGNVFEEKKE---- : 45
                G   g           6           r

```

```

                *           80           *           100           *           120
AT1G75290 : GKTIVQNFKDFGVTLHGDLNDHESLVKAIK--QADVVIISTVG-----SMQILD : 92
AT1G75280 : GKTIVQSFKDLGVTLHGDLNDHESLVKAIK--QVDVVIISTVG-----SMQILD : 92
AT1G75300 : SKTIQSFKDLGVTLHGDLNDHESLVKAIK--QVDVVIISTIG-----HKQIFD : 92
AT1G19540 : AQLVERFKDLGVTLILYGLSLSDKESLVKAIK--QVDVVISAVGRF-----QTEILN : 91
AT4G39230 : SSTIENFKNLGVQFLLGDLDDHTSLVNSIK--QADVVIISTVG-----HSLIGH : 91
AT4G34540 : S-----LSDAGVTLLKGSLEDEGSLAEAVS--KVDVVISATP-----SKHVLD : 89
AT4G13660 : VQLLSFKRLGALHVEGFSFDHQSLVSAVK--QVDVVVSAMSGVHFR-----THNIPV : 101
AT1G32100 : VQLFLSFKKLGARIVEGFSFDHQSLVSAVK--LVDVVVSAMSGVHFR-----SHNIPV : 101
TcLAR : ASTIKFLQDRGATVIYGSITDKEFMKVLKEHKEIVVVISAVG-----GGSIDL : 107
VvLAR : ANIFKALEDKGAIVYGLINEQEAMEKILKEHEIDIVVISTVG-----GESILD : 98
LcLAR1-1 : AATIVKSFQDRGAKVIHGVINDELKELMVKILKDYEIDVVISLVG-----GGNIMD : 100
LcLAR1-2 : AATIVKSFQDRGAKVIHGVINDELKELMVKILKDYEIDVVISLVG-----GGNIMD : 100
MtLAR : AATIKTFQEKGAIVYGVVNNKEFVEMILKKEYEIDTVISATG-----AESILD : 101
DuLAR : AVIIKTFQDKGAKVIYGVINDKECMEKILKEYEIDVVISLVG-----GARILD : 100
MtDFR : HLLLELPGAKGKLSLWKADLGEEGSFDFAIKGCTGVFHVATPMDFEKDPENEMIKPTIKG : 106
AtDFR : HLLDLPNAKTLTLTWKADLSEEGSYDDAINCGDGVFHVATPMDFEKDPENEVIKPTVNG : 106
ZmA1 : PLMDLPGATERLSIWKADLAEEGFSFHDAIRGCTGVFHVATPMDFLSKDPENEVIKPTVEG : 111
AtBAN : HLRKQLQ-ELGDLKIFKADLTDEDSFESSFGCEYIFHVATPINFKSEDPEKDMIKPAIQG : 110
LcANR : HLLVQLQ-SLGLNLNIFGGELTVENDFDTPIEGSELVFLATPVNFASEDPENDMIKPAIKG : 110
TcANR : HLVTLQ-KLGLDKIFRADLTDEGSFDVPIAGCDLVFHVATPVNFASEDPENDMIKPAIKG : 108
VvANR : HLLLELQ-ELGDLKIFRADLTDELSFEAPIAGCDFVHVATPVHFASEDPENDMIKPAIQG : 109
CsANR : HLLALK-GSGNLKIFRADLTDEQSFDTPIVAGCDLVFHVATPVNFASEDPENDMIKPAIQG : 108
ZmA2 : APRIPVVDISPLLDSSSQQQQRDECEAVRAAAADWGMHTAGHGI PAELMDRLRAAGTA : 114
PfANS : GPQLPTIDLEEMD--SRDEEGRKKCHEELKKAATDWCVMHLINHGIPPEELDRVKAAGKE : 107
AtANS : GPQVPTIDLKNIET--SDDEKIRENCIEELKKASLDWGMHLINHGIPADLMERVKAGEE : 101
MtANS : GPQVPTIDLKEIN--SSDEIVRGKCKEELKKAEEWGMHLINHGISDDLINRLKAGET : 101
TCANS : GPQVPTIDLKEID--SEDEVRERCRQELKRAATEWGMHLINHGISDELMERVKAGQK : 103

```

```

*           140           *           160           *           180
AT1G75290 : QTKIIISAIKEAGNVKREFLPSEFGMDVDKS-SAVEPAKSAFGRKLOTRRDIEAEGIPVITYL : 151
AT1G75280 : QTKIIISAIKEAGNVKREFLPSEFGVDVDR-SAVEPAKSAFAGKIQIRRTIEAEGIPVITYA : 151
AT1G75300 : QTKIIISAIKEAGNVKREFLPAEFGIDVERT-SAVEPAKSLFAGKVQIRRAIEAEGIPVITYV : 151
AT1G19540 : QTNIIIDAIKESGNVKREFLPSEFGNDVDR-VAIEPTLSEFITKAQIRRAIEAAKIPVITYV : 150
AT4G39230 : QYKIIISAIKEAGNVKREFLPSEFGNDVDRV-FTVEPAKSAYATKAKIRRTIEAEGIPVITYV : 150
AT4G34540 : QKLLVVRVIKQAGSIKREFLPAEYGANPKDT-QVSDLDHDFYSKKSEIRHMIESEGIPIVITYI : 148
AT4G13660 : QLKLVAAIKEAGNVKREFLPSEFGMDPSRMGHAMPPGSETFDQKMEIRNAIKAAAGISHTYL : 161
AT1G32100 : QLKLVVAIKEAGNVKREFLPSEFGMDPPRMGHALPPGRETFDQKMEVRQAIEAAGIPVITYV : 161
TcLAR : QFNLEIARNVDTVKREFLPSEFGHDTDRA-DPVEPGLTMYEQKRQIRROVEKSGIPVITYI : 166
VvLAR : QIALVKAMKAVGTIKREFLPSEFGHDVNR-DPVEPGLNMYREKRRVROLVEESGIPVITYI : 157
LcLAR1-1 : QRTLVDIAIKSVKTVKREFLPSEFGHDTDRA-NPVEPGLTMYKPKRLIRRLIEESGIPVITYI : 159
LcLAR1-2 : QRTLVDIAIKSVKTVKREFLPSEFGHDTDRA-NPVEPGLTMYKPKRLIRRLIEESGIPVITYI : 159
MtLAR : QLTLVVEAMKSIKTIKREFLPSEFGHDVDR-DPVEPGLAMYKQKRLVRRVIEESGVPVITYI : 160
DuLAR : QLTLLEAIKSVKTIKREFLPSEFGHDVDR-DPVEPGLTMYKPKRLVRAVEEYGIPIVITNI : 159
MtDFR : VLDIMKACLKAKTVRREFLFTSSAGTLNVTEQKP--LWDESCWSDVEFCRRVKMTGVMYF : 164
AtDFR : MLGIMKACVKAKTVRREFLFTSSAGTVNVEEHQKN--VYDENDWSDLEFIMSKKMTGVMYF : 164
ZmA1 : MISIMRACKKAGTVRREFLFTSSAGTVNLEERQP--VYDEESWTDVDFCRRVKMTGVMYF : 169
AtBAN : VINVLKSKLKSIVKRVLYTSSAAAVSINNLSTGIVMNEENWTDVEFLTEKPFNIGYP : 170
LcANR : VLNVLKSCARAK-VKRVLLTSSAAAVTIGELKGTDLVMDENWTDVEFLSNAKPPVIGYP : 169
TcANR : VLNVLKACAKAKTVKRVLLTSSAAAVSINTLKGTDLVLEKDWTDVEFLSSAKPPVIGYP : 168
VvANR : VVNVMKACTRAKSVKRVLLTSSAAAVTINQLDGTGLVVDEKNWTDIEFLTSAKPPVIGYP : 169
CsANR : VVNVLKACAKAGTVKRVLLTSSAAAVSINKLNGTGLVMDENWTDTEFLNSAKPPVIGYP : 168
ZmA2 : FFALPVQDKEAYANDPAAGRLQGYGSRLATNTCGQREWEDYLFHLVHPDGLADHALWPAY : 174
PfANS : FFELPVEEKKAYANDQAAGNVQGYGSKLANNASGQLEWEDYFFHCVYPEHKTDLSTIPKT : 167
AtANS : FFSLSVEEKKAYANDQATGKIQGYGSKLANNASGQLEWEDYFFHFLAYPEEKRDLSIPKT : 161
MtANS : FFELPVEEKKAYANDQSSGKIQGYGSKLANNASGQLEWEDYFFHCVYPEDKRDLSIPKT : 161
TCANS : FFELSVEEKKAYANDQTLGKIQGYGSKLANNASGQLEWEDYFFHFLVYPEDKRDLSIPQT : 163

```

6 r

```

*           200           *           220           *           240
AT1G75290 : VTNYFAGYYLPTLVQLEPGLTS-----PPRDVKVIFGDGNVQVE----- : 190
AT1G75280 : VTGCGGGYYLPTLVQFEPGLTS-----PPRDKVITLGDGNAK----- : 188
AT1G75300 : VSNCSAGFYLRLLQFESGLISHFRDKAIIIFGDKNVPDRDKVITLGDGNAK----- : 202
AT1G19540 : VSGCFAGLFVPCLGQCHRLRLS-----PPRDKVSIIYDTGNGK----- : 187
AT4G39230 : SCNFAGYFLPTLAQ--PGATS-----APRDKVIIVLGDGNPK----- : 185
AT4G34540 : CCGFLMRVLLPSLVQ--PGLQS-----PPTDKVTVFVGDGNVK----- : 183
AT4G13660 : VGACFAAYFAGNLSQMGV-----LFPKPKVVDIYGDGNVK----- : 196
AT1G32100 : VGACFAAYFAGNLSQMT-----LLPKPKVNIYGDGNVK----- : 196
TcLAR : CCNSIAAWPYHDNTHPADVL-----PPLDRFKIYGDGTVK----- : 201
VvLAR : CCNSIASWPYNNIHPSEVL-----PPTDFQIYGDGNVK----- : 192
LcLAR1-1 : CCNSIASWPYHDNCHPSKVP-----PPVDQFLIYGDGTVK----- : 194
LcLAR1-2 : CCNSIASWPYHDNCHPSKVP-----PPVDQFLIYGDGGSVK----- : 194
MtLAR : CCNSIASWPYDNCPSQLP-----PPLDQLHIYGHGNVK----- : 195
DuLAR : CCNSIASWPYDNCPSQVP-----PPMDQFOIYGDGNTK----- : 194
MtDFR : VSKTLAEQEAWKFAKEHNMDFIT-----IIPPLVVGPFILPTMPPS----- : 205
AtDFR : VSKTLAEKAAWDFAEKGLDFIS-----IIPPLVVGPFISASMPSS----- : 205
ZmA1 : VSKTLAEKAALAYAAEHGLDLVT-----IIPPLVVGPFISASMPSS----- : 210
AtBAN : ISKVLAEKTAWEFAKENKINLVT-----VIPALIAGNSLLSDPPSS----- : 211
LcANR : ASKTLAEKAAWKFAEENHIDLIT-----VIPSLLITGPSLTPDIPSS----- : 210
TcANR : ASKTLAEKAAWKFAEENNIDLIT-----VIPSLLMTGPSLTPDVPS----- : 209
VvANR : ASKTLAEKAAWKFAEENNIDLIT-----VIPTLMAGSSLTSDVPS----- : 210
CsANR : LSKTLAEKAAWKFAEENNINLIT-----VIPTLMAGPSLTADVPS----- : 209
ZmA2 : PPDYIAATRDFGRTRDLASTLLAILSMGL-----LGTDRGDALEKALTTTTTRTAADD : 228
PfANS : PPDYIPATSEYAKQLRALATKILSVLSIG-----LGLEKG-RLEKEVG-----GAE : 212
AtANS : PSDYIEATSEYAKCLRLATKVFKALSVG-----LGLEPD-RLEKEVG-----GLE : 206
MtANS : PADYTKVTSEYAKELRVLASKIMEVLSLE-----LGLEGG-RLEKEAG-----GME : 206
TCANS : PSDYTEVTSEYARQLRVLASKIILSALSIC-----LGLEEG-RLEKEVG-----GLE : 208

```

```

*           260           *           280           *           300
AT1G75290 : -YFIAVINKEEDIAAYTIKAVDDPRTLNKTLYINPPNNTLSMNEIVTLWEKKIKGKSVKEKI : 249
AT1G75280 : ----AVINKEEDIAAYTIKAVDDPRTLNKILYIKPSNNTLSMNEIVTLWEKKIKGKSEKKT : 244
AT1G75300 : ----VVINKEEDVAAYMIKAVDDLRLTNKTLYISPPNNTLSMNEMVTLWEKKIKGKSEKKT : 258
AT1G19540 : ----AIVNTEEDIVAYTLKAVDDPRTLNKILYIHPPNYIVSQNDMVGLWEKKIKGKTEKKT : 243
AT4G39230 : ----AVFNKEEDIGTYTINAVDDPRTLNKILYIRPPMNTYSFNDLVSLWENKIKGKTEKKT : 241
AT4G34540 : ----AVFVNDVDVAEFTIKTIIDDPRTLNKTLYLSPPGNICSMNDLVLEWEGKIEKKTTEKKT : 239
AT4G13660 : ----VVFVDEDDMAKYTAKTLNDPRTLNKTVYVRPTDNILTQMELVQIWEKLTKEKTEKKT : 252
AT1G32100 : ----VVFVDEDDIAKYTAKTLNDPRTLNKTVNIRPPDNVLTQLELVQIWEKLTGKETEKKT : 252
TcLAR      : ----AYFVAGTDIGKFTIMSIEDDRTLNKTVHFQPPSNLLNINEMASLWEEKIGRTTPRV : 257
VvLAR      : ----AYFVAGTDIGKFTMKTVDDVRTLNKSVHFRPSCNCLNINELASVWEKKIGRTTPRV : 248
LcLAR1-1   : ----AYFVDGNDIGKFTMKATIDDIRTRNKNVHFRPPSNVCSINELASLWEKIIIGRTTPRA : 250
LcLAR1-2   : ----AYFVDGNDIGKFTMKATIDDIRTRNKNVHFRPPSNVCSINELASLWEKIIIGRTTPRA : 250
MtLAR      : ----AYFVDGYDIGKFTMKVDDERTINKSVHFRPSTNVCYSMNELASLWENKIARKTTPRA : 251
DuLAR      : ----AYFIDGNDIGKFTMKTIIDDIRTRNKNVHFRPSSNVCYSINELASLWEKKIGRTTPRF : 250
MtDFR      : ---LITALSPITGNEAHYSIIKQGO-----FVHLDLDCDAHIFLFEHMEVEGGRYTCSSA : 255
AtDFR      : ---LITALSPITRNEAHYSIIRQGO-----YVHLDLDCNAHIFLYEQAAAGRYTCSSA : 255
ZmA1       : ---LITALALITGNAPHYSIIKQVQ-----LIHLDLDCDAEIFLFENPAAAGRYTCSSA : 260
AtBAN      : ---LSLSMSFITGKEMHVTGLKEMQKLSGSI SFVHVDLARAHLFLAEKETASGRYTCSSA : 268
LcANR      : ---VGLATSLITGNDFLINAMKGMQLLSGSI SITHVEDVCRAHIFVAEKOSASGRYTCSSA : 267
TcANR      : ---IGLATSLLSGNEFLVNALKGMQMLSGSI SITHVEDVCRAHVFLAEKESASGRYTCSSA : 266
VvANR      : ---IGLAMSLITGNEFLINGMKGMLSGSVSIAHVEDVCAHIFVAEKESASGRYTCSSA : 267
CsANR      : ---IGLAMSLITGNEFLINGLKGMLSGSISISHVEDVCRAHVFLAEKESASGRYTCSSA : 266
ZmA2       : DLLLQKINYYPKCPQPELALGVEAHTDVSALSFILHNGVPGLOVLHGARWVTARHEPGT : 288
PfANS      : DLIVQMKINYPKCPQPELALGVEAHTDVSALTFILHNMVPGLOLFYEDKWVTAKCVPNS : 272
AtANS      : ELLLQMKINYPKCPQPELALGVEAHTDVSALTFILHNMVPGLOLFYEGKWVTAKCVPDS : 266
MtANS      : ELLLQMKINYPKCPQPELALGVEAHTDVSALTFILHNMVPGLOLFYEGKWVTAKCVPDS : 266
TCANS      : ELLLQMKINYPKCPQPELALGVEAHTDVSALTFILHNMVPGLOLFYEGKWITAKCVPNS : 268
6

```

```

*           320           *           340           *           360
AT1G75290 : YMSEE-QIFKSIQESPVBFNVLLSINHAFVVKGDQTNFTIEPSFGFEASELYP--DIKYT : 306
AT1G75280 : HLPPE-QIFKSIQESPIPINVLLSINHAFVNGD-TNISIEPSFGVEASELYP--DVKYT : 300
AT1G75300 : HISEE-QIFKSIQ---VVIDVFKSINHAFVVKGDQTSFTIEPWFGEASVLYP--DVKYT : 312
AT1G19540 : YVSEE-ELDKTIQESKPPMDFLVGLIHTILVKSDFTSFTIDPSPFGVEASELYP--EVKYT : 300
AT4G39230 : YVPEE-QLKQIIESPPLNVMLSLCHCVFVKGGHTSFEIEPSFGVEASELYP--DVKYT : 298
AT4G34540 : FATEN-QLKKIKETPYPDNMEMVFIYSVFIKGDHTYFDIESCGGVNGTELYP--DVRYM : 296
AT4G13660 : YVSGN-DFADIEDKEISHQAGLGHFYHIYEGCLTDHEVGD--DEEATKLYP--DVKYK : 307
AT1G32100 : NIAAQ-DFLANIEQMEI BHQAGIGHFYHIFYEGCLTDHEVGE--DEEASLYP--DVKYK : 307
TcLAR      : TITEE-DLQMAKEMRIPQSVVAALTHDIFINGCQINFSLDKPTDVEVCSLYP--DTPER : 314
VvLAR      : TVTED-DLAAAAGENIIPQSVVAAFTHDIFIKGCQVNF SIDGPEDVEVTTLYP--EDSER : 305
LcLAR1-1   : TVSAE-DLAAAANCI PRSIVAAFTHDIFINGCQINF SIEGPNDIEIGTLYP--DEKER : 307
LcLAR1-2   : TVSAE-DLAAAANCI PRSIVAAFTHDIFINGCQINF SIEGPNDIEIGTLYP--DEKER : 307
MtLAR      : IVSED-DLGGIAANCI PESVVASITHDIFINGCQVNF KIDGIHDVEISTLYP--GESER : 308
DuLAR      : TVTAD-KLGAHAENIIPESIVSFTHDIFINGCQVNF SIDEHSDVEIDTLYP--DEKER : 307
MtDFR      : CEANI-HDAKLINTKYBEYNIPTKFNNI PDELELVRFSSKKIKDLGFEFKYS-LEDMYT : 313
AtDFR      : HDATI-LTISKFLRPKYPEYNVPSTFEGVDENLKSIEFSSKKLTDMGFNFKYS-LEEMEI : 313
ZmA1       : HDVTI-HGLAAMLRDRYPEYDVPQRFPGIQDDLQPVRFSSKKLQDLGFTFRYKTLDEMED : 319
AtBAN      : YNTSV-PEADFLIQRYPKYNVLSFEFEGLSIPK-LTSSQKLINEGFRFEYG-INEMYD : 325
LcANR      : HNTSV-PEAKFLNKRYPOYKVSTEFNDFPAKAK-LIISPEKLIKEGFSFKYG-VEEIED : 324
TcANR      : VNSSV-PEAKFLNQRYPEFKVPTDFGDFPSKAK-LIISSDKLINEGFSFKYG-IEEIID : 323
VvANR      : ANTSV-PEAKFLSKRYPOYKVPTDFGDFPPKSK-LIISSEKLVKEGFSFKYG-IEEIID : 324
CsANR      : VSTSV-PEAKFLNKRYPEYNVPDFGDFPSKAK-LIISSEKLTKEGFSFKYG-IEEIID : 323
ZmA2       : IIVHVGDAEILSNGRYT SVLHRGLVNREAVRISWVVFCEPPPDSVLLHPLPELVTEGHP : 348
PfANS      : IIMHIGDTEILSNGKYKSILHRGLVNKEKVRISWAVFCEPPKEKIVLQPLPETVSEVP : 332
AtANS      : IVMHIGDTEILSNGKYKSILHRGLVNKEKVRISWAVFCEPPKDKIVLQPLPEMVSVEVP : 326
MtANS      : ILMHIGDTEILSNGKYKSILHRGLVNKEKVRISWAVFCEPPKEKIIILQPLPELVTEKEP : 326
TCANS      : IIMHIGDTEILSNGKYKSILHRGLVNKEKVRISWAVFCEPPKEKIIILQPLPETVSETEP : 328

```

```

*           380           *           400           *           420
AT1G75290 : SIDEYLSYFALGTSLNT----- : 323
AT1G75280 : SVDEYLSYFA----- : 310
AT1G75300 : SIDEYLSQFT----- : 322
AT1G19540 : SVDEFLNRFI----- : 310
AT4G39230 : TVDEILNQYV----- : 308
AT4G34540 : TVSEFLDTLL----- : 306
AT4G13660 : RMDEYLKLFV----- : 317
AT1G32100 : RMDDYLRMFL----- : 317
TcLAR    : TINECFEDEFKAKIIDNAKAVS-----KPAASNNAI : 344
VvLAR    : TVEECFGEYIVKIEE-----KQPTADSAI : 329
LcLAR1-1 : CLEECFKDFVPMTHDMNVHVG-----TTEIN : 333
LcLAR1-2 : CLEECFKDFVPMTHDMNVHVG-----TTEIK : 333
MtLAR    : SLEDCFESFVAMAADK-IHKG-----ENGVT : 333
DuLAR    : SLDDCYEDFVPMVHDK-IHAGKSGEIKIKDGRPLVQTGTIEEINKDIKTLVETQPNEEIK : 366
MtDFR    : EAIDTCTEKGLLPK-----FVKSTNK----- : 334
AtDFR    : ESIEtCRQKGF LPVSLSYQSI SEIKTKNENIDVKTGDGLTDGMKPCNKtETGtGERTDA : 373
ZmA1     : AAIRTCQEKGLIPL-----ATAAGGDGFASVRAPGTEATIGA----- : 357
AtBAN    : QMIEYFESKGLIKAK----- : 340
LcANR    : QTLEYLKTKGALKN----- : 338
TcANR    : QTVEYMNKAGLLK----- : 336
VvANR    : ESVEYFKAKGLLQN----- : 338
CsANR    : QSVEYFKAKGILKN----- : 337
ZmA2     : ARFTPRtFKQHLDRKLFKKKQQHKAKAEEDGGNGDHRHEPPPQTN----- : 395
PfANS    : BRFPPrtFAQHlKHKLFrKtDGDLDekPtY----- : 362
AtANS    : AKFPPrtFAQHIEHKLFgKEQEELVSEKND----- : 356
MtANS    : ARFPPrtFAQHlHhKLFrKDEEEKDDPKK----- : 356
TcANS    : PLFPPrtFAQHlHhKLFrKtQDGLSN----- : 354

```

```

*
AT1G75290 : ----- : -
AT1G75280 : ----- : -
AT1G75300 : ----- : -
AT1G19540 : ----- : -
AT4G39230 : ----- : -
AT4G34540 : ----- : -
AT4G13660 : ----- : -
AT1G32100 : ----- : -
TcLAR    : FVPTAKPGALPITAICT : 361
VvLAR    : ANTGPVVMRQVTATCA : 346
LcLAR1-1 : NNRKSLVEVAPITAMG- : 349
LcLAR1-2 : NNRKSLVEVAPITAMG- : 349
MtLAR    : GGTkALVEVPITASC- : 349
DuLAR    : KDMKALVEAVPISAMG- : 382
MtDFR    : ----- : -
AtDFR    : PMLAQQMCA----- : 382
ZmA1     : ----- : -
AtBAN    : ----- : -
LcANR    : ----- : -
TcANR    : ----- : -
VvANR    : ----- : -
CsANR    : ----- : -
ZmA2     : ----- : -
PfANS    : ----- : -
AtANS    : ----- : -
MtANS    : ----- : -
TcANS    : ----- : -

```

## VITA

Yi Liu

### Education:

Doctor of Philosophy

The Pennsylvania State University

8/2001-5/2010

Integrative Biosciences (IBIOS)

Bachelor of Science

Huazhong Agricultural University

9/2000-7/2004

Biological Science

### Professional Experience:

Genentech, Inc., DMPK Department

South San Francisco, CA

6/2008-9/2008

Research intern, R&D Group

DuPont Qualicon

Wilmington, DE

Research intern, R&D Group

6/2006-9/2006

Penn State University, Biology Department,

University Park, PA, USA

Teaching assistant for Bio230W

Fall 2005, 2007