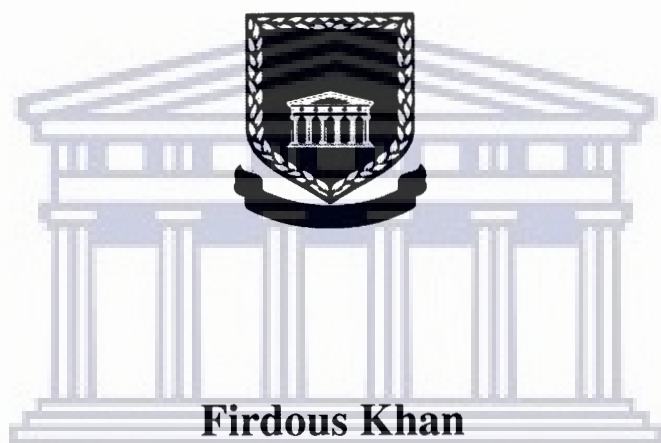


# **Regulatory attributes of the carotenoid biosynthetic pathway**

**in *Arabidopsis thaliana* under abiotic stress.**



**Firdous Khan**

UNIVERSITY *of the*

---

WESTERN CAPE

Thesis presented in fulfillment of the requirements for the

Magister Scientiae

at the South African National Bioinformatics Institute

Faculty of Natural Sciences, University of the Western Cape

Advisor: **Prof. Alan Christoffels**

March 2, 2012

# Declaration

I declare that “**Regulatory attributes of the carotenoid biosynthetic pathway in *Arabidopsis thaliana* under abiotic stress.**” is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.



A handwritten signature in black ink, appearing to read 'A. Khan', is written over a horizontal line.

Signed March 2, 2012



UNIVERSITY of the  
WESTERN CAPE

The logo of the University of the Western Cape, featuring a stylized classical building with columns and a pediment, is positioned behind the text.

# Abstract

---

Carotenoids are tetraprenoid (C40) molecules synthesized in plants, fungi, bacteria and algae, via the carotenoid biosynthetic pathway (CBP). Some carotenoids are readily converted to vitamin A (VA) in humans, e.g.  $\beta$ -carotene,  $\alpha$ -carotene and  $\beta$ -cryptoxanthin 1,2. Vitamin a deficiency (VAD) affect millions especially children under the age of five. The CBP in plants is a key source of pro-vitamin A and is vital to the biofortification of staple crops such as maize, rice and sorghum, could alleviate the global VAD problem. However the incomplete understanding of regulation of the pathway is a limiting factor to predictably control carotenoid content at the systems level. Previous studies have shown that growth conditions, such as light, play a major role in the biosynthesis of carotenoids. A systems biology approach was therefore used to analyse microarray data sets derived from *A. thaliana* grown under various conditions and treated with different stimuli.

Thirty two genes have previously been identified as being involved in the CBP. These genes were found to be highly differentially expressed depending on stress type. All stimuli including drought, cold, heat, osmotic, oxidative and salt but wounding had a significant influence on the CBP genes. Gene expression induced by abiotic stress occurred 30 min after exposure. These findings are indicative that an immediate systemic signal is sent to the rest of the plant in response to stress. A correlation analyses revealed strongly positive correlation between PSY and

---

its co-expressed genes, suggesting they share a common regulatory mechanism. Promoter content analyses identified 20 enriched TFBMs among carotenoid genes. The most prevalent TFBMs found in the promoter regions of the CBP genes show a 1.25-3 fold increase in prevalence with a  $p$ -value  $< 0.05$ . Similar GO terms are enriched for CBP genes and their co-expressed genes. These findings indicate that carotenoid biosynthetic pathway genes and their co-expressed genes are involved in similar metabolic pathways and functional processes. This study identified cold, drought and heat to influence carotenoid gene expression and has led to the identification of molecular switches that can be modulated to control the biosynthetic pathway.

Four motifs without any GO annotation and no specific known motif in plant databases were identified using MEME suite. In this study I propose that these predictions might be novel motifs and could be specific to carotenoid genes, and may be directly involved in the regulation of carotenoid biosynthesis.

These findings may lead to a better understanding of the underlying regulatory mechanisms involved in the biosynthesis of carotenoids. Furthermore, these findings may assist in establishing ways of enhancing the production of carotenoids, especially pro-vitamin A, in *Arabidopsis thaliana*.

---

**Keywords:** Vitamin A, Pro-vitamin A, carotenoid biosynthetic pathway, *Arabidopsis thaliana*, transcription factor binding motif, promoter, gene expression, correlation, abiotic stress, microarray, regulatory networks, clustering.

# Dedication



*Dedicated to my Mum and Dad*

*Zubayda Khan*

UNIVERSITY *of the*  
&  
WESTERN CAPE

*Abubaker Khan*

# Acknowledgement

*Make it a habit to tell people thank you. To express your appreciation, sincerely and without the expectation of anything in return. Truly appreciate those around you, and you'll soon find many others around you. Truly appreciate life, and you'll find that you have more of it.*

Ralph Marston.

All praise and glory be to the Almighty ALLAH (saw) my creator without whom none of this would be possible. To my supervisor Prof. Alan Christoffels thank you very much for the opportunities you have given me throughout the past two years and for believing in me. Dr Samson Muyanga I am indebted to your enthusiasm towards my progress and for believing in me. It has been a privilege to work under your supervision. Dr Musa Nur Gabere words cannot describe how thankful I am for the time and effort you spent on me to give me an invaluable learning experience, I could not have asked for a better teacher. Dr Ashley Pretorius I am eternally grateful to you for always being a pillar of strength to me. You were not only a mentor to me but a best friend that I will cherish till the end, Thank you. Stanley Mbandi Kimbung thank you so much for the endless hours you spent making sure my script writing was up to scratch, I know I was not the easiest of students to teach. Maryam Dale, Ferial and Dr Junaid Gamieldien shukran and thank you to all of you for making the past two years at SANBI very memorable I have learned many many life lessons from you all. My dearest friend Kavisha thank you so much for the gift of

---

friendship I will cherish it always and for ever. To the rest of the SANBI crew old and new I am thankful that I have had the opportunity to get to know each and every one of you, you have all made the past two years really special. To my family Zubayda, Gadija and Mogammad Nur Khan shukran to all of you for the love, motivation and challenges you all have given me, as well as all the sacrifices you had to make I know it was not an easy two years but you guys stuck by me no matter what and I am eternally grateful. Each and every one of you have had a part in sculpturing the person I am and for pushing me to become the best I can be I Thank You all.....



# Nomenclature

## Acronyms

hrs	hours
rthf	roots expression at time point 0.5hrs
shhf	shoots expression at time point 0.5hrs
rt1	roots expression at time point 1hr
sh1	shoots expression at time point 1hr
rt3	root expression at time point 3hrs
sh3	shoot expression at time point 3hrs
rt6	root expression at time point 6hrs
sh6	shoot expression at time point 6hrs
rt12	root expression at time point 12hrs
sh12	shoot expression at time point 12hrs
rt24	root expression at time point 24hrs
sh24	shoot expression at time point 24hrs
TAIR	The Arabidopsis Information Resource



---

$\alpha$	alpha
$\beta$	Beta
$\gamma$	Gamma
$\delta$	Delta
$\epsilon$	Epsilon
$\zeta$	Zeta
TFBS	transcription factor binding site
TFBM	Transcription factor binding motif
VAD	Vitamin A deficiency
VA	Vitamin A
PSY	Phytoene Synthase
PDS	Phytoene desaturase
CRTISO	carotene cis trans isomerase
LCY $\beta$	Lycopene beta cyclase
LCY $\epsilon$	lycopene epsilon cyclase
LUT5	lutein-deficient 5
LUT1	lutein-deficient 1
ZDS	Zeta carotene desaturase
$\beta$ OHase1	beta-carotene hydroxylase 1
$\beta$ OHase2	beta-carotene hydroxylase 2

---

C	Cytosine
G	Guanine
GO	Gene ontology
CBP	Carotenoid biosynthetic pathway
GGPP	Geranylgeranyl phosphate pathway
ACT	Arabidopsis co-expression tool
ATHENA	Arabidopsis thaliana expression network analyses
NASCarray	Nottingham Arabidopsis Stock Centre
STRING	Search tool for the retrieval of interacting genes/proteins
n	number
bp	base pairs
DB	Database
%PBS	percentage of promoters bound in the subset
#of GS	number of genes present in subset with bound promoters
%PBG	percentage of promoters bound in the genome
#GG	number of genes in genome with bound promoter
FC	fold change
%CG	percentage of promoters bound in Carotenoid genes

---

<i>A. thaliana</i>	<i>Arabidopsis thaliana</i>
3D	three dimensional
2D	two dimensional
HSP	heat shock protein
MRE	metal response
DNA	Deoxyribose nucleic acid
RNA	Ribonucleic acid
mRNA	messenger Ribonucleic acid
A	Adenine
T	Thymine
#CG	number of carotenoid genes with bound promoter
% in GG	percentage of promoters bound in genome
#GG	number of genes in genome with bound promoter
ABA	Abscisic acid
cDNA	Complementary Deoxyribose nucleic acid

# Contents

<b>1 Literature Review and Introduction</b>	<b>1</b>
1.1 Carotenoids	1
1.2 Carotenoid biosynthetic pathway and the production of carotenes	3
1.3 Gene regulation	4
1.4 Regulation of gene expression in plants in response to abiotic stress	9
1.5 Gene expression profiling using microarrays	12
1.6 Co-expression and correlation	15
1.7 Promoter content analyses	16
1.8 Rationale and focus	20
1.9 Structure of the thesis	22
<b>2 Methodology</b>	<b>24</b>
2.1 Putative condition identification	24
2.2 Co-expression and co-correlation analyses	27

## CONTENTS

xii

---

2.3 Promoter content analyses . . . . .	33
<b>3 Results</b>	<b>39</b>
3.1 Identification of putative conditions affecting carotenoid gene expression . . . . .	40
3.2 Co-expression and correlation analyses of carotenoid and co-expressed genes . . . . .	58
3.3 Promoter content analyses, functional annotation and <i>de novo</i> motif discovery . . . . .	61
<b>4 Discussion</b>	<b>70</b>
4.1 Identification of putative conditions affecting carotenoid gene expression . . . . .	70
4.2 Co-expression and correlation analyses . . . . .	74
4.3 Promoter content analyses, functional annotation and <i>de novo</i> motif discovery . . . . .	77
<b>5 Conclusion and future work</b>	<b>81</b>
5.1 Research contributions and limitations . . . . .	81
5.2 Future work . . . . .	84
<b>Appendix</b>	<b>86</b>
<b>A Supplementary material for Chapter 2</b>	<b>86</b>
<b>B Supplementary material for Chapter 3</b>	<b>96</b>
<b>References</b>	<b>104</b>

# List of Figures

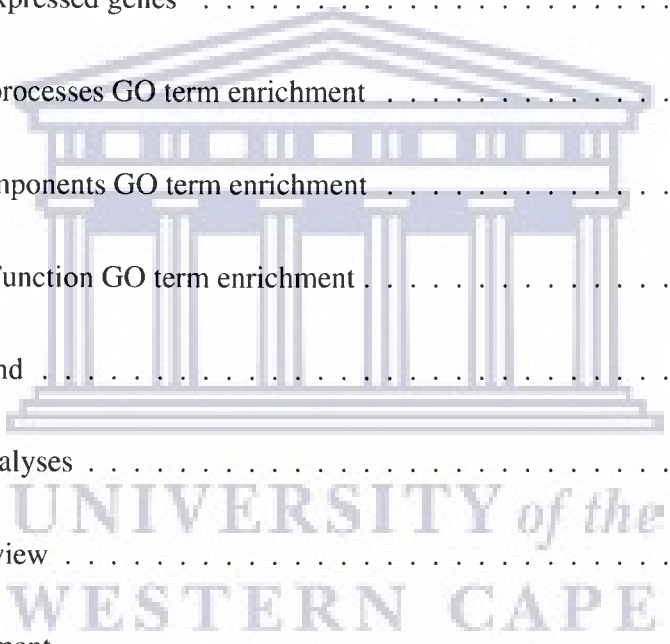
1.1	Structure formula for carotenes. . . . .	2
1.2	Carotenoid biosynthetic pathway in plants. . . . .	4
1.3	Pre-initiation complex . . . . .	7
1.4	Human heat shock protein . . . . .	8
1.5	Microarray procedure . . . . .	13
1.6	<b>A diagram of methodologies used in the thesis.</b> . . . . .	23
2.1	Practical protocols . . . . .	29
2.2	<i>De novo</i> motif prediction . . . . .	36
3.1	Heat map expression profiles of CBP genes under cold stress. . . . .	45
3.2	Heat maps expression profiles of CBP genes under drought stress. . . . .	46
3.3	Heat maps expression profiles of CBP genes under heat stress. . . . .	47
3.4	Heat maps expression profiles of CBP genes under osmotic stress. . . . .	48

**LIST OF FIGURES**

**xiv**

---

3.5	Expression profiles for Drought . . . . .	50
3.6	Expression profiles for Cold . . . . .	52
3.7	Expression profiles for Heat . . . . .	54
3.8	Expression profiles for Osmotic . . . . .	56
3.9	Correlation analyses . . . . .	59
3.10	List of co-expressed genes . . . . .	60
3.11	Biological processes GO term enrichment . . . . .	62
3.12	Cellular components GO term enrichment . . . . .	62
3.13	Molecular Function GO term enrichment . . . . .	63
3.14	TFBM legend . . . . .	64
3.15	Enriched analyses . . . . .	65
3.16	Motif Overview . . . . .	67
3.17	Motif alignment . . . . .	68
3.18	Motif alignment . . . . .	69
B.1	Expression profiles for Oxidative . . . . .	97
B.2	Expression profiles for Salt . . . . .	98
B.3	Expression profiles for wounding . . . . .	99
B.4	Venn diagrams of LUT1 & LUT5 . . . . .	100



---

B.5	Venn diagrams of $\beta$ OHase1, $\beta$ OHase2, CRTISO & ZDS . . . . .	101
B.6	Venn diagrams of LCY $\epsilon$ , LCY $\beta$ , PSY & PDS . . . . .	102



UNIVERSITY *of the*  
WESTERN CAPE

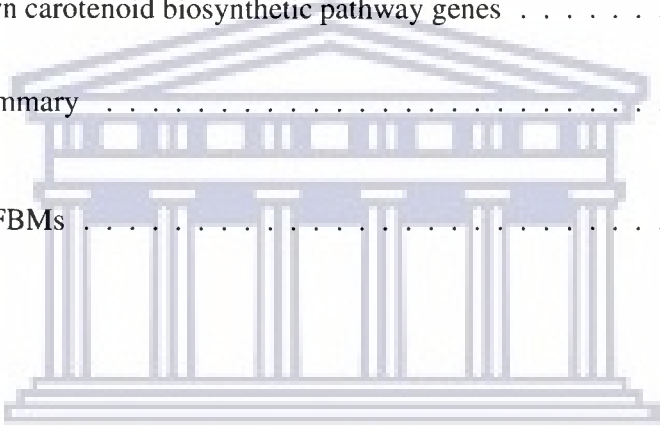


# List of Tables

3.1 List of known carotenoid biosynthetic pathway genes . . . . . 41

3.2 Heatmap summary . . . . . 44

A.1 Predicted TFBMs . . . . . 93

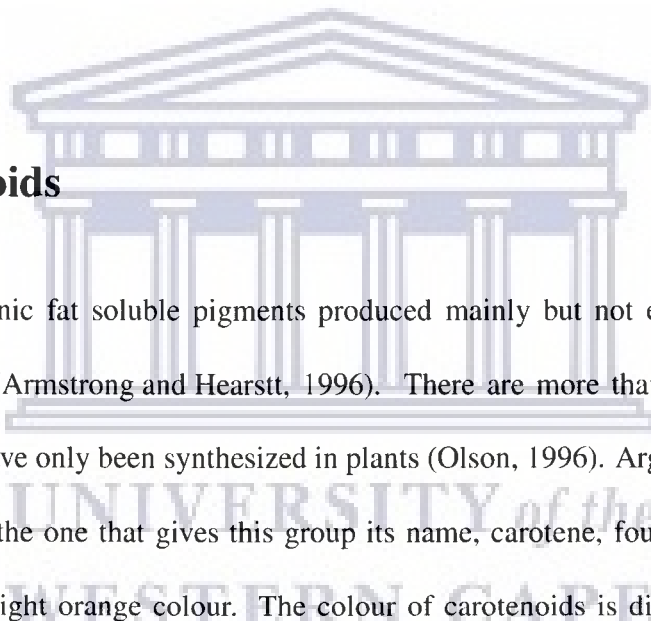


UNIVERSITY *of the*  
WESTERN CAPE

# Chapter 1

## Literature Review and Introduction

### 1.1 Carotenoids



Carotenoids are organic fat soluble pigments produced mainly but not exclusively in photosynthetic organisms (Armstrong and Hearstt, 1996). There are more than 600 known natural carotenoids and all have only been synthesized in plants (Olson, 1996). Arguably the most well-known carotenoid is the one that gives this group its name, carotene, found in carrots and responsible for their bright orange colour. The colour of carotenoids is directly linked to their structure in that carotenoids are characterized by a large (35-40 carbon atoms) polyene chain, sometimes terminated by rings (Figure 1.1). The double carbon-carbon bonds interact with each other in a process called conjugation (Marrs, 1996). As the number of double bonds increases, the wavelength of the absorbed light increases, giving the compound an increasingly red appearance (Armstrong and Hearstt, 1996; Park *et al.*, 2002).

In photosynthetic organisms, carotenoids play a vital role in the photosynthetic reaction centre (Mayer *et al.*, 1999). They participate in the energy-transfer process and also protect the reac-

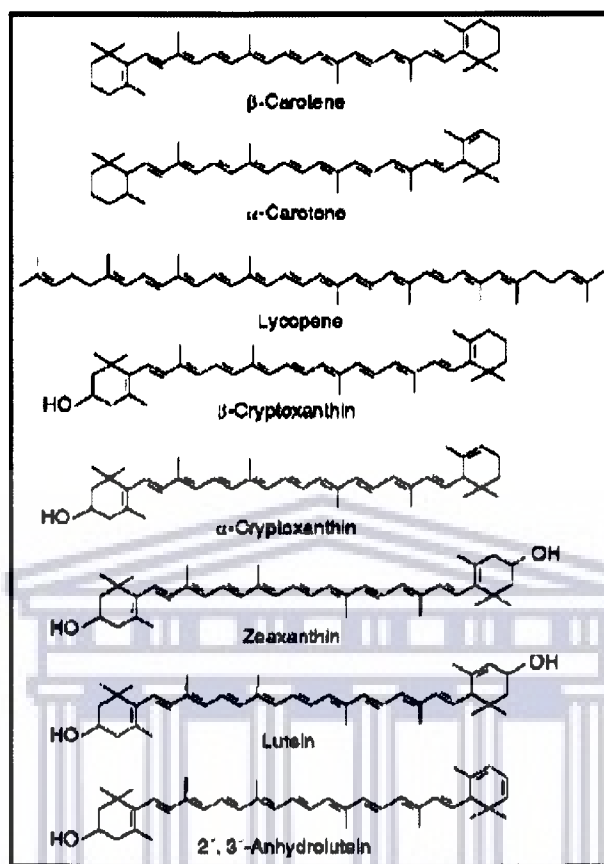


Figure 1.1: **Structural formulae of common Carotenes and Xanthophylls of carotenoids (Cunningham *et al.*, 1996).**  $\beta\beta$ -rings are shown in the structure of  $\beta$ -carotene. Single  $\beta$  ring is present in  $\alpha$ -carotene structure. Xanthophylls contains a hydroxyl (OH) group at either one end or both ends of the chemical structure

tion centre from auto-oxidation. In non-photosynthetic organisms, carotenoids have been linked to oxidation-preventing mechanisms (Armstrong and Hearstt, 1996; Olson, 1996). Carotenoids where some of the double bonds have been oxidized, such as lutein and zeaxanthin, are known as xanthophylls; whereas the un-oxidized carotenoids such as  $\alpha$ -carotene,  $\beta$ -carotene and lycopene are known as carotenes (Olson *et al.*, 1993).

## 1.2 Carotenoid biosynthetic pathway and the production of carotenes

In the initial stage of the carotenoid biosynthetic pathway, geranylgeranyl pyrophosphate (GGPP) is converted to phytoene using the enzyme phytoene synthase (PSY) (Isaacson *et al.*, 2002). Two units of the C<sub>20</sub> compound GGPP are aggregated together by the use of PSY to form the C<sub>40</sub> compound phytoene (Armstrong and Hearst, 1996). It is this enzymatic step that is known to be rate limiting in tissues and developmental stages of various plant species (Park *et al.*, 2002).

Phytoene which is unable to absorb light at visible wavelengths is not a true pigment in the sense that it undergoes four consecutive desaturation steps (Isaacson *et al.*, 2002). The first two steps in the pathway are performed by PDS and the latter two steps are performed by ZDS (Li *et al.*, 2009). The red pigment lycopene is produced from this reaction, which is the main pigment in red tomatoes (Armstrong and Hearst, 1996). Lycopene is also produced in cyanobacteria and in plants by these desaturase reactions, and it is known as pro-lycopene (Bartley *et al.*, 1999). A further enzymatic step is necessary to produce the all-trans-lycopene. The enzyme carotene isomerase (CRTISO) converts lycopene from *cis-trans* lycopene, which is thus the main substrate for downstream reactions such as the bifurcation of the CBP pathway (Cunningham, 2002; Isaacson *et al.*, 2002; Park *et al.*, 2002).

Bifurcation of the CBP pathway leads to either (i) the synthesis of  $\alpha$ -carotene involving two different cyclases or (ii)  $\beta$ -carotene production catalysed by  $\beta$ -carotene cyclase (LCY $\beta$ ) in two consecutive cyclization reactions (Cunningham *et al.*, 1996; Cunningham, 2002). Knockout mutants in the  $\epsilon$ -ring cyclization step (LUT2) results in the accumulation of higher levels of  $\beta$ -

carotene and violaxanthin (Villamor and Fawzi, 2005). The only difference between  $\alpha$ -carotene and  $\beta$ -carotene is the position of a double bond in one of the end rings (Figure 1.2) (Park *et al.*, 2002).

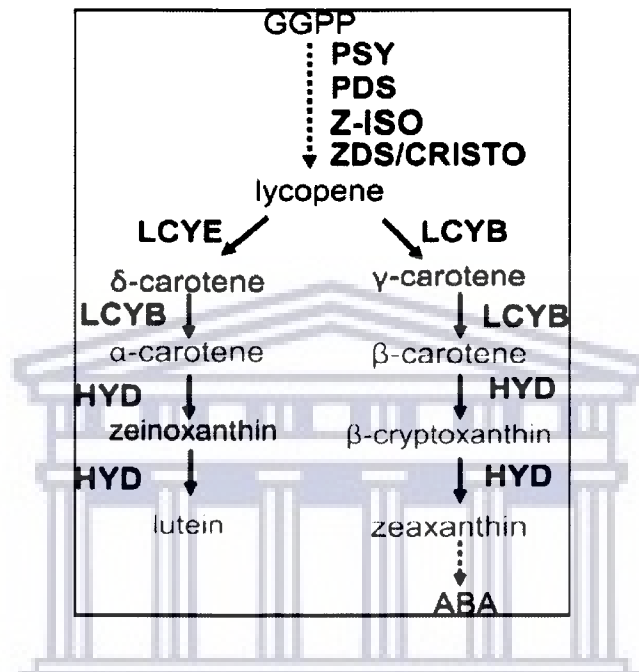


Figure 1.2: A simplified version of the carotenoid biosynthetic pathway in plants. Enzymatic reactions are represented by arrows, dashed lines represent multiple enzymatic steps (Mathews and Wurtzel, 2007).

### 1.3 Gene regulation

Gene expression of all genes may vary according to one or more of the following:

- the developmental stage,
- external stimuli such as heat, drought, cold, wounding, oxidation and osmosis,
- location of the cell, and
- cell type (root cells are different from shoot cells, different from stem cells, and so forth).

Regulation of gene expression occurs at a multitude of levels within a cell and is therefore important for the transcription of DNA (Dillon and Festenstein, 2002). DNA is said to be transcribed within the nucleus of the cell into mRNA, which is then spliced in higher organisms (Chow *et al.*, 1977; Arabidopsis and Initiative, 2000). Splicing occurs to remove introns and bring exons together. Thereafter translation of mRNA into protein sequences occur in the cytoplasm and the protein is then folded into a functional 3D structure (Mayer *et al.*, 1999; Meyer, 2000; Meier *et al.*, 2008).

### **1.3.1 Transcriptional regulation**

Transcription is the first leading step leading to gene expression (Koch, 1996; Kane *et al.*, 2000). When a cell receives an external stimulus, a protein signalling cascade transmits the message from a surface receptor to the nucleus of the cell (Logemann *et al.*, 1995; Melhus *et al.*, 1998). Within the nucleus, the chromatin unwinds in order for the regulator proteins to attach to the chromatin and as a result more or less transcription of the gene will occur. (Cold *et al.*, 2000). The transcribed mRNA molecule, which is a copy of the specific gene, is exported to the cytoplasm where translation and folding of proteins occur (Mayer *et al.*, 1999; Seki *et al.*, 2002).

Li *et al.*, (2009) has shown that the transcription process is integral to the level of gene expression. From empirical evidence by Struhl (1995), transcription is seen to be one of the strongest and most versatile stages of regulation of expression levels. The way transcription occurs is by the attachment of transcription factors onto the DNA. Once these factors are attached, they either block, enhance or initiate the transcription of a gene in the vicinity (Rao *et al.*, 2000; Tabata *et al.*, 2000). The transcription factors usually bind to locations in the DNA that are char-

acterised by the presence of short motifs known as the TFBS (Wilhelm and Thomashow, 1993; Molina and Grotewold, 2005).

Upstream of a gene close to the transcription start site there is a core promoter. In plants, it is roughly five to fifteen base pairs long. Core promoter motifs such as the TATA-box (3-6 base pairs long) are well conserved between different species suggesting that their sequence is crucial to the regulation of genes (Shahmuradov *et al.*, 2005; Molina and Grotewold, 2005).

### **1.3.2 Regulation in the promoter region**

Upstream of the core promoter, is the general promoter region, which contains regulatory information important for determining when and where these genes are to be transcribed (Li *et al.*, 2009). There are many other DNA and protein elements present such as enhancers, activators, sigma factors and transcription factors, that are crucial to the regulation of promoters and transcription (Figure 1.3) (Hegde *et al.*, 2000).

In general, the core promoter is necessary to start transcription (Molina and Grotewold, 2005). The enhancers are not always necessary, however, it may be utilized during the exposure to aspecific stimuli (Dillon and Festenstein, 2002). Therefore, at the location of regulatory elements, specific transcription factors are bound based on the "lock and key" mechanism because they encompass a complementary component that fits well into the DNA structure of nucleotides (Cold *et al.*, 2000). RNA polymerase then binds to the DNA and starts transcribing, resulting in an mRNA copy of the gene, which is produced linearly from 5' to 3'. New proteins such as heat shock proteins (HSP) are then formed which controls the expression level of different genes (Rao *et al.*, 2000).

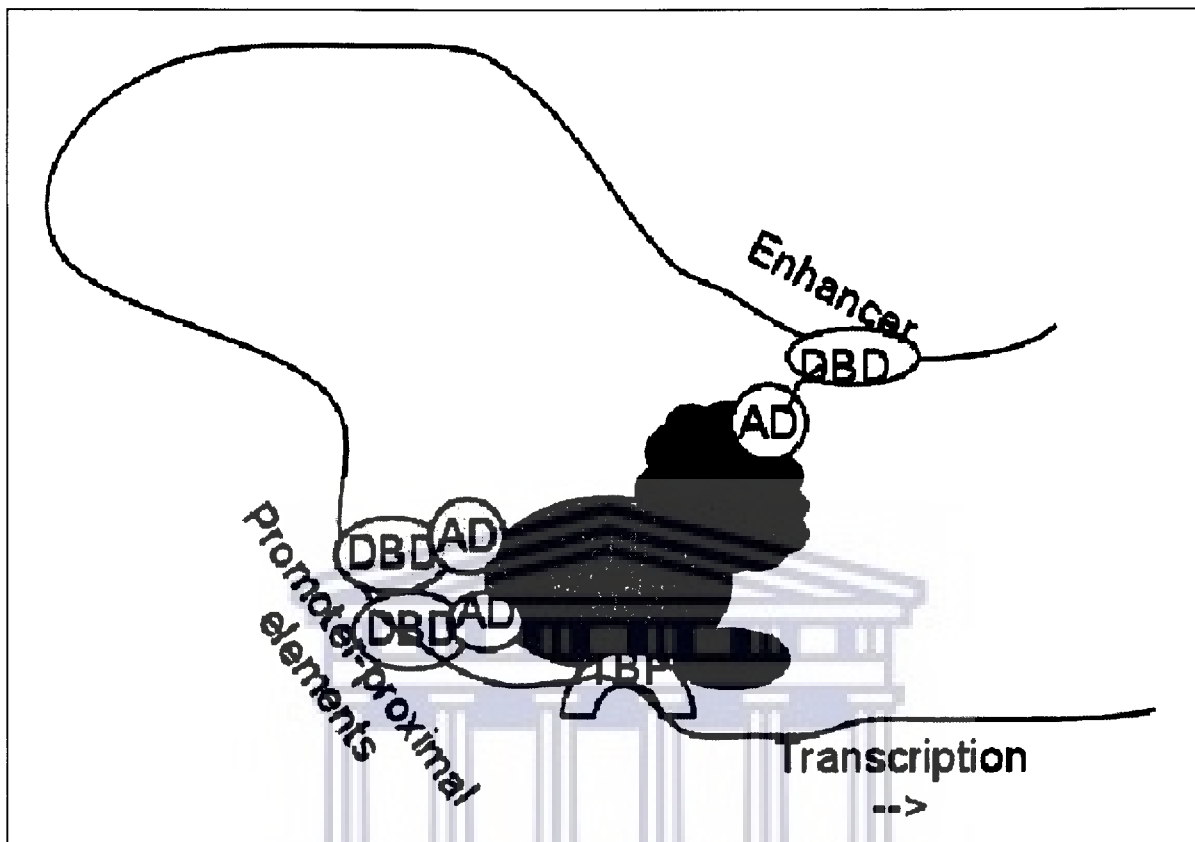


Figure 1.3: **Diagram of the pre-initiation complex present at the time of transcription.** This image shows the pre-initiation complex. RNA polymerase II is only able to start transcription by binding to the pre-initiation complex. RNA polymerase II is bound to a group of proteins that assemble around the start point of transcription, the proteins provide a base for the binding of polymerase II. The enhancer proteins and promoter proteins, as shown in the figure, help to stabilize the pre-initiation complex long enough for RNA polymerase II to bind. Without these enhancer proteins and promoter proteins, it is less likely that transcription will occur (Rao *et al.*, 2000).

Heat shock proteins (HSP) are a class of functionally related proteins involved in the folding and unfolding of other proteins. HSPs are found in virtually all living organisms, from bacteria to humans. Their expression is increased when cells are exposed to elevated temperatures or other stress (De Maio, 1999). This increase in expression is transcriptionally regulated. The dramatic up regulation of the heat shock proteins is a key part of the heat shock response and is induced primarily by heat shock factor (HSF) (Wu, 1995).



## 1.4 Regulation of gene expression in plants in response to abiotic stress

Gene expression profiling is the measurement of the expression of thousands of genes at once, to create a global picture of cellular function. These profiles can, for example, distinguish between cells that are actively dividing, or show how the cells react to a particular stimuli. Many experiments of this sort measure the expression from an entire genome simultaneously, that is, every gene transcript present in a particular cell.

Monitoring the genome-wide expression levels of thousands of genes concurrently under different conditions requires robust, large-scale experimental tools. DNA Microarray technology captures genome-wide gene expression profiles, informing us about the differential expression of genes under certain environmental conditions. Microarray technology enables monitoring of cell-, tissue- and developmental stage-specific gene expression profiles and simultaneous quantitative analyses of expression levels of genes (DeRisi *et al.*, 1996; Baldwin *et al.*, 1999; Van Hal *et al.*, 2000; Ye *et al.*, 2002), which could aid in the understanding of the involvement of multiple genes in the particular biological processes or signalling pathways.

Microarray technology has been used in the plant research field for gene studies in recent years (Wisman and Ohlrogge, 2000). Schena *et al.* (1995) studied the expression of 45 Arabidopsis genes printed on an array and detected low expression levels of these genes in response to abiotic stress. Reymond *et al.* (2000) analyzed 150 wounded and insect-related genes of Arabidopsis, which are genes that are up regulated during wounding as a result of insect feeding. Grike *et al.* (2000) analyzed genes related to seed development in Arabidopsis. Grike *et al.*

(2000) found that 25% and 10% of the genes tested showed a 2 and 10 times increase in expression respectively during *Arabidopsis* seed development. All these results suggest that cDNA microarrays are useful in identifying new genes, as well as in the study of expression profiling of the tissue-specific or environment responsive genes.

Kilian and others (2007) observed large differences between the expression levels of differentially expressed genes in response to abiotic stress, where the number of genes up regulated genes exceeded the number of genes down regulated genes. Fewer genes elicited a response after exposure to drought stress, this led to the conclusion that plants recovered relatively fast from drought. Cold and osmotic stress displayed similar responses as observed for drought albeit at a lower expression level. The rest of the stimuli including oxidative stress, salt and wounding caused only transient changes. Furthermore, Kilian *et al.* (2007) looked at the transcription factors that were at play during the exposure of the above mentioned stresses. Four of the 9 genes that were up regulated after only 30min exposure to cold, drought and UVB collectively, were bonafide transcriptional regulators and included compliment signalling components such as Ca<sup>2+</sup> (Kilian *et al.*, 2007)

A recent paper by Meier *et al.*, (2011) observed a number of genes expression being highly correlated with PSY expression, the driver gene of the carotenoid biosynthetic pathway. The top 50 co-expressed genes had *r*-values ranging between 0.84 and 0.91. Where the *r*-value is the expression correlation coefficient. PSY was found to be co-expressed with genes encoding proteins that have critical functional roles in photosynthetic machinery. Similarly co-expression analyses revealed that expression of all nuclear genes that are known or predicted to function at each of the individual steps in the CBP are highly correlated to PSY. The high degree of

co-expression between the MEP pathway, phytochrome pathway and ALA biosynthesis strongly suggests that transcription of these pathways are regulated by a common mechanism (Meier *et al.*, 2011).

Heat maps revealed that the transcription of PSY and its co-expressed genes is modulated in a uniform manner in response to various environmental conditions. Meier *et al.*,(2011) found this to be consistent with the high expression values of the co-expressed genes in roots. Functional annotation of the co-expressed genes revealed a number of significantly enriched GO terms. For the category "biological processes", terms such as "photosynthesis", "plastid organisation" and "biogenesis" were enriched. For the category molecular function, terms such as "tetraprenoid metabolic process" and "carotenoid biosynthesis" were enriched and for cellular components "chloroplasts", "thylakoid parts" and "plastid parts" were enriched. Additionally Meier *et al.* (2011) looked at carotenoid gene expression under osmotic stress and found that a more immediate response was elicited in roots where the stress was applied, than in shoots. Specific genes such as ZDS,  $\beta$ OHase 1,  $\beta$ OHase 2, ABA1, VDE and NCED3 had an early and sustained increase in expression in response to osmotic stress. In addition the researchers showed a reduction in expression of carotenoid genes in shoot tissue between 3 and 6 hours and they show a continued decrease for the entire 24hr period. They also observed a strong and transient induction in  $\beta$ OHase 2 and ZDS, two known core carotenoid genes, between 3-12hrs in shoot tissue. Promoter content analyses revealed 2 *cis*-elements (GBOX and AuxRE) to be enriched in the promoters of the co-expressed genes and were proposed as candidate regulatory elements that aid in the regulation of their transcription of the co-expressed genes.

## 1.5 Gene expression profiling using microarrays

Gene expression profiling is a means of measuring the expression of thousands of genes at once to get a global understanding of cellular function. Studying these profiles may lead to the detection of genes that are differentially regulated in response to a specific treatment or cells that are actively dividing. Expression profiling studies report genes that are differentially regulated and that shows statistically significant differences in gene expression under defined experimental conditions.

Microarrays are tools used for analysing gene expression, consisting of glass slides or tiny membranes, probes and multiple numbers of genes arranged in an ordered manner (Ramaswamy and Golub, 2002). Microarray technology exploits the capability of a given cDNA molecule to specifically hybridize to the template from which it has originated (Lee *et al.*, 2000). Using this tool, scientists are able to determine the expression levels of thousands of genes within a cell (Kane *et al.*, 2000).

### 1.5.1 The Importance of Microarrays

Microarrays are useful when one wants to survey a large number of genes quickly or when the sample that needs to be analysed is small (Churchill, 2002). This technology can be used to assay gene expression in a single sample or to compare the gene expression in two different cell types or tissue samples such as healthy or diseased tissue (Hegde *et al.*, 2000). Because a microarray can be used to examine the expression of hundreds or thousands of genes, it promises to revolutionize the way scientists examine gene expression (Ramaswamy and Golub, 2002).

an increased likelihood of functional relation as a result of the guilt by association principle (Mayer *et al.*, 1999).

### 1.5.2 Future prospects of microarrays

As more and more information accumulates, scientists will be able to use microarrays to answer more complex questions and perform more challenging experiments. With these new advances, researchers will be able to functionally categorise genes based on similarities in expression patterns in comparison to patterns of known genes. This is mainly due to the fact that genes that share similar expression profiles tend to share similar functional annotations. Ultimately, such advanced studies promise to reveal new patterns of gene expression and to expand the sizes of current existing gene families. Furthermore, since gene products interact with an array of other gene products our limited understanding of these gene interactions will become much more clearer through this type of analyses. Microarrays may also enable scientists to examine much larger datasets rapidly and thus decrease the time needed for identification of key genes involved in various biological processes.

Microarray based expression profiling has been remarkably successful at elucidating the spatio-temporal patterns of mRNA transcripts within cells and tissues, however there are a number of shortcomings to the existing technology. Both sensitivity and specificity can be low with microarrays. Accuracy can also be negatively affected by the low dynamic range of existing microarray technology. Perhaps more importantly, microarrays restrict the expression profiling data to specific annotations and content.

Digital expression profiling using next generation sequencing (NGS) promises to reduce or in

some cases eliminate these weaknesses. NGS has enabled sequencing of DNA at unprecedented speed and thereby has revolutionised genomics. Next-Gen technologies facilitate whole genome and transcriptome sequencing and targeted resequencing, and are applicable to a wide variety of scientific investigations. NGS offers extremely high sensitivity and accuracy which is in contrast with that seen from microarrays. NGS is thus seen as the way forward for expression profiling in the near future.

## 1.6 Co-expression and correlation analyses of carotenoid genes and their co-expressed genes

Genes are seen to be co-expressed when they share a similar expression profile under a specific stimulus or when they are expressed at the same time in the same model organism as determined by multiple experiments (Chen *et al.*, 2010). Gene co-expression, can imply the presence of a functional linkage between genes (Meier *et al.*, 2011). Co-expression analysis has uncovered gene regulatory mechanisms in model organisms such as *Escherichia coli* and yeast. Recently, accumulation of Arabidopsis microarray data has facilitated a genome-wide inspection of gene co-expression profiles in this model plant. In this study Manfield *et al.*, (2006) used a network analyses approach that provided an intuitive way to represent complex co-expression patterns between many genes (Manfield *et al.*, 2006). Co-expression network analysis is a powerful approach for data-driven hypothesis construction and gene prioritization, and provides novel insights into the system-level understanding of plant cellular processes (Obayashi *et al.*, 2007).

## 1.7 Promoter content analyses and functional enrichment

In higher eukaryotes, gene transcription is controlled by a variety of mechanisms such as chromatin modifications or degradation via complementary miRNAs. Gene promoters and their *cis*-regulatory element composition, however, are the initial checkpoints for transcriptional gene activities and define the potential spatiotemporal expression of a gene (Howell *et al.*, 2009). Identifying and characterising transcription factor binding sites is a prerequisite to understanding regulation of individual genes and their functions within regulatory networks. To overcome experimental limitations, computational methods have been developed as time- and cost-effective complements for large-scale motif discovery. These include mapping of known motifs and the identification of *de novo* motifs (Wang *et al.*, 2009).

Transcription factors interact with specific DNA elements, protein elements and other factors and the basal transcriptional machinery to regulate the expression of target genes. In plants, transcriptional regulation is mediated by more than 1500 transcriptional factors; each of these factors controls the expression of tens or even thousands of target genes in complex signalling networks (Li and Tompa, 2006).

Microarray gene expression data can help to identify groups of co-expressed genes. Clusters of such co-expressed genes are assumed by Meier *et al.* 2011 to be co-regulated and upstream sequences of these genes are likely to share common DNA motifs (Sandve and Drablø s, 2006). Presumed upstream regulatory regions of arbitrary length can be selected to identify candidate DNA motifs (Sharma *et al.*, 2011). Because of their importance we studied motifs that were over represented amongst the promoters of carotenoid genes and their co-expressed genes to

get a global understanding of the regulatory modules involved in the regulation of carotenoid biosynthetic pathway. As a result of the incomplete understanding of regulation it is clear that novel discoveries should be a priority, as a result *de novo* motif prediction will add value to the basic knowledge available to fully understand the regulatory modules that exist in complex organisms (Ettwiller *et al.*, 2007). With a broader spectrum of knowledge, more in depth studies can be undertaken and more complex questions can be answered.

### 1.7.1 Available computational tools for expression profiling and promoter content analyses

There are many online tools available for research in the area of gene expression profiling, co-expression analyses, correlation analyses and promoter content analyses. Databases housing microarray data for *Arabidopsis thaliana* under the influence of various environmental stimuli include the Information Arabidopsis Resource (TAIR) (<http://www.arabidopsis.org/index.jsp>) (Huala *et al.*, 2001) and the Nottingham Arabidopsis Stock Centre's (NASCArray) (Craigon *et al.*, 2004).

TAIR is a database containing genetic and molecular information about *Arabidopsis thaliana*, including complete genome sequences, gene product information as well as gene structures. Other information includes metabolic data, genome maps, genetic and physical maps, gene expression, publications as well as information about the Arabidopsis research community (Huala *et al.*, 2001).

NASCArray (<http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl>) is a database containing over 400 arrays of *Arabidopsis thaliana* experimental data under various biotic and abiotic



stimuli. The website contains pre-calculated values for the expression of genes of interest under various experimental conditions. Either multiple or single experiments can be viewed by the user depending on the biological question.

### **STRING database**

STRING (<http://string-db.org/>) the online database for the identification of protein-protein interactions focuses on functional protein association. Protein-protein interaction networks are an important ingredient for the system-level understanding of cellular processes. Such networks can be used for filtering and assessing functional genomics data and for providing an intuitive platform for annotating structural, functional and evolutionary properties of proteins, all this is possible by interrogating the protein protein interactions between genes. Exploring the predicted interaction networks can suggest new directions for future experimental research and provide cross-species predictions for efficient interaction mapping.

### **ATTED-II database**

ATTED-II (<http://atted.jp>) is a database of gene co-expression in Arabidopsis that can be used to prioritize genes and to infer functionality of genes for studies targeted at understanding the underlying regulatory modules of co-expressed genes (Obayashi *et al.*, 2007, 2009). ATTED-II provides a resource for expression networks offering the following functionality:

- It has a new measure for gene co-expression, to enable the retrieval of functionally related genes more accurately,
- It contains click-able maps for all gene networks in order to enhance step-by-step naviga-

tion,

- It includes information about protein-protein interactions,
- It identifies conserved patterns of co-expression, and
- It shows and connects Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information to identify functional modules (Obayashi *et al.*, 2007, 2009).

#### **ACT database**

ACT (<http://www.arabidopsis.leeds.ac.uk/act/index.php/>) uses large microarray datasets from the Nottingham Arabidopsis Stock Centre. This database stores pre-calculated co-expression results for 22,800 genes based on data from over 400 arrays. It allows for the identification of gene co-expression patterns across single or multiple arrays depending on the user's needs (Manfield *et al.*, 2006).

#### **ATHENA database**

ATHENA (<http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl>) is an online web tool that aids in the understanding of the regulatory networks that control plant gene expression. It enables one to visualize the promoter regulatory sequences in *Arabidopsis thaliana*. Athena contains over 30000 predicted promoter sequences. The visualisation tool enables the inspection of key regulatory elements in multiple sequences (?).

## 1.8 Rationale and focus

In 2007 a study by Kilian and others observed large differences in expression levels between differentially regulated genes. From the studies done on drought stress it was found that very few genes indicated a reaction to drought and this led them to the conclusion that plants recovered relatively fast from drought as an increase in expression profiles were observed. They also found that majority of the up and down regulated genes were specific for drought, cold and UVB stress. Similarly these genes were also responsive to salt wounding and osmotic stress and up regulated genes were shared to a high degree amongst the stresses in both roots and shoots. Furthermore, Kilian *et al.* (2007) looked at the transcription factors that were active during the exposure to these stresses. They found that 4 of the 9 genes that were up regulated after only 30min under cold, drought and UVB were well characterised transcriptional regulators. These four factors identified were found to complement signalling components such as Ca<sup>2+</sup> (Kilian *et al.*, 2007).

A systems biology approach has successfully been used in predicting regulatory mechanisms in eukaryotes (Meier *et al.*, 2008). Meier *et al.* have shown that when sets of genes are co-expressed under various stimuli, it is likely that they share common regulatory mechanisms, particularly common transcription factor binding sites and specific motifs ( $\approx 6-8$  bases) (Meier *et al.*, 2011)). Therefore, the identification of common promoter signatures encoding carotenoid genes and their co-expressed genes in the CBP can be used to build putative gene networks that are predicted to play a key role in elevating carotenoid levels with particular relevance to food crops.

However, investigations by Kilian *et al.*, (2007) and more recently Meier *et al.*, (2011), looked at overall gene expression in Arabidopsis. The gene expression in the carotenoid pathway was

not studied in isolation. Similarly, regarding regulatory modules, Kilian *et al.*, (2007) and Meier *et al.*, (2011) looked at global regulators involved in stress tolerance and isoprenoid production respectively. As regards to stress factors, Meier *et al.*,(2011) focused on the effect of osmotic stress on ABA, GA and carotenoid biosynthesis. Kilian, *et al.*, (2007), however, looked at the global effect of cold, drought, UV-B, salt, wounding and osmotic stresses in *Arabidopsis*.

During the preparation of this thesis a similar albeit slightly different approach was published by Meier *et al.* (2011). In this study I investigate the transcriptional effects of environmental factors influencing carotenoid biosynthesis, as well as the underlying transcriptional regulatory mechanism involved in carotenoid synthesis in *Arabidopsis thaliana*. In Meier *et al.*, very broad conclusions were made regarding the development of *Arabidopsis thaliana* and its underlying regulatory mechanisms without investigating the carotenoid biosynthetic pathway and its associated genes in isolation. Not much is known, therefore, about carotenoid biosynthesis and which genes are directly involved in the biosynthetic pathway of carotenoids.

The aims of this study are:

1. To compile a list of known carotenoid genes with a direct involvement in  $\beta$  carotene production,
2. To identify genes that are co-expressed and have correlated differential expression with known carotenoid genes, across various conditions and stresses.
3. To discover putative *cis*-elements and transcription factors with a key role in the transcription regulation pathway of carotenoid biosynthetic genes.
4. To use a systems biology approach to mine microarray data to identify conditions affecting

carotenoid gene expression.

## 1.9 Structure of the thesis

The rest of the thesis is organised as follows. Chapter 2 details the methods used to study the regulatory attributes of the carotenoid biosynthetic pathway in *Arabidopsis thaliana* under abiotic stress. A flow chart representing the flow of information and the various tools used is shown in Figure 1.6.

In Chapter 3 I present the results obtained from analyses tools and provide detailed descriptions of the findings.

In Chapter 4 I discuss the relevance and significance of results obtained, and draw conclusions from the results which contribute to a better understanding of the underlying regulatory mechanism of the carotenoid biosynthetic pathway.

In Chapter 5 I summarise the key findings and main conclusions from this study and propose future avenues to enhance and extend this research.

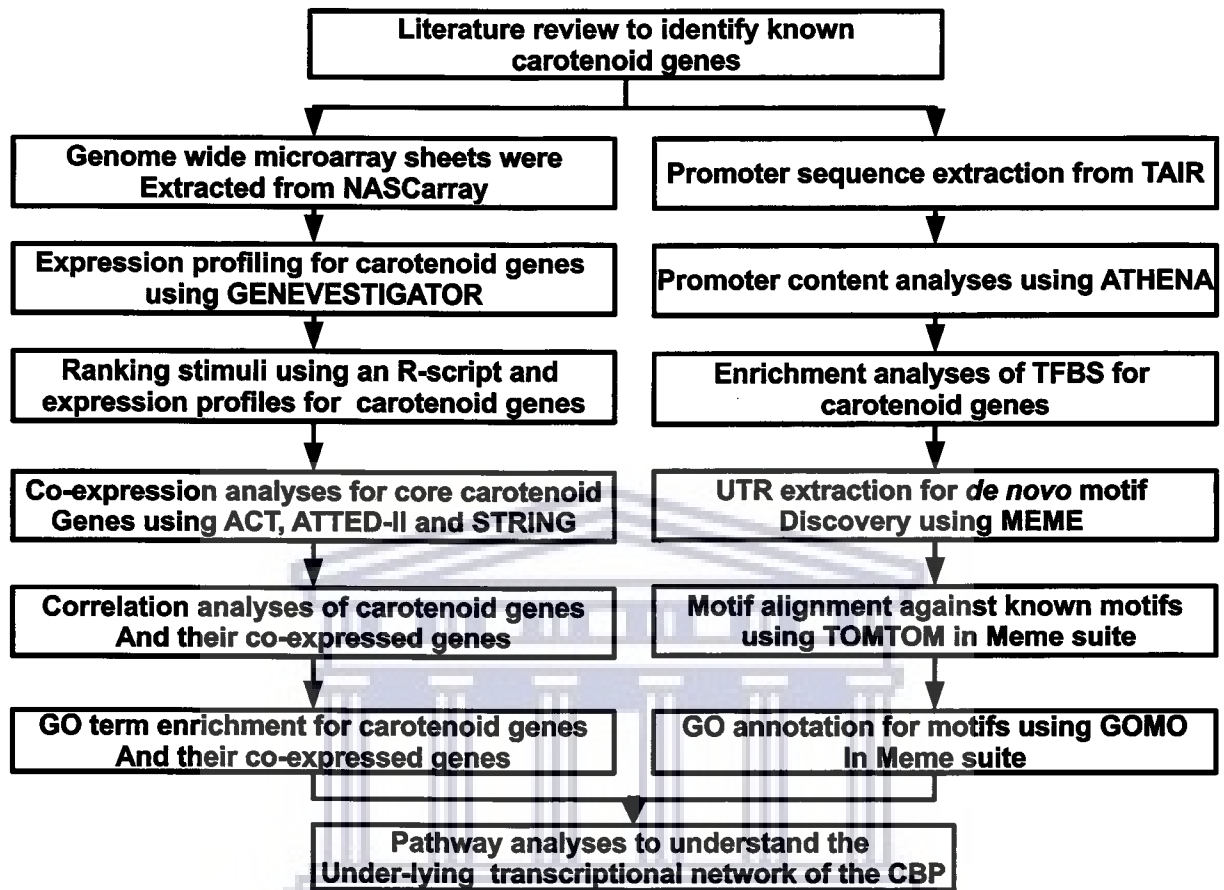
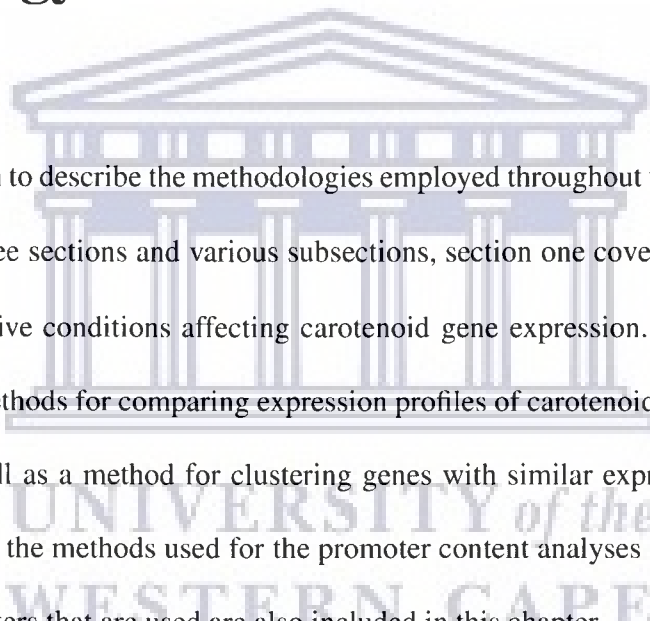


Figure 1.6: A diagram of methodologies used in the thesis.

# Chapter 2

## Methodology



In this Chapter we aim to describe the methodologies employed throughout this thesis. The chapter is divided into three sections and various subsections, section one covers the methodologies used to identify putative conditions affecting carotenoid gene expression. The second section addresses different methods for comparing expression profiles of carotenoid genes under various abiotic stresses as well as a method for clustering genes with similar expression profiles. The final section describes the methods used for the promoter content analyses section of this thesis. Details on the parameters that are used are also included in this chapter.

### **2.1 Identification of conditions affecting carotenoid gene expression**

#### **2.1.1 Extraction of data using literature**

A list of 32 experimentally verified carotenoid genes was compiled using of literature. These genes include those previously identified as being potentially linked to the carotenoid biosyn-

thetic pathway (CBP) and genes that are carotenoid producing. Genes from the non-mevalonate pathway (MEP), mevalonate, geranylgeranyl phosphate pathway (GGPP) and flavonoid pathway were also included in the list. As they have previously been shown to be interlinking pathways that are involved in carotenoid biosynthesis process. These pathways are also known for their involvement in the production of pro-vitamin A in *Arabidopsis thaliana* (Li *et al.*, 2009).

### 2.1.2 Identification of environmental stimuli that affect carotenoid gene expression

Literature was skillfully mined to determine which of the stresses were most highly influential in the survival of the *Arabidopsis thaliana* species and critical to carotenoid biosynthesis. From the extensive experiments done by investigators such as Kilian *et al.* (2007) and Meier *et al.* (2008) it was clear that drought, cold and UVB stress affecting *Arabidopsis thaliana*, had a greater effect on the gene expression and were directly involved in carotenoid biosynthesis. To investigate the gene expression profiles of core carotenoid genes, expression data derived from microarray experiments were retrieved from NASCarray Files -137, 138, 139, 140, 141, 145 and 146, containing the expression of the whole genome under various abiotic stresses, were retrieved from the Arabidopsis information resource (TAIR) and NASC- array (the Nottingham Arabidopsis Stock Centre's microarray database) (Huala *et al.*, 2001; Craigan *et al.*, 2004). The stresses include: cold, heat, osmotic, oxidative, drought, light, wounding, and salt stress and were a larger dataset compared to that used by Meier *et al.* (2011)

The retrieved files were text-mined to extract the relevant information pertaining to the known carotenoid genes using a perl script (col.pl in appendix A) and this information was stored as a tab



delimited file. These files then became the dataset for further analysis. The complete genome files were used as our background datasets. The following ten core carotenoid genes with TAIR id's (alias), AT5G17230 (PSY), AT4G14210 (PDS), AT1G06820 (CRTISO), AT1G10230 (LCY $\beta$ ), AT4G25700 ( $\beta$ OHase1), AT5G52570 ( $\beta$ OHase2), AT1G31800 (LUT5), AT3G53130 (LUT1), AT3G04870 (ZDS) and AT5G57030 (LCY $\epsilon$ ) were selected from the list of 32 genes. These genes were selected because they are genes that are serially represented in the carotenoid biosynthetic pathway. PSY, one of the core carotenoid genes, is the driver gene of the carotenoid biosynthetic pathway and thus influences carotenoid gene expression directly. This is therefore the primary reason for selecting these genes for analyses in this section. The above mentioned genes will be used as a reference to investigate the effect of various stimuli on carotenoid gene expression at different time points ranging between 0.5hrs- 24hrs.

### 2.1.3 Data processing

The data downloaded from TAIR was then formatted using a perl script (extract.pl in appendix A). New tab delimited text files were created for the whole genome. This file contained relevant columns of the original data sheet including gene ids, *p*-values and expression values for each of the time points varying from 0.5-24hrs. The relevant tab delimited files were then imported into a MySQL database. This was done by running another script on the data (ave.pl in appendix A). This script calculated the mean value for the replicates of each gene at 6 different time points namely 0.5, 1, 3, 6, 12 and 24hrs for the whole genome. The average *p*-value for the replicates was also calculated and placed in the tab delimited file along with the mean expression value for the whole genome. Information for the 32 known carotenoid genes were extracted and placed in

new tab delimited files for further analyses.

### 2.1.4 Expression profiling

An expression fold change value was calculated for the core carotenoid genes under various stress conditions as mentioned in section 2.2.3, with respect to the control samples from NASCarray-137 in two tissue types (root and shoot).

A fold change is a ratio of the measured value for an experiment sample to the value of a control sample.

The file for the control samples is available at:

<http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=137>.

The fold change values were calculated by dividing the expression of the core carotenoid gene divide by the expression value of those genes in the control sample. An R-script (foldchange appendix A) was used to plot graphs of the fold change expression values at different time points (0.5, 1, 3, 6, 12, 24hrs) in two tissue types (roots and shoots) for all the above mentioned abiotic stress treatments.

## 2.2 Co-expression and co-correlation analyses

There are two methods of undertaking a co-expression analysis, namely a guide-gene approach and a non-targeted approach (Figure 2.1).

In a guide-gene approach a single gene is used as a driver gene at any given time. Publicly available databases are mined using the driver gene as a guide. Results obtained from mining the

co-expression databases are visualised. A second guide gene is selected and a similar process is followed thereafter the results from both guide genes are compared and evaluated with regard to the hypothesis.

In a non-targeted approach a list of genes i.e. test dataset, is used as the driver of the experiment. Multiple genes are taken into consideration. Co-expression databases are mined to identify genes that are co-expressed with the entire test dataset instead of a single guide-gene at a time as in the guide-gene approach. The results obtained are then visualised and evaluated based on the hypothesis.

For the purpose of our co-expression analysis a combination of the two approaches were used to identify genes that were co-expressed and co-correlated with carotenoid biosynthetic pathway genes. This can be compared to the guide gene approach used by Meier *et al.* (2011)

### **2.2.1 Co-expression analyses using STRING, ATTEDII and ACT co-expression databases**

In this section, the methodologies employed for the co-expression analyses of carotenoid biosynthetic pathway genes using three publicly available databases will be reported.

Protein sequences for the ten core carotenoid genes were retrieved using the Ensembl plants database

(<http://plants.ensembl.org/index.html>). Protein coding sequences of the ten core carotenoid genes were used as a query to search for known and putative protein-protein interactions between core carotenoid genes and their co-expressed genes using the STRING database. This tool was used to generate gene networks of the protein-protein interactions between co-expressed genes and

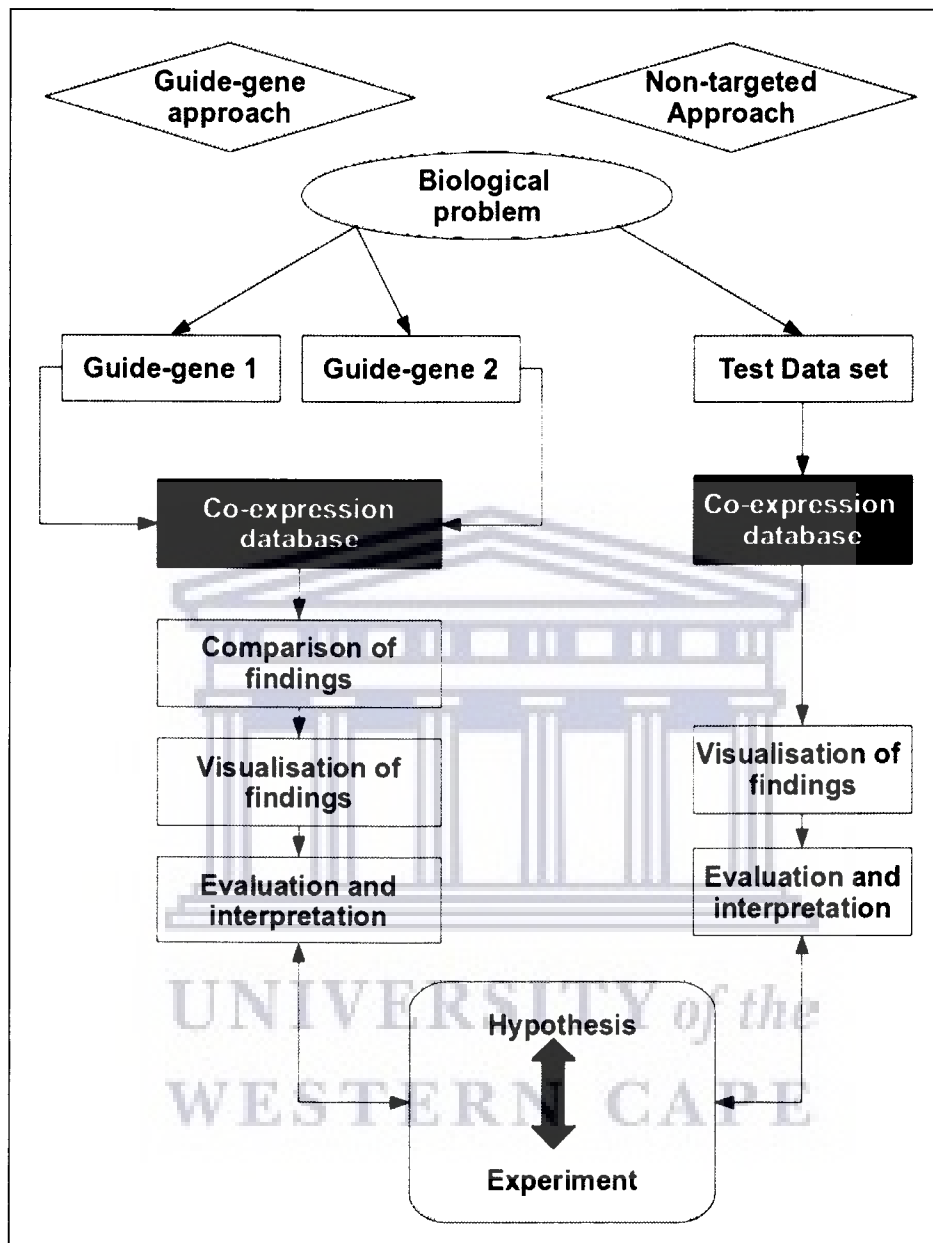


Figure 2.1: **Practical protocols of co-expression analysis.** Left: guide-gene approach, in which co-expression profiles between and within selected guide genes are first investigated. Right: non-targeted approach, in which the modular structure is extracted from the entire network according to the topology of the links.

each of the 10 core carotenoid biosynthetic genes (Snel *et al.*, 2000). The core carotenoid genes were used as driver genes to produce expression networks. To produce each of the expression

networks, parameters were judiciously chosen as follows: (i) a confidence level of 0.7, (ii) a network depth of 4 and (iii) restricting to show only the top 50 interactions between the core carotenoid genes and their co-expressed genes. These networks showed the association between the known genes and possible new uncharacterised genes present in the expression networks (Mering, 2003).

Gene ids were used as a query in ATTED-II (<http://atted.jp>) to identify genes that were co-expressed with the core carotenoid genes. All the parameters for this search were maintained at default. default setting are sufficient as co-expression is based on mutual rank (MR), that is calculated as the geometric mean of the correlation rank of gene A to gene B and of gene B to gene A. Lists of coexpressed genes based on MR values are provided along with a weighted Pearson's correlation coefficient (PCC) for each gene pair. Expression networks were produced showing the level of co-expression with reference to the correlation coefficient.

Gene ids of the ten core carotenoid genes were used to extract probe ids using ACT the Arabidopsis co-expression tool ( <http://www.arabidopsis.leeds.ac.uk/act/index.php>). The probe ids were used as a query to identify genes that were co-expressed with core carotenoid genes. The parameters used for this search included, selecting ATH1-22K arrays, limiting the output to the top 50 genes co-expressed with core carotenoid genes and finally ranking them in descending (positively correlated) order with regard to the  $r$ -value coefficient.

Each of the databases produced a list of co-expressed genes for each of the core carotenoid genes. These lists of co-expressed genes were used for further analyses.

Each of the co-expressed gene lists produced from the three databases for each of the core carotenoid genes were scanned using a statistical program, R-script, to identify which genes were

commonly co-expressed for each gene across web tools. Venn diagrams were plotted to represent the amount of co-expressed genes shared between the various web tool. Only the genes that were common for at least two web tools were considered as being truly co-expressed. We now possessed a co-expressed gene list for each of the core carotenoid genes. These lists were then scanned using another R-script which then searched across the 10 gene lists for genes that were enriched across the lists. Each gene was allocated a score that ranged from 0-10, indicating the amount of gene lists that a specific gene occurs in. Only genes with an occurrence score of more than 50% were added to the final co-expressed gene list. A list of 86 potentially co-expressed genes were generated and these were used for further analyses.

## 2.2.2 Co-correlation analyses using Pearson correlation measure

The correlation analyses is based on Pearson correlation which measures the strength and direction of a linear relationship between the variables  $X$  and  $Y$ . The measure indicates whether there is negative or positive correlation and usually ranges from -1 to 1. The closer the correlation is to  $\pm 1$ , the closer to a perfect linear relationship. The association between two variables can be described using the following indicators:

- -1.0 to -0.7: strong negative association.
- -0.7 to -0.3: weak negative association.
- -0.3 to +0.3: little or no association.
- +0.3 to +0.7: weak positive association.
- +0.7 to +1.0: strong positive association.

Both positive and negative correlation are of biological significance. According to the formula below 2.1, it indicates that the Pearson correlation coefficient,  $r(X, Y)$ , effectively normalizes the magnitude of the expression vector. This equation was used to calculate the correlation coefficient in the online web tool ACT. That is, for genes which have a relatively moderate expression pattern, even if the expression levels are dramatically different, they will be identified as having a similar expression response (the Pearson correlation coefficient is close to one).

$$r(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}} \quad (2.1)$$

where

- $X, Y$  are variables,
- $\bar{X}, \bar{Y}$  are averages for  $X$  and  $Y$  respectively,
- $X_i$  and  $Y_i$  are elements in  $X$  and  $Y$  respectively.

## 2.3 Promoter content analyses, functional enrichment and *de novo* motif prediction

### 2.3.1 Promoter extraction of carotenoid genes and their co-expressed genes

The promoter sequences of 1000bp upstream and 200bp down stream from the transcription start site for all 32 carotenoid genes and their co-expressed genes were extracted using ATHENA (*Arabidopsis thaliana* expression network analyses). TAIR gene ids were used as an input for analyses with ATHENA. The promoter regions of all 32 carotenoid genes were visualized using ATHENA visualization tools. Predicted transcription factor binding motifs were identified and represented by coloured lines visible in the promoter regions of the carotenoid genes. ATHENA identified CpG islands as well and these are represented by the aqua boxes in the compact view (?).

### 2.3.2 *De novo* motif discovery using MEME Suite

Promoter regions 1000bp upstream and 200bp down stream were extracted from TAIR for all the carotenoid genes. Using a perl script (UTR.pl in appendix A) the promoter proximal regions were extracted and saved in FASTA format. The list of promoter proximal regions were now the unaligned sequences as indicated in Figure 2.2. The file with the unaligned sequences was used as an input for MEME (multiple Em for motif elicitation). To submit a job to meme the following command needs to be used in the command line: `meme <dataset> [optional arguments]`. For the purpose of our analyses the following parameters were used, the `<dataset>` file containing



sequences in FASTA format. The optional arguments included *-dna* indicating that the sequences used were in DNA alphabet, *-nmotifs* were set to 10 motifs to be identified and *-o* indicating the location to which the results need to be written. The rest of the parameters of the software remained at default. The output of the job from *meme* was called *meme\_results* and served as an input for the rest of the analyses.

The MEME results contained a *.html* file for visualization for the user, *.xml* file and a *.txt* file. Each of these files contained exactly the same information, however the different file formats were needed as input for Average Motif Affinity (AMA,<sup>3</sup>), Gene Ontology for Motifs (GOMO,<sup>4</sup>) and Motif comparison tool (TOMTOM,<sup>2</sup>) programs respectively. The flow of information is depicted in Figure 2.2.

Identified enriched *de novo* motifs were next analyzed by TOMTOM in the MEME suite for comparison against a database of known motifs. TOMTOM uses a list of motifs created by *meme* and compares it against known nucleotide databases using Pearson correlation. The format of the command needed for analyses in TOMTOM was as follows: `tomtom [options] <query file> <target file> +`, where the options were the parameters, the query file was the *.txt* file from *meme* and the target file was the database of reference which in this case was the JASPAR\_CORE\_DATABASE. All the parameters used for the TOMTOM analyses were set to default except the threshold and verbosity which was set to 1 and 1 respectively.

The program scores a set of DNA sequences given a DNA-binding motif, treating each position in the sequence as a possible binding event. The score is calculated by averaging the likelihood ratio scores for all feasible binding events to the given sequence and to its reverse strand. The binding strength at each potential site is defined as the likelihood ratio of the site

under the motif versus under a zero-order background model provided by the user. The format of the command used to create the *.cismf* file needed for analyses in GOMO is as follows: `ama [options] <motif file> <sequence file> [<background file>]`, where options indicate the parameters, motif file represents the *.xml* file from the meme analyses, the sequence file is *.na* file in the GOMO database repository and the background file is represented by the *.na.bile* file found in the GOMO database folder. All the parameters for this analysis were set to the programs default settings.

The purpose of GOMO is to identify possible roles (Gene Ontology terms) for DNA binding motifs. GOMO returns a list of GO-terms that are significantly associated with target genes of the motif, sorted by *q*-value (minimum false discovery rate). Gene Ontology provides a controlled vocabulary to describe gene and gene product attributes in any organism. The command format for analyses in GOMO was as follows: `gomo [options] <go-map file> <scoring file> +`, where options represent the parameters, the go map file represents the Path to the optional Gene Ontology DAG to be used for highlighting the specific terms in the gomo *.xml* output and the scoring file represents the *.cismf* file generated by AMA previously. All the parameters for this analysis were set to default.

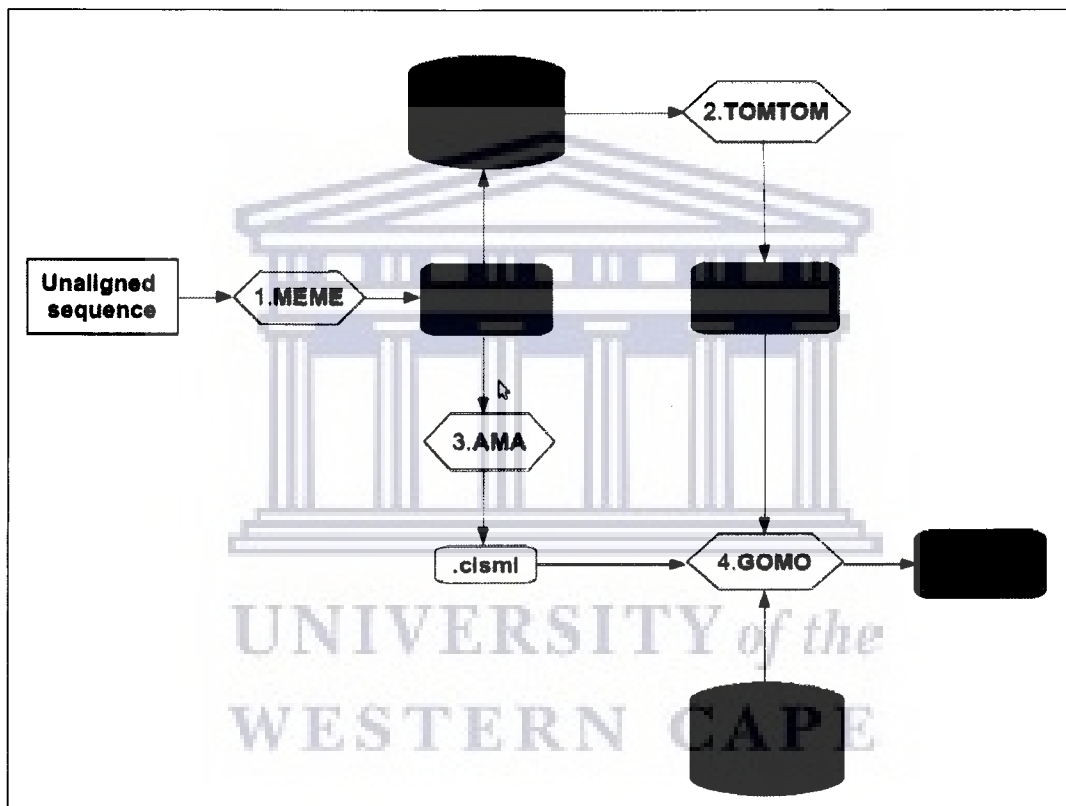


Figure 2.2: Graphical representation of the work flow for *de novo* motif prediction.

### **2.3.3 Identification and enrichment analyses of transcription factor binding motifs**

Predicted transcription factor binding motifs (TFBMs) were identified for both the known 32 carotenoid genes and their 86 co-expressed genes. A list of potential TFBMs along with the frequency of the TFBMs present within in the test set (known carotenoid genes) and the entire genome are available in appendix A. The  $p$ -value indicating the significance of each of the TFBMs are also present in the in appendix A.  $P$ -values are calculated according to the hypergeometric distribution. A  $p$ -value cut-off of  $10^{-4}$  is used and is corrected according to the Bonferroni correction method for errors. For further information on the Bonferroni correction method visit the following url ([http://en.wikipedia.org/wiki/Bonferroni\\_correction](http://en.wikipedia.org/wiki/Bonferroni_correction)).

The list of predicted TFBMs were scanned for potential TFBMs that were significantly enriched within our subset in comparison to the entire *Arabidopsis thaliana* genome. A percentage fold change enrichment value was calculated using the following equation:

$$FC = \frac{\% \text{ present in subset}}{\% \text{ present in the genome}} \quad (2.2)$$

### **2.3.4 GO term enrichment and Functional annotation**

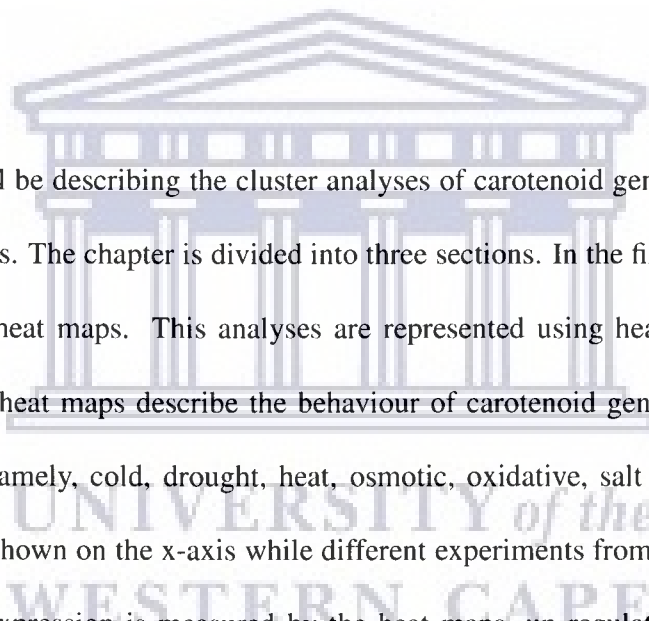
The goal of the GO term enrichment analyses was to identify dynamic controlled vocabularies that can be used to describe the roles of genes and gene products in all organisms. The three organizing principles of GO are molecular function, biological process and cellular component.(Swarbreck *et al.*, 2008).

GO terms are used as attributes of gene products in order to facilitate uniform queries across various databases. The controlled vocabularies of terms are structured to allow both attribution and querying to be at different levels of granularity. GO annotations identified through TAIR, are annotations that have been manually curated by curators of TAIR and TIGRs arabidopsis annotation. The annotations made by TAIR are made using a combination of manual and computational methods (Consortium, 2006).



# Chapter 3

## Results



In this chapter we will be describing the cluster analyses of carotenoid genes under various environmental conditions. The chapter is divided into three sections. In the first section the results will be displayed as heat maps. This analyses are represented using heat maps as shown in Figures 2.2A-D. The heat maps describe the behaviour of carotenoid genes under various environmental stimuli namely, cold, drought, heat, osmotic, oxidative, salt and wounding. The carotenoid genes are shown on the x-axis while different experiments from each stress is on the y-axis. Differential expression is measured by the heat maps, up regulation is shown by red blocks, down regulation is shown by green blocks and genes that show little or no effect under the influence of the applied stress is shown in black. The second section provides a detailed list of the results obtained from various computational tools and gives insight in the type of genes that are co-expressed and co-correlated with carotenoid biosynthetic genes. The final section we will be describing the results obtained from the Arabidopsis thaliana network analyses tool and MEME suite. This section is further subdivided into four subsections. Subsection one covers the overall representation of the promoter region and the identification of predicted TFBM's present

in the promoters of the carotenoid genes. The results are represented by compact views of the promoter regions extracted from ATHENA. Subsection two cover the identification of enriched TFBMs, within promoters of carotenoid biosynthetic pathway genes. Subsection three focuses on GO term enrichment amongst carotenoid genes and their co-expressed genes. The findings are represented by Pie charts of the three GO term categories, Biological processes, molecular function and cellular components. The final subsection demonstrates the results for the promoter content analyses and the identification of novel motifs.

## **3.1 Identification of putative conditions affecting carotenoid gene expression**

### **3.1.1 Identification of known carotenoid genes**

A total of 32 known carotenoid biosynthetic genes were identified through a literature search. The gene identifiers, gene descriptions, their aliases and five prospective biosynthetic pathways are summarized in Table 3.1. These pathways include the MEP (non mevalonate), mevalonate, GGPP, carotenoid biosynthetic and flavonoid pathways. The CBP genes are split into two groups (Table 3.1). The group highlighted in green are the core carotenoid biosynthetic genes. These genes were selected because they are genes that are linearly represented in the carotenoid biosynthetic pathway. The group in blue, contains a group of degradative enzymes involved in carotenoid biosynthesis.

Table 3.1: 32 known carotenoid biosynthetic pathway genes derived from literature. Yellow blocks represent genes involved in MEP, GGPP and Mevalonate pathways. Green blocks represent the core carotenoid genes. The blue blocks are carotenoid genes that act as degradative enzymes.

<b>AT5G11380</b>	<b>1-deoxy-D-xylulose 5-phosphate</b>	<b>DXPS3</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT4G15560</b>	<b>1-deoxyxylulose 5-phosphate synthase</b>	<b>DXPS2</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT5G62790</b>	<b>1-deoxy-D-xylulose 5-phosphate reductoisomerase</b>	<b>DXR</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT2G26830</b>	<b>4-(cytidine-5'-phospho)-2-C-methyl-D-erythritol kinase</b>	<b>ATCDPMEK</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT1G63970</b>	<b>2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase</b>	<b>MECPS</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT5G60600</b>	<b>Hydroxy-2-methyl 1-2-(E)-butenyl 4-diphosphate synthase</b>	<b>HDS</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT4G34350</b>	<b>Isoprenoid diphosphate reductase</b>	<b>ISPH</b>	<b>MEP Non-mevalonate pathway</b>
<b>AT3G02780</b>	<b>Isopentenyl diphosphate isomerase 2</b>	<b>IPP2</b>	<b>Mevalonate pathway</b>
<b>AT4G38460</b>	<b>Geranylgeranyl reductase</b>	<b>GGR</b>	<b>GGPPS</b>
<b>AT2G23800</b>	<b>Geranylgeranyl pyrophosphate synthase 2</b>	<b>GGPS2</b>	<b>GGPPS</b>
 <p><b>UNIVERSITY of the WESTERN CAPE</b></p>			
<b>AT2G26170</b>	<b>MORE AUXILLARY BRANCHING 1</b>	<b>MAX1</b>	<b>Flavanoid pathway</b>



### **3.1.2 Cluster analyses of carotenoid genes under various environmental conditions**

The behaviour of carotenoid genes under various environmental stimuli namely cold, drought, heat, osmotic, oxidative, salt and wounding were represented as heat maps. The data represented in the heat maps are summarized in table 3.2.

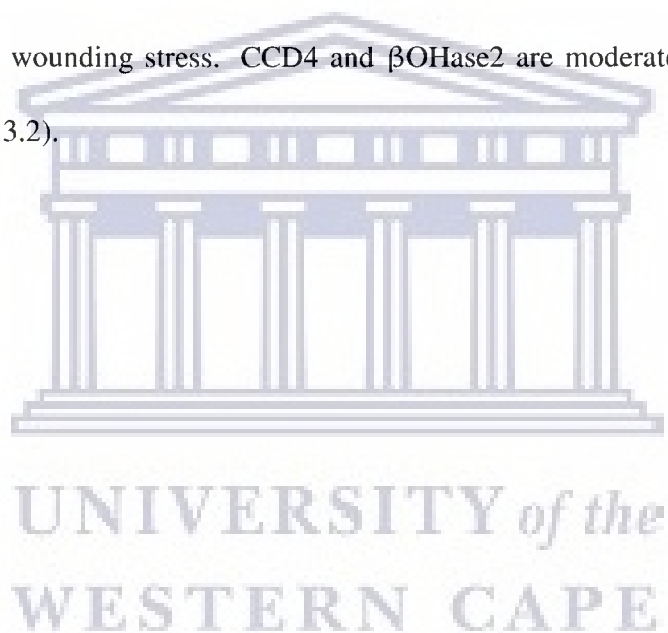
From the heat maps it is clear that stimuli differ with regard to their effect on the differential expression of carotenoid genes. Under cold stress NCED3,  $\beta$ OHase2 and ZEP are strongly up regulated, while HDS, ISPF, DXS2 and CCD4 are strongly down regulated. Carotenoid genes such as LCY $\beta$ , ZDS, PNPase, and ISPH are moderately up regulated, while MAX1, IPP2, ZISO, PDS, CCD1, VDE, LCY $\epsilon$ , LUT5,  $\beta$ OHase1 and PSY are moderately down regulated. The rest of the carotenoid genes show no effect under the influence of cold stress (Figure 3.1 and Table 3.2). During drought stress, NCED3 is strongly up regulated and PSY, LUT5, VDE, LCY $\beta$ , LCY $\epsilon$ , DXS2, LUT1 and CCD4 are strongly down regulated. Some carotenoid genes including PNPase, DXR, GGPS1,  $\beta$ OHase2 and ESPE are moderately down regulated under the influence of drought stress (Figure 3.2 and Table 3.2).

When plants are subjected to heat stress, an array of responses prevail. The genes CCD4 and LCY $\beta$  are strongly up regulated while most of the genes except  $\beta$ OHase2 and LUT1 are moderately up regulated. MAX1 however is moderately down regulated under the influence of heat stress. The rest of the carotenoid genes show no real effect under the influence of heat stress, (Figure 3.3 and Table 3.2).

During osmotic stress NCED3 and  $\beta$ OHase1 are strongly up regulated while LUT1 and ISPE

### **3.1.2 Cluster analyses of carotenoid genes under various environmental conditions 43**

are strongly down regulated. Carotenoid genes such as PSY, ZEP, VDE,  $\beta$ OHase2 and LCY $\beta$  are moderately up regulated while LUT5, GGPS1, DXS2, PNPase, DXR, LCY $\beta$  and LCY $\epsilon$  are moderately down regulated. The rest of the carotenoid genes show no effect under the influence of osmotic stress (Figure 3.4) and Table 3.2. When carotenoid genes are exposed to salt stress NCED3 is strongly up regulated whereas MAX1 is strongly down regulated. The rest of the carotenoid genes show no effect during the exposure to salt stress. However when carotenoid genes are placed under wounding stress all of the carotenoid genes except CCD4 and  $\beta$ OHase2 show no effect under wounding stress. CCD4 and  $\beta$ OHase2 are moderately down regulated, (Figure 3.4 and Table 3.2).



### 3.1.2 Cluster analyses of carotenoid genes under various environmental conditions 44

Table 3.2: Overview of the differentially expressed carotenoid genes under different conditions namely, cold, drought, heat, osmotic, oxidative, salt and wounding.

Condition	Expression		
	Up-regulated	Down-regulated	Expression status
<b>Cold treatment</b>	NCED3, $\beta$ OHase2, and ZEP	HDS, ISPE, CLA1/DXS2 and CCD4	Strongly expressed
	B-LCY, ZDS, PNPase and ISPH	MAX1, IPP2, Z-ISO, PDS, CCD1 VDE, LCY $\epsilon$ , LUT5, $\beta$ OHase1 and PSY	Moderately expressed
<b>Heat treatment</b>	CCD4 and LCY $\beta$	No clear pattern	Strongly expressed
	Almost all genes except $\beta$ OHase2 NCED5 & LUT1	MAX1	Moderately expressed
<b>Osmotic stress</b>	NCED3, & B $\beta$ OHase1	LUT1 & ISPE	Strongly expressed
	PSY, ZEP, $\beta$ OHase2 VDE, LCY $\epsilon$ ,	LUT5, VDE, E-LCY, LCY $\beta$ , DNS2, PNPase, DXR & GGPS1	Moderately expressed
<b>Salt stress</b>	NCED3	MAX1	Strongly expressed
<b>Wounding</b>	N/A	N/A	Strongly expressed
	N/A	CCD4 & $\beta$ OHase2	Moderately expressed
<b>Drought stress</b>	NCED3	PSY, LUT5, VDE, LCY $\epsilon$ , LCY $\beta$ , CLA1, LUT1, CCD4	Strongly expressed
	N/A	PNPase, DXR, GGPS1, $\beta$ OHase2, ESPE	Moderately expressed

The expression status of carotenoid genes are indicated in the last column and is taken from the heat maps. Specific targeted genes are mentioned under either the up or down regulated columns

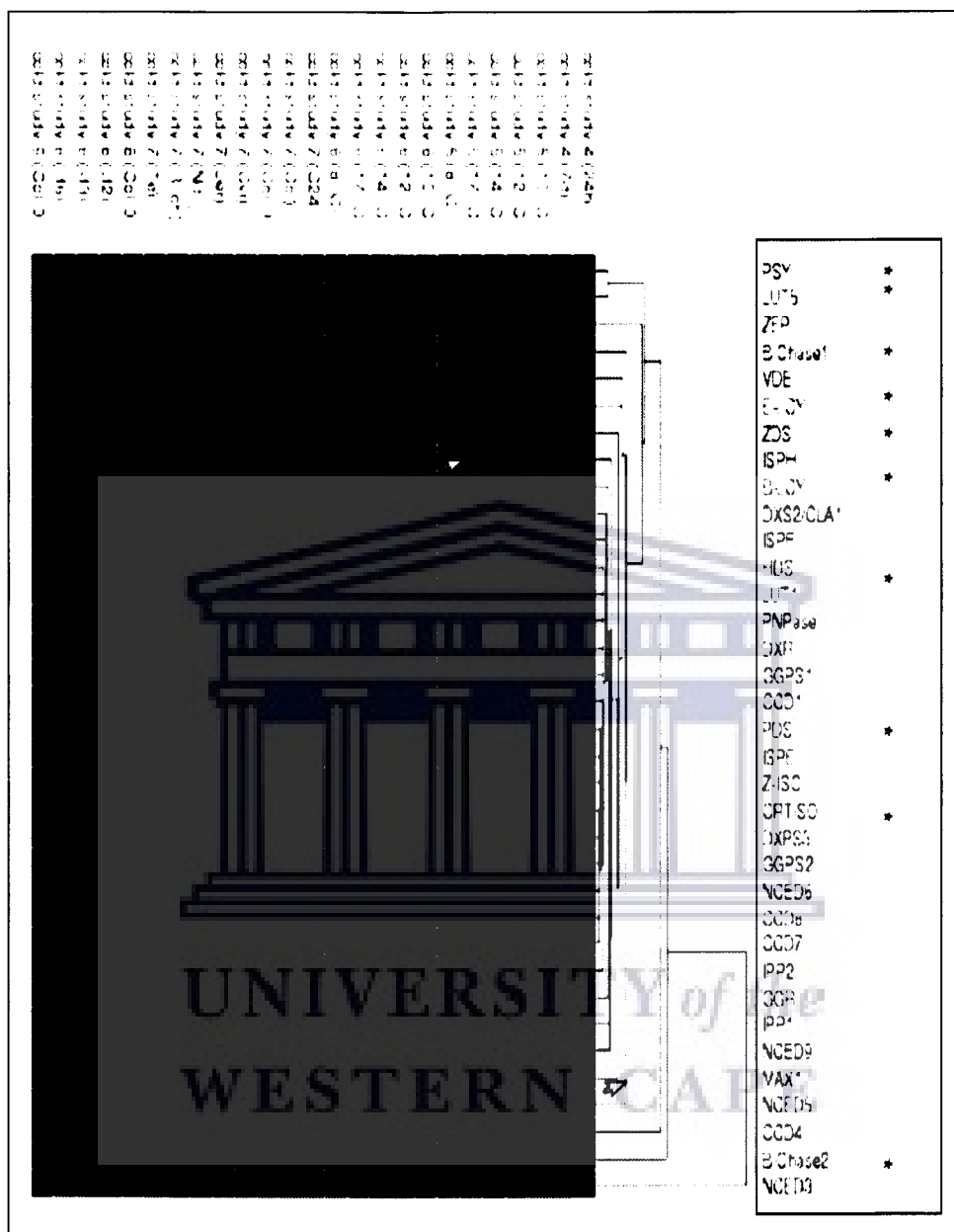


Figure 3.1: Heat map expression profiles of CBP genes across multiple experiments under cold stress.

Up regulation is shown in red, down regulation in green and no change is shown in black. The Figure shows clustering profiles under multiple cold stress. Asterix(\*) indicate the core carotenoid genes and their position on the heat map and their location in the respective clusters. Genes that are affected in the same manner under each of the stresses are clustered together.

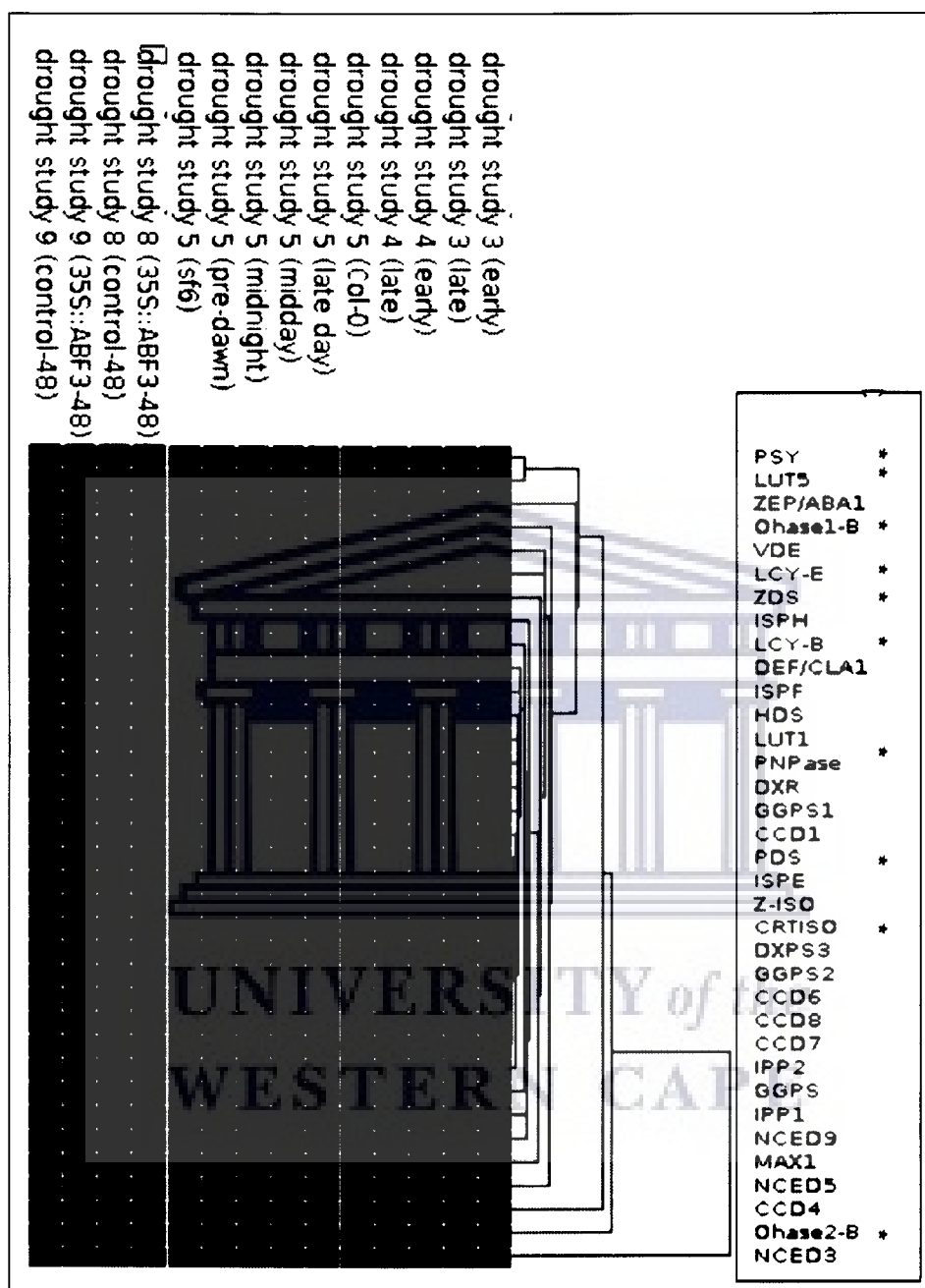


Figure 3.2: Heat maps expression profiles of CBP genes across multiple experiments under drought stress.

Up regulation is shown in red, down regulation in green and no change is shown in black. The Figure shows clustering profiles under drought stress. Asterisk (\*) indicate the core carotenoid genes and their position on the heat map and their location in the respective clusters. Genes that are affected in the same manner under each of the stresses are clustered together.

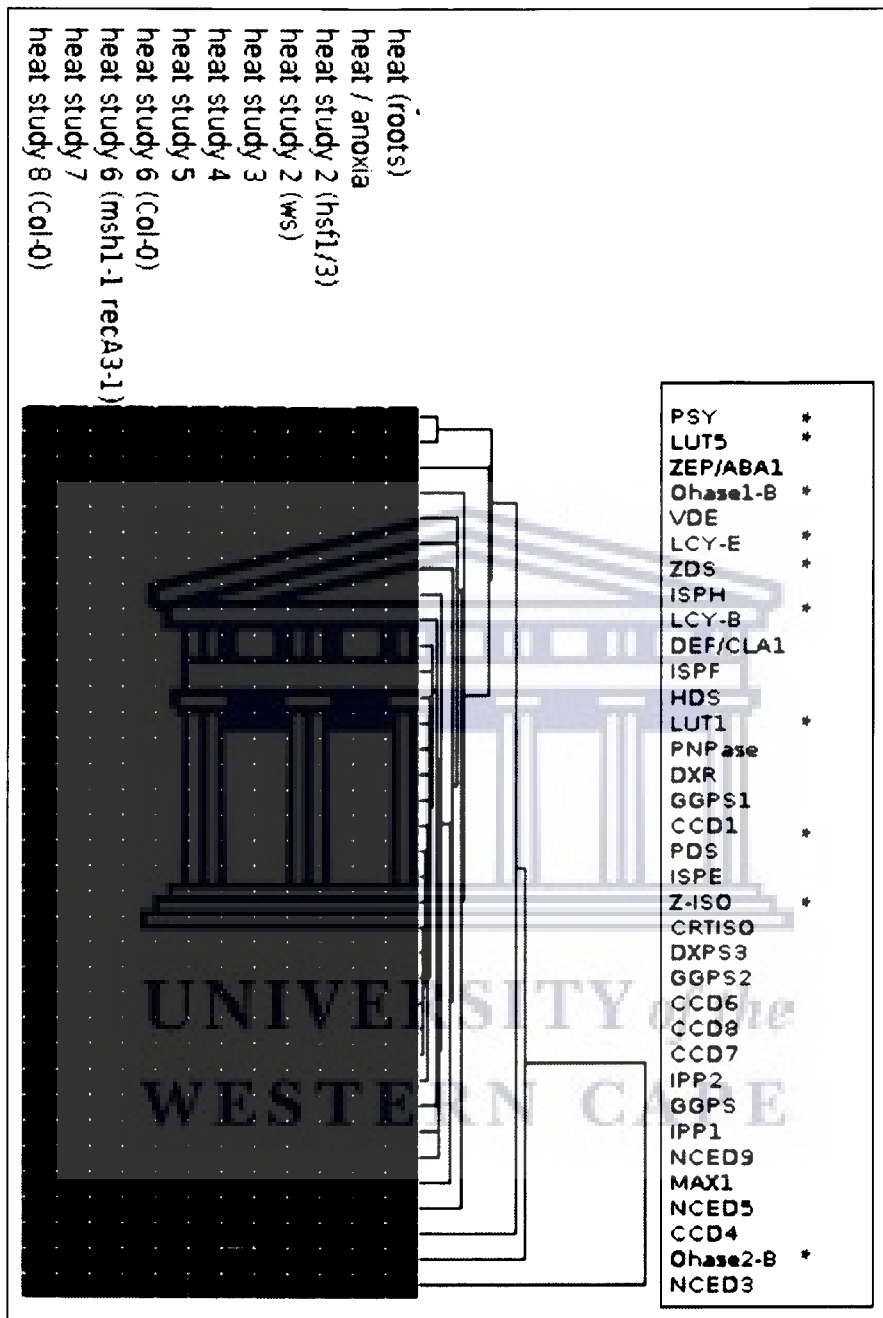


Figure 3.3: Heat maps expression profiles of CBP genes across multiple experiments under heat stress.

Up regulation is shown in red, down regulation in green and no change is shown in black. The figure shows clustering profiles of CBP genes under multiple heat stress studies. Asterix (\*) indicate the core carotenoid genes and their position on the heat map and their location in the respective clusters. Genes that are affected in the same manner under each of the stresses are clustered together.

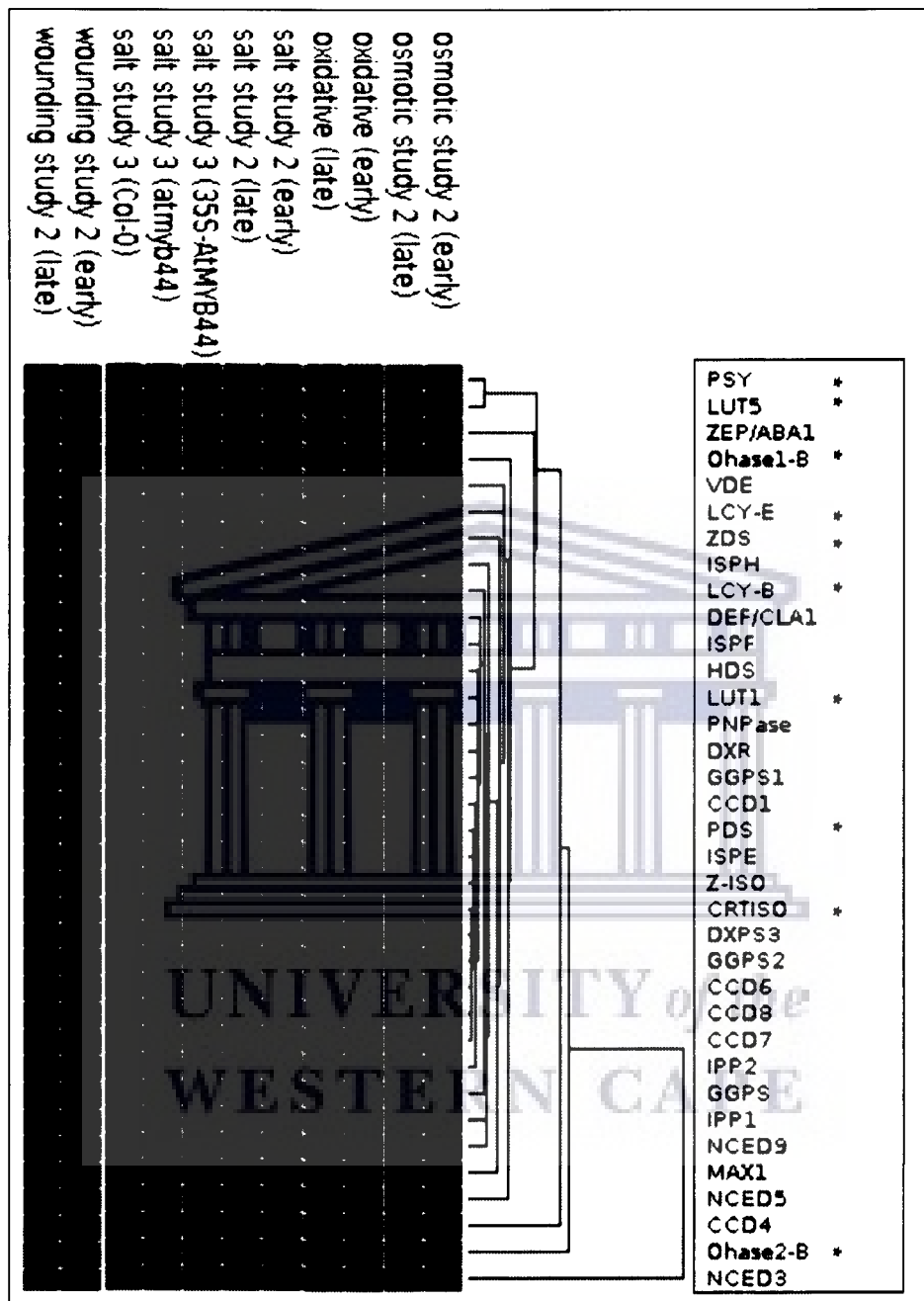


Figure 3.4: Heat maps expression profiles of CBP genes across multiple experiments under osmotic, oxidative, salt and wounding stress. Up regulation is shown in red, down regulation in green and no change is shown in black. The figure shows clustering profiles of CBP genes under osmotic, salt, oxidative and wounding stress. Asterix (\*) indicate the core carotenoid genes and their position on the heat map and their location in the respective clusters. Genes that are affected in the same manner under each of the stresses are clustered together.

### 3.1.3 Expression profiling of core carotenoid genes

Expression profiles were generated for core carotenoid genes under various stimuli i.e. drought, cold, heat, osmotic, oxidative, salt and wounding stress at different time points ranging from 0.5hrs-24hrs in both roots and shoots.

#### Drought Stress

The expression fold change of core carotenoid genes in shoots and roots at different time points under drought stress are shown in Figure 3.5. CRTISO and ZDS display an approximate 4-fold change in the expression of the shoot tissue throughout the time points in comparison to the expression of core carotenoid genes in control sample of *Arabidopsis thaliana* (Figure 3.5) (red bar graph).  $\beta$ OHase2, the gene responsible for converting  $\beta$  carotene to zeaxanthin via cryptoxanthin shows a 2-3 fold change in expression during time points 0.5hrs-3hrs. Thereafter a steep decline in fold change expression is observed and a fold change of less than 1 is prevalent from 6hrs-24hrs in the shoot tissue. LUT1, the lutein deficient gene in *Arabidopsis* shows an insignificant fold change throughout the time points except at time point 24hrs, where a fold change of 2 is noted in shoot tissue. LCY $\epsilon$  and LCY $\beta$  only shows an increase in fold change expression after 12hrs and continues to increase as time of exposure increases. In root tissue there is a completely different expression profile, only LCY $\epsilon$ , LCY $\beta$  and ZDS shows a 2-fold change in expression across time points. LUT5 shows a constant insignificant fold change across time points and the rest of the core carotenoid genes show a fold change of 1 throughout the time points (Figure 3.5) (blue bar graph).



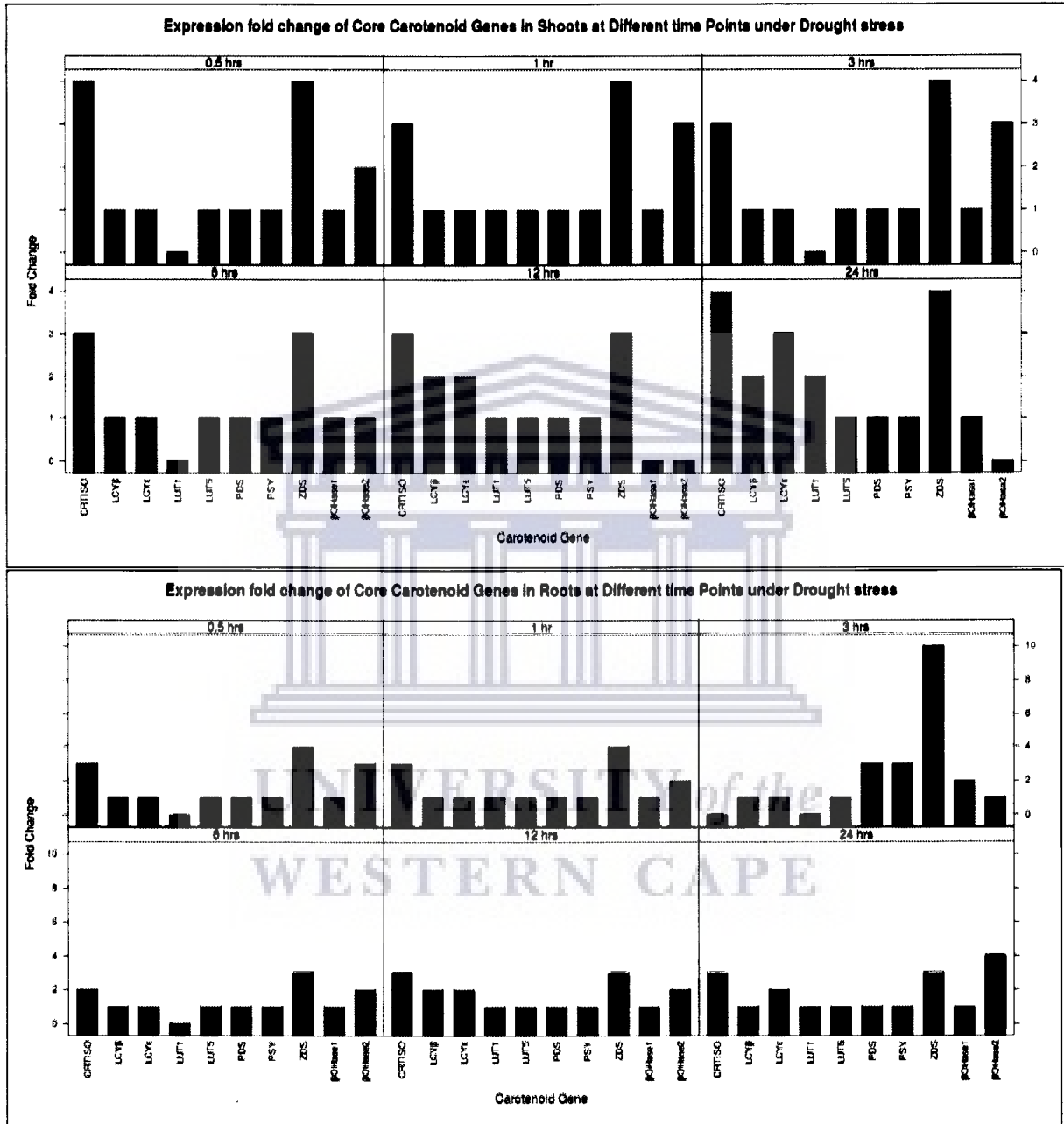


Figure 3.5: Expression profiles of core carotenoid genes under drought stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types: Shoots shown by red bar graphs and roots shown by blue bar graphs.

### Cold Stress

The expression profile of core carotenoid genes in shoots and roots at different time points under cold stress are shown in Figure 3.6. In shoot tissue, CRTISO, ZDS and  $\beta$ OHase2 show a fold change in expression of higher than 2 across time points. At 3hrs ZDS shows a 9-fold change in expression whereas CRTISO shows an insignificant fold change at 3hrs as shown in Figure 3.6 (red bar graph). LCY $\epsilon$  shows a 2-3 fold change in expression but this is only prevalent after 12 hrs. PSY and PDS shows a 3-fold change at 3hrs and then drops to 1 for the rest of the time points. The Lutein deficient gene LUT1 again shows an insignificant fold change in expression throughout the time points. The rest of the core carotenoid genes show a 1-fold change in expression under cold stress as shown in Figure 3.6 (red bar graph). In root tissue, LCY $\beta$  and LCY $\epsilon$  show a 2-fold change in expression across time points.  $\beta$ OHase2 shows a 7-fold change in expression after 24 hrs. The rest of the genes show a 1-fold change in expression throughout the time points as shown in Figure 3.6 (blue bar graph).



UNIVERSITY *of the*  
WESTERN CAPE

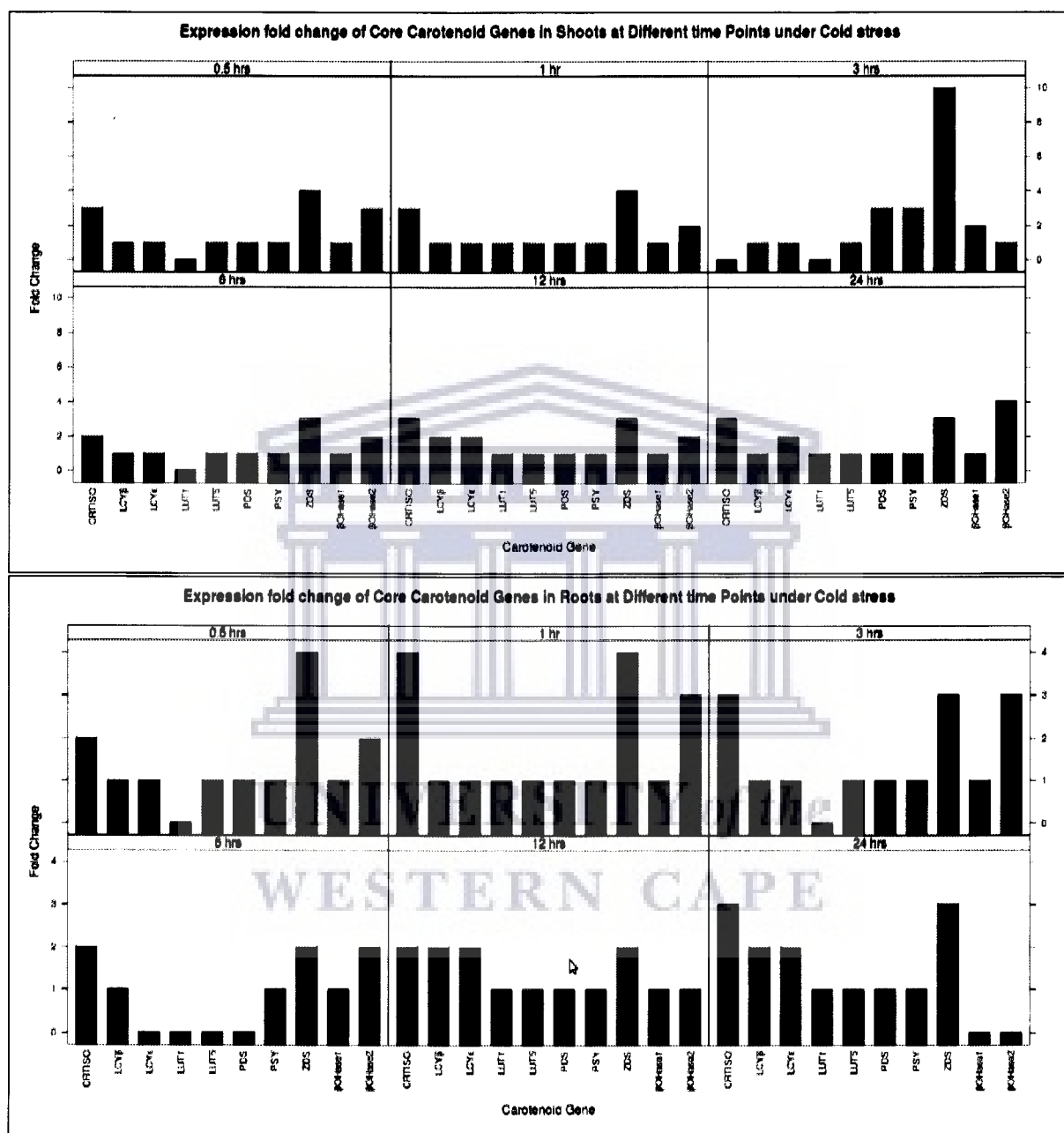
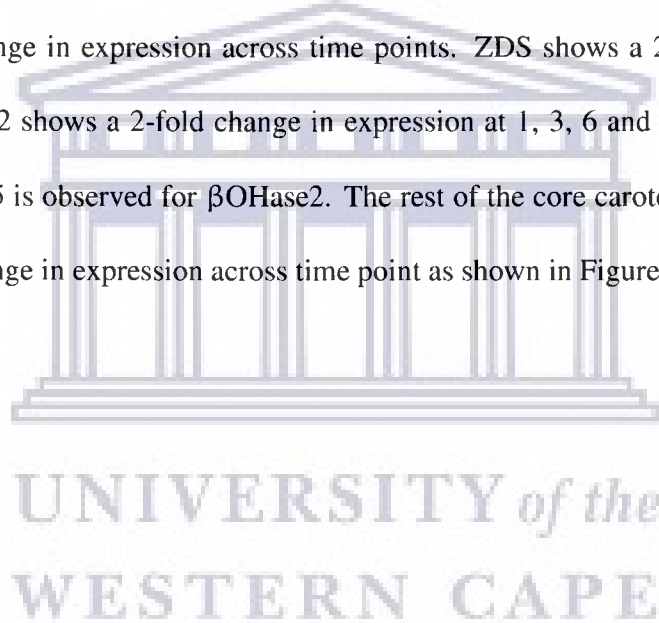


Figure 3.6: Expression profiles of core carotenoid genes under cold stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types: Shoots shown by red bar graphs and roots shown by blue bar graphs.

### Heat Stress

The expression fold change of core carotenoid genes in shoots and roots at different time points under heat stress are shown in Figure 3.7. In shoot tissue, CRTISO, ZDS and  $\beta$ OHase2 show a 2-4 fold change in expression across time points. LCY $\beta$  and LCY $\epsilon$  show a 2-fold change in expression from 12hrs onwards. At 24hrs  $\beta$ OHase1 and  $\beta$ OHase2 show an insignificant fold change in expression. The rest of the carotenoid genes show a 1-fold change in expression across all time points as shown in Figure 3.7 (red bar graph). In root tissue LCY $\beta$  and LCY $\epsilon$  show a 2 or more fold change in expression across time points. ZDS shows a 2-fold change across time points.  $\beta$ OHase2 shows a 2-fold change in expression at 1, 3, 6 and 12 hrs. Thereafter a fold change below 0.5 is observed for  $\beta$ OHase2. The rest of the core carotenoid genes show an insignificant fold change in expression across time point as shown in Figure 3.7 (blue bar graph)



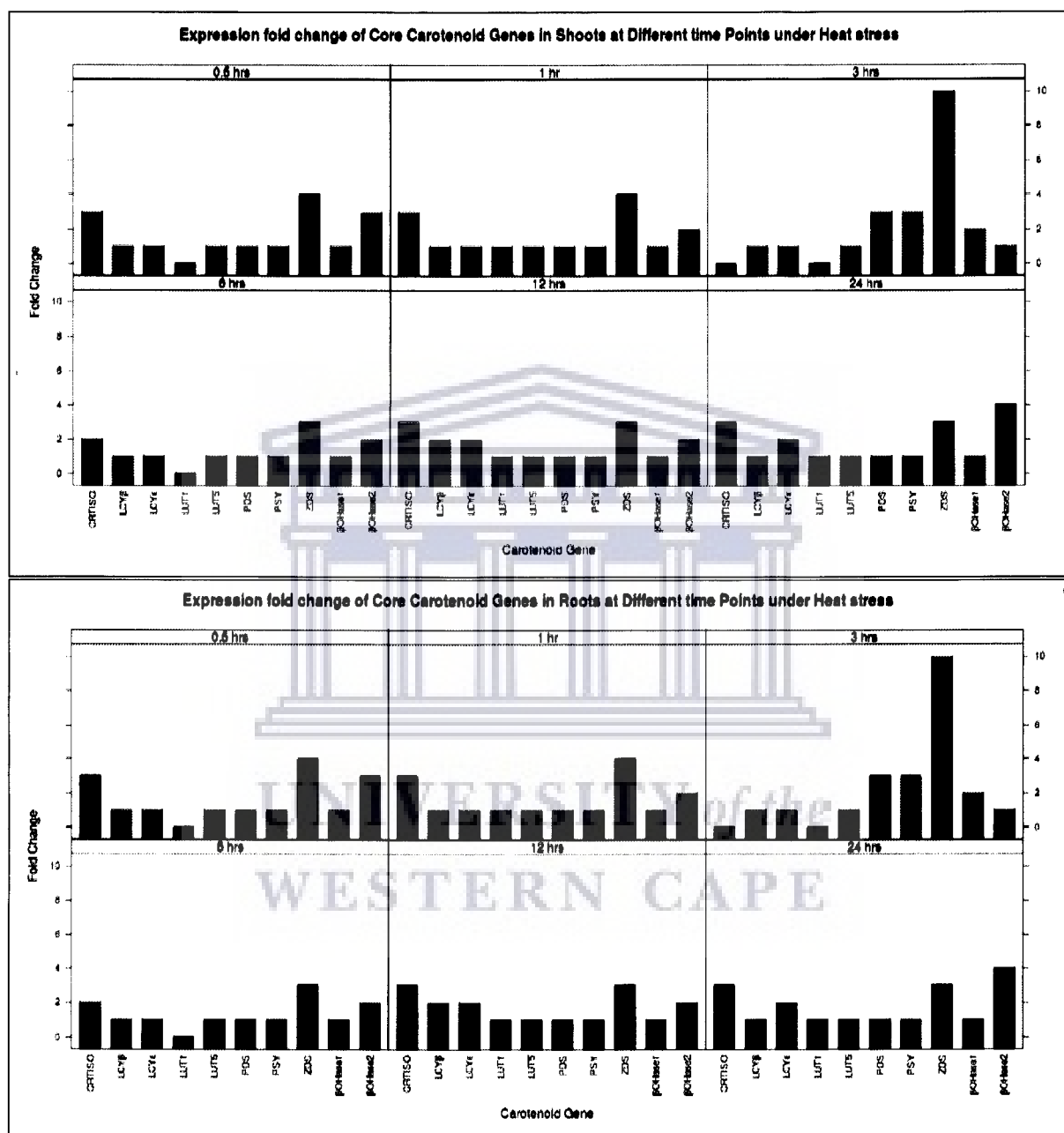


Figure 3.7: Expression profiles of core carotenoid genes under heat stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types: Shoots shown by red bar graphs and roots shown by blue bar graphs.

### Osmotic Stress

The expression fold change of core carotenoid genes in shoots and roots at different time points under osmotic stress are shown in Figure 3.8. In shoot tissue ZDS has a 5, 4, 3, 4, 5 and 6-fold change in expression at 0.5hrs, 1hr, 3hrs, 6hrs, 12hrs and 24hrs respectively. CRTISO and  $\beta$ OHase2 have a 2-fold change in expression at all time points except time points 1hr and 12hrs, where an increase of 3-fold is observed. After 12hrs of exposure CRTISO has a 3-fold change in expression and  $\beta$ OHase2 has an insignificant fold change in the expression. The lutein deficient gene LUT1 has an insignificant fold change in expression across all time points. The rest of the core carotenoid genes show a 1-fold change in expression across time points under osmotic stress, (Figure 3.8) (red bar graph). In root tissue LCY $\beta$ , LCY $\epsilon$ , PSY and ZDS has a 2 or more fold change in expression. At 0.5hrs-1hr the lowest fold change in expression is observed and it increases at time point 3hrs.  $\beta$ OHase1 and  $\beta$ OHase2 has a 2-fold change but after 6hrs it increases to a 3-fold change and continues to increase as the time of exposure increases. CRTISO shows an insignificant fold change in expression across time points under osmotic stress which is completely different to the reaction in shoots. The rest of the core carotenoid genes show an insignificant fold change in expression as shown in Figure 3.8 (blue bar graph).

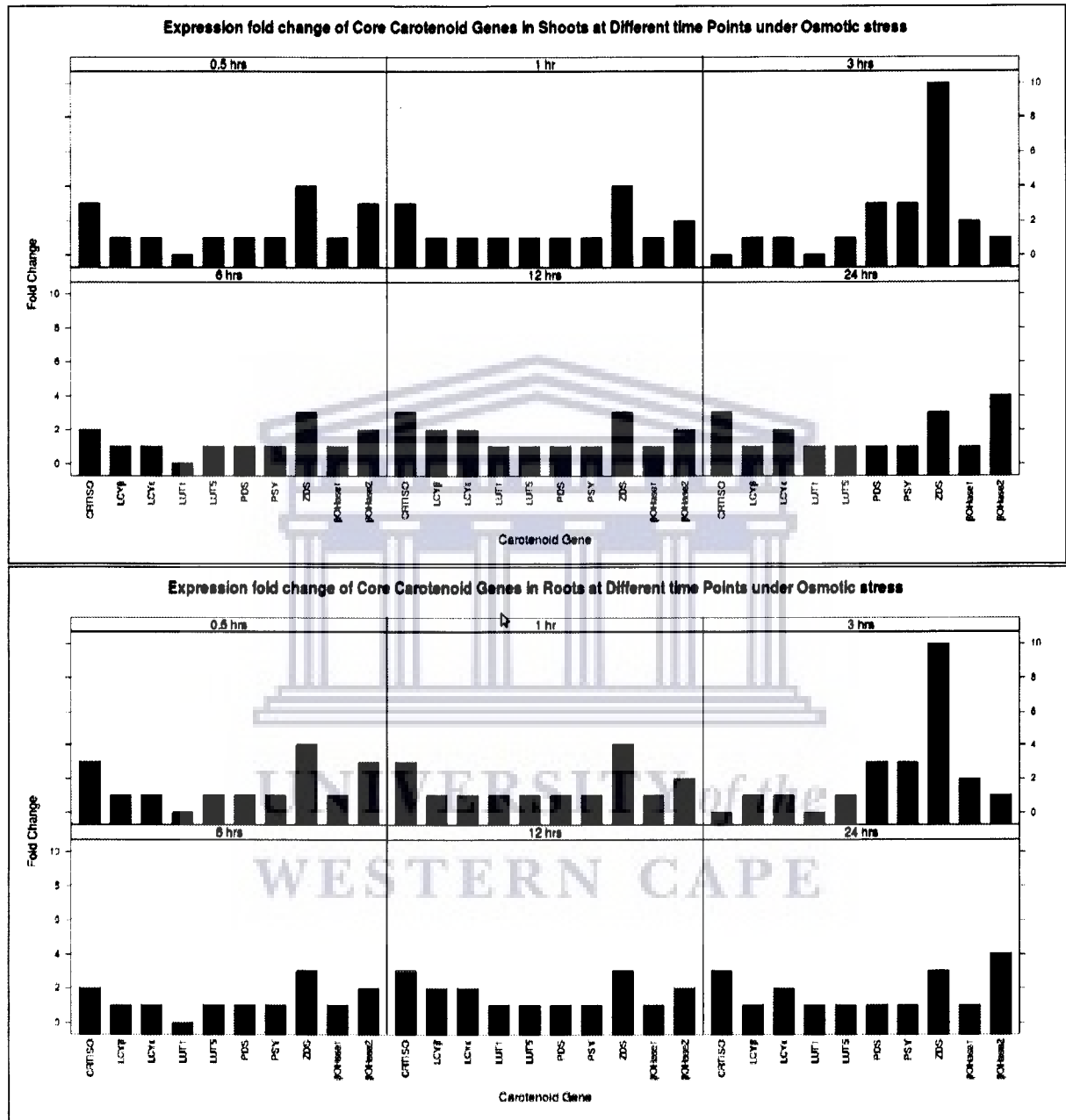


Figure 3.8: Expression profiles of core carotenoid genes under osmotic stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types: Shoots shown by red bar graphs and roots shown by blue bar graphs.

### Oxidative, Salt and Wounding Stress

The expression fold change of core carotenoid genes in shoots and roots at different time points under oxidative, salt and wounding stress are shown in Figure B.1, B.2, B.3 in A. CRTISO and ZDS show a 2-3 fold change in expression, except at 24hrs where ZDS shows an increase in expression to the value of 4-fold under oxidative stress.  $\beta$ OHase2 shows a 2 and 3-fold change in expression at 0.5hrs and 24hrs respectively. After 6hrs of exposure to oxidative stress  $\beta$ OHase2 show a decrease in fold change, after 6 hrs- 12hrs an insignificant fold change prevails. LCY $\beta$  and LCY $\epsilon$  show a 2-3 fold change in expression after 12hrs and 24hrs of exposure to oxidative stress. The lutein deficient gene remains insignificant throughout the time points except at time point 12hrs where an increase in expression to the value of 2-fold prevails. PSY and PDS shows an increase in expression to the value of 2-fold, although the change is only prevalent after 24hrs. The rest of the carotenoid genes show an increase to the value of 1-fold across time points during the exposure of oxidative stress, (Figure B.1, B.2, B.3 appendix B) (red bar graph). In root tissue LCY $\beta$ , LCY $\epsilon$  and ZDS shows an increase in expression of 2 or more fold under oxidative stress. PSY shows an increase of 2-fold after 12hr and 24hrs.  $\beta$ OHase2 shows an increase in expression to the value of 2-fold after 24hrs, the rest of the core carotenoid genes show a 1-fold increase across the time points and under the exposure of oxidative stress (Appendix figure B.1in, B.2, B.3B) (blue bar graph).

A general trend indicates that CRTISO, ZDS and  $\beta$ OHase2 showed similar expression patterns across all stimuli. ZDS had the highest fold change in expression, ranging 2-9 fold across stimuli and time points, both in roots and shoots. Shoot expression intensities were higher than that of roots.



## 3.2 Co-expression and correlation analyses of carotenoid and co-expressed genes

### 3.2.1 Co-expression analyses of core carotenoid genes

The co-expression analyses for the ten core carotenoid genes are represented by Venn diagrams as shown in Figure B.4, B.5, and B.6 in the Venn diagram section in B in appendix. From the Venn diagrams, it is clear that certain gene lists share genes. Between 30 and 200 genes are shared amongst the gene lists extracted from ACT and ATTED-II. The list generated from STRING does not share any genes with the lists from both ACT and ATTED-II. CRTISO has the least number of genes shared equating to only 39 genes. However, PSY, the driver gene of the carotenoid biosynthetic pathway, shares 96 genes between the two lists generated from the online tools ACT and ATTEDII and contains almost all the genes from the co-expressed gene list (Figure B.6) in appendix B). A list of co-expressed genes can be found in appendix B.

### 3.2.2 Co-correlation analyses of carotenoid genes and their co-expressed genes

A scatter plot showing the Pearson correlation ( $r$ -values) points of the co-expressed genes in relation to two core carotenoid genes, PSY and LCY $\beta$  against the entire *Arabidopsis thaliana* genome was plotted as shown by Figure 3.9. In the figure red dots indicate the position of the co-expressed genes associated with the core carotenoid biosynthetic pathway genes.

Co-expressed genes are positively co-correlated with PSY and LCY $\beta$  and these genes are

indicated by the red dots on the scatter plot and are labelled PSY and LCY respectively. The co-expressed genes are localized in the top right hand corner of the graph (Figure 3.9). PSY is shown on the  $x$ -axis and LCY is shown on the  $y$ -axis which represents LCY $\beta$ . To show a co-correlation of co-expressed genes, the Pearson correlation coefficient is taken into consideration. The correlation coefficient ranges from -1 to 1, therefore, as the coefficient tends to 1 the closer the co-expressed genes are correlated to the core carotenoid genes PSY and LCY. All the co-expressed genes have a correlation coefficient of 0.7 and above. Therefore, the co-expressed genes show a strong positive correlation to PSY, which is the driver gene of the carotenoid biosynthetic pathway. A detailed layout of the interaction between core carotenoid genes and their co-expressed genes in CBP is represented in Figure 3.10 where correlation values are indicated in brackets.

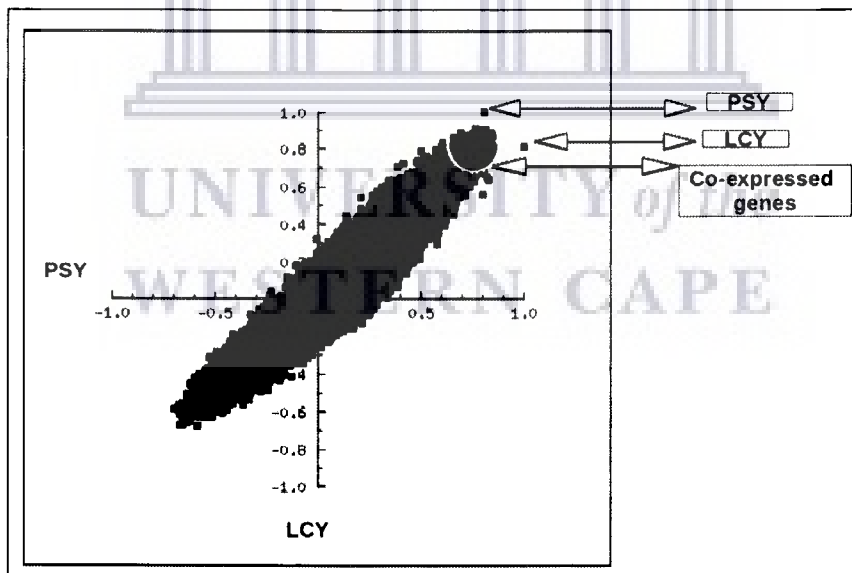


Figure 3.9: Scatter plot shows expression correlation ( $r$ -value) of all Arabidopsis genes relative to PSY and LCY (LCY $\beta$ ). Red dots are co-expressed genes associated with the carotenoid biosynthetic pathway genes. Co-expressed genes show a strong positive correlation to PSY the driver gene of the carotenoid biosynthetic pathway

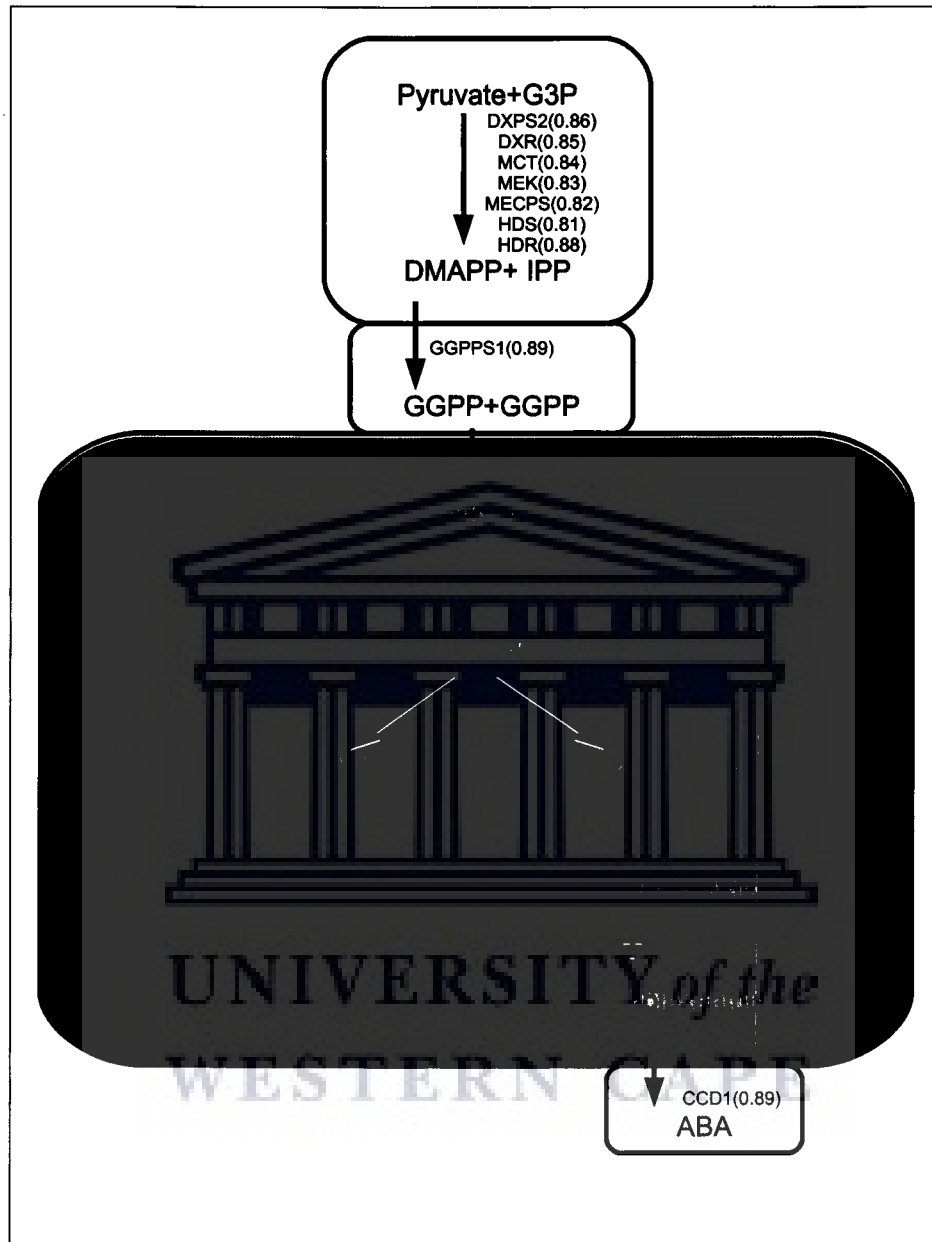


Figure 3.10: **Diagram of the CBP detailing the level of co-expression between core carotenoid genes and their co-expressed genes.** The pathways represented include the MEP pathway in the top most block in yellow, the GGPP pathway in the middle yellow block and the carotenoid pathway in the light blue block. Reaction substrates and products are represented in bold blue. Genes co-expressed with each of the core carotenoid genes are indicated in red next to bold black arrows which indicate the direction of the enzymatic reactions. The numbers in parentheses next to each of the co-expressed genes represent the expression correlation or the correlation coefficient ( $r$ -value). All  $r$ -values  $> 0.7$  are listed in the figure. A list of all 86 co-expressed genes are available in B

### 3.3 Promoter content analyses, functional annotation and *de novo* motif discovery

#### 3.3.1 GO term enrichment and functional annotation of carotenoid genes and their co-expressed genes

GO term enrichment was done in order to identify which functional annotations could be identified for both carotenoid genes and their co-expressed genes. By determining which GO terms were enriched for the two groups of genes, a link could be made to identify involvement in biological processes, molecular functions or cellular processes. This will thus give a clear understanding of the functional relatedness of carotenoid genes and their co-expressed genes.

GO annotations were extracted for three categories namely, biological processes, molecular function and cellular components. Under the category "biological processes" six GO terms were enriched for both the carotenoid genes and their co-expressed genes. These GO terms include "response to biotic and abiotic stimulus", "response to stress", "developmental processes", "transport", "cell organization" and "biogenesis" "an electron transport" or "energy pathways". For the category "cellular components" GO terms "chloroplasts", "plastids", "cytosol", "mitochondria" and "plasma membrane" were enriched for both carotenoid genes and their co-expressed genes. For the category "molecular function" only three GO terms were enriched for both sets of genes these included "transferase activity", "protein binding" and "kinase activity". These results are available in Figures 3.11, 3.12 and 3.13

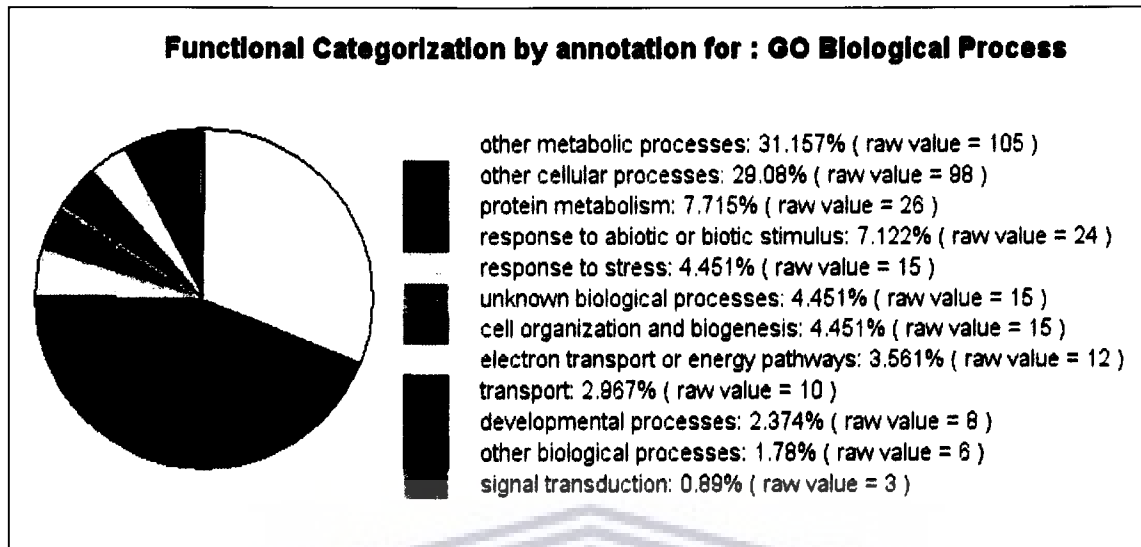


Figure 3.11: Pie chart representation of GO terms enriched for co-expressed genes in comparison to the entire *Arabidopsis thaliana* genome under the category biological processes.

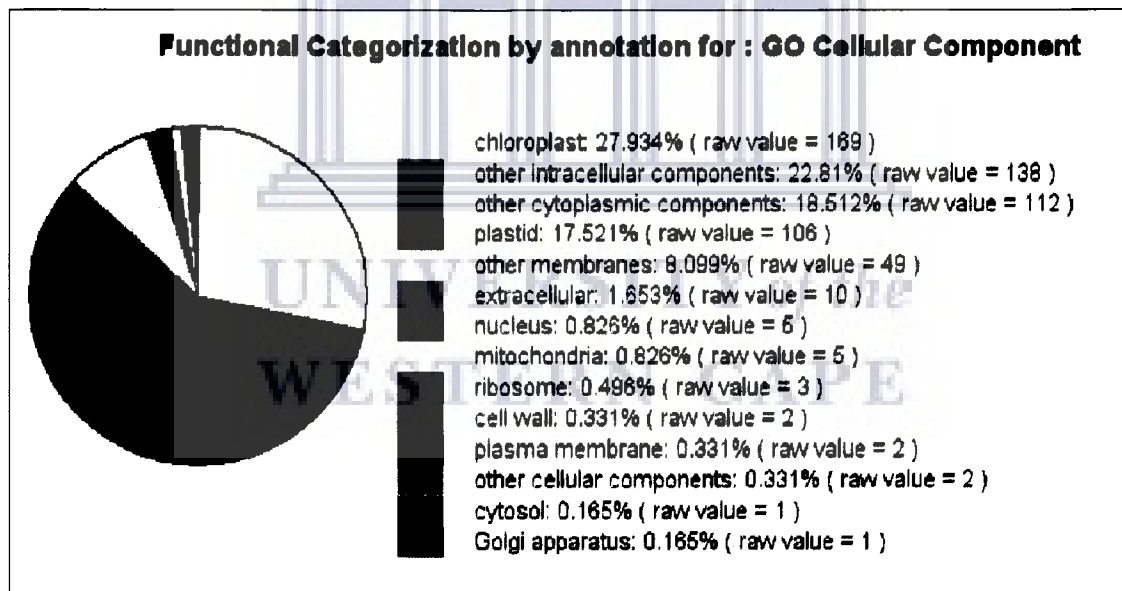


Figure 3.12: Pie chart representation of GO terms enriched for co-expressed genes in comparison to the entire *Arabidopsis thaliana* genome under the category cellular components.

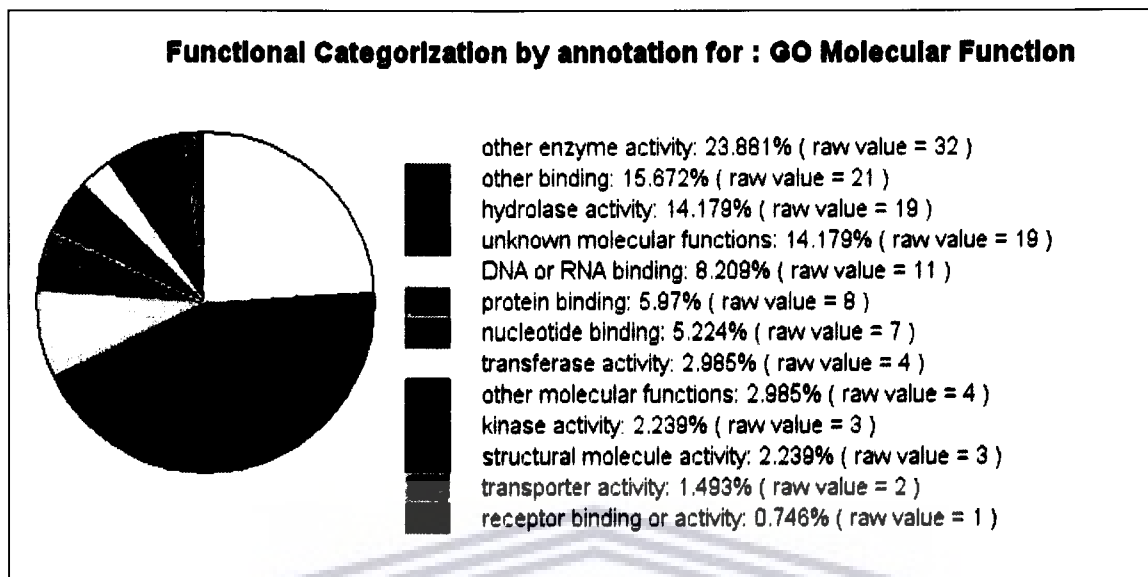


Figure 3.13: Pie chart representation of GO terms enriched for co-expressed genes in comparison to the entire *Arabidopsis thaliana* genome under the category Molecular function.

### 3.3.2 Identification of TFBMs present in promoters of carotenoid genes

Significance of transcription factor binding motifs were calculated using a hypo-geometric distribution and binding motifs with a  $p$ -value of  $<0.05$  were selected. ATHENA the online web tool for expression and network analyses allows for the visualisation of promoter regions in one of two ways, compact or cartoon displays. Both displays visualize transcription factor (TF) binding sites, transcription start sites and predicted CpG islands. The compact display provides a simple and intuitive view of the promoter sequences, while the cartoon display includes more detail about the fine structure of the promoter and does improve the illustration of displaying overlapping TF binding sites. For this section a compact visualization was used.

An enrichment analyses was done to determine which TFBMs were enriched within promoters of carotenoid genes in comparison to the entire *Arabidopsis thaliana* genome. TFBMs and

P-value	Motif name	# gs	#gg	P-value	Motif name	# gs	#gg
0.2345	ABEs binding site motif	2	2	0.1479	ABRE binding site motif	3	3
0.0537	ABRE-like binding site motif	10	12	0.0472	ACGTABREMOTIFA2QSEM	8	8
0.2923	AGATCONSENSUS	1	1	0.4468	AGCBOXNPGLB	1	1
0.3059	AGL2ATCONSENSUS	1	1	0.5418	ARF binding site motif	10	12
0.1213	ATHB1 binding site motif	2	4	0.6309	ATHB2 binding site motif	3	3
0.2247	ATHB5ATCORE	2	4	0.3045	ATHB6 binding site motif	2	2
0.6316	AtMYB2_BS_in_RD22	3	3	0.4797	AtMYC2_BS_in_RD22	10	13
0.8784	BoxII promoter motif	9	15	0.4346	CACGTGMOTIF	5	10
0.6063	CARGCWGAT	17	52	0.1295	CATG promoter motif	4	8
0.4900	CCA1 binding site motif	8	11	0.5545	DRE core motif	6	6
0.8561	DREB1A/CBE3	1	1	0.8753	EveningElement promoter moti	1	1
0.4519	GADOWNAT	3	3	0.0653	GAREAT	20	25
0.1040	GATA Motif	1	1	0.3494	GBF1/2/3_BS_in_ADH1	1	2
0.1723	GBOXLERBCS	2	2	0.8240	GCC-box promoter motif	1	1
0.3003	Gap-box Motif	4	4	0.7239	Hexamer promoter motif	2	2
0.7670	I-box promoter motif	9	11	0.9366	L1-box promoter motif	2	2
0.5008	LEAFYATAG	3	3	0.3963	LTRE promoter motif	2	2
0.3887	MYB binding site promoter	9	12	0.7917	MYB1 binding site motif	1	1
0.8918	MYB1AT	21	48	0.5026	MYB1LEPR	5	5
0.9036	MYB2AT	5	5	0.4302	MYB3 binding site motif	2	2
0.3605	MYB4 binding site motif	22	43	0.4797	MYCATERD1	10	13
0.3496	RAV1-B binding site motif	4	5	0.3647	SV40 core promoter motif	6	6
0.9876	T-box promoter motif	9	12	0.4595	TATA-box Motif	25	83
0.9404	TELO-box promoter motif	1	1	0.5896	TGA1 binding site motif	1	1
0.5896	UPRMOTIFAT	1	1	0.3731	W-box promoter motif	20	32
0.5160	Z-box promoter motif	1	1				

Figure 3.14: List of all the motifs identified in promoters of the 10 core carotenoid genes. Where column one contains the  $p$ -values associated with the significance of motifs. Column two represents the motif name. Column three and four represents the number of genes containing the motif where # gs is the amount of genes in the subset and # gg is the number of genes in the genome.

their frequency of occurrence within the promoter regions of the carotenoid biosynthetic genes are displayed in appendix A. The  $p$ -value associated with the prediction is also present within in the table. In this study twenty key TFBMs have been identified as being specific to promoters of carotenoid genes. These TFBMs have been shown to play a role in the regulation of the carotenoid biosynthetic pathway.

From the enrichment analyses a definite pattern emerges namely, the GBOXLERBCS, ATHB1, AGATCONSENSUS, GATA and GBF transcription factor binding motifs show a 3-fold increase in prevalence amongst carotenoid genes in comparison to the entire *Arabidopsis thaliana* genome.

Motifs such as ATHB6, ABF's, ATHB5ATCORE and ABRE show a 2 to 2.25 fold increase. The rest of the transcription factor binding motifs have an increase of 1.2 to 1.7 fold increase in comparison to the *A. thaliana* genome.

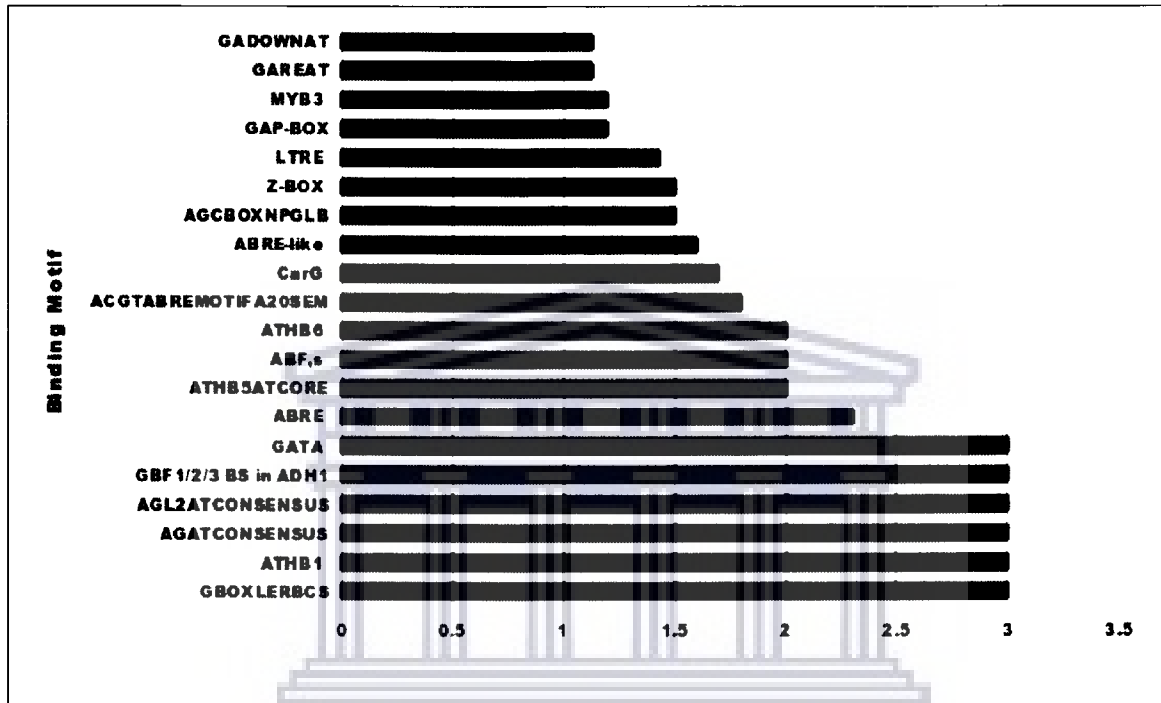


Figure 3.15: Enrichment analyses of motifs bound in promoters of core carotenoid genes. Transcription factor binding motifs for CBP genes relative to all the genes in *Arabidopsis thaliana* genome. On the x-axis is the motif name and on the y-axis is the value of enrichment

### 3.3.3 *de novo* Motif Discovery

MEME motifs are represented by position-specific probability matrices that specify the probability of each possible letter appearing at each possible position in an occurrence of the motif. These are displayed as "sequence LOGOS", containing stacks of letters at each position in the motif. The total height of the stack is the "information content" of that position in the motif in bits. The height of the individual letters in a stack is the probability of the letter at that position multi-



plied by the total information content of the stack. Ten motifs were identified as being significant amongst carotenoid genes as shown in Figure 3.16. The first column in the figure contains the motif identifier and the second column contains the *E*-value. The two motif logos present for each motif represent the sequence of the motif and the reverse complement of the motif.

Identified motifs were scanned against known databases to determine the similarity of the predicted motifs to known motifs that were previously identified. Column one illustrates the motif number, column two includes the motif logos that represents the specific motif and column three contains the genes that have been found to encompass these motifs.

Gene ids of the respective genes containing a specific motif are available in Figure 3.17 along with the motif logo, number of matches and the motif identifier. The name (or number) identifying the motif in the input files is present in column one. The number of term predictions is presented in column two and the top 5 specific GO annotation predictions for the motifs identified in MEME are displayed in the third column. The GO terms associated with the respective motifs are located in column three and are accompanied by the relevant functional category i.e. Biological process, molecular function and cellular component. Five of the six motifs that had GO terms attached to them contained the GO term "transcription factor activity". Common GO terms for the category "cellular components" include "chloroplast", "plasma membrane", "chloroplast stroma". For the category "biological processes" common GO terms include, "regulation of transcription", "positive regulation of transcription and translation" (Figure 3.18)

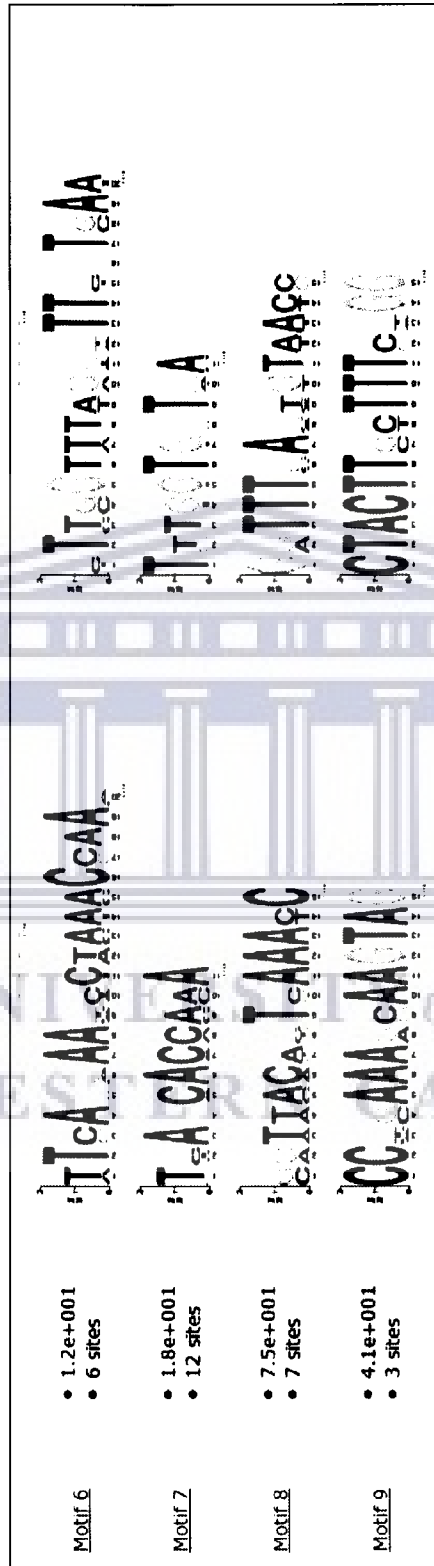


Figure 3.16: List of identified motifs shared amongst carotenoid genes, and not identified in carotenoid genes in any other plant TFBM database queried. Where column 1 contains the motif name, column 2 contains the *E*-value associated with the motif prediction, column 3 contains the motif logo and column 4 contains the logo for the reverse complement of the motif.

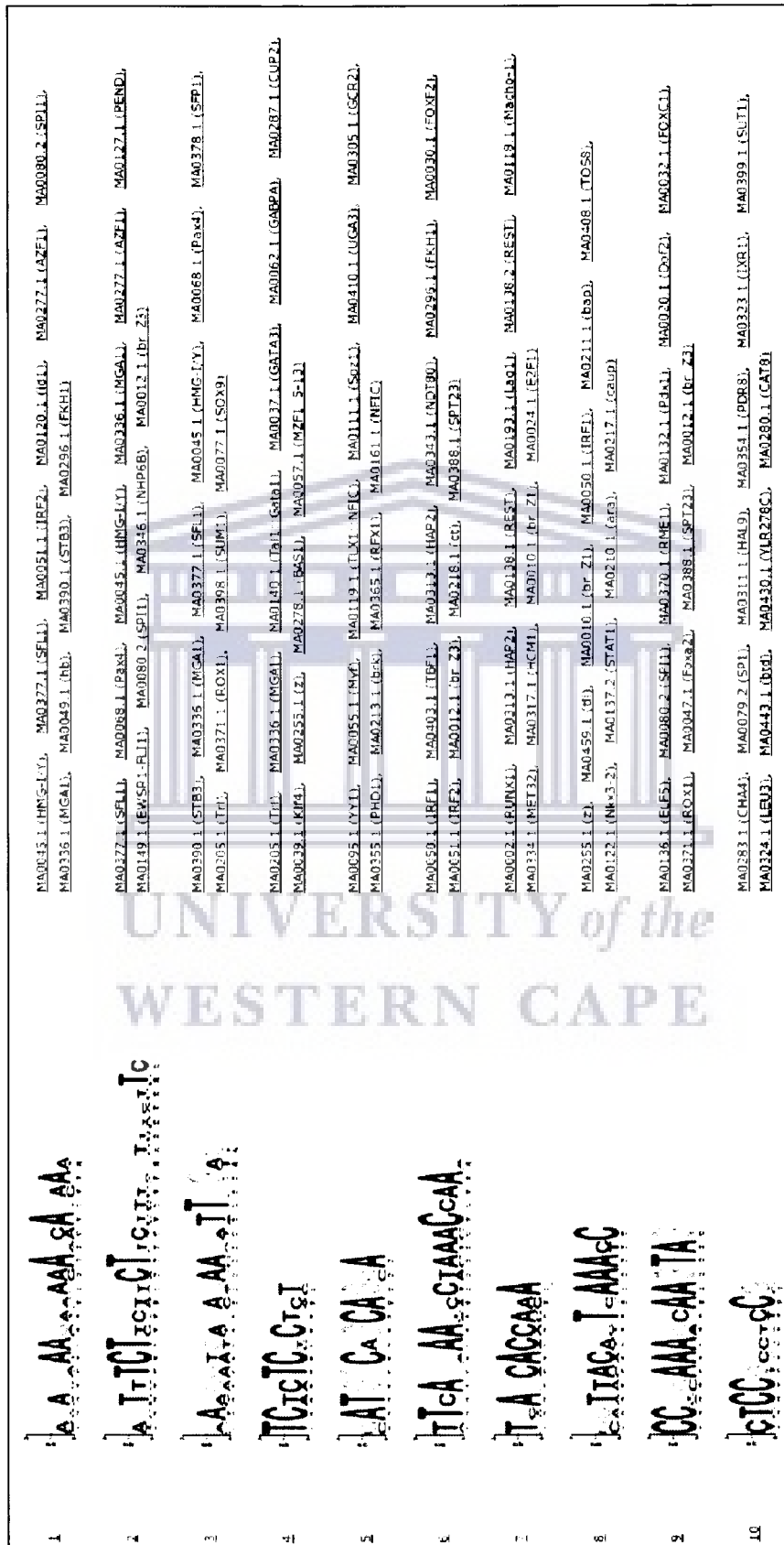


Figure 3.17: Overview of gene identifiers containing the respective motifs identified through MEME. Where column 1 contains the motif number, column 2 contains the motif logo and column 3 contains the gene ids of genes containing the motif.

Motif	Predictions	Top 5 specific predictions
1	30	MF transcription factor activity CC plasma membrane BP transmembrane receptor protein tyrosine kinase signaling pathway MF protein serine/threonine kinase activity BP regulation of transcription, DNA-dependent
10	45	CC mitochondrion CC chloroplast stroma MF structural constituent of ribosome MF RNA binding CC ribosome
2	38	MF transcription factor activity CC plasma membrane BP protein amino acid phosphorylation CC chloroplast BP transmembrane receptor protein tyrosine kinase signaling pathway
3	3	MF transcription factor activity BP regulation of transcription CC anchored to membrane
4	73	MF transcription factor activity BP regulation of transcription, DNA-dependent BP positive regulation of transcription MF transcription activator activity MF microtubule motor activity
5	23	BP translation MF structural constituent of ribosome CC chloroplast thylakoid membrane CC chloroplast stroma CC mitochondrion
6	0	
7	0	
8	0	
9	0	

Figure 3.18: Overview of gene identifiers containing the respective motifs identified through MEME. Where column 1 contains the motif name, column 2 contains the total number of GO terms predicted and column 3 contains the top 5 GO terms predicted.

# Chapter 4

## Discussion

### 4.1 Identification of putative conditions affecting carotenoid gene expression

The basic underlying mechanisms involved in the synthesis and accumulation of vital minerals may be understood by studying the complex biochemical pathways of a plant such as the carotenoid biosynthetic pathway. The synthesis of minerals such as pro-vitamin A can be aided by setting up a comprehensive knowledge base from biofortification studies. This knowledge base can be useful in the process of increasing valuable nutrients in crops and in turn will be of great value to the developing world. This is especially true for countries in Africa as their staple diets are what they solely rely on for their nutrient intake. The identified genes of interest and the information gained can be used to make informed decisions when aiming to biofortify crops. The new found wealth of knowledge will also lead to the identification of tissues specific to development of plants and will maximise the production of valuable nutrients such as pro-vitamin A under harsh environmental conditions.

A list of 32 known carotenoid genes were identified as being genes involved in the carotenoid biosynthetic pathway (CBP). These genes were identified from literature and only genes that were experimentally validated as being involved in CBP and the synthesis of carotenoids were taken into consideration.

The heat maps represent the expression profiles of all the known carotenoid genes. Differential expression of these genes are presented under the treatment of various abiotic and biotic stresses. These stresses include, but are not limited to, cold, drought, heat, osmotic, oxidative, salt and wounding. Most of the carotenoid genes are down regulated under the influence of the above mentioned stresses. This indicates that carotenoid biosynthesis is switched off at the onset of stress and therefore if these genes can be altered to become resistant to a particular stress then a definite increase in carotenoid content may be achieved. From the heat maps, it is clear that genes that have similar differential expression profile are clustered together. This was evident under all the above mentioned stress factors. Genes such as LCY $\epsilon$  and ZDS are clustered together share similar expression profiles with core carotenoid genes under the selected environmental stresses.

An important point to note is that at the onset of stress, PSY, the driver gene of the carotenoid biosynthetic pathway was completely switched off. It was neither up nor down regulated and this indicated that the carotenoid biosynthesis is halted at the onset of stress. NCED3, a carotenoid gene as well as a gene implicated in drought tolerance, is highly up regulated across stimuli. This suggests a strong link between neighbouring pathways and the carotenoid biosynthetic pathway. A recent study by Meier *et al.*, (2011) showed strong links between the carotenoid pathway, ABA pathway, phytochrome pathway and the photosynthetic pathway, when looking at the co-expression and promoter content of carotenoid genes and their co-expressed genes.

CRTISO was found to be highly expressed in shoot tissue in comparison to root tissue by Cazzonelli *et al.* (2010). This gene was identified as being a giberellin associated responsive element as well as a light responsive element. By being a light responsive element many processes and interactions come into play. For example, CRTISO is unique to carotenoid biosynthesis and is responsible for the desaturation of 15-cis phytoene to all trans-lycopene (Lee *et al.*, 2007). Its expression is also seen as tissue specific as it is more highly expressed in shoots than in roots according to Cazzonelli *et al.* (2010), and this is evident in our study as well. CRTISO is more highly expressed in tissue that is rapidly dividing and growing i.e. tissue such as shoots. This gene is also highly expressed in photosynthetic and floral male reproductive tissues. As a result of shoots being a photosynthetic tissue, it is therefore, more likely that expression levels of CRTISO are found to be higher in shoots than in roots throughout our study. This suggests that CRTISO is therefore essential for defining plant development and architecture.

Gene expression profiling was carried out to see the effect of various stimuli on carotenoid biosynthetic pathway genes. The initial aim of the gene expression profiling experiment was to determine which of the stimuli were most affecting carotenoid biosynthetic genes. A method of ranking these genes was determined, that is using fold change. The calculated fold change expression values were plotted in order to determine a ranking order for abiotic stress factors affecting carotenoid biosynthesis.

From these analyses, it is clear that certain stimuli have a larger effect on carotenoid genes than others. Drought and cold stress exposure for a period of 24hrs shows a continuous up regulation of core carotenoid genes in both root and shoot tissue. In the shoot tissue, a increase of 4-fold was observed for drought stress, while for roots a 3 fold expression was observed. Similar

patterns were seen in cold stress, however it is important to note that a 9 and 7-fold increase was seen in shoots and roots respectively which is much higher than the fold change expression for drought stress. These findings are concurrent with those found by Kilian *et al.* (2007). It is however important to note that Kilian and others investigated the overall gene expression for arabidopsis and did not investigate carotenoid biosynthesis directly.

An increase in expression was also observed amongst the other abiotic stress factors heat, osmotic, oxidative, salt and wounding, although it was not to the degree that was observed for cold and drought stress. Under heat, oxidative and wounding stress the expression levels observed in the roots was nearly insignificant, indicating an almost identical expression pattern to that of the control samples. This finding is different to the finding by Kilian and others (2007) as they found a significant increase in expression for all stimuli including wounding. Here we show that wounding does not have significant effect on carotenoid biosynthetic gene expression, suggesting that wounding does not significantly affect the expression of carotenoid genes. Meier *et al.* (2011) showed a significant effect in genes exposed to osmotic stress, whereas our finding indicated otherwise. This could be due to the fact that I looked at only the core carotenoid genes and a global picture was not established in this study.

In general, gene expression induced as a result of abiotic stress occurs at a rapid pace. The first changes in expression was observed only 0.5hrs after stress was applied across all conditions. These immediate changes were seen in both roots and shoot tissue. This observation leads to the idea that an immediate systemic signal from the roots and shoots is transferred to the rest of the plant organs.

The overall gene expression varies considerably when looking at all the stimuli. A striking



example is the elevated levels of gene expression under the continuous application of drought and salt stress for 24hrs. It is clear that the plant is capable of coping with high salt conditions and dehydration for a long period of time. These findings are indicative that the plant is able to identify ionic stressors and responds accordingly at a global gene expression level.

## 4.2 Co-expression and correlation analyses of carotenoid genes and their co-expressed genes

### 4.2.1 Co-expression analyses

Three online web tools namely, STRING, ATTED-II and ACT were used to identify genes that were co-expressed with known carotenoid genes. The core carotenoid genes AT5G17230 (PSY), AT4G14210 (PDS), AT1G06820 (CRTISO), AT1G10230 (LCY $\beta$ ), AT4G25700 ( $\beta$ OHase1), AT5G52570 ( $\beta$ OHase2), AT1G31800 (LUT5), AT3G53130 (LUT1), AT3G04870 (ZDS) and AT5G57030 (LCY $\epsilon$ ) were used as driver genes in each of the instances. Eighty six co-expressed genes were identified based on their correlation  $r$ -values and protein protein interactions. From the three gene lists obtained from each of the online tools, it is clear that for all the core carotenoid genes, between 50-300 genes were shared between the lists derived from ACT and ATTED-II. The list from STRING however, does not share any genes with either of the other two lists (ACT and ATTED-II). This finding is the complete opposite of what we had hoped for.

The reason for this can be explained in three parts. Firstly, STRING (version 8) is based primarily on protein-protein interaction between genes. Secondly, STRING takes the protein-protein

interactions between core carotenoid genes and their co-expressed genes into consideration rather than the correlation coefficient, which ACT and ATTED-II uses. Thirdly, after further investigation it was found that this version of STRING used did not contain any co-expression data for the carotenoid biosynthetic pathway genes. Therefore, the co-expressed gene lists retrieved for the core carotenoid genes were based on the involvement in secondary pathways that are more effectively characterized in comparison to the carotenoid pathway. The identified lists were compared to determine which genes were co-expressed in common to all of the core carotenoid genes. A list of 86 genes were identified, common to the gene lists. This list formed the catalogue of co-expressed genes for core carotenoid biosynthetic pathway genes.

#### 4.2.2 Correlation analyses

PSY was used as a driver gene in expression correlation analyses against the entire *Arabidopsis thaliana* genome for expression (transcript levels) averaged across 400 stimuli conditions and developmental stages making our study more robust than Meier *et al.*, (2011) which only focused on 320 stimuli. LCY $\beta$  was identified as being the top ranked correlated carotenoid pathway gene. The correlation  $r$ -values for the entire *Arabidopsis* genome were plotted against PSY and LCY $\beta$  as shown in Figure 3.9. A group of 86 co-expressed genes shown in this figure illustrates a strong positive correlation between PSY, the driver gene of the carotenoid biosynthetic pathway and co-expressed genes indicated by red dots. These results suggest the presence of global transcriptional regulators that control co-expression. A similar study was carried out by Meier *et al.*, (2011), where they observed a list of 50 genes co-expressed with only PSY the driver gene of the carotenoid biosynthetic pathway correlation analyses. The rest of the core carotenoid genes

were not tested in their study.



UNIVERSITY *of the*  
WESTERN CAPE

## 4.3 Promoter content analyses, functional annotation and *de novo* motif discovery

This section is divided into three subsections. Subsection one and two will cover promoter content and functional enrichment analyses where the 32 known carotenoid gene ids were used as the input for ATHENA and TAIR. Subsection three focusses on *de novo* motif discovery and uses the sequences of the promoter proximal regions as input for MEME suite.

### 4.3.1 Enrichment analyses

The promoter visualisation tool allowed for identification of the specific location of predicted transcription factors. The significance of each of the transcription factors was calculated using a hyper-geometric distribution and this selected only those transcription factors which were enriched within the test dataset to be identified. Among the identified TFBMs 20 were significantly (ATHENA  $p$ -value  $<0.05$ ) enriched amongst the carotenoid biosynthetic pathway genes. These factors were thus seen as candidate regulatory elements that may co-ordinate carotenoid biosynthesis

The most prevalent transcription factors present in promoters of carotenoid genes are involved in light (GAPBOX, GBOX, ABRE, GATA, ABF's) ABA (ABRE-like, ACGT ABRE Motif A20SEM, Z-BOX, ABF's) and were found to be development responsive. Other factors include GA (GAREAT and GADONAT) cold, hypoxia and hormone response and are in agreement with the heat maps. The transcription factors identified through the enrichment analyses identified specific stimuli that were involved in carotenoid biosynthesis. Light stress is one of

the stimuli that was tagged as having a significant effect on carotenoid biosynthesis. Light is an important factor with regard to photosynthesis and it is one of the key elements for this process as it aids in the production of various end products within the plant. The G-box is the most highly enriched TFBS within the carotenoid subset and found upstream of all light induced carotenoid genes, and is specifically induced under the stimulation of light stress. This finding is consistent with findings from Meier *et al.*, (2011), where a GBOX was identified as being enriched amongst the promoters of co-expressed genes. These findings indicate that the G-box is key to the regulation of both carotenoid genes as well as their co-expressed genes, however, it is important to note that the GBOX is not the only motif with the capability of regulating carotenoid biosynthesis.

### 4.3.2 GO annotation and functional enrichment

GO annotation and functional enrichment revealed that both carotenoid genes and co-expressed genes share commonly enriched terms throughout the three categories namely biological processes, molecular function and cellular components. Terms such as response to abiotic or biotic stimulus and development processes suggest that these genes may share an underlying regulatory mechanism as genes in both carotenoid and co-expressed groups have similar GO terms enriched. This finding shows that carotenoid genes and their co-expressed genes are tightly co-expressed and validates our previous finding from correlation analyses above. Under the category of cellular components, GO annotations such as chloroplasts, cytosol and plastids were most prevalent. This finding is to be expected as carotenoid biosynthesis is localised to the regions of the plastid and cytosol. Plastid and cytosolic activity is of high importance for the biosynthesis of carotenoids.

These results also suggest that genes involved in the carotenoid biosynthetic pathway are chlorophyll dependent.

### 4.3.3 *de novo* motif discovery

A total of 10 motifs were identified amongst the promoter proximal regions of the carotenoid genes. The associated *E*-value for each motif is statistically significant ( $E$ -value < 1) indicating that these motif predictions are highly accurate. These predicted motifs were aligned against known motifs from various databases and were found to be structurally similar with an offset of 2-7bp. To further determine if our motifs were valid predictions and to identify *de novo* predictions we scanned the motifs against the plant database in GOMO of the MEME suite to identify GO terms associated with each motif. Six of the 10 motifs namely 1, 10, 2, 3, 4 and 5 had GO annotations such as transcription factor activity, transcriptional regulation and chloroplasts attached to them, suggesting that the motifs identified were involved in similar regulatory processes and that they share biological function. The rest of the significant GO terms associated with these six motifs are present in Figure 3.18. The 4 remaining motifs had no significant GO term attached to them after being scanned against the plant database in GOMO. This suggests that these predictions might be *de novo* motifs. The 4 remaining motifs were also scanned against the JASPAR core database of known *cis* elements and were found to be structurally similar but not identical with an offset of 2-7bp to known motifs namely, IRF1, IRF2, TBF1, HAP2, NDT80, FKH1, SPT23, FOXF2, brZ3 and ct. This finding further suggests that these 4 motifs are novel (Tanaka *et al.*, 1993; Pozner *et al.*, 2000; Noh *et al.*, 2004; Lin *et al.*, 2005; Hollingsworth, 2008; Cui *et al.*, 2009). The known motifs mentioned above are all involved in the transcription process

# Chapter 5

## Conclusion and future work

### 5.1 Research contributions and limitations

Throughout this thesis a number of topics were discussed and researched in chapter 1. A comprehensive compilation of literature was done on the area of carotenoid research. Chapter two covered the analyses of microarray data with regards to carotenoid gene expression to identify environmental stimuli and genes that were key to carotenoid gene expression and carotenoid biosynthesis. Chapter three covered an encompassing co-expression analysis and co-correlation analyses to identify genes that were co-expressed and correlated with carotenoid genes. Chapter 4 covered a promoter content analyses of carotenoid genes as well as their co-expressed genes to identify putative elements affecting carotenoid gene expression.

#### 5.1.1 Expression profiling

Thirty two genes have been identified as being involved in the carotenoid biosynthetic pathway. Some of the genes involved in the carotenoid biosynthetic pathway are also involved in three

other pathways namely, MEP, GGPP and mevalonate pathway. Genes from these pathways have been implicated in the carotenoid biosynthetic process.

Carotenoid genes are strongly differentially expressed depending on the stress type. NCED3 was the most strongly stress-inducible gene in almost all stress conditions except heat and wounding treatments. NCED3 expression contributes to ABA biosynthesis and stress tolerance. Other important up-regulated carotenoid genes include  $\beta$ Oase2 and ZEP. PSY, LUT5, VDE, LCY $\epsilon$  and ISPE are mostly down-regulated genes during stress. All of the conditions except wounding stress induces a significant influence on the carotenoid biosynthetic pathway genes.

Gene expression induced by abiotic stress occurs at a rapid pace. The first changes in the gene expression were seen on 0.5hrs after the application of stress. Elevated expression levels were prevalent in both root and shoot tissue. These findings are indicative that an immediate systemic signal is sent to the rest of the plant in the form of stress. These signals thus allow the plant to respond accordingly at a global gene expression level.

### **5.1.2 Correlation and co-expression analyses**

Genes co-expressed with core carotenoid genes were identified from three publicly available databases. A list of eighty six genes were identified to be co-expressed with carotenoid genes.

A co-correlation between two core carotenoid genes and the list of 86 co-expressed genes reveals a tightly positive correlation to PSY the driver gene of the carotenoid biosynthetic pathway. These results suggest that core carotenoid genes and their co-expressed genes share common regulatory elements.



### 5.1.3 Promoter content analyses and functional enrichment

Promoter content analyses of carotenoid genes and their co-expressed genes allowed for an array of transcription factor binding motifs (TFBM) prevalent in promoters of carotenoid genes to be identified. The most enriched TFBMs found in the promoter regions of the carotenoid biosynthetic pathway genes show a 1.25-3 fold increase in prevalence with a  $p$ -value of  $< 0.05$ . These TFBMs are involved in a number of responses. For example the motifs GAPBOX, GBOX, ABRE, GATA and ABF's are involved in light responsiveness, ABRE-like, ACGT ABRE Motif A20SEM, Z-BOX, ABF's are involved in the ABA process. Other important factors include GA (GAREAT and GADOWNAT), cold, hypoxia and hormone treatments, which is in agreement with the heat map results.

GO annotation was done to determine if there were any underlying functions that were shared between carotenoid genes and their co-expressed genes. Similar GO terms were found to be enriched for carotenoid biosynthetic pathway genes and their co-expressed genes, using the *Arabidopsis thaliana* genome as a background. These findings are consistent across all three categories namely, biological processes, molecular function and cellular components. These results indicate that carotenoid biosynthetic pathway genes and their co-expressed genes are involved in similar functions. These findings are also in accordance with the enrichment analyses.

#### 5.1.4 *de novo* motif discovery

Meme identified 10 motifs present in the UTRs of carotenoid genes. The associated  $E$ -value for each motif is very low  $E < 1$  indicating that these motif predictions are highly accurate. Six

of the 10 motifs had GO annotations such as "transcription factor activity", "transcriptional regulation" and "chloroplasts" attached to them, suggesting that the motifs identified were involved in similar regulatory processes and that they share biological function. The 4 remaining motifs had no significant GO term attached to them after being searched against the plant database in GOMO. This suggests that these predictions might be novel motifs. The known motifs are involved in transcription and because our predicted motifs are structurally similar it could suggest that our motifs are implicated in the same biological processes. A further suggestion could be made that these 4 predicted motifs are specific to carotenoid genes and are directly involved in the regulation of carotenoid biosynthesis.

## 5.2 Future work

Future prospects of this project encompasses many areas some of which include: experimentally validating the finding from the *in-silico* studies, in particular detecting mutants of the 4 novel motifs. Although much work has been covered within the realms of this study a lot still needs to be done to fully understand the mechanism by which genes are regulated within the carotenoid biosynthetic pathway. Furthermore promoter motif analyses of the 86 co-expressed genes will lead to a more comprehensive understanding of the underlying mechanisms involved in controlling the CBP and will allow for the further characterisation of the carotenoid biosynthetic pathway. Another aspect of expanding this project would include biofortification of staple crops (maize and sorghum) within Africa, in order to increase carotenoid content. Biofortification along with advances in plant science such as biotechnology can provide novel traits for breeding that are not currently available. It is for this reason that biofortification can be directly attributed

to an expansion in nutrigenomics and thus exemplifies the power and potential benefits of this process. By doing this, production and conversion of vitamins will increase and therefore the global problem of vitamin deficiencies such as VAD may be addressed.



# Appendix A

## Supplementary material for Chapter 2

### Chapter Perl scripts

#### col.pl

```
#!/usr/bin/perl
#####
#Author: Firdous Khan
#Christoffels lab
#South African National Bioinformatics Institute
#University van wes-kaapland
#####
#
#Generates a tab delimited file of microarray data signals and gene name
#####

use strict;
use warnings;

#check if a file input exit for sanity purposes
if (@ARGV !=1)
{
    print STDERR "perl column.pl <filename>\n";
    exit(1);
}

#Assigns file to a variable
my $infile = $ARGV[0];

#Associates a file handle to file and opens file
open (IF, $infile) || die "cannot open $infile:$!";

#Takes each line in file as and array
my @line = <IF>;

open(OF, ">signal.csv");
```

```

#Closes files so that it can be accessed by other program
close IF;

#loops over each line in file
foreach my $line (@line)
{
#Removes the new line character at the end of each line
chomp $line;
#Splits each line into an array using tab meta-character
my @col = split(/\t/, $line);
#print the following lines by the indices
print OF $col[0]."\t".$col[1]."\t".$col[14]."\t"
        . $col[18]."\t".$col[22]."\t".$col[26]."\t".$col[30]."\t"
        . $col[34]."\t".$col[38]."\t".$col[42]."\t".$col[46]."\t"
        . $col[50]."\t".$col[54]."\t".$col[58]."\t".$col[62]
        . "\t".$col[66]."\t".$col[70]."\t".$col[74]."\t".$col[78]
        . "\t".$col[82]."\t".$col[86]."\t".$col[90]."\t"
        . $col[94]."\t".$col[98]."\t".$col[102]."\t".$col[106]."\t"."\\n";
}

close OF;

```

**ave.pl**

```

#!/usr/bin/perl
#####
#
#Author: Firdous Khan
#Christoffels lab
#South African National Bioinformatics Institute
#University van wes-kaapland
#####
#
#Generates a tab delimited file of microarray data signals and gene name
#####
use strict;
use warnings;

#check if a file input exit for sanity purposes
if (@ARGV !=1)
{
    print STDERR "perl ave.pl <filename>\n";
    exit(1);
}

#Assigns file to a variable
my $infile = $ARGV[0];

#Associates a file handle to file and opens file
open (IF, $infile) || die "cannot open $infile:$!";
open(OF, ">signal.csv");

#loops over each line in file
while (<IF>) {
my $line = $_;
#print "...$line....\n";
#Removes the new line character at the end of each line

```

```

chomp $line;
#Splits each line into an array using tab meta-character
my @col = split(/\t/, $line);
next unless ($col[2] =~ /AT/);
#print the following lines by the indices

#prints gene name
my $gname = ($col[2]);

#ave replicates for shoots 0.5h
my $ave_sh_0_5 = ($col[14]+$col[18])/2;
my $pval_sh_0_5 = ($col[16]+$col[20])/2;
#ave replicates for roots 0.5h
my $ave_rt_0_5 = ($col[22]+$col[26])/2;
my $pval_rt_0_5 = ($col[24]+$col[28])/2;

#ave replicates for shoots 1h
my $ave_sh_1 = ($col[30]+$col[34])/2;
my $pval_sh_1 = ($col[32]+$col[36])/2;
#ave replicates for roots 1h
my $ave_rt_1 = ($col[38]+$col[42])/2;
my $pval_rt_1 = ($col[40]+$col[44])/2;

#ave replicates for shoots 3h
my $ave_sh_3 = ($col[46]+$col[50])/2;
my $pval_sh_3 = ($col[48]+$col[52])/2;
#ave replicates for roots 3h
my $ave_rt_3 = ($col[54]+$col[58])/2;
my $pval_rt_3 = ($col[56]+$col[60])/2;

#ave replicates for shoots 6h
my $ave_sh_6 = ($col[62]+$col[66])/2;
my $pval_sh_6 = ($col[64]+$col[68])/2;
#ave replicates for roots 6h
my $ave_rt_6 = ($col[70]+$col[74])/2;
my $pval_rt_6 = ($col[72]+$col[76])/2;

#ave replicates for shoots 12h
my $ave_sh_12 = ($col[78]+$col[82])/2;
my $pval_sh_12 = ($col[80]+$col[84])/2;
#ave replicates for roots 12h
my $ave_rt_12 = ($col[86]+$col[90])/2;
my $pval_rt_12 = ($col[88]+$col[92])/2;

#ave replicates for shoots 24h
my $ave_sh_24 = ($col[94]+$col[98])/2;
my $pval_sh_24 = ($col[96]+$col[100])/2;
#ave replicates for roots 24h
my $ave_rt_24 = ($col[102]+$col[106])/2;
my $pval_rt_24 = ($col[104]+$col[108])/2;

print OF $gname."\t". $ave_sh_0_5."\t".$pval_sh_0_5."\t".

```

```

    save_rt_0_5."\t". $pval_rt_0_5."\t". $save_sh_1."\t"
    . $pval_sh_1."\t". $save_rt_1."\t". $pval_rt_1."\t".
    $save_sh_3."\t". $pval_sh_3."\t". $save_rt_3."\t"
    . $pval_rt_3."\t". $save_sh_6."\t". $pval_sh_6."\t".
    $save_rt_6."\t". $pval_rt_6."\t". $save_sh_12."\t"
    . $pval_sh_12."\t". $save_rt_12."\t". $pval_rt_12."\t"
    . $save_sh_24."\t". $pval_sh_24."\t"
    . $save_rt_24."\t". $pval_rt_24."\n";
}
close OF;

```

### extract.pl

```

#!/usr/bin/perl
#####
#
#Author: Firdous Khan
#Christoffels lab
#South African National Bioinformatics Institute
#University van wes-kaapland
#####
#
# Generates a tab delimited file of mean signals in roots and shoots
# for coex genes
#####

##reading in file and removing the ".
##also sending it to a new file called coex id
cat new_genes.csv | sed 's/"//g' > coex.id.txt

### Reads in the coex.id.txt file and searches the carvl db in cold
###table for all ids in the coex.id.txt file
./mysqlfetch.sh coex.id.txt cold carvl > coex_cold.txt

###takes the output file and only select lines starting with "AT" and
###places it in a new file called coex_cold.2
grep 'AT' coex_cold.txt > coex_cold.2.txt

###the same can be done for all stimuli

```

## R script

### Fold\_change.r

```

#####
#Author: Firdous Khan
#Christoffels lab
#South African National Bioinformatics Institute
#University van wes-kaapland
#####

#####
#
# Plots the Fold change bar graphs
#

```

```

#
#####

##converting csv-txt and removing unnecessary delimiters

sed 's/"//g' firdous_foldchange_cold_shoots.csv > firdous_foldchange_cold_shoots.txt
rm firdous_cold_time_shoots.csv

library(lattice)

?getwd

getwd()

input=read.table(file=file.choose(), header=T, sep="\t")

/home/firdous/firdous_R_plots_work/firdous_foldchange_cold_shoots.txt

attach(input)

barchart(Gene.Name ~ FC | Time, data=input, col = "blue",
         main="Expression of Core Carotenoid Genes under Cold
         Stress at different time points", ylab="Carotenoid Gene",
         xlab="FC")

barchart(FC ~ Gene.Name | Time, data=input, col="red", horizontal=FALSE,
         scales=list(x=list(rot=90)), main="Expression fold change of Core
         Carotenoid Genes in shoots at different time points under Cold stress
         ", xlab="Carotenoid Gene", ylab="FC")

barchart(FC ~ Gene.Name | Time, data=input, col="blue",
         horizontal=FALSE, scales=list(x=list(rot=90)), main="Expression fold
         change of Core Carotenoid Genes in roots at different time points
         under Cold stress ", xlab="Carotenoid Gene", ylab="FC")

barchart(Gene.Name ~ FC | Time, data=input, col="purple", layout
         =c(6,1), main="Expression of Core Carotenoid Genes under Cold Stress
         at Different Time points", ylab="Gene", xlab="FC")

barchart(FC ~ Gene.Name | Time, data=input, col="green",
         horizontal=FALSE, scales=list(x=list(rot=90)), layout=c(1,6),
         main="Expression fold change of Core Carotenoid Genes under Cold
         Stress at different Time points in shoots", xlab="Carotenoid Gene", ylab="FC")

barchart(Gene.Name ~ Expression | Time, data=input, col="purple",
         layout =c(6,1), main="Expression of Core Carotenoid Genes under Cold
         Stress at Different Time points", ylab="Gene", xlab="FC")

rm(input)

#####
#
#Author: Firdous Khan
#Christoffels lab
#South African National Bioinformatics Institute
#University van wes-kaapland
#####

```



```
#####
#
#           Plots the Fold change bar graphs           #
#
#####

##converting csv-txt and removing unnecessary delimiters

####Shoots
sed 's/"//g' firdous_foldchange_cold_shoots.csv > firdous_foldchange_cold_shoots.txt
sed 's/,\\t/g' firdous_foldchange_cold_shoots.txt > firdous_foldchange_cold_shoots

####Roots
sed 's/"//g' firdous_foldchange_cold_roots.csv > firdous_foldchange_cold_roots.txt
sed 's/,\\t/g' firdous_foldchange_cold_roots.txt > firdous_foldchange_cold_roots

##Initialize R
R
library(lattice)

### Make sure you are in the directory where the files are ##
input=read.table(file=file.choose(), header=T, sep="\\t")

##file name shoots
firdous_foldchange_cold_shoots

###file name roots
firdous_foldchange_cold_roots

/home/firdous/firdous_R_plots_work/
attach(input)

### Plotting the graph for shoots
barchart(FC ~ Gene.Name | Time, data=input, col="red",
horizontal=FALSE, scales=list(x=list(rot=90)), main="Expression fold
change of Core Carotenoid Genes in shoots at different time points
under Cold stress ", xlab="Carotenoid Gene", ylab="Fold Change")

### Plotting the graph for roots
```

```
barchart(FC ~ Gene.Name | Time, data=input, col="blue",  
horizontal=FALSE, scales=list(x=list(rot=90)), main="Expression fold  
change of Core Carotenoid Genes in roots at different time points  
under Cold stress ", xlab="Carotenoid Gene", ylab="Fold Change")
```

```
rm(input)
```

## promoter content analyses



Table A.1: Predicted TFBMs present amongst the carotenoid gene list in comparison to the entire *A. thaliana* genome. Abbreviations used in the table are as follows %PBS-percentage of promoters bound in the subset, # of GS-number of genes present in subset with bound promoters, %PBG- percentage of promoters bound in the genome, #GG-number of genes in genome with bound promoter.

Motif Name	% PBS	# of GS	% PBG	# GG	p-value
MYB4 binding site motif	79%	23	73%	21988	0.302
TATA-box Motif	75%	22	80%	24133	0.8
GAREAT	65%	19	59%	17796	0.31
W-box promoter motif	65%	19	68%	20552	0.707
MYB1AT	65%	19	79%	24041	0.98
CARGCW8GAT	58%	17	64%	19288	0.794
Ibox promoter motif	44%	13	43%	13187	0.529
BoxII promoter motif	41%	12	47%	14399	0.812
ABRE-like binding site motif	41%	12	24%	7326	0.032
AtMYC2 BS in RD22	34%	10	40%	12278	0.811
ACGTABREMOTIFA2OSEM	34%	10	17%	5199	0.019
MYCATERD1	34%	10	40%	12278	0.811
ARF binding site motif	34%	10	41%	12347	0.817
MYB2AT	31%	9	33%	9978	0.664
CCA1 binding site motif	31%	9	32%	9777	0.635
MYB binding site promoter	31%	9	33%	10099	0.68
DRE core motif	31%	9	25%	7742	0.32
T-box promoter motif	27%	8	57%	17350	0.999
SV40 core promoter motif	24%	7	22%	6784	0.491
CACGTGMOTIF	20%	6	18%	5513	0.444
ATHB2 binding site motif	17%	5	15%	4572	0.457
GADOWNAT	17%	5	10%	3207	0.191
MYB1LEPR	13%	4	21%	6397	0.892
ABRE binding site motif	13%	4	5%	1665	0.073
L1-box promoter motif	13%	4	20%	6023	0.86
RAV1-B binding site motif	13%	4	13%	4063	0.563
Hexamer promoter motif	13%	4	10%	3277	0.39
CArG promoter motif	10%	3	9%	2712	0.492
AtMYB2 BS in RD22	10%	3	14%	4462	0.825
LTRE promoter motif	10%	3	6%	1894	0.274
CBOXLERBCS	10%	3	3%	926	0.058
LEAFYATAG	10%	3	12%	3727	0.715
Gap-box Motif	10%	3	13%	3985	0.758
ABFs binding site motif	10%	3	3%	1152	0.098
ABREATRD22	6%	2	2%	895	0.213
GBF1/2/3 BS in ADH1	6%	2	1%	551	0.098
MYB3 binding site motif	6%	2	6%	2086	0.607
DREB1A/CBF3	6%	2	8%	2619	0.732
TELO-box promoter motif	3%	1	11%	3379	0.968
ATHB5ATCORE	3%	1	4%	1289	0.719
GCC-box promoter motif	3%	1	7%	2224	0.892
AGATCONSENSUS	3%	1	1%	485	0.376
TGA1 binding site motif	3%	1	3%	1117	0.666
ACE promoter motif	3%	1	0%	44	0.041
ATHB1 binding site motif	3%	1	2%	885	0.579
AGCBOXNPGLB	3%	1	2%	758	0.523
MYB1 binding site motif	3%	1	7%	2197	0.889
Z-box promoter motif	3%	1	3%	1047	0.642
RY-repeat promoter motif	3%	1	4%	1474	0.767
ATHB6 binding site motif	3%	1	5%	1622	0.799
UPRMOTIFIAT	3%	1	3%	1117	0.666

UTR.pl

```
#!/usr/bin/perl #
##### #
#Author: Firdous Khan #
#Christoffels lab #
#South African National Bioinformatics Institute #
#University van wes-kaapland #
```

```
#####
#
#Generates a tab delimited file of microarray data signals and gene name #
#####

use strict;
use warnings;

open (IN, "$ARGV[0]");
open (OUT, ">", "$ARGV[0]_out");

LINE:
while (<IN>) {
    my $line = $_;
    chomp ($line);

    if ($line =~ /^(\>)/) {
        print OUT $line.'_5UTR'."\n";
        next LINE;
    }

    else {
        $line =~ s/^(.{2000})//;
        print OUT $line."\n";
    }
}

```

### mean\_1.pl

```
#!/usr/bin/perl #####
#####
#
#Author: Firdous Khan #####
#Christoffels lab #####
#South African National Bioinformatics Institute #####
#University van wes-kaapland #####
#####
#
#Generates a tab delimited file of mean signals and spot id and probename #
#####
use strict;
use warnings;

#Absolute path of infile2
my $infile2 = "/home/user/signal.csv";

open (IF2, $infile2) || die "cannot open $infile2:$!";
open(OF2, ">mean.csv");
my @line2 = <IF2>;
foreach my $line2(@line2)
{

```

```
#Removes the new line character at the end of each line
chomp $line2;
#Splits each line into an array using tab meta-character
my @col2 = split(/\t/, $line2);
my $spotid = shift@col2;
my $probename = shift@col2;
my $len = scalar@col2;
my $total = 0;
while (my $num = shift@col2)
{
$total += $num ;
}
my $mean = $total/$len;
print OF2 $spotid."\t".$probename."\t".$mean."\n";
}

close OF2
```



UNIVERSITY *of the*  
WESTERN CAPE

# Appendix B

## Supplementary material for Chapter 3

Fold change graphs



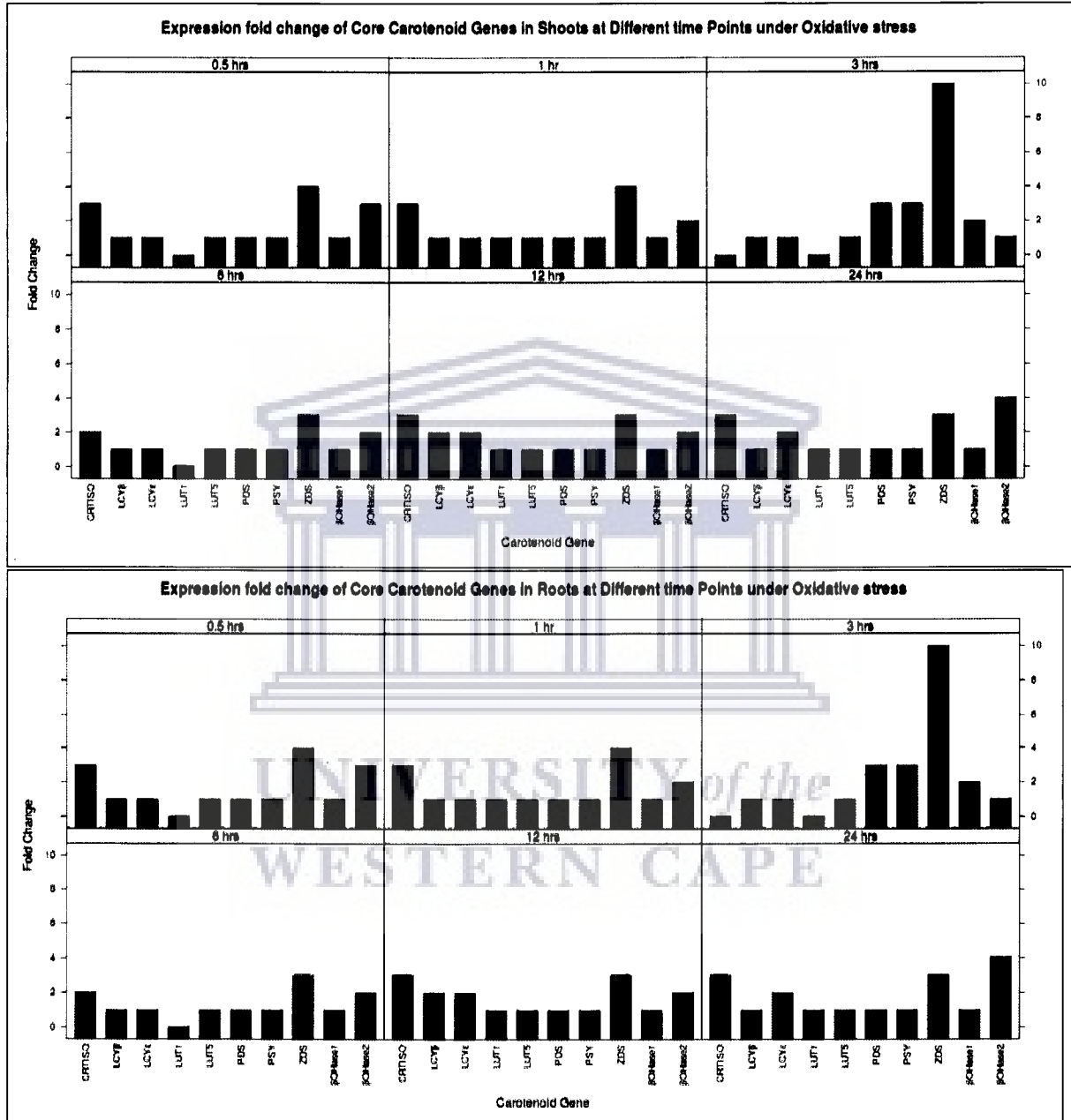


Figure B.1: Expression profiles of core carotenoid genes under oxidative stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types Shoots shown by red bar graphs and roots shown by blue bar graphs.

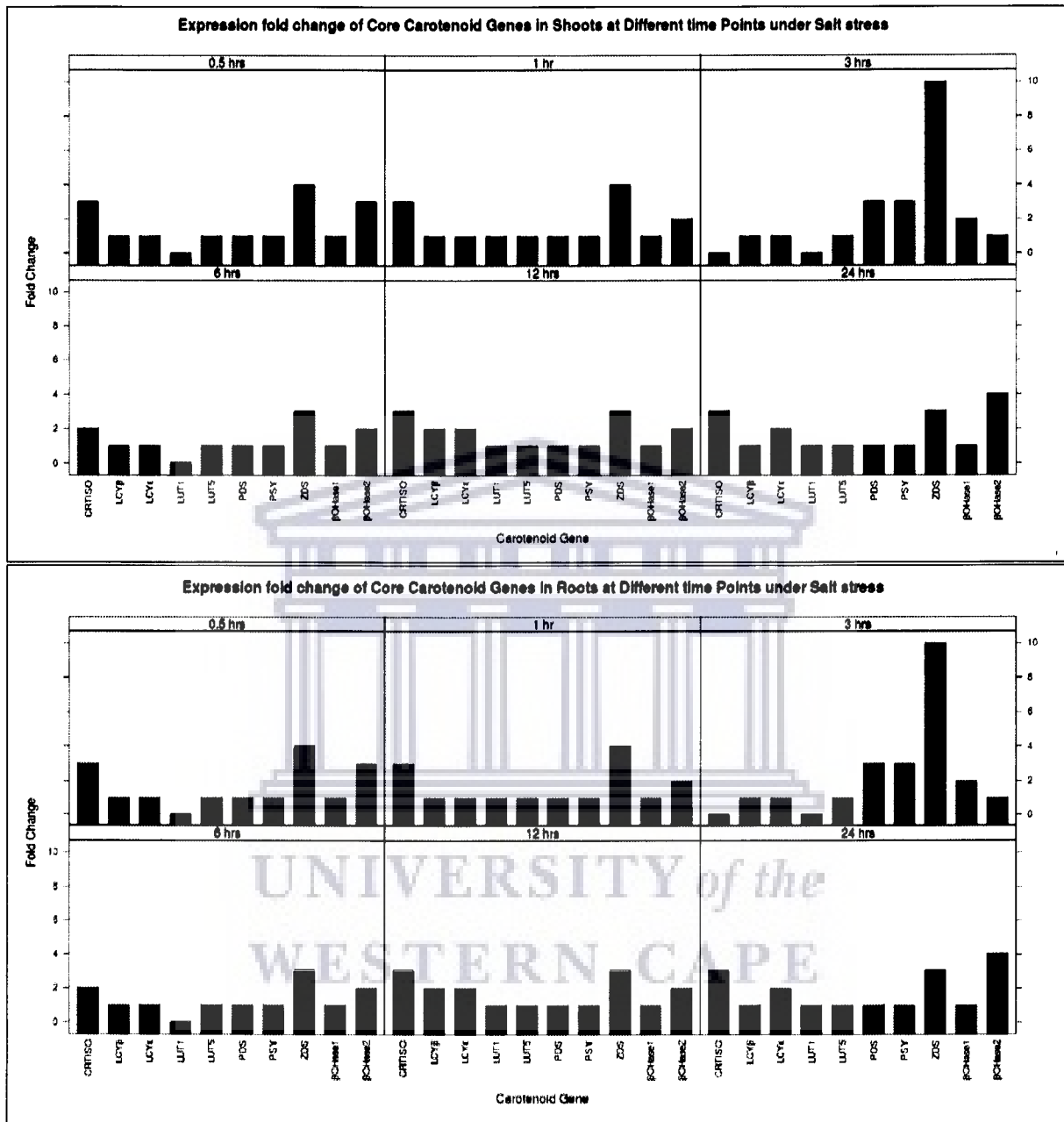


Figure B.2: Expression profiles of core carotenoid genes under salt stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types Shoots shown by red bar graphs and roots shown by blue bar graphs.



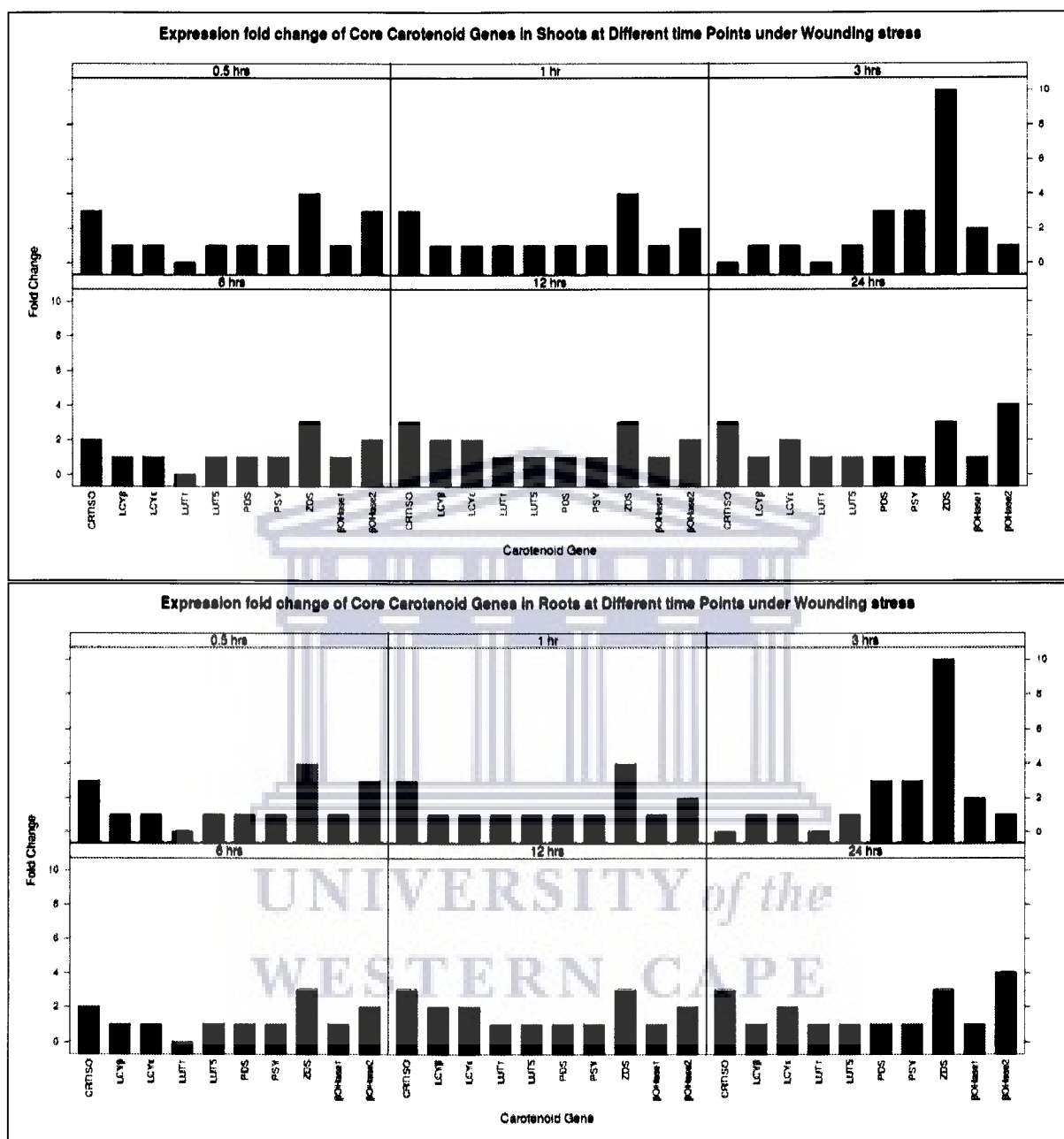


Figure B.3: Expression profiles of core carotenoid genes under wounding stress. Six time points 24hrs, 12hrs, 6hrs, 3hrs, 1hrs, 0.5hrs were taken into consideration as well as two tissue types Shoots shown by red bar graphs and roots shown by blue bar graphs.

## Venn diagrams

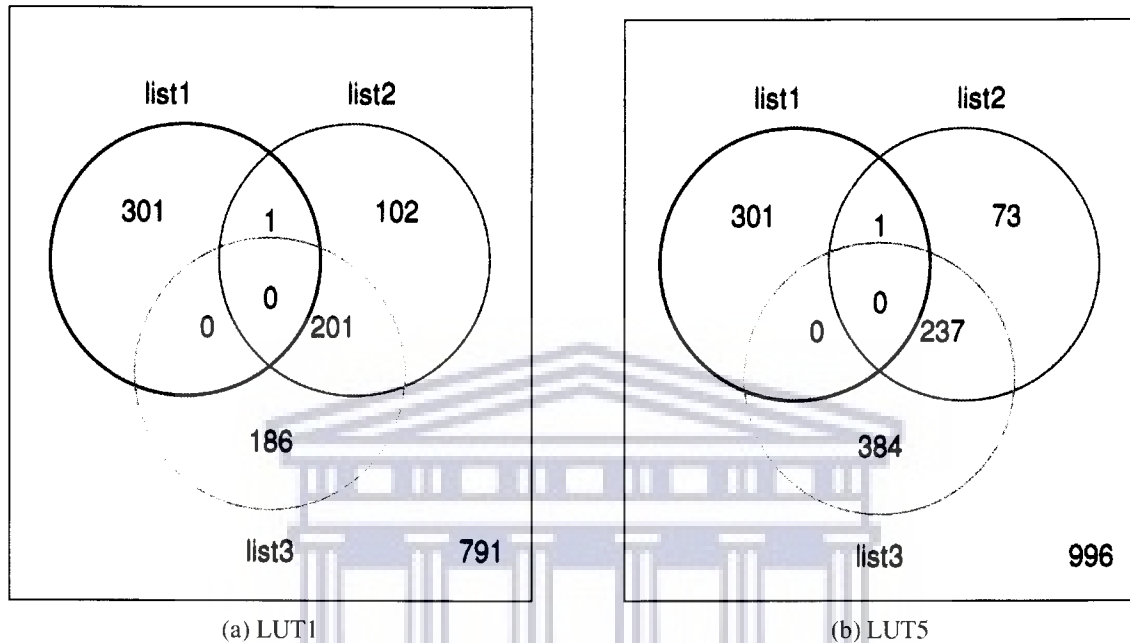


Figure B.4: **Venn diagrams of LUT1, LUT5 and their co-expressed genes.** Each of the coloured circles represent a co-expressed gene list 3 (yellow) co-expressed genes produced using ACT, list 2 (light blue) co-expressed list produced by ATTED-II list 1 (dark blue) co-expressed genes produced by STRING. Overlapping of circle show the number of genes shared between respective lists.

UNIVERSITY of the  
WESTERN CAPE

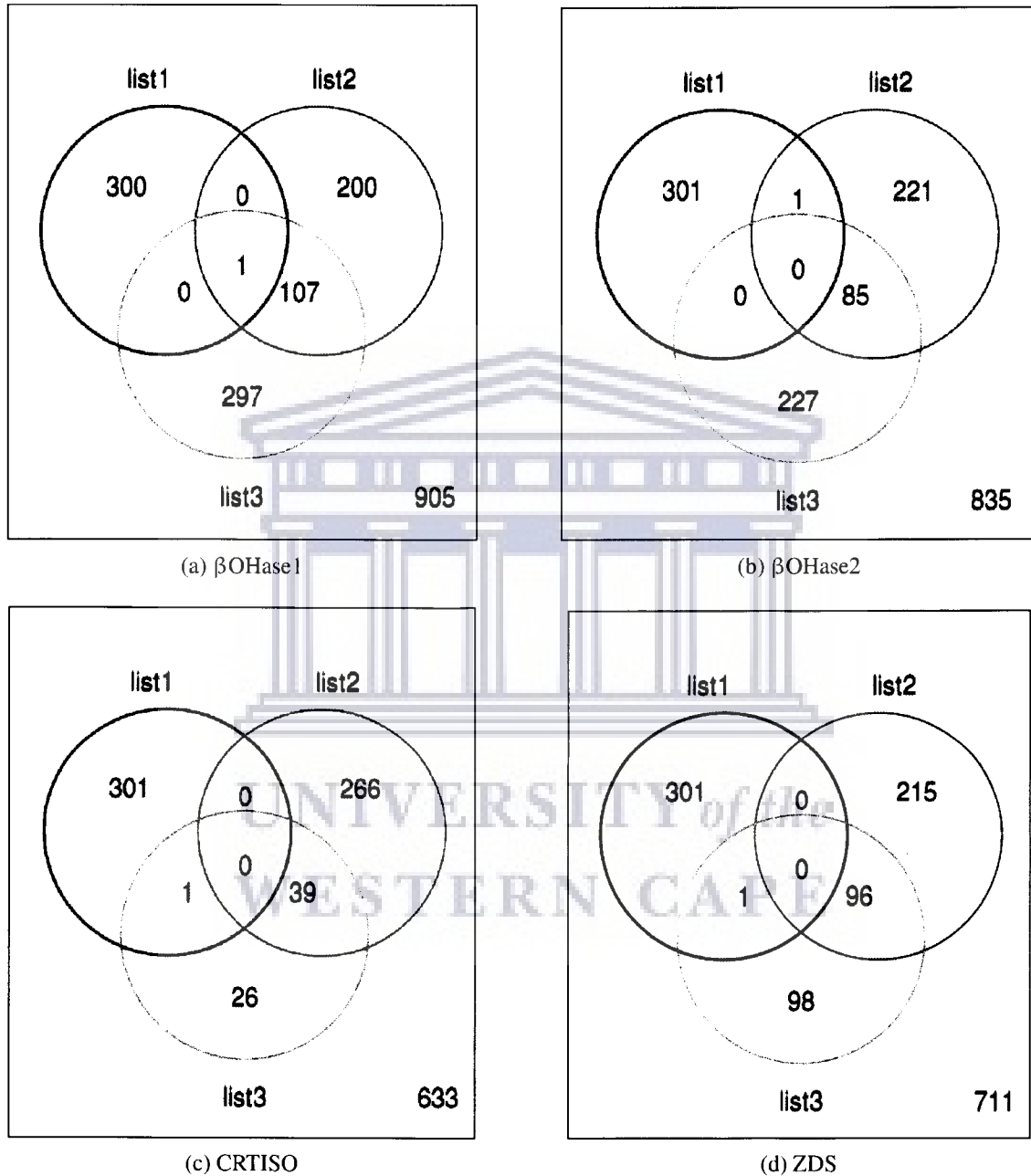


Figure B.5: Venn diagrams of  $\beta$ OHase1,  $\beta$ OHase2, CRTISO, ZDS and their co-expressed genes. Each of the coloured circles represent a co-expressed gene list 3 (yellow) co-expressed genes produced using ACT, list 2 (light blue) co-expressed list produced by ATTED-II list 1 (dark blue) co-expressed genes produced by STRING. Overlapping of circle show the number of genes shared between respective lists.

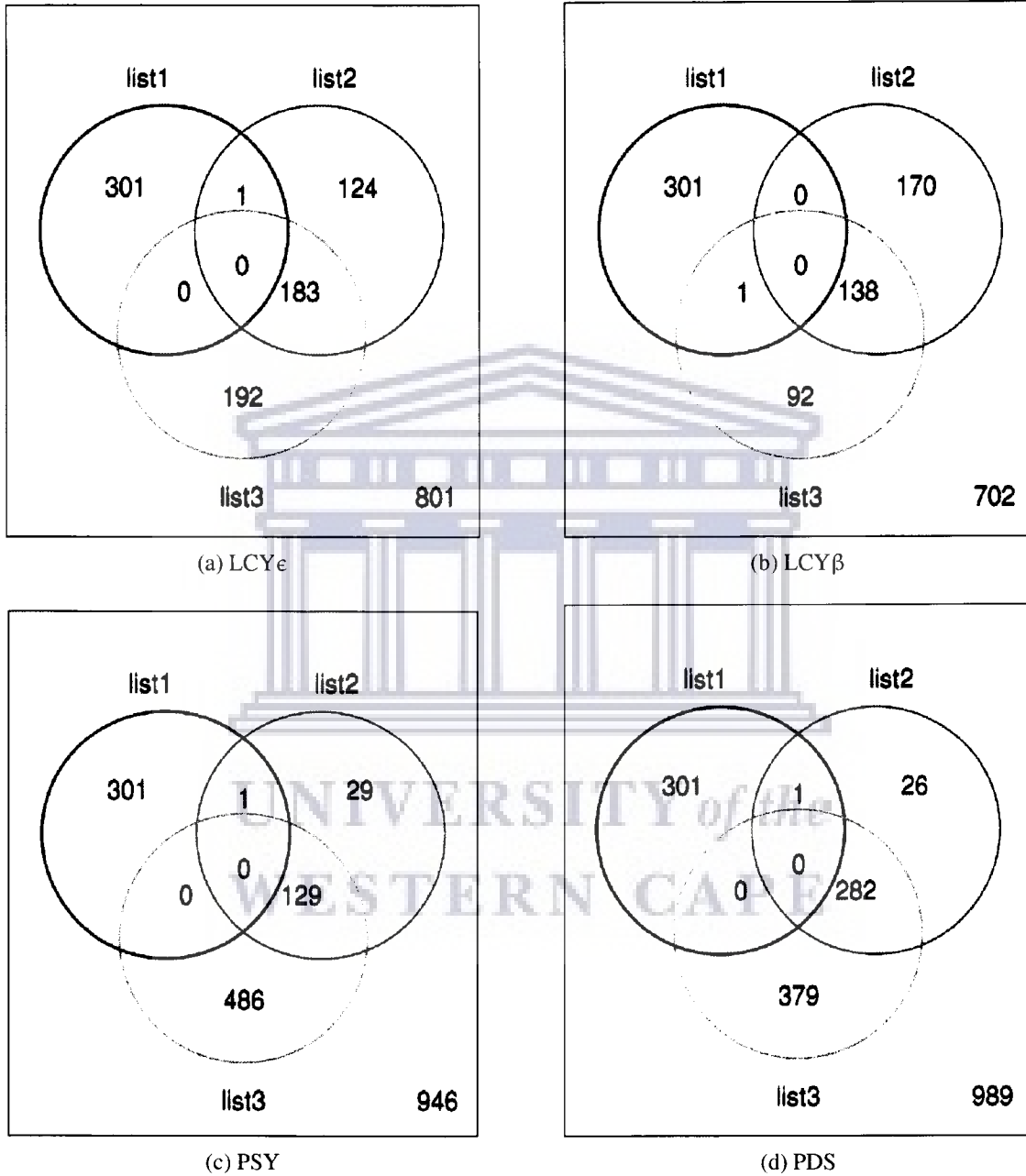


Figure B.6: Venn diagrams of  $LCY\epsilon$ ,  $LCY\beta$ , PSY, PDS and their co-expressed genes. Each of the coloured circles represent a co-expressed gene list 3 (yellow) co-expressed genes produced using ACT, list 2 (light blue) co-expressed list produced by ATTED-II list 1 (dark blue) co-expressed genes produced by STRING. Overlapping of circle show the number of genes shared between respective lists.

## List of co-expressed genes

Probe Set GeneID Annotation

262410\_at AT2G04039 expressed protein

262645\_at AT1G02750 elongation factor Tu family protein similar to elongation factor G SP:F34811 [Glycine max (Soybean)]

261488\_at AT1G14345 expressed protein contains one transmembrane domain

256115\_at AT1G16850 uridylyltransferase-related similar to [Protein-P11] uridylyltransferase (P11 uridylyl-transferase; [Uridylyl removing enzyme] (UTase); SP:Q9AC53; [Caulobacter crescentus]

251784\_at AT2G55330 photosystem II reaction center PsbP family protein contains Pfam profile PF01789: PsbP

265073\_at AT1G55480 expressed protein

263676\_at AT1G039340 expressed protein

245877\_at AT1G26220 GCN5-related N-acetyltransferase (GNAT) family protein low similarity to SP:IP09453

Ribosomal-protein-alanine acetyltransferase (EC 2.3.1.129) [Escherichia coli];  
 \ contains Pfam profile PF00563: acetyltransferase, GNAT family

259625\_at AT1G42970 glyceraldehyde-3-phosphate dehydrogenase B, chloroplast (GAPB) / NADP-dependent glyceraldehydephosphate dehydrogenase subunit B identical to SP:IP25857

Glyceraldehyde 3-phosphate dehydrogenase B, chloroplast precursor (EC 1.2.1.13)  
 (NADP-dependent glyceraldehydephosphate dehydrogenase subunit B) [Arabidopsis thaliana]

253283\_at AT4G34090 expressed protein

262418\_at AT1G00220 thioredoxin x nearly identical to thioredoxin x SB:AAF19502 GI:6539616 from [Arabidopsis thaliana]

262954\_at AT1G54500 rubredoxin family protein similar to SP:IP00270 Rubredoxin (Rd) [Desulfovibrio gigas]; contains Pfam profile PF00301: Rubredoxin

262483\_at AT1G17220 translation initiation factor IF-2, chloroplast, putative similar to SP:FW19971IF2C\_EHAVO

Translation initiation factor IF-2, chloroplast precursor (FvIF2cp) [Phaseolus vulgaris]

249007\_at AT5G44450 expressed protein

257311\_at AT3G28570 phosphate transporter family protein contains Pfam profile: PF01384 phosphate transporter family

245701\_at AT5G04140 glutamate synthase (GLU1) / ferredoxin-dependent glutamate synthase (Fd-GOGAT 1) identical to ferredoxin-dependent glutamate synthase precursor [Arabidopsis thaliana] GI:3869251

255540\_at AT4G01800 preprotein translocase secA subunit, putative similar to preprotein translocase secA subunit, chloroplast [precursor] SP:Q9SY10 from [Arabidopsis thaliana]; non-consensus GA donor splice site at exon 4

264394\_at AT1G11860 aminomethyltransferase, putative similar to aminomethyltransferase, mitochondrial precursor SP:Q49840 from [Flaveria anomala]

245806\_at AT1G45474 chlorophyll a-b binding protein, putative (Lhca5) identical to Lhca5 protein [Arabidopsis thaliana] GI:4741942; contains Pfam profile: PF00504 chlorophyll a-b binding protein; similar to light-harvesting complex protein GI:22752 from [Pinus sylvestris]

262377\_at AT1G73110 ribulose biphosphate carboxylase/oxygenase activase, putative / RuBisCO activase, putative similar to ribulose biphosphate carboxylase/oxygenase activase, chloroplast precursor (RuBisCO activase, RA) [Oryza sativa] SWISS-PROT:P93431

263761\_at AT2G21330 fructose-bisphosphate aldolase, putative strong similarity to plastidic fructose-bisphosphate aldolase (EC 4.1.2.13) from Nicotiana paniculata (NPALD1P) [GI:4827251], Oryza sativa, PIR2:702057 [SP:Q40677]

247816\_at AT5G08260 expressed protein

249120\_at AT5G43750 expressed protein

261788\_at AT1G15980 expressed protein

258025\_at AT3G19460 D-3-phosphoglycerate dehydrogenase, putative / 3-PGDH, putative similar to SP:Q04130 from [Arabidopsis thaliana]

AT3G19460 sodium hydrogen antiporter, putative similar to NhaD [Vibrio parahaemolyticus] gi131237281;JBA25994; Na<sup>+</sup>/H<sup>+</sup> antiporter (NhaD) family member, PMID:9150543

255809\_at AT4G10300 expressed protein

262369\_at AT1G73040 expressed protein

AT1G73070 leucine-rich repeat family protein contains leucine rich-repeat (LRR) domains  
 Pfam:PF00560, INTERPRO:IPR001611; contains similarity to receptor-like protein kinase INRPK1 [Ipomoea nil] gi14495542;JBAAB36553

250073\_at AT5G17170 rubredoxin family protein contains Pfam profile PF00301: Rubredoxin

259098\_at AT3G04790 ribose 5-phosphate isomerase-related similar to ribose 5-phosphate isomerase GI:18654317 from [Spinacia oleracea]

264584\_at AT1G05140 membrane-associated zinc metalloprotease, putative similar to Hypothetical zinc metalloprotease A11391 [SP:Q8Y064] [strain PCC 7120] [Anabaena sp.]; Similar to Synechocystis hypothetical protein (sb1D90408); contains Pfam PF00595: PDZ domain (Also known as DHR or GLGF); contains TIGRFAM TIGR00054: membrane-associated zinc metalloprotease, putative

250531\_at AT5G08650 GTP-binding protein LepA, putative

249878\_at AT5G1120 photosystem II stability/assembly factor, chloroplast (HCF136) identical to SP:Q82660 Photosystem II stability/assembly factor HCF136, chloroplast precursor [Arabidopsis thaliana]

260704\_at AT1G37470 glycine cleavage system H protein, mitochondrial, putative similar to SP:IP25855 glycine cleavage system H protein L, mitochondrial precursor [Arabidopsis thaliana]; contains Pfam profile PF01597: Glycine cleavage H-protein

261053\_at AT1G01320 tetrapeptide repeat (TPR)-containing protein low similarity to SP:IP46825 Kinexin light chain (KLC) [Loligo pealeii]; contains Pfam profile PF00515: TPR Domain

255719\_at AT1G32080 membrane protein, putative contains 12 transmembrane domains; similar to yohK [GI:405873] [Escherichia coli]

265415\_at AT2G20890 expressed protein

258755\_at AT3G11950 UbiA prenyltransferase family protein contains Pfam profile PF01040: UbiA prenyltransferase family

256076\_at AT1G48060 expressed protein

251885\_at AT3G54050 fructose-1,6-bisphosphatase, putative / D-fructose-1,6-bisphosphate 1-phosphohydrolase, putative / FBPase, putative strong similarity to fructose-1,6-bisphosphatase [Brassica napus] GI:289367; identical to SP:IP25851 Fructose-1,6-bisphosphatase, chloroplast precursor (EC 3.1.3.11) (D-fructose-1,6-bisphosphate 1-phosphohydrolase) (FBPase) [Arabidopsis thaliana]; contains Pfam profile PF00316: fructose-1,6-bisphosphatase

259140\_at AT3G10230 lycopene beta cyclase (LYC) identical to lycopene beta cyclase GI:1399183;GB:AA853337 [Arabidopsis thaliana]

267430\_at AT2G34860 chaperone protein dnaJ-related contains Pfam PF00684 : DnaJ central domain (4 repeats); similar to Chaperone protein DnaJ (Heat shock protein 40) [SP:Q9UXE9] [Methanosarcina thermophila]

264442\_at AT1G27480 lecithin:cholesterol acyltransferase family protein / LACT family protein similar to LCAT like lysophospholipase (LLPL) [Homo sapiens] GI:4589720; contains Pfam profile PF02450: Lecithin:cholesterol acyltransferase [phosphatidylcholine:sterol acyltransferase]

251118\_at AT3G63410 chloroplast inner envelope membrane protein, putative (APG1) similar to SP:IP23529 37 kDa inner envelope membrane protein, chloroplast precursor (E37) [Spinacia oleracea]; contains Pfam profile PF01209: methyltransferase, Ubie/COQ5 family

256049\_at AT1G07010 calcineurin-like phosphoesterase family protein contains Pfam profile: PF00149 calcineurin-like phosphoesterase

259981\_at AT1G76450 oxygen-evolving complex-related SP:Q9S720; contains a PsbP domain

249247\_at AT5G42310 pentatricopeptide (PTP) repeat-containing protein contains Pfam profile PF01535: PTP repeat

245354\_at AT4G17600 lili3 protein identical to Lili3 protein [Arabidopsis thaliana] gi147419661;JBA028780

250563\_at AT5G03050 expressed protein predicted protein, Arabidopsis thaliana

251243\_at AT3G61870 expressed protein hypothetical protein - Synechocystis sp. (strain PCC 6803), PIR:275859

262878\_at AT1G44770 expressed protein

262169\_at AT1G74730 expressed protein

267247\_at AT2G30170 expressed protein  
 248440\_at AT5G51110 expressed protein  
 251744\_at AT3G56010 expressed protein  
 256088\_at AT1G20910 immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family protein identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 1, chloroplast precursor (Ppiase) (Rotamase) (SP:Q9LM71) (Arabidopsis thaliana); similar to SP|P25138 FK506-binding protein (Peptidyl-prolyl cis-trans isomerase) (Ppiase) (EC 5.2.1.8) (Rotamase) (Neisseria meningitidis); contains Pfam PF00254; peptidyl-prolyl cis-trans isomerase, FKBP-type  
 257172\_at AT3G23700 S1 RNA-binding domain-containing protein contains Pfam domain, PF00575: S1 RNA binding domain  
 248242\_at AT5G53590 aldo/keto reductase family protein contains Pfam profile PF00248: oxidoreductase, aldo/keto reductase family  
 251762\_at AT3G59800 sedoheptulose-1,7-bisphosphatase, chloroplast / sedoheptulose-bisphosphatase identical to SP|P46283 Sedoheptulose-1,7-bisphosphatase, chloroplast precursor (EC 3.1.3.37) (Sedoheptulose-bisphosphatase) (SBPASE) (SED1, 7) (P2AZE) (Arabidopsis thaliana)  
 258566\_at AT5G07020 proline-rich family protein  
 245116\_at AT2G41660 thioredoxin reductase, putative / NADPH dependent thioredoxin reductase, putative The last 2 exons encode thioredoxin. There is an EST match to exons 5-7, and the distance between exon 7 and exon 8 is only 90bp. It is unlikely this is two separate genes, but more likely a hybrid protein.  
 253549\_at AT4G30500 expressed protein  
 255088\_at AT430350 DnaJ heat shock N-terminal domain-containing protein similar to P|Q45552 Chaperone protein dnaJ (Bacillus stearothermophilus); contains Pfam profile PF00226: DnaJ domain  
 253387\_at AT4G33010 glycine decarboxylase [decarboxylating], putative / glycine decarboxylase, putative / glycine cleavage system P-protein, putative strong similarity to SP|P49361 Glycine decarboxylase [decarboxylating] A, mitochondrial precursor (EC 1.4.4.2) (Flaveria pringlei); contains Pfam profile PF02347: Glycine cleavage system P-protein  
 250256\_at AT5G12650 elongation factor family protein contains Pfam profiles: PF00009 elongation factor Tu GTP binding domain, PF00679 elongation factor G C terminus, PF03144 elongation factor Tu domain 2  
 256015\_at AT1G19150 chlorophyll A-B binding protein, putative / LHCI type II, putative very strong similarity to PSI type II chlorophyll a/b-binding protein  
 Inca241 GI:541565 from (Arabidopsis thaliana); contains Pfam profile: PF00504 chlorophyll A-B binding protein  
 262104\_at AT1G02910 tetratricopeptide repeat (TPR)-containing protein contains Pfam profile PF00513: TPR Domain  
 253139\_at AT4G35450 ankyrin repeat family protein / AFT protein (AFT) contains ankyrin repeats, Pfam:PF00023; identical to cDNA AFT protein (AFT) GI:3478699  
 255720\_at AT1G32060 phosphoribulokinase (PRK) / phosphoenolpyruvate kinase nearly identical to SP|P25697 Phosphoribulokinase, chloroplast precursor (EC 2.7.1.19) (Phosphoenolpyruvate kinase) (PRK) (Arabidopsis thaliana)  
 258495\_at AT3G02690 integral membrane family protein similar to PecM protein (GI:5852331) (Vossella indigofera) and PecM protein (SP:F42194) (Erwinia chrysanthemi)  
 247694\_at AT5G59750 riboflavin biosynthesis protein, putative similar to SP|P50955 Riboflavin biosynthesis protein riba [Includes: GTP cyclohydrolase II (EC 3.5.4.25); 3,4-dihydroxy-2-butanone 4-phosphate synthase (DHBP synthase) (Actinobacillus pleuropneumoniae); contains Pfam profiles PF00925: GTP cyclohydrolase II, PF00926: 3,4-dihydroxy-2-butanone 4 phosphate synthase  
 262142\_at AT1G65230 expressed protein  
 261422\_at AT1G15730 expressed protein  
 247700\_at AT5G59250 sugar transporter family protein similar to Dxylose-H<sup>+</sup> symporter from Lactobacillus brevis GI:2895856, sugar porter family protein 2 (Arabidopsis thaliana); GI:14585701; contains Pfam profile PF00093: major facilitator superfamily protein  
 247259\_at AT5G64290 oxoglutarate/malate translocator, putative similar to SWISS-PROT P1041364 2-oxoglutarate/malate translocator, chloroplast precursor (Spinach) (Spinacia oleracea)  
 248962\_at AT5G45660 FK506-binding protein 1 (FKBP1) identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 3, chloroplast precursor (Ppiase) (Rotamase) (SP:Q9SCY2) / FK506 binding protein 1 (GI:1595744) (Arabidopsis thaliana); contains Pfam PF00254; peptidyl-prolyl cis-trans isomerase, FKBP-type  
 251157\_at AT3G63140 mRNA-binding protein, putative similar to mRNA binding protein precursor (GI:26453355) (Lycopersicon esculentum)  
 255046\_at AT4G09650 ATP synthase delta chain, chloroplast, putative / H(+)-transporting two-sector ATPase, delta (OSCP) subunit, putative similar to SP|P32980 ATP synthase delta chain, chloroplast precursor (EC 3.6.3.14) (Nicotiana tabacum); contains Pfam profile PF00213: ATP synthase F1, delta subunit  
 262473\_at AT1G50250 cell division protein ftsH homolog 1, chloroplast (FTSH) (FTSH) identical to UP:Q39102 Cell division protein ftsH homolog 1, chloroplast precursor (EC 3.4.24.-) (Arabidopsis thaliana)  
 253208\_at AT4G34820 expressed protein  
 250613\_at AT5G07240 calmodulin-binding family protein contains Pfam profile PF00612: IQ calmodulin-binding motif  
 AT5G07250 rhomboid family protein contains PFAM domain PF01694, Rhomboid family  
 246736\_at AT5G27540 expressed protein hypothetical protein alr170c - Synechocystis sp., PIR:Q75312  
 249574\_at AT5G38520 hydrolase, alpha/beta fold family protein low similarity to hydrolase [Bacteroides sp. BDF63] GI:14196240; contains Pfam profile PF00561: hydrolase, alpha/beta fold family  
 261206\_at AT1G12800 S1 RNA-binding domain-containing protein contains Pfam domain, PF00575: S1 RNA binding domain  
 256542\_at AT1G42550 expressed protein

# References

- Arabidopsis, T. and Initiative, G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.
- Armstrong, G. A. and Hearst, J. E. (1996). Serial review carotenoids 2. *Plant Cell*, 10(February).
- Baldwin, D., Crane, V., and Rice, D. (1999). A comparison of gel-based, nylon filter and microarray techniques to detect differential rna expression in plants. *Current Opinion in Plant Biology*, 2(2):96–103.
- Bartley, G. E., Scolnik, P. a., and Beyer, P. (1999). Two *Arabidopsis thaliana* carotene desaturases, phytoene desaturase and zeta-carotene desaturase, expressed in *Escherichia coli*, catalyze a poly-cis pathway to yield pro-lycopene. *European journal of biochemistry / FEBS*, 259(1-2):396–403.
- Braam, J., Sistrunk, M. L., Polisensky, D. H., Xu, W., Purugganan, M. M., Antosiewicz, D. M., Campbell, P., and Johnson, K. a. (1997). Plant responses to environmental stress: regulation and functions of the *Arabidopsis* TCH genes. *Planta*, 203 Suppl:S35–41.
- Chen, W.-H., de Meaux, J., and Lercher, M. J. (2010). Co-expression of neighbouring genes in *Arabidopsis*: separating chromatin effects from direct interactions. *BMC genomics*, 11:178.

- Chow, L. C., Gelinis, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger rna. 1977. *Reviews in Medical Virology*, 10(6):362–371; discussion 355–356.
- Churchill, G. a. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature genetics*, 32 Suppl(december):490–5.
- Cold, T., Harbor, S., McCombie, W. R., Bastide, M. D., Habermann, K., Parnell, L., and Dedhia, N. *et.al.* (2000). The complete sequence of a heterochromatic island from a higher eukaryote. The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems Arabidopsis Sequencing Consortium. *Cell*, 100(3):377–86.
- Consortium, G. O. (2006). The Gene Ontology (GO) project in 2006. *Nucleic acids research*, 34(Database issue):D322–6.
- Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic acids research*, 32(Database issue):D575–7.
- Cui, H.-Y., Lestradet, M., Bruey-Sedano, N., Charles, J.-P., and Riddiford, L. M. (2009). Elucidation of the regulation of an adult cuticle gene Acp65A by the transcription factor Broad. *Insect molecular biology*, 18(4):421–9.
- Cunningham, F. X. (2002). Regulation of carotenoid synthesis and accumulation in plants. *Pure and Applied Chemistry*, 74(8):1409–1417.
- Cunningham, F. X., Pogson, B., Sun, Z., McDonald, K. A., DellaPenna, D., and Gantt, E. (1996). Functional Analysis of the [beta] and [epsilon] Lycopene Cyclase Enzymes of Ara-



- bidopsis Reveals a Mechanism for Control of Cyclic Carotenoid Formation. *Plant Cell*, 8(September):1613–1626.
- De Maio, A. (1999). Heat shock proteins: facts, thoughts, and dreams. *Shock (Augusta, Ga.)*, 11(1):1–12.
- DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cdna microarray to analyze gene. *Nature Genetics*, (14):457–460.
- Dillon, N. and Festenstein, R. (2002). Unravelling heterochromatin: competition between positive and negative factors regulates accessibility. *Trends in genetics : TIG*, 18(5):252–8.
- Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. (2007). Trawler : de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods*, 11(June):2–4.
- Hegde, P., Qi, R., Abernathy, K., Gay, C., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N., and Quackenbush, J. (2000). BioFeature A Concise Guide to cDNA Microarray Analysis BioFeature. *Biotechniques*, 29(3).
- Hollingsworth, N. M. (2008). Deconstructing meiosis one kinase at a time: polo pushes past pachytene. *Genes & development*, 22(19):2596–600.
- Howell, K. a., Narsai, R., Carroll, A., Ivanova, A., Lohse, M., Usadel, B., Millar, a. H., and Whelan, J. (2009). Mapping metabolic and transcript temporal switches during germination in rice highlights specific transcription factors and the role of RNA instability in the germination process. *Plant physiology*, 149(2):961–80.

- Huala, E., Dickerman, a. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., and Huang, W. *et. al.* (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic acids research*, 29(1):102–5.
- Isaacson, T., Ronen, G., Zamir, D., and Hirschberg, J. (2002). Cloning of tangerine from tomato reveals a carotenoid isomerase essential for the production of beta-carotene and xanthophylls in plants. *The Plant Cell*, 14(2):333–342.
- Kane, M. D., Jatko, T. a., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research*, 28(22):4552–7.
- Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., DAngelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The atgenexpress global stress expression data set: protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses. *The Plant Journal*, 50(2):347–363.
- Koch, K. E. (1996). Carbohydrate-Modulated Gene Expression in Plants. *Annual review of plant physiology and plant molecular biology*, 47:509–540.
- Lee, M. L., Kuo, F. C., Whitmore, G. a., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):9834–9.

- Li, F., Tsfadia, O., and Wurtzel, E. T. (2009). The phytoene synthase gene family in the Grasses. *4(3):208–211.*
- Li, N. and Tompa, M. (2006). Analysis of computational approaches for motif discovery. *Algorithms for molecular biology : AMB*, 1:8.
- Lin, L.-H., Lee, H.-C., Li, W.-H., and Chen, B.-S. (2005). Dynamic modeling of cis-regulatory circuits and gene expression prediction via cross-gene identification. *BMC bioinformatics*, 6:258.
- Logemann, E., Parniske, M., and Hahlbrock, K. (1995). Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proceedings of the National Academy of Sciences of the United States of America*, 92(13):5905–9.
- Manfield, I. W., Jen, C.-h., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M., and Westhead, D. R. (2006). Arabidopsis Co-expression Tool ( ACT ): web server tools for microarray-based gene expression analysis. *Search*, 34:504–509.
- Marrs, K. a. (1996). the Functions and Regulation of Glutathione S-Transferases in Plants. *Annual review of plant physiology and plant molecular biology*, 47:127–158.
- Matthews, P. D. and Wurtzel, E. T. (2007). *Biotechnology of food colorant production In Food Colorants*. CRC Press.
- Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, a., Stiekema, W., Entian, K. D., and Terry, N. *et. al.* (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, 402(6763):769–77.

- Meier, S., Gehring, C., MacPherson, C. R., Kaur, M., Maqungo, M., Reuben, S., Muyanga, S., Shih, M.-D., Wei, F.-J., and Wanchana, S. *et. al.* (2008). The promoter signatures in rice leaf genes can be used to build a co-expressing leaf gene network. *Rice*, 1(2):177–187.
- Meier, S., Tzfadia, O., Vallabhaneni, R., Gehring, C., and Wurtzel, E. T. (2011). A transcriptional analysis of carotenoid, chlorophyll and plastidial isoprenoid biosynthesis genes during development and osmotic stress responses in *Arabidopsis thaliana*. *BMC systems biology*, 5(1):77.
- Melhus, H., Michaelsson, K., Kindmark, A., and Bergstrom, R. (1998). Excessive Dietary Intake of Vitamin A Is Associated with Reduced Bone Mineral Density and Increased Risk for Hip Fracture Methods Bone Mineral Density Study. *Europe*.
- Mering, C. V. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261.
- Meyer, P. (2000). Transcriptional transgene silencing and chromatin components. *Plant molecular biology*, 43(2-3):221–34.
- Molina, C. and Grotewold, E. (2005). Genome wide analysis of *Arabidopsis* core promoters. *BMC genomics*, 6:25.
- Noh, B., Lee, S.-h., Kim, H.-j., Yi, G., and Shin, E.-a. *et. al.* (2004). Divergent Roles of a Pair of Homologous Jumonji / Zinc-Finger Class Transcription Factor Proteins in the Regulation of *Arabidopsis* Flowering Time. *October*, 16(October):2601–2613.
- Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides co-expressed gene networks for *Arabidopsis*. *Nucleic acids research*, 37(Database issue):D987–91.

- Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., and Ohta, H. (2007). ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic acids research*, 35(Database issue):D863–9.
- Olson, J. A. (1996). Symposium : Prooxidant Effects of Antioxidant Vitamins Benefits and Liabilities of Vitamin A and Carotenoids1 ' 2. *Biochemistry*, pages 1208–1212.
- Olson, J. A., Olin, W., and Atwater, T. (1993). Special Article 1992 Atwater Lecture . The irresistible carotenoids and vitamin A14 fascination of. *Clinical Nutrition*.
- Park, H., Kreunen, S. S., Cuttriss, A. J., Dellapenna, D., and Pogson, B. J. (2002). Identification of the Carotenoid Isomerase Provides Insight into Carotenoid Biosynthesis , Prolamellar Body Formation , and Photomorphogenesis. *Society*, 14(February):321–332.
- Pozner, A., Goldenberg, D., Negreanu, V., Le, S.-y., Elroy-stein, O., Levanon, D., and Groner, Y. (2000). Transcription-Coupled Translation Control of AML1 / RUNX1 Is Mediated by Cap- and Internal Ribosome Entry Site-Dependent Mechanisms. *Society*, 20(7):2297–2307.
- Ramaswamy, B. S. and Golub, T. R. (2002). BIOLOGY OF NEOPLASIA DNA Microarrays in Clinical Oncology. *Society*, 20(7):1932–1941.
- Rao, M. V., Lee, H., Creelman, R. a., Mullet, J. E., and Davis, K. R. (2000). Jasmonic acid signaling modulates ozone-induced hypersensitive cell death. *The Plant cell*, 12(9):1633–46.
- Sandve, G. K. and Drablø s, F. (2006). A survey of motif discovery methods in an integrated framework. *Biology direct*, 1:11.

- Regulatory Factor 1 ( IRF-1 ) and IRF-2 , Regulators of Cell Growth and the Interferon System. 13(8):4531–4538.
- Van Hal, N. L., Vorst, O., Van Houwelingen, A. M., Kok, E. J., Peijnenburg, A., Aharoni, A., Van Tunen, A. J., and Keijer, J. (2000). The application of dna microarrays in gene expression analysis. *Journal of Biotechnology*, 78(3):271–280.
- Villamor, E. and Fawzi, W. W. (2005). Effects of Vitamin A Supplementation on Immune Responses and Correlation with Clinical Outcomes. *Society*, 18(3):446–464.
- Wang, X., Haberer, G., and Mayer, K. F. X. (2009). Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC genomics*, 10:284.
- Wilhelm, K. S. and Thomashow, M. F. (1993). Arabidopsis thaliana cor15b, an apparent homologue of cor15a, is strongly responsive to cold and ABA, but not drought. *Plant molecular biology*, 23(5):1073–7.
- Wisman, E. and Ohlrogge, J. (2000). Resources and Opportunities Arabidopsis Microarray Service Facilities 1. *Society*, 124(December):1468–1471.
- Wu, C. (1995). Heat shock transcription factors: structure and regulation. *Annual Review of Cell and Developmental Biology*, 11:441–469.
- Ye, S. Q., Lavoie, T., Usher, D. C., and Zhang, L. Q. (2002). Microarray, sage and their applications to cardiovascular diseases. *Cell Research*, 12(2):105–115.