Systems biology-based study of provitamin A carotenoid biosynthesis in *Arabidopsis thaliana*

By

Oren Tzfadia

A dissertation submitted to the Graduate Faculty of Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy,

The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Biology in satisfaction of the dissertation requirement for the Doctor of Philosophy

Dr. Eleanore T. Wurtzel

_ _

Date Chair of Examining Committee

Dr. Laurel A. Eckhardt

_ _

Date Executive Officer

Dr. Ron Shamir

_ _

Dr. Avi Maayan

_ _

Dr. Dwight Kincaid

_ _

Dr. Haiping Chang

_ _

Dr. Erich Grotewold

_ _

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Systems biology-based study of provitamin A carotenoid biosynthesis in *Arabidopsis thaliana*

By

Oren Tzfadia

Adviser – Dr. Eleanore Wurtzel

Due to their great nutritional and health value, understanding the regulatory mechanisms and recognizing new points of control in the carotenoid pathway can be the goal of breeding plans for increasing carotenoids accumulation in crop plants. Systems biology is an inter-disciplinary field, which integrates computational models and tools with molecular biology and different types of data including *in silico* transcriptomics, co-expression correlation, metabolomics, proteomics and phylogenetic information in order to develop hypotheses with statistically sound robustness. In the first steps of my work I describes the sequential use of freely available databases to explore the regulation of carotenoid biosynthesis in *Arabidopsis* during chloroplast development. The findings suggested that coordinated transcriptional regulation of genes along the isoprenoid-related biosynthesis pathways, play a major role in coordinating the synthesis of functionally related, chloroplast-localized isoprenoid-derived compounds. Next I aspired to find candidate genes that are participating in or regulating the carotenoid pathway. A model was developed to integrate several types of high-throughput data, in order to optimize candidate gene ranking in an effort to best define associated genes for a specific studied pathway. The candidate ranking was achieved by using an iterative algorithm (called MORPH), which is built on implementation of machine learning techniques.

Application of the method on several biological pathways in *Arabidopsis* proved the ability of the algorithm to capture experimentally proven gene candidates related to known biological pathways. The robustness of the predictions provided by MORPH creates an exciting research methodology to explore regulation of biological pathways in plants. Although the development of the computational algorithm was initially triggered by the specific needs of our laboratory, namely, for close analysis of the carotenoid pathway, the algorithm is suitable for almost any biological pathway in plants. Moreover the method could be applied to any other model system that has enough available high-throughput data.

ACKNOWLEDGMENTS

First, I would like to thank my parents (and sister) for raising me with enormous amounts of love, dedication and freedom and to whom I owe my not-so-bad set of genes that made me who I am. **Ima, Aba, I love you so much and I want to dedicate this work to you!**

I would also like to thank to: **Prof. Elli Wurtzel** for being my adviser and providing me with almost absolute freedom to explore and express my skills. For letting me pursue my interest and become a Bioinformatician. For being patient and flexible, and allowing me to overcome all obstacles that came in the way. To NIH for providing funding support to E.T.W. To **Prof. Chris Gehring** for exposing me to the wonderful world of BioInformatics and inspiring me choose this path in my career. I want to thank all my committee members for their support and in specific thanks to **Prof. Ron Shamir** for hosting me in his lab and guiding me in the past year, which served as a huge spring board to my work. I want to also thank to **Didi Amar** for his amazing work and a great collaboration. Toda Abuya! I want to say a huge thanks to **Dr. Louis Bradbury** for being a friend, a brother and a mentor. I owe my Ph.D to you man! Could not do it without you (and Thom York and the Auction House). I want to thank also all my lab mates – **Dr. Abby Cuttriss**, **Dr. Maria Shumskaya**, **Rena Quinlan**, **Chucho Beltran** and **Yao Xiaoling**. And to all my NY hommies and in specific to **James Lendemer**, **Aman Gill**, **Vinson Doyle**, **Dan Kulakowski** and dearest **Cuppy (Lisa Offringa)** for being such recreational and inspirational mates! To my seven brothers (and one sister) from another mothers – **Yaniv Cohen**, **Eran Bahari**, **Ronen Golovinski**, **Shay Maller**, **Ziv Klienman**, **Eyal Avisar** and **Tamar Tzur**. You brought me here and you were here with me always in mind, heart and soul. I want to thank my (other) parents and sisters: **Bilha**, **Zohar**, **Shani**, **Nir**, **Gideon** and **Tal**. I love you so much!

Last I want to thank the three loves of my life: Inbalula. I can't say enough how much I love you and how amazing you are. Thank you for being mine. "Can you read my mind?". **Geffen** for being my precious little 'decent with modifications'. And **Mokalul** for being the purest soul on four.

*And a special thank to **Charles Darwin**. You showed us the light. And we shall follow!*

Funding: This research was supported by a grant (to Eleanore Wurtzel) from the United States National Institutes of Health (GM081160).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

List of abbreviations:

ABA - abscisic acid

ACT – *Arabidopsis* co-expression tool

BR – Brassinosteroid

BTH- [Benzo-(1,2,3)-thiadiazole-7-carbothioic acid S-methyl ester

CBRG – Carotenoid biosynthesis-related genes

CCD - Carotenoid cleavage enzymes

CHX- Cycloheximide

DMAPP - Dimethylallyl diphosphate

ExPr - Expressed protein

FDR - False discovery rate

GO - Gene ontology

K – Potassium

IPP - Isopentenyl diphosphate

MEP - Methylerythritol 4-phosphate pathway

NCED - Nine-cis-epocaratenoid dioxygenase

PAIR - Predicted *arabidopsis* intercom resource

PhQ - Phylloquinones

PIFs – Phytochrome-interacting transcription factors

PSY0.6 – Probes that had an r-value $\geq$0.6 with *PSY*

PSY50 – 50 probes most co-expressed with *PSY* in the ACT data

PSY25 – 25 probes most co-expressed with *PSY* in the ACT data

PQ - Plastoquinone

Rt – Roots

St – Shoots

TFs - Transcription factor

CHAPTER 1. THE CAROTENOID BIOSYNTHETIC PATHWAY

1.1 Background

Carotenoids are natural pigments biosynthesized by all plants, cyanobacteria, some fungi and bacteria (Britton et al., 2004) and recently even in aphids (Moran and Jarvik, 2010). Carotenoids are essential for plant survival, as they serve as photoprotectors preventing oxidative damage, absorb light during photosynthesis (Niyogi, 2000) and serve as precursors to apocarotenoids such as abscisic acid (ABA) (Nambara and Marion-Poll, 2005) and strigolactones central to plant architecture (Booker et al., 2004; Akiyama et al., 2005). Due to their great nutritional and health value, a major growth has been witnessed in demand for an interest in the carotenoid industry, including the animal feed (chicken and salmon farming), pharmaceutical, cosmetic, and food and dietary supplement industries. The global market for carotenoids is estimated at about $900 million per year (Giuliano, 2008). Carotenoids are important for both human and animal health, as they serve as antioxidants and protect against certain diseases (van den Berg et al., 2000; Fraser and Bramley, 2004). Human and animals can only accumulate vitamin A by including pro-vitamin A carotenoids in their diet. It is estimated that vitamin A deficiency affects ~250 million children globally, primarily in developing countries (Underwood and Arthur, 1996; West, 2002). These effects include visual impairment or even permanent blindness, increased susceptibility to certain illnesses and increased risk of maternal transmission of viruses, such as HIV (Semba et al., 1994).

1.2 Carotenogenesis in plants

Understanding the regulatory network of carotenoid production in a plant model system, will allow for translation of the information to species bearing greater agricultural and pharmaceutical value. Flux through the carotenoid pathway is dependent on both the expression of carotenoid-related genes and the expression of genes encoding enzymes associated with substrates of upstream pathways.

The plastidial methylerythritol 4-phosphate pathway (MEP) functions upstream and generates prenyl diphosphate precursors required for carotenoid biosynthesis (Cordoba et al., 2009). The 1-deoxy-D-xylulose 5-phosphate synthase (*DXS*) enzyme is the first enzyme along the MEP pathway and catalyzes the synthesis of DXP from pyruvate and gLCYEraldehyde 3-phosphate. The prenyl diphosphate precursors, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are the end products of the MEP pathway. In plastids, a geranylgeranyl diphosphate synthase (*GGPPS*) enzyme then catalyzes the sequential addition of three molecules of IPP to one molecule of DMAPP, resulting in the formation of the poly-isoprenoid GGPP.

The phytoene synthase (*PSY*) gene encodes the first enzyme dedicated to the carotenoid pathway, which catalyses the condensation of two geranylgeranyl diphosphate molecules to form 15-*cis*-phytoene (Cuttriss, 2011)(Figure 1.1). Phytoene desaturase (*PDS*) then introduces two sets of double bonds in 15-*cis*-phytoene to form 9,15,9'-tri-*cis*-ζ-carotene, the substrate for the recently discovered isomerase, *Z-ISO*, zeta carotene isomerase (*Z-ISO*) (Chen et al., 2010). *Z-ISO* isomerizes the *PDS* enzymatic product to form 9,9'-di-*cis*- ζ-carotene (Chen et al., 2010), the substrate for the second desaturase, ζ-carotene desaturase (*ZDS*) (Fiore et al., 2006; Dall'Osto et al., 2007). The *ZDS* product is then isomerized by carotene isomerase (*CrtISO*) to form all-*trans* lycopene (Park et al., 2002). Lycopene can then be channeled along two distinct pathways. The β-cyclase (*LCYB*) enzyme converts lycopene into β-carotene, while the dual action of ε-cyclase (*LCYE*) and *LCYB* results in the formation of α-carotene. The α- and β- carotenes can then be hydroxylated to form α- and β- branch xanthophylls which are essential components of the photosynthetic apparatus in higher plants where they function in photosystem assembly, light harvesting and photoprotection (Park et al., 2002). The *CYP97A* (*LUT5*) and *CYP97C* (*LUT1*) enzymes function to hydroxylate the β- and ε- ring of α-carotene, respectively, to form lutein, which is the most abundant carotenoid in plant photosynthetic tissues (von Lintig et al., 1997). The non-heme-di-iron β-carotene hydroxylases (*HYD*), and a heme-binding cytochrome P450 (*CYP97A/LUT5*), exhibit redundant activity in hydroxylating the β-rings of β-carotene to form zeaxanthin, which can subsequently be epoxydated by zeaxanthin epoxidase (*ZEP*/*ABA*1) to form violaxanthin (Quinlan et al., 2007). Violaxanthin can then be further processed to form

neoxanthin by a reaction requiring neoxanthin synthase (*NXS/ABA4*) (Dall'Osto et al., 2007). Substrate availability for the pathway is also affected by protein flux into competing pathways, such as those leading to the gibberellins and brassinosteroids hormones (Rodriguez-Villalon et al., 2009). Recent expression profiling of maize isoprenoid and the carotenoid pathway genes revealed multiple bottlenecks for biosynthesis of isoprenoids and carotenoids (Welsch et al., 2003). Induction of carotenoid accumulation demonstrated a positive correlation with transcript levels of *PSY1* (Li et al., 2008) and of *DXS3*, *DXR*, *HDR* and *GGPPS1* (isoprenoid pathway genes) at specific temporal stages of endosperm development. In contrast, both *CrtISO* and *ZEP*, enzymes involved in depletion of carotenoids to form abscisic acid, showed a negative correlation between transcript levels and seed carotenoid content (Welsch et al., 2003).

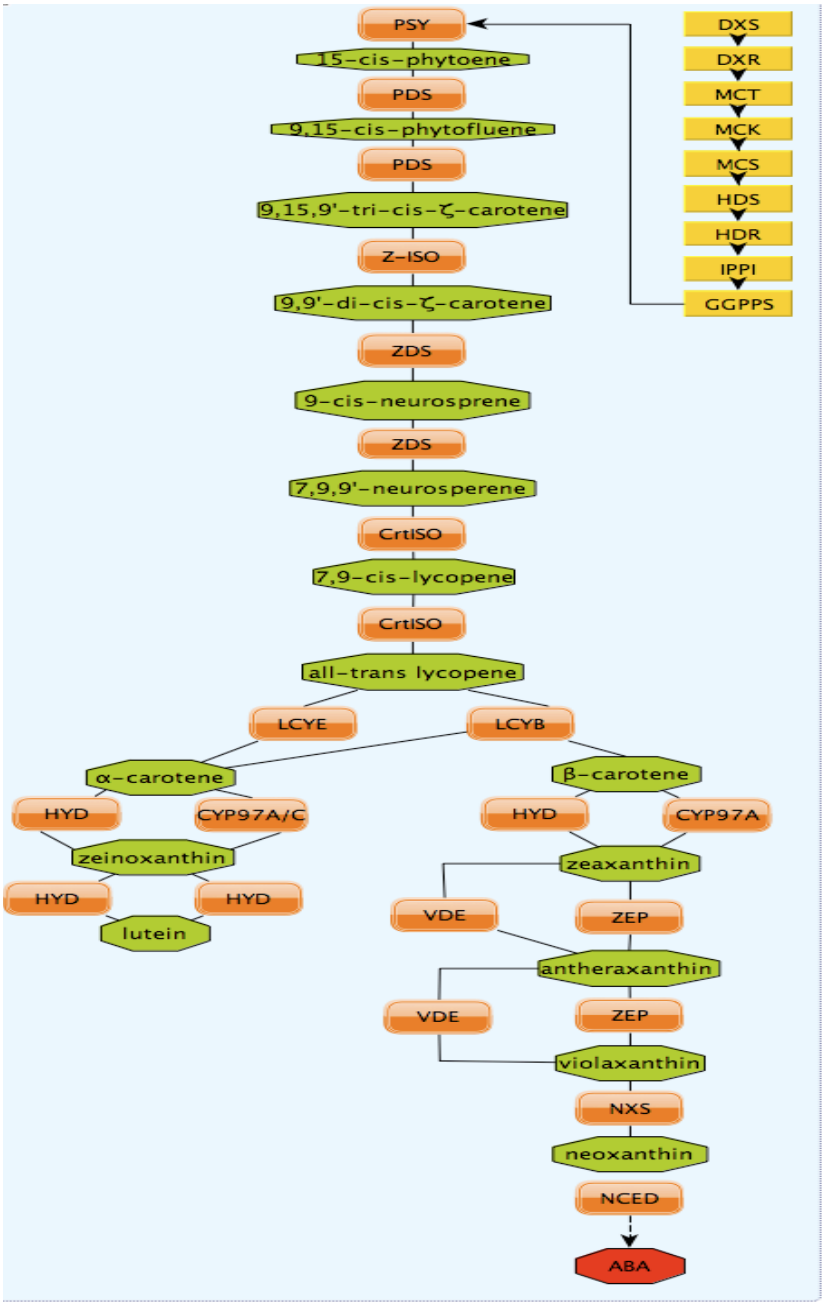


Figure 1.1 The carotenoid biosynthetic pathway in plants (orange nodes) and its upstream MEP pathway (yellow nodes). phytoene synthase (*PSY*); Phytoene desaturase (*PDS*); 15-*cis* ζ-carotene isomerase (*Z-ISO*); ζ-carotene desaturase (*ZDS*); carotene isomerase (*CrtISO*); β-cyclase (*LCYB*); ε-cyclase (*LCYE*); *CYP97A* and *CYP97C* enzymes function to hydroxylate the β- and ε- ring of α-carotene, respectively; non-heme-di-iron β-carotene hydroxylases (*HYD1* and *HYD2*); heme-binding cytochrome P450 (*CYP97*); zeaxanthin epoxidase (*ZEP*); neoxanthin synthase (*NXS*). Edges connect enzymes and metabolites (green nodes) to illustrate enzyme substrates. All shown enzyme functions have been verified biochemically.

1.3 *Other pathways intersecting with the carotenoid pathway*

Apart from feeding the carotenoid pathway, GGPP stands at a key metabolic junction, as it serves as a precursor for several other chloroplast-localized isoprenoid-derived pathways, such as tocopherol and gibberellins. GGPP also feeds into pathways that produce components that are critical to the photosynthetic apparatus, including plastoquinone (PQ) and phylloquinone (PhQ) (Von Lintig *et al*., 1997; Park *et al*., 2002) (see Figure 1 in Appendix II). PQ and PhQ function in the photosynthetic electron transfer reactions, essential to photo-morphogenesis and carotenoid biosynthesis (Carol and Kuntz, 2001). The PQ redox state is directly responsible for the desaturation reactions of *PDS* and *ZDS*. In this redox chain, electrons are transferred from phytoene and ζ-carotene (via *PDS* and *ZDS*) to create a PQ pool to molecular oxygen via plastid terminal oxidase (*PTOX*) (Carol and Kuntz, 2001). PQ reduction is catalyzed by the NADH dehydrogenase (Ndh) complex (Burrows et al., 1998; Endo et al., 2008), which is responsible for preventing over-reduction of stroma under stress conditions. *Ndh* mutant plants produce yellow-orange fruits, due to a carotenoid deficiency (Nashilevitz et al., 2010).

Carotenoid degradation via carotenoid cleavage dioxygenase (CCD and NCED) converts provitamin A and nonprovitamin A carotenoids to form apocarotenoids. In *Arabidopsis*, the cleavage enzymes are encoded by a complex nine-gene family (Tan et al., 2003). This family can be divided into two classes, based on their cleavage substrate specificity. One subfamily includes the CCD enzymes that cleave 9, 10 double bonds, while the second includes the NCEDs (9-cis epoxycarotenoid dioxygenases), that cleave 11, 12 double bonds. The CCD group is also responsible for producing the recently discovered carotenoid-derived strigolactone plant hormones. The strigolactones inhibit shoot branching, and can be secreted from plant roots to promote mycorrhizal fungi recruitment, beneficial for plant yield (Booker et al., 2004). In photosynthetic tissues, these apocarotenoids function as stress response alleviators (Koltai, 2011b).

Little is known about the regulation of the carotenoid biosynthesis pathway in crop species. Since the carotenoids are produced in different types of plastids, the related regulatory mechanisms are expected to be complex. Moreover, each plant species

contains a different number of gene copies. For example, PSY is encoded by a single copy gene in *Arabidopsis* and but three copies can be found in maize and other grasses (add the maize PSY3 paper as a reference). Thus, the regulatory mechanisms controlling this biosynthetic pathway are expected to differ between plant species.

1.4 Regulation of the carotenoid biosynthesis pathway

Metabolic engineering is the approach taken when striving to reduce vitamin A deficiency worldwide, by improving levels of provitamin A carotenoids in the endosperm of food staples such as corn, wheat, and rice (Wurtzel and Grotewold, 2006). Metabolic engineering of the carotene biosynthetic pathway has been achieved in "golden rice" and tomato, although the results have not always been predictable (see review by Giuliano et al., 2008; Vallabhaneni and Wurtzel, 2009). While the biochemistry of carotenogenesis has been extensively studied, little is known about the regulation of carotenogenic gene expression in higher plants (Wurtzel and Grotewold, 2006; Cuttriss, 2011). In order to manipulate crop grass endosperms to obtain increased carotenoid accumulation, it is essential to uncover the spatial and temporal behavior of carotenoid biosynthesis-related genes (CBRG) governing carotenoid accumulation (Li et al., 2008a; Cuttriss, 2011). Investigations of important control points of the pathway have been conducted in our lab using different maize inbred lines (Harjes et al., 2008; Vallabhaneni et al., 2009; Vallabhaneni and Wurtzel, 2009).

Very little is known with respect to transcription factors regulating the carotenoid pathway. A study done by Welsch and co-workers identified the *RAP2.2* transcription factor that binds to the ATCTA *cis* element of *PSY* and *PDS* promoters. However, when *RAP2.2* was over-expressed, transcription of the above genes was decreased, leading to reduced levels of carotenes in callus and root tissue (Welsch et al., 2007).

Quantitive Trait Loci (QTL), studies with recombinant inbred lines have pointed to more than 30 locations associated with carotenoid composition and accumulation (Wong et al., 2004; Chander et al., 2007). The number of identified QTLs is larger than the number of the known structural genes encoding for carotenoid pathway genes (Wurtzel, 2004; Cuttriss, 2011). This fact suggests that there are additional genes that might

participate in regulating carotenoid production. Little is known about the localization of carotenoid biosynthetic enzymes and their post-transcriptional. There is established data that supports the existence of post-transcriptional regulation of the level of the upstream MEP pathway proteins: *DXS*. Since the downstream products of *DXS* are extremely important for the plants physiological development, this regulation appears to be conserved among plants (Cordoba et al., 2009). Further deep understanding of protein import and interactions within the plastid will enable targeted manipulation of biosynthesis and more effective breeding strategies. Furthermore, understanding of allelic differences and recognizing new points of control in the carotenoid pathway can also be the goal of breeding plans for increasing carotenoids accumulation in crop plants. Carotenoids sequestration is another aspect of the pathway that could be targeted in order to manipulate carotenoids levels in plants. For example the *Or* gene which been shown to control by inducing chromoplasts formation which creates a metabolic sink for carotenoids deposit and sequestration (Zhou et al., 2008).

1.5 Systems biology of the carotenoid biosynthetic pathway

This research dissertation takes systems level approach, to enhance our understanding of the carotenoid biosynthesis pathway. Such techniques are expected to support development of strategies for 'predictive metabolic engineering' of the pathway (Wurtzel and Grotewold, 2006). While systems biology exploits the enormous amount of available high-throughput data, appropriate analysis must be performed, leading to educated predictions with regards to cellular biological processes. Systems biology enables the integration of different types of data including *in silico* transcriptomics, co-expression correlation, metabolomics, proteomics and phylogenetic information in order to develop hypotheses with statistically sound robustness.

Although several studies in recent years have developed excellent approaches and tools that provide critical insight into basic principles of biological networks (Atias et al., 2009; Mutwil et al., 2009; Mutwil et al., 2011), a need for a model that relates new genes to specific biological pathways in plants, still exists. In this dissertation, I will present my research of systems biology of carotenoid biosynthesis in the most studied plant model system, *A. thaliana*. To maximize the potential of carotenoid biosynthesis and

accumulation in plants, I herein suggest the use of various large-scale analyses to predict the genes that participate in or regulate the carotenoid pathway.

This work begins with an inquisitive look at the correlation between co-expressed genes along the carotenoid biosynthetic pathway. I then zoom out of the carotenoid pathway to study the expression of the carotenoid biosynthesis related genes (CBRG) (Supplementary Table 1.5), in relation to the rest of the *Arabidopsis* genome (see Chapter 2). I also probed public microarray data sets that induced differential expression of the CBRG, with focus on the carotenoids most co-expressed genes (see Chapter 2). In addition to the general co-expression network delineated in this work, I have also sketched co-expression networks describing the correlation between CBRG expression and that of sub-gene-groups, such as known transcription factor-encoding genes (see Chapter 2).

A collaborative effort with the lab of Dr. Chris Gehring (The King Abdullah University of Science and Technology, Saudi Arabia) yielded a publication (Meier et al., 2011) that describes the sequential use of freely available databases to explore the regulation of carotenoid biosynthesis in Arabidopsis during chloroplast development. The findings described in this publication suggest that coordinated transcriptional regulation of genes along the isoprenoid-related biosynthesis pathways play a major role in coordinating the synthesis of functionally related, chloroplast-localized isoprenoid-derived compounds. (see Appendix II).

Knowledge about a certain metabolic pathway covers only tiny portion of genes that play a role in the actual biological processes of the cell. My initial objective was to assign genes to the carotenoid biosynthetic pathway, and to address the information gaps common in many metabolic pathways. In a collaborative project with the laboratory of Professor Ron Shamir (Tel Aviv University, Israel), I aimed to develop a generic systems biology model for extending metabolic pathways. A model was developed to integrate several types of high-throughput data to optimize candidate gene ranking in an effort to best define associated genes for a specific studied pathway. The candidate ranking was achieved by using an iterative algorithm that is built on implementation of machine learning techniques (see Chapter 3).

I choose to focus on the most resource-rich plant model system – *Arabidopsis thaliana,* to develop a systems-level approach to biosynthetic pathway delineation. It is very easy to justify this choice, though *Arabidopsis* is neither an edible plant nor a target crop for metabolic engineering. The most high-throughput data on plants is available for *Arabidopsis*, in the form of thousands of microarray experiments, co-expression data, metabolic networks and protein-protein interaction maps. *Arabidopsis*-based network models are expected to provide information translatable to other systems, such as rice and maize, which bear greater economical and nutritional value (Mutwil et al., 2011).

Chapter 2. Systems-level computational approaches to reveal gene co-expression networks for the carotenoid biosynthetic pathway

2.1 Motivation

Eukaryotic cellular processes require the participation of multiple gene products, and co-expression of large gene sets responsive to specific stimuli (Eisen et al., 1998; Tamayo et al., 1999; Tavazoie et al., 1999). Similarly, metabolic pathways, which often include gene co-expression clusters, require such multi-gene participation. Collectively, these studies suggest that the coordinated transcription of functionally related genes plays a major role in synchronizing cellular responses and coordination of functionally related metabolic pathways.

Arabidopsis thaliana was the first plant species to have its complete genome sequenced and has avaliable data for thousands of microarray experiments. Therefore, it presents an ideal model for studying global transcriptional responses. These data summarize genome-wide transcriptional responses to a broad range of experimental conditions that encompass developmental stages, responses to stress, chemicals and hormones and mutant behavior. In addition, a plethora of analysis tools are available for the *Arabidopsis* model and can aid in identifying modules of co-expressing genes. Furthermore, genome-wide sequence data allows for analysis of promoter regulatory regions and for identification of putative regulatory elements.

In an attempt to extend our understanding of regulation of carotenogenesis and its coordination with the synthesis of other components of the photosynthetic apparatus in higher plants, I performed a genome-wide co-expression analysis using *PSY* as the key reference gene. *PSY* was selected as the driver gene for this analysis as it is the first dedicated enzyme of carotenogenesis and its transcriptional regulation has been described to be a major driving force for carotenoid production (Toledo-Ortiz et al., 2010). Identification of genes highly co-expressed with *PSY* will highlight genes that play key roles in carotenogenesis or in the synthesis of other functionally related compounds.

In this chapter, I propose the sequential use of computational-based tools to integrate profiles of promoters responsible for the expression of related genes, to

determine gene co-expression correlations and to analyze data collected from large-scale screening of conditioned microarray experiments.

2.2 Analysis of the PSY promoter in Zea mays

At early stages of my work I participated in an effort to uncover the rate-controlling steps limiting predictability of metabolic engineering in cereal crops in Poaceae (the Grass family). Functionalization of the first committed carotenoid biosynthesis step, which is mediated by phytoene synthase (*PSY*), was studied. In the grasses, *PSY* is encoded by three genes. I analyzed promoter elements of maize *PSY1* in comparison to those of grass orthologs (sorghum and rice), in efforts to identify the elements regulating different expression patterns of *PSY1* across alleles and species (Figure 2.1).

Both maize and rice *PSY1* promoters shared *cis*-acting element arrangements, with the exception of a few transposon insertions in the maize *PSY1* (*Y1*) promoter. The *ins2* insertion, at 300bp upstream to the transcription start site in maize *PSY1* promoters, has been associated with yellow endosperm (i.e carotenoid accumulation) (Buckner et al., 2006; Li et al., 2009b). Promoter analysis demonstrated that this insertion pushed back cis-acting elements in the maize *PSY1* promoter, leading to reduced *PSY1* induction by light. These results suggest that the yellow *Y1* is a gain of function mutation, which came at the expense of photoregulation in green photosynthetic tissue. Therefore *PSY1* regulation in yellow endosperm maize differs from *PSY1* regulation in the white endosperm maize progenitor, teosinte. In contrast, the white endosperm *y1* allele has maintained photoregulation in green tissue (Li et al., 2009a; Li et al., 2009b)(Figure 2.1).

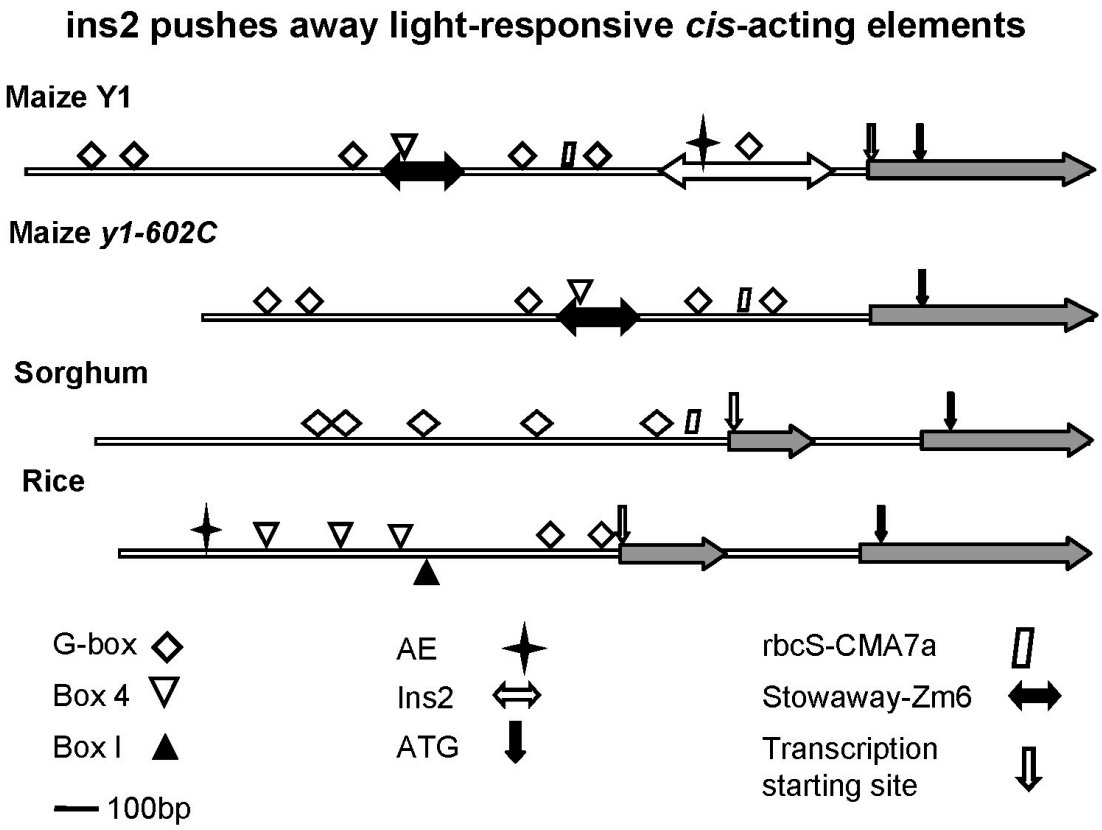


Figure 2.1. *ins2* transposon in maize *Y1* push back light-responsive *cis*-acting elements. Light-responsive cis-acting elements within maize *Y1* and *y1-602C* allele, sorghum and rice *PSY1* 5' upstream regions were predicted with PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html). Transposon *ins2* inserts 300bp upstream of transcript starting site in maize *Y1* but not in allele *y1-602;* AE – ABA Element. Adopted from Li et al., 2009b; see Appendix I.

2.3 Gene expression along a given pathway must be in synchrony if the pathway is regulated on a global level

Since the first committed gene of the *PSY* pathway is considered the rate-limiting step of carotenoid biosynthesis, I applied it as the driving gene in expression correlation analyses against the entire genome (see supplementary Table 2.3 for the top 50 genes co-expressed with *PSY*). Transcript levels averaged across 322 stimuli and developmental stages identified *LCYB* as the CBRG most tightly correlating with *PSY* expression. The correlation for the entire genome was then plotted against those of *PSY* and *LCYB* (Figure 2.2). Group A genes (Figure 2.2, red dots) represent isoprenoid and carotenoid pathway genes that strongly correlated with *PSY* and *LCYB* expression. These results suggest the presence of global transcriptional regulators that control co-expression of the CBRG.

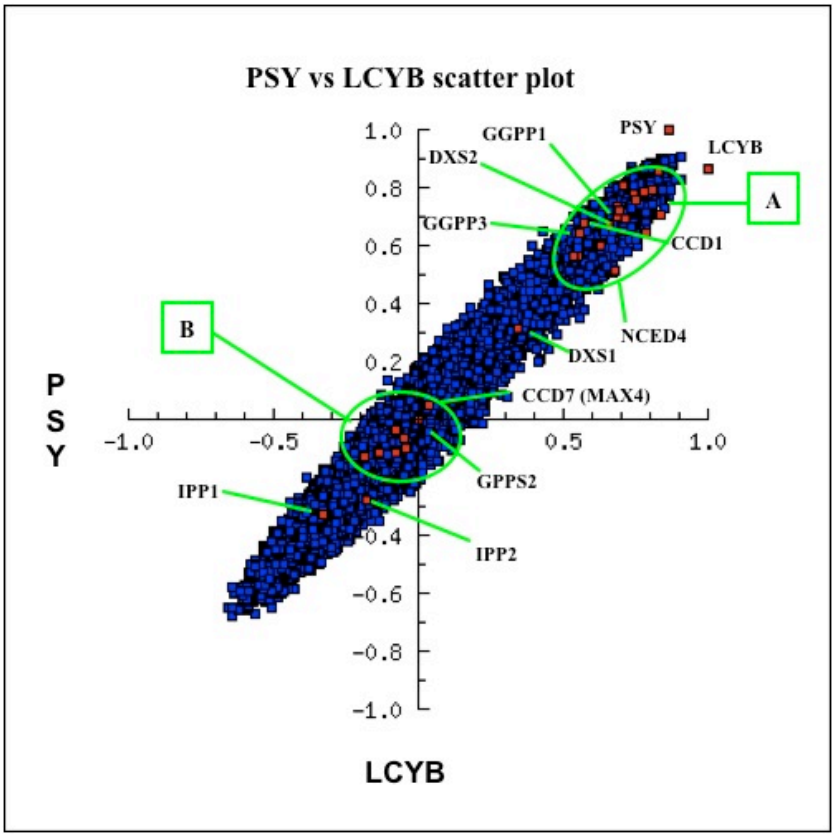


Figure 2.2. **Co-correlation scatter plot illustrating the expression of all Arabidopsis genes relative to *PSY* and *LCYB*** expression. The red dots represent the carotenoid associated genes listed in Table 1. Select genes relating to carotenoid biosynthesis are labeled. Co-expression levels to *PSY* are in Table 2.1

Table 2.1. Pearson Co-expression correlation values of genes associated with carotenoid biosynthesis that are marked in Figure 2.2.

Genes positively correlating with PSY and LCYB expression (area A)		**Genes not correlating with PSY and LCYB expression (area B)**	
LYCOPENE BETA CYCLASE (LCYB) AT3G10230	0.86	NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 5 (NCED5) AT1G30100	0.06
LYCOPENE EPSILON CYCLASE (LCYE) AT5G57030	0.81	CAROTENOID CLEAVAGE DIOXYGENASE 7 (CCD7) AT2G44990	0.05
PHYTOENE DESATURASE (PDS) AT4G14210	0.79	NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 9 (NCED9);NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 9 (NCED9) AT1G78390	-0.1
2C-METHYL-D-ERYTHRITOL 2,4-CYCLODIPHOSPHATE SYNTHASE (MCS) AT1G63970	0.78	CAROTENOID CLEAVAGE DIOXYGENASE 8 (CCD8);MORE AXILLARY BRANCHING 4 (MAX4); (CCD8)	-0.11
4-HYDROXY-3-METHYLBUT-2-ENYL DIPHOSPHATE SYNTHASE (HDS) AT5G60600	0.78	NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 2 (NCED2);NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 2 (NCED2) AT4G18350	-0.12
1-DEOXY-D-XYLULOSE 5-PHOSPHATE REDUCTOISOMERASE (DXR) AT5G62790	0.77	NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 3 (NCED3) AT3G14440	-0.13
CAROTENOID ISOMERASE (CRTISO) AT1G06820	0.72		
ZEAXANTHIN EPOXIDASE (ZEP) AT5G67030	0.71		
4-(CYTIDINE 5'-PHOSPHO)-2-C-METHYL-D-ERITHRITOL KINASE (MCK) AT2G26930	0.70		
1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE 2 (DXS)AT4G15560	0.69		
CYTOCHROME P450-TYPE MONOOXYGENASE 97A3 (CYP97A) AT1G31800	0.68		
LUTEIN DEFICIENT 1 (LUT1);CYTOCHROME P450 97C1 (CYP97C) AT3G53130	0.67		

4-HYDROXY-3-METHYLBUT-2-ENYL DIPHOSPHATE REDUCTASE (HDR) AT4G34350	0.64
BETA CAROTENOID HYDROXYLASE 1 (HYD1) AT4G25700	0.60
BETA CAROTENOID HYDROXYLASE 2 (HYD2) AT5G52570	0.60
2-C-METHYL-D-ERYTHRITOL 4-PHOSPHATE CYTIDYLTRANSFERASE (MCT) AT2G02500	0.57
VIOLAXANTHIN DE-EPOXIDASE 1 (VDE) AT1G08550	0.56

2.4 Co-expression correlation network of the CBRG

The Arabidopsis Co-expression Tool (ACT) (http://www.arabidopsis.leeds.ac.uk/) (Manfield et al., 2006) was used in order to identify *Arabidopsis* genes that are most co-expressed with genes encoding enzymes along the carotenoid biosynthetic pathway. This tool uses hybridization signal intensities from microarray experiments to calculate a Pearson correlation coefficient (r-value), which is a scale-invariant measure of expression similarity. The analysis was performed across all the 322 available Ath1 22K microarrays in the NASC/GARNet dataset (see methods section in Meier et al., 2011; Appendix II). All expression microarray information in the dataset was collected after standardized labeling, normalization against background, standardization and analysis, thus providing homogeneous and comparable experimental data. The data contain probe sets that recognize 21,891 *Arabidopsis* genes. The arrays included in this analysis covered a broad range of experimental samples, such as various tissue types, developmental stages, mutants and abiotic and biotic treatments.

The Arabidopsis co-expression correlation values were downloaded from the ACT database (http://www.arabidopsis.leeds.ac.uk/). I aimed to collect co-expression correlation files for CBRG genes and their top co-expressed genes. This required computational solution. To this end, I devised a web-crawling RUBY script (see 'ACRcrawler.rb' in methods section, 2.8.1) designed to retrieve all files containing the r-values of the each *Arabidopsis* gene relative to each of the CBRG and to the upstream MEP pathway genes. The r-values of the whole genome relative to other gene groups, such as the downstream apocarotenoids cleavage genes including the nine-cis-epocaratenoid dioxygenase family (NCEDs) and carotenoid cleavage enzymes (CCDs), were also extracted. Next a PERL script (see 'grabRowsByRValueAndTFs.pl' in methods section, 2.8.1) was used to parse the data files, using various r-value thresholds, and to extract the genes that most highly co-expressed with the CBRG. These gene lists (see Table 2.2) were then used to build the CBRG co-expression network using the open source tool, Cytoscape (http://www.cytoscape.org/) (Shannon et al., 2003).

Initially the CBRG co-expression network (r-value threshold > 0.7 ; chosen semi-arbitrary, based on multiple observations of co-expression networks) contained 1463

nodes (genes) and 9866 edges (a co-expression relationship). The large and complex network was then trimmed down by defining a higher r-value threshold ($\geq$ 0.85) (see Figure 2.3 and see Table 2.2 for a full list of gene IDs in this network). The new sub-network contained only 264 nodes and 412 edges. Despite the stringent co-expression threshold, many of the carotenoid and MEP genes shared co-expressed genes, consistent with the global co-expression results obtained from the scatter plot representing whole genome co-expression in relation to *PSY* and *LCYB* (Figure 2.2).

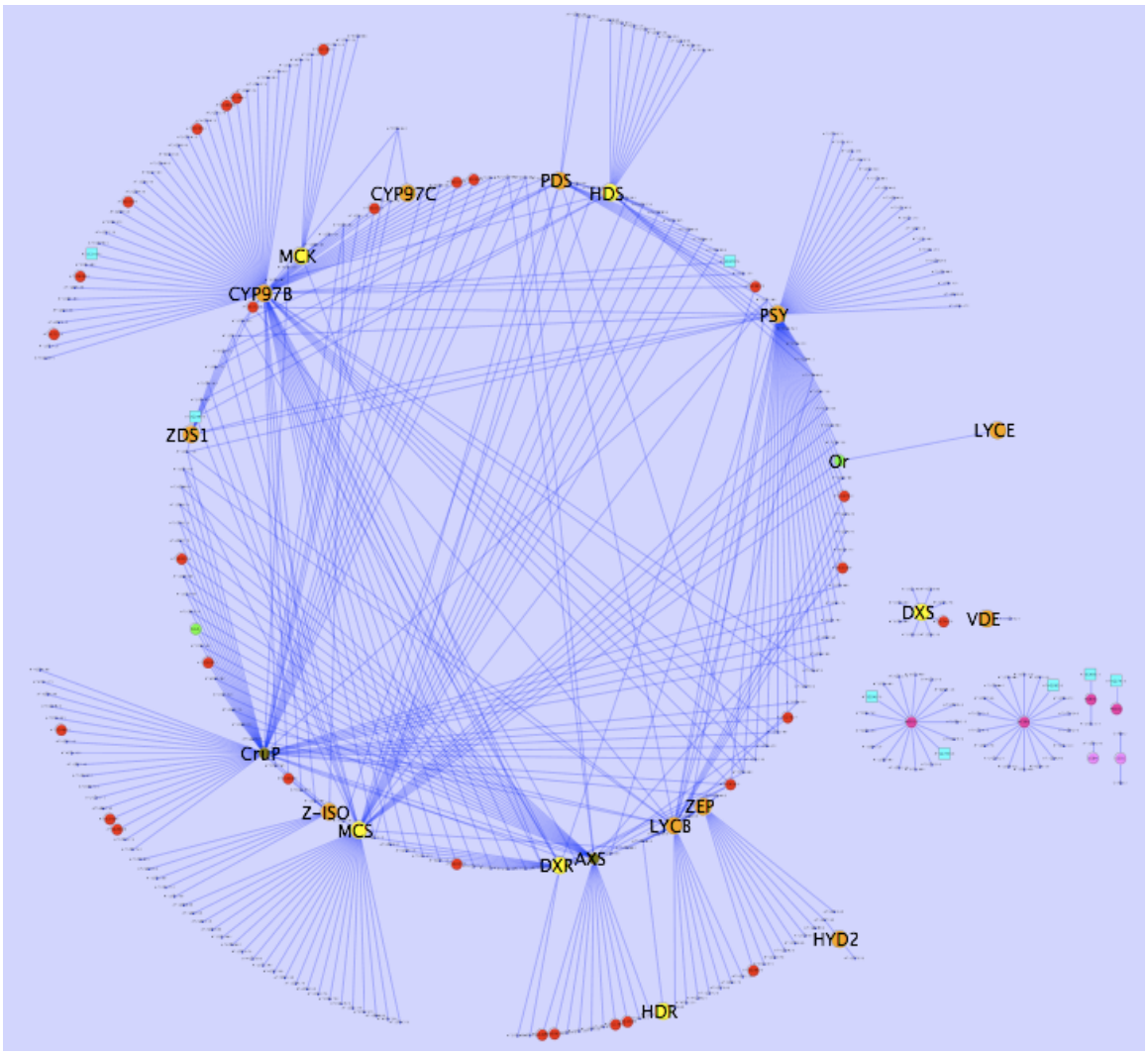


Figure 2.3. Co-expression network of the carotenoid biosynthesis related genes (CBRG) and genes most highly co-expressed with them (r-value > 0.85). The carotenoid genes nodes are colored in orange. MEP pathway gene nodes are colored in yellow. NCED gene nodes are colored in purple; Table 2.2 contains a full list of the genes participating in the above network. Nodes colored in red are predicted as candidates for the Carotenoid Core pathway based on my MORPH algorithm (see Chapter 3).

Table 2.2. The CBRG genes that appear in Figure 2.3 and their co-expressed genes with r-value > 0.85.

Carotenoid pathway gene name	**Carotenoid pathway gene AGI**	**Co-expressed genes with r-value > 0.85**	**Co-expressed gene annotation** (see section 2.8.1 and Supplementary Table 2.8.1).
PSY	AT5G17230	AT1G64680	expressed protein
		AT3G04790	ribose 5-phosphate isomerase-related similar to ribose-5-phosphate isomerase GI:18654317 from [*Spinacia oleracea*]
		AT3G54050	Encodes a chloroplastic fructose 1,6-bisphosphate phosphatase. also known as HCEF1 (High Cyclic Electron Flow 1)
		AT5G08650	GTP-binding protein LepA
		AT5G04140	glutamate synthase (GLU1) / ferredoxin-dependent glutamate synthase (Fd-GOGAT 1) identical to ferredoxin-dependent glutamate synthase precursor [*Arabidopsis thaliana*] GI:3869251
		AT1G01320	tetratricopeptide repeat (TPR)containing protein low similarity to SP|P46825 Kinesin light chain (KLC) {*Loligo pealeii*}; contains Pfam profile PF00515: TPR Domain
		AT1G55480	expressed protein
		AT5G43750	expressed protein
		AT1G45474	chlorophyll A-B binding protein
		AT1G16880	uridylyltransferase-related similar to [Protein-PII] uridylyltransferase (PII uridylyl- transferase) (Uridylyl removing enzyme) (UTase)(SP:Q9AC53) [*Caulobacter crescentus*]
		AT5G44650	expressed protein
		AT1G73110	ribulose bisphosphate carboxylase/oxygenase activase
		AT1G14345	expressed protein contains one transmembrane domain
		AT1G54500	rubredoxin family protein similar to SP|P00270 Rubredoxin (Rd) {*Desulfovibrio gigas*}; contains Pfam profile PF00301: Rubredoxin
		AT5G17170	rubredoxin family protein contains Pfam profile PF00301: Rubredoxin
		AT1G26220	GCN5-related N-acetyltransferase (GNAT) family protein low similarity to SP|P09453 Ribosomal-protein-alanine acetyltransferase (EC 2.3.1.128) {*Escherichia coli*}; contains Pfam profile PF00583: acetyltransferase
		AT5G36790	phosphoglycolate phosphatase
		AT4G34090	expressed protein
		AT2G04039	expressed protein
		AT3G11950	UbiA prenyltransferase family protein contains Pfam profile PF01040: UbiA prenyltransferase family
		AT3G55330	photosystem II reaction center PsbP family protein contains Pfam profile PF01789: PsbP

		AT3G04870	zeta-carotene desaturase (ZDS1) / carotene 7
		AT4G10300	expressed protein
		AT5G23120	photosystem II stability/assembly factor
		AT3G26570	phosphate transporter family protein contains Pfam profile: PF01384 phosphate transporter family
		AT1G73060	expressed protein
		AT1G76450	oxygen-evolving complex-related SP:Q9S720; contains a PsbP domain
		AT1G18060	expressed protein
		AT1G05140	membrane-associated zinc metalloprotease
		AT1G62750	elongation factor Tu family protein similar to elongation factor G SP:P34811 [*Glycine max* (Soybean)]
		AT2G20890	expressed protein
		AT3G10230	lycopene beta cyclase (LYC) identical to lycopene beta cyclase GI:1399183|GB:AAB53337 [*Arabidopsis thaliana*]
		AT1G32470	glycine cleavage system H protein
		AT1G32080	membrane protein
		AT1G17220	translation initiation factor IF-2
		AT1G50320	thioredoxin x nearly identical to thioredoxin x GB:AAF15952 GI:6539616 from [*Arabidopsis thaliana*]
		AT2G21330	fructose-bisphosphate aldolase
		AT2G34860	chaperone protein dnaJ-related contains Pfam PF00684 : DnaJ central domain (4 repeats); similar to Chaperone protein dnaJ (Heat shock protein 40) (SP:Q9UXR9) {*Methanosarcina thermophila*}
		AT3G63410	chloroplast inner envelope membrane protein
		AT5G58260	expressed protein
		AT3G14415	(S)2-hydroxy-acid oxidase
		AT5G36700	phosphoglycolate phosphatase
		AT3G19490	sodium hydrogen antiporter
		AT1G42970	gLCYEraldehyde-3-phosphate dehydrogenase B
		AT1G09340	expressed protein
		AT1G07010	calcineurin-like phosphoesterase family protein contains Pfam profile: PF00149 calcineurin-like phosphoesterase
		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)
		AT4G01800	preprotein translocase secA subunit
		AT1G77090	thylakoid lumenal 29.8 kDa protein identical to SP|O49292|TL30ARATH (*Arabidopsis thaliana*);contains a PsbP domain AF370571; SIMILAR TO GI:13926195F22K20.16

		AT1G15980	expressed protein
		AT3G19480	D-3-phosphogLCYErate dehydrogenase
		AT5G42310	pentatricopeptide (PPR) repeat-containing protein contains Pfam profile PF01535: PPR repeat
		AT1G73070	leucine-rich repeat family protein contains leucine rich-repeat (LRR) domains Pfam:PF00560
		AT1G11860	aminomethyltransferase
		AT3G14420	(S)2-hydroxy-acid oxidase
PDS	AT4G14210	AT1G29700	expressed protein
		AT2G30960	expressed protein
		AT5G38520	hydrolase
		AT2G30950	FtsH protease (VAR2) identical to zinc dependent protease VAR2 GI:7650138 from [*Arabidopsis thaliana*]
		AT3G51510	expressed protein
		AT2G21960	expressed protein
		AT5G51110	expressed protein
		AT2G20890	expressed protein
		AT5G58330	malate dehydrogenase [NADP]
		AT5G27560	expressed protein hypothetical protein slr1702 - Synechocystis sp.
		AT3G56010	expressed protein
		AT4G36530	Hydrolase
		AT3G08010	expressed protein
		AT1G64770	expressed protein
		AT1G77090	thylakoid lumenal 29.8 kDa protein identical to SP|O49292|TL30ARATH (*Arabidopsis thaliana*);contains a PsbP domain AF370571; SIMILAR TO GI:13926195F22K20.16
		AT3G07670	SET domain-containing protein similar to ribulose-1
		AT2G42130	expressed protein contains weak hit to Pfam PF04755: PAPfibrillin
Z-ISO	AT1G10830	AT2G43560	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family protein identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 2
		AT1G11750	ATP-dependent Clp protease proteolytic subunit (ClpP) identical to ATP-dependent Clp protease proteolytic subunit GI:2827888 from [*Arabidopsis thaliana*]; contains Pfam profile PF00574: Clp protease; contains TIGRfam profile TIGR00493: ATP-dependent Clp protease
		AT5G06290	2-cys peroxiredoxin
		AT5G45390	ATP-dependent Clp protease proteolytic subunit (ClpP4) identical to nClpP4 GI:5360593 from [*Arabidopsis thaliana*]
		AT5G52970	thylakoid lumen 15.0 kDa protein SP:Q9LVV5; similar to unknown protein (pir |S77462)

		AT5G45680	FK506-binding protein 1 (FKBP13) identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 3
		AT5G42765	expressed protein
		AT1G67700	expressed protein
		AT3G25805	expressed protein
		AT1G01080	33 kDa ribonucleoprotein
		AT5G27560	expressed protein hypothetical protein slr1702 - Synechocystis sp.
ZDS	AT3G04870	AT5G17230	phytoene synthase
		AT2G24820	Rieske [2Fe-2S] domain-containing protein similar to Rieske iron-sulfur protein Tic55 from *Pisum sativum* [gi:2764524]; contains Pfam PF00355 Rieske [2Fe-2S] domain
		AT5G53580	aldo/keto reductase family protein contains Pfam profile PF00248: oxidoreductase
		AT1G05140	membrane-associated zinc metalloprotease
		AT1G62750	elongation factor Tu family protein similar to elongation factor G SP:P34811 [*Glycine max* (Soybean)]
		AT5G58330	malate dehydrogenase [NADP]
		AT1G17220	translation initiation factor IF-2
		AT3G56010	expressed protein
		AT1G64770	expressed protein
		AT2G24830	zinc finger (CCCH-type) family protein / D111/G-patch domain-containing protein contains Pfam profiles PF01585: G-patch domain
CRTISO1	AT1G06820	*	
LCYB	AT3G10230	AT4G35250	vestitone reductase-related low similarity to vestitone reductase [*Medicago sativa* subsp. sativa] GI:973249
		AT1G64680	expressed protein
		AT5G08650	GTP-binding protein LepA
		AT2G30170	expressed protein
		AT1G14345	expressed protein contains one transmembrane domain
		AT5G17230	phytoene synthase
		AT5G08050	expressed protein predicted protein
		AT1G26220	GCN5-related N-acetyltransferase (GNAT) family protein low similarity to SP|P09453 Ribosomal-protein-alanine acetyltransferase (EC 2.3.1.128) {*Escherichia coli*}; contains Pfam profile PF00583: acetyltransferase
		AT2G04039	expressed protein
		AT5G02120	thylakoid membrane one helix protein (OHP) identical to one helix protein GI:3283057 from [*Arabidopsis thaliana*]
		AT3G54660	gluthatione reductase
		AT5G57960	GTP-binding family protein similar to SP|P25519 GTP-

			binding protein hflX {*Escherichia coli*}
		AT5G51110	expressed protein
		AT4G28025	expressed protein
		AT3G26570	phosphate transporter family protein contains Pfam profile: PF01384 phosphate transporter family
		AT5G38660	expressed protein similar to unknown protein (pir||S75762)
		AT1G18060	expressed protein
		AT1G62750	elongation factor Tu family protein similar to elongation factor G SP:P34811 [*Glycine max* (Soybean)]
		AT1G17220	translation initiation factor IF-2
		AT1G16720	expressed protein
		AT5G07020	proline-rich family protein
		AT1G07010	calcineurin-like phosphoesterase family protein contains Pfam profile: PF00149 calcineurin-like phosphoesterase
		AT1G77090	thylakoid lumenal 29.8 kDa protein identical to SP|O49292|TL30ARATH (*Arabidopsis thaliana*);contains a PsbP domain AF370571; SIMILAR TO GI:13926195F22K20.16
		AT5G42070	expressed protein similar to unknown protein (dbj|BAA92898.1)
		AT2G34460	flavin reductase-related low similarity to SP|P30043 Flavin reductase {*Homo sapiens*}
		AT5G64840	ABC transporter family protein
		AT2G26080	glycine dehydrogenase [decarboxylating]
LCYE	AT5G57030	AT1G15980	expressed protein
HYD1	AT4G25700	*	
HYD2	AT5G52570	AT1G07180	pyridine nucleotide-disulphide oxidoreductase family protein contains similarity to alternative NADH-dehydrogenase GI:3718005 from [*Yarrowia lipolytica*]
		AT1G17050	geranyl diphosphate synthase
CYP97A	AT1G31800	*	
CYP97B	AT4G15110	AT1G63610	expressed protein
		AT1G68590	plastid-specific 30S ribosomal protein 3
		AT4G01690	protoporphyrinogen oxidase (PPOX) identical to SP|P55826
		AT5G05740	peptidase M50 family protein / sterol-regulatory element binding protein (SREBP) site 2 protease family protein contains Pfam PF02163: Sterol-regulatory element binding protein (SREBP) site 2 protease
		AT3G55800	sedoheptulose-1
		AT1G32550	ferredoxin family protein similar to ferredoxin from *Synechocystis* sp. [GI:48019]; contains Pfam profile PF00111 2Fe-2S iron-sulfur cluster binding domain
		AT5G38520	Hydrolase
		AT5G57930	expressed protein

		AT4G31850	pentatricopeptide (PPR) repeat-containing protein contains Pfam profile PF01535: PPR repeat
		AT4G29060	elongation factor Ts family protein similar to SP|P35019 Elongation factor Ts (EF-Ts) {*Galdieria sulphuraria*}; contains Pfam profiles PF00627: UBA/TS-N domain
		AT4G18370	protease HhoA
		AT1G74730	expressed protein
		AT5G52100	dihydrodipicolinate reductase family protein weak similarity to dihydrodipicolinate reductase [*Corynebacterium glutamicum*] GI:311768; contains Pfam profiles PF01113: Dihydrodipicolinate reductase N-terminus
		AT5G44650	expressed protein
		AT5G06290	2-cys peroxiredoxin
		AT1G54500	rubredoxin family protein similar to SP|P00270 Rubredoxin (Rd) {*Desulfovibrio gigas*}; contains Pfam profile PF00301: Rubredoxin
		AT5G54600	50S ribosomal protein L24
		AT4G15510	photosystem II reaction center PsbP family protein contains PsbP domain PF01789; identical to SP:O23403 (*Arabidopsis thaliana*)
		AT1G78180	mitochondrial substrate carrier family protein contains Pfam profile: PF00153 mitochondrial carrier protein
		AT4G18480	magnesium-chelatase subunit chlI
		AT3G50685	expressed protein
		AT3G51510	expressed protein
		AT3G61870	expressed protein hypothetical protein - *Synechocystis* sp. (strain PCC 6803)
		AT1G20810	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family protein identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 1
		AT1G08520	magnesium-chelatase subunit chlD
		AT3G55330	photosystem II reaction center PsbP family protein contains Pfam profile PF01789: PsbP
		AT4G15110	cytochrome P450 97B3
		AT5G46580	pentatricopeptide (PPR) repeat-containing protein contains similarity to 67kD chloroplastic RNA-binding protein
		AT5G03940	signal recognition particle 54 kDa protein
		AT2G35410	33 kDa ribonucleoprotein
		AT1G50900	expressed protein
		AT5G30510	30S ribosomal protein S1
		AT4G17560	ribosomal protein L19 family protein similar to plastid ribosomal protein L19 precursor [*Spinacia oleracea*] gi|7582403|gb|AAF64312
		AT3G15110	expressed protein
		AT5G51110	expressed protein

		AT5G52970	thylakoid lumen 15.0 kDa protein SP:Q9LVV5; similar to unknown protein (pir |S77462)
		AT1G76450	oxygen-evolving complex-related SP:Q9S720; contains a PsbP domain
		AT1G14030	ribulose-1
		AT3G12780	phosphogLCYErate kinase
		AT2G37660	expressed protein
		AT5G45680	FK506-binding protein 1 (FKBP13) identical to Probable FKBP-type peptidyl-prolyl cis-trans isomerase 3
		AT2G03420	expressed protein
		AT2G20890	expressed protein
		AT5G58330	malate dehydrogenase [NADP]
		AT1G05190	ribosomal protein L6 family protein Similar to Mycobacterium RlpF (gb|Z84395). ESTs gb|T75785
		AT4G34190	stress enhanced protein 1 (SEP1) identical to stress enhanced protein 1 (SEP1) GI:7384978 from [*Arabidopsis thaliana*]
		AT5G42765	expressed protein
		AT4G24750	expressed protein
		AT3G51820	chlorophyll synthetase
		AT1G32080	membrane protein
		AT4G09010	L-ascorbate peroxidase
		AT3G23700	S1 RNA-binding domain-containing protein contains Pfam domain
		AT5G62840	phosphogLCYErate/bisphosphogLCYErate mutase family protein contains Pfam profile PF00300: phosphogLCYErate mutase family
		AT4G13670	peptidoglycan-binding domain-containing protein similar to spore cortex-lytic enzyme prepeptide (GI:1644192) [*Bacillus cereus*]; contains Pfam PF01471: Putative peptidoglycan binding domain; contains Pfam PF00684 : DnaJ central domain (4 repeats)
		AT3G52380	33 kDa ribonucleoprotein
		AT1G01080	33 kDa ribonucleoprotein
		AT1G71720	S1 RNA-binding domain-containing protein contains Pfam domain
		AT1G17220	translation initiation factor IF-2
		AT2G36990	RNA polymerase sigma subunit SigF (sigF) / sigma-like factor (SIG6) identical to RNA polymerase sigma subunit SigF [*Arabidopsis thaliana*] GI:7209640; contains Pfam profiles PF04545: Sigma-70
		AT2G27680	aldo/keto reductase family protein contains Pfam profile PF00248: oxidoreductase
		AT3G47650	bundle-sheath defective protein 2 family / bsd2 family similar to bundle sheath defective protein 2 [*Zea mays*]

			GI:4732091
		AT3G63410	chloroplast inner envelope membrane protein
		AT1G55370	expressed protein
		AT2G35370	glycine cleavage system H protein 1
		AT4G04350	leucyl-tRNA synthetase
		AT3G62030	peptidyl-prolyl cis-trans isomerase
		AT1G03630	protochlorophyllide reductase C
		AT1G70200	RNA recognition motif (RRM)containing protein contains INTERPRO:IPR000504 RNA-binding region RNP-1 (RNA recognition motif) domain
		AT3G16000	matrix-localized MAR DNA-binding protein-related similar to matrix-localized MAR DNA binding protein MFP1 GI:1771158 from [*Lycopersicon esculentum*]
		AT2G41680	thioredoxin reductase
		AT4G17600	lil3 protein identical to Lil3 protein [Arabidopsis thaliana] gi|4741966|gb|AAD28780
		AT5G11450	oxygen-evolving complex-related 23 kDa polypeptide of water-oxidizing complex of photosystem II
		AT3G07670	SET domain-containing protein similar to ribulose-1
		AT3G63490	ribosomal protein L1 family protein ribosomal protein L1
		AT3G01480	peptidyl-prolyl cis-trans isomerase
		AT3G56910	expressed protein
		AT4G24770	31 kDa ribonucleoprotein
		AT3G26060	peroxiredoxin Q
		AT2G42130	expressed protein contains weak hit to Pfam PF04755: PAPfibrillin
CYP97C	AT3G53130	AT5G62840	phosphogLCYErate/bisphosphogLCYErate mutase family protein contains Pfam profile PF00300: phosphogLCYErate mutase family
ZEP	AT5G67030	AT1G64860	RNA polymerase sigma subunit SigA (sigA) / sigma factor 1 (SIG1) identical to sigma factor SigA [*Arabidopsis thaliana*] GI:5478439
		AT5G35970	DNA-binding protein
		AT1G07180	pyridine nucleotide-disulphide oxidoreductase family protein contains similarity to alternative NADH-dehydrogenase GI:3718005 from [Yarrowia lipolytica]
		AT2G29650	inorganic phosphate transporter
		AT3G54660	gluthatione reductase
		AT3G21670	nitrate transporter (NTP3) nearly identical to nitrate transporter [*Arabidopsis thaliana*] GI:4490323; contains Pfam profile: PF00854 POT family
		AT4G02920	expressed protein
		AT1G18060	expressed protein
		AT1G79600	ABC1 family protein contains Pfam domain
		AT5G58870	FtsH protease

		AT2G34460	flavin reductase-related low similarity to SP|P30043 Flavin reductase {*Homo sapiens*}
		AT3G01060	expressed protein
		AT5G64840	ABC transporter family protein
VDE	AT1G08550	AT1G20020	ferredoxin-NADP(+) reductase

* Annotations are based on the TAIR.7 release since the ACT tool was not updated since 2009 see Supplementary Table 2.8.1 for full list of annotations for all probes represented in the ACT files.

In order to measure the expression levels of the genes that are highly expressed with *PSY,* three groups of genes co-expressed with *PSY* were categorized. The first group included the top 25 genes demonstrating expression correlating with that of *PSY* (*PSY 25*) (first 25 genes in SupplementaryTable2.3.1), with a Pearson correlation coefficient (r-value) ranging from 0.91 to 0.87. The second group comprised the top 50 gene probes that most correlated with *PSY* expression (*PSY 5*0) (first 50 genes in SupplementaryTable2.3.1). The last group included 1108 probes (~ 4.3% of the whole genome) with r-values $\geq$0.6 (*PSY 0.6*). In addition ~600 probes, 2.6% of the entire genome, had an r-value $\geq$0.7 (see Supplementary Table 2.3). The resulting distribution of genes co-expressed with *PSY* indicates that *PSY* is co-expressed with only small percentage of select genes in the *Arabidopsis* genome.

Genes highly co-expressed with *PSY* were hypothesized to bear associated functional roles that could be further delineated by displaying co-expression of these gene groups while probing numerous specific-stimuli microarray datasets. To test this hypothesis, the *PSY 25* and *PSY 0.6* probe lists were extracted. In addition, a list of 1000 genes that do not display co-expression with *PSY* (-0.1< r-value < 0.1) (averaged non correlated genes; *av non correl*) was generated. I then plotted the fold change in *PSY* transcripts levels, and the average fold change of *PSY 25*, *PSY 0.6* and *av non correl*, as recorded in various microarray experiments (Figure 2.4). The experiments providing the data tested responses to different light regimes (Figure 2.4A), and biotic and abiotic stresses (Figure 2.4B and C, respectively), as well as a time series experiment evaluating responses to continuous white light illumination (Figure 2.4D). In Appendix II (see Figure 2.4), a time-course analysis of gene responsiveness to osmotic stress in shoot and root tissues is presented. This analysis revealed some interesting tissue-specific expression response patterns. Thus, I also chose to examine the changes of transcript levels among the 50 probes that were most co-expressed with *PSY* (*PSY50)*, in the osmotic stress experiment (Figure 2.5). Although the average expression of all the 50 co-expressed genes displayed milder expression than *PSY*, trends in transcript level changes remained very similar, supporting the proposed hypothesis that the global co-expression of genes with PSY is also reflected in small scale specific microarray experiments.

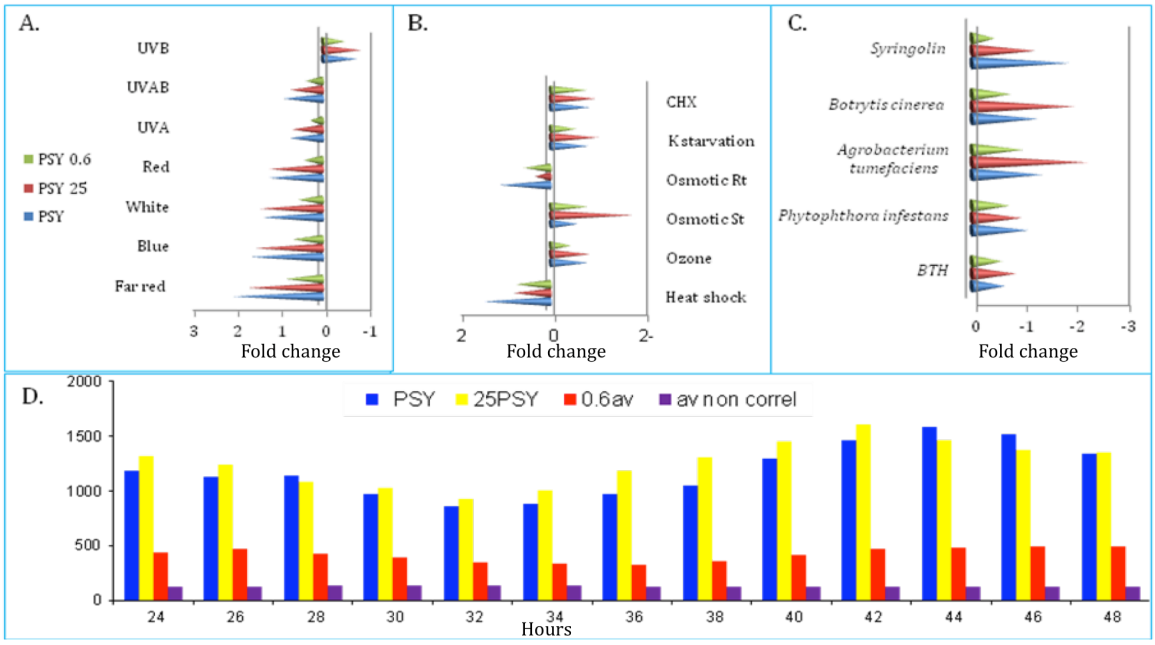


Figure 2.4 Transcript levels of *PSY*, as measured in various microarray experiments, in comparison to the average transcript levels of the 25 genes whose expression most correlated with that of *PSY* (PSY 25) and in comparison to average transcript levels of the 1108 genes with r-value $\geq$ 0.6 (PSY 0.6) extracted from the CBRG co-expression data from the Arabidopsis co-expression tool (ACT). (For full lists of genes see SupplementaryTable2.3.2.xlsx) **A.** Gene expression following plant challenge with various light regimes**. B.** Gene expression following plant challenge with abiotic stress conditions. CHX- Cycloheximide (inhibitor of protein biosynthesis); K – potassium; Osmotic Rt – osmotic stress in roots; Osmotic St – osmotic stress in Shoots; **C.** Gene expression following plant challenge with biotic stress conditions. Syringolin, the product of the activity of a mixed non-ribosomal peptide/polyketide synthetase, is secreted by *Pseudomonas syringae*; BTH- [benzo-(1,2,3)-thiadiazole-7-carbothioic acid S-methyl ester] from *Uromyces appendiculatus* **D.** Time-series analysis of the gene expression following plant challenge with continuous white light illumination. The X-axis units in A-C is fold change (log 2), and in D is hours of illumination. The Y-axis units in D are the raw detection intensity calls. '*av non correl*' – group of 1000 genes with (-0.1< r-value < 0.1) with *PSY*. Details of the microarray experimental conditions are presented in Appendix II in Supporting Information S3.

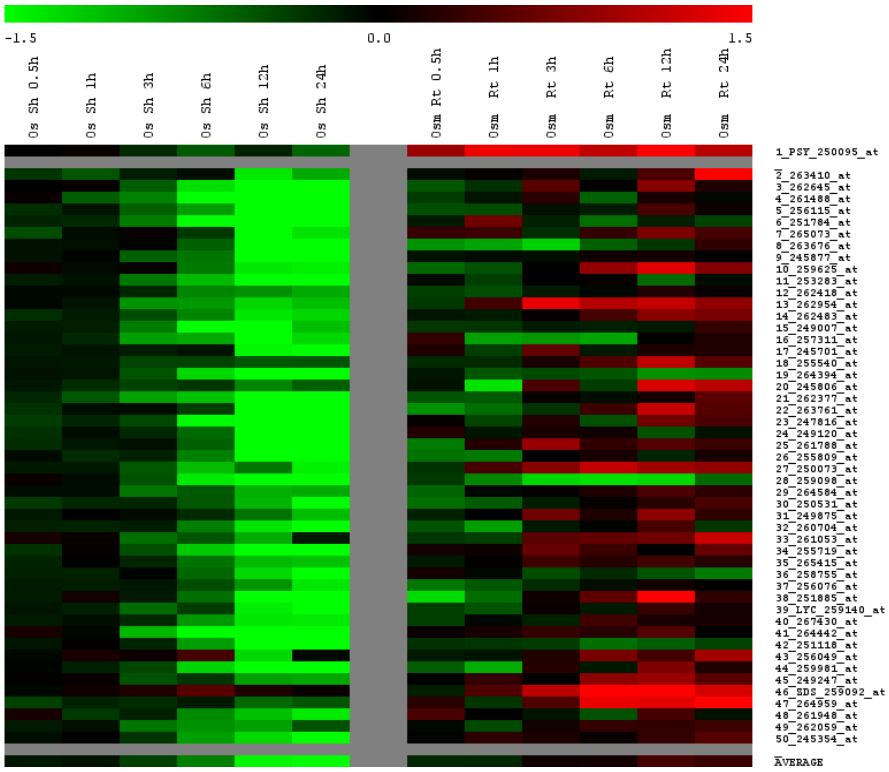


Figure 2.5. Time-course experiment illustrating the effect of osmotic stress on expression of the *PSY50* in root and shoot tissue. The numbers on the right represents the probes of the top 50 genes co-expressed with *PSY*. Fold-change (log2) in gene expression was measured in root and shoot tissue at the indicated time points following continuous osmotic stress application (mannitol) to root tissue (accession number: ME00327). Details of the microarray experimental conditions are presented in Appendix II, supporting Information S3 and http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3123201/?tool=pubmed). List of ATG IDs for the corresponding probes is in Supplementary Table2.3.1.xlsx. See Supplementary Table 2.4 for the raw data used to generate this heat map.

Because the co-expression correlation analysis is performed across data collected from multiple tissues and in response to a broad range of experimental conditions, co-expression analysis provides a general measure of co-expression patterns between CBRG and other genes in the genome. The topology of the co-expression network does reflect physiologically significant gene-gene relations by using different co-expression correlation coefficient thresholds (Obayashi and Kinoshita, 2009; Vandepoele et al., 2009).

2.5 Functional analysis of the CBRG co-expression correlation network

In order to determine if there was any statistically significant biological process associated with genes co-expressing with the CBRG expression correlated genes, a Gene Ontology (GO) analysis was performed using the Cytoscape plug-in "BiNGO" (version 2.42) (Maere et al., 2005). The CBRG network expression profile was compared to its expected frequency in the complete genome. The 264 genes whose expression was highly correlated with *CBRG* (Figure 2.3, r-value > 0.85) were selected for this analysis. The BiNGO tool uses Benjamini & Hochberg false discovery rate (FDR) correction accounting for multiple testing. FDR represents the proportion of false positive hits among all significant hypotheses, by controlling the expected proportion of null hypotheses that were mistakenly rejected. The GO terms analysis identified a number of significant enrichments in functional terms associated with the CBRG network (Supplementary Table 2.5). Among the biological processes that were significantly enriched were genes associated with the terms carotenoid (p-val = 1.07E-13) and tetraterpenoid metabolic (p-val = 1.07E-13) processes, photosynthesis (p-value = 1.50E-20), plastid organization (p-value = 2.43E-11) and chlorophyll biosynthesis (p-value = 1.04E-7). The significant enrichment for carotenoid and chlorophyll related terms in the CBRG co-expression network, provides support for the global regulatory mechanisms proposed to control expression of the CBRG and its related pathways, as implied by the CBRG co-expression network (Figure 2.3).

2.6 Identification of genes encoding transcription factors whose expression correlates with that of CBRG

Even in well-studied organisms, such as *Arabidopsis,* it is estimated that only a small percentage of the > 2300 predicted transcription factors (TFs) have been functionally characterized. Moreover the TFs identification is typically restricted to well characterized responses (Riechmann et al., 2000; Tompa et al., 2005). This biased analysis does not represent the true complexity of the regulatory networks in the context of competing metabolic pathways. To some extent, it is possible to identify a single TF that participates in the control of a whole biosynthetic pathway. For example, overexpression of *CRY2* (the blue light photoreceptor) in the tomato, increased flavonoid as well as carotenoid levels in leaves and fruits (Giliberto et al., 2005). Gantet and Memelink have used TFs as a metabolic engineering tool to manipulate biosynthetic pathways (Gantet and Memelink, 2002). In the present experiment, I defined co-expression correlation gene networks that highlight the TF genes whose expression most tightly correlates with that of the CBRG. The CBRG-TFs co-expression network was prepared by screening the carotenoid co-expression network for TF-encoding genes (list of 1923 genes downloaded from DATF -Database of Arabidopsis Transcription Factor - http://datf.cbi.pku.edu.cn/download/datf.id_locus), that were co-expressed with CBRG with r-value thresholds of $\geq$0.6, $\geq$0.7 and $\geq$0.8 (Figure 2.6) (see Supplementary Table 2.6 for a full list of the CBRG and MEP pathway genes and their best co-expressed TFs).

It is been generally assumed that transcription factors regulate gene expression via highly organized modules arranged in a hierarchic, pyramid-like design, controlled by a complex network of biological processes. In recent years, it had become increasingly clear that metabolism functions as a highly integrated network (Sweetlove et al., 2008). Therefore, specific TFs can operate on multiple regulatory levels, with different impacts on individual pathways. Thus, changes in the expression of a specific TF can influence flux between pathways that share common metabolites and intersect.

Co-expression of a TF and the CBRG does not necessary mean that the TF directly binds and regulates the CBRG promoters. These TFs may function at higher junctions along the CBRG regulatory hierarchy and control expression of functionally related

processes. Thus, caution must be taken when studying the reguatory role of TFs in a specific biological pathway. While validation of TF-promoter interactions remains a major challenge in molecular biology, the results presented in this section (Figure 2.5) can aid in generating hypotheses with regards to the role of highly co-expressed TFs in CBRG expression profiles and may help resolve the hierarchical location of the carotenoid pathway in the context of the entire regulatory network. It is also important to take into account that since some of the TFs are global regulators, mutations in the encoding genes are often lethal. One must also make note that there are still many un-annotated, expressed proteins in the *Arabidopsis* genome, which may function as TFs and would be overlooked in this TF-CBRG co-expression network.

Close inspection of the CBRG-TF (Figure 2.6A; see SupplementaryTable2.6.xlxs for full list) co-expression network reveals some interesting co-expression relationships. For example, 12 TFs shown to be correlated with at least 13 MEP and carotenoid pathway genes (r-value > 0.6) (Figure 2.6B) and only the carotenoid genes (Figure 2.6C). The network was further trimmed down by increasing the r-value threshold (r-value > 0.8), which allowed for fine resolution visualization of the correlation between TF expression and the MEP and carotenoid pathways (Figure 2.6D). This sub-network includes AT2G6830, a circadian rhythm repressing-associated TF that is highly co-expressed with *HYD2*. It has been shown that several genes from the CBRG and MEP pathway genes are controlled by circadian rhythm elements in their promoter regions (Covington et al., 2008).

In addition AT1G49010, a MYB super-family-related TF demonstrated an r-value > 0.8 with *PSY* and *LCYE* expression, along with an r-value > 0.6 for additional pathway genes. TFs from this super-subfamily (~198 genes in *Arabidopsis*), have been reported to function in response to various plant hormones (Yanhui et al., 2006), suggesting the connection between the requirements for proper carotenoid flux during responses to hormonal cues. AT2G24830 exhibited an r-value > 0.8 with six genes, four of which belonged to the carotenoid pathway, including *PSY, PDS, ZDzS* and *LCYB,* and two of which were related to the MEP pathway (Figure 2.6E). AT2G24830 encodes a zinc finger protein that belongs to the CCCH family, which has been recently reported to play a role

in stress tolerance in *Arabidopsis* (Wang et al., 2008). As we have discussed in Appendix II, it seems that most of the CBRG are regulated as a result of both biotic and abiotic stress.

Recent mutant analysis revealed that expression levels of the isoprenoid genes and CBRG in dark grown plants are controlled by Phytochrome Interacting transcription Factors (PIFs) which are subsequently degraded by light-activated phytochromes (Toledo-Ortiz *et al*., 2010). PIFs facilitate a strong, coordinated induction of an array of genes and lead to large increases in carotenoid and chloroplast biosynthesis (Toledo-Ortiz *et al*., 2010). However, PIFs did not come up in the CBRG-TF co-expression network; only one of the five known PIFs (*PIF5*) had an r-value > 0.6 with two MEP pathway genes (*DXS* and *HDR*; r-value 0.63 and 0.65, respectively) and with one CBRG gene (*ZEP*; r-value 0.64). The low correlation between PIF and CBRG expression, in general and with *PSY*, in particular (r-values ranged -0.3 - 0.4), was not surprising, as it has been shown that PIFs inhibit *PSY* expression, in the dark, by directly binding to its promoter (Toledo-Ortiz *et al*., 2010). These negative regulators suggest we should not neglect investigating negative co-expression in addition to the positive co-expression trends while trying to uncover regulatory mechanisms that are reflected in co-expression maps.

The topology of the CBRG-TF co-expression network illustrates the limitations in interpreting co-expression relationship caused by complex TF-driven regulatory mechanisms. On the one hand, the CBRG-TF co-expression network included known interactions between the CBRG and plant hormones under various stress conditions (especially ABA and drought stress). And also the involvement of circadian rhythm TFs in the CBRG co-expression network is not surprising. On the other hand, the same co-expression network included links to other co-expression links that are too general to allow us to reach clear biologically relevant assumptions.

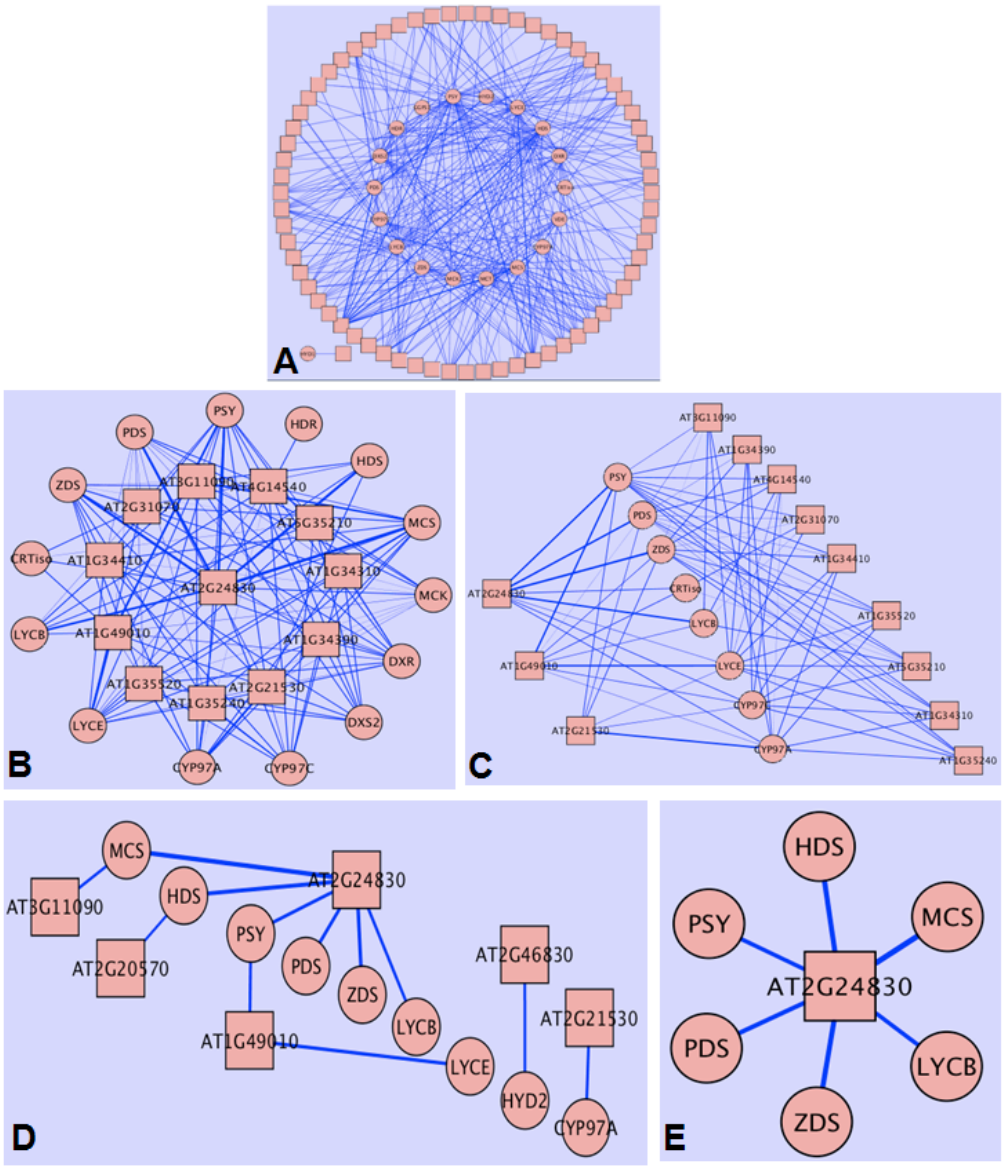


Figure 2.6. Co-expression correlation network of the CBRG and TF genes most tightly co-expressed with them. A) The carotenoid and MEP pathway genes (circle nodes) with the co-expressed TF genes (square nodes) of r-values ≥ 0.6. B) The carotenoid and MEP pathway genes that share at least 10 TFs (r-value ≥ 0.6). C) Carotenoid pathway genes sharing at least 10 TFs (r-value ≥ 0.6). D) The carotenoid and MEP pathways genes with the co-expressed TF genes of r-value ≥ 0.8. E) TF genes that are most shared among carotenoid and MEP pathway genes (r-value ≥ 0.8). In panels A-C the edge thickness represents the increasing Pearson correlation coefficient values. For full list of the TFs IDs and their annotation see Supplementary Table 2.6.

2.7 CBRG – Microarray stimuli-specific transcription analysis

As the co-expression correlation analysis is performed across multiple tissues and in response to a broad range of experimental conditions, the co-expression map provides a generalized measure of correlation between the expression of CBRG and other genes in the genome. A wide *in silico* expression inspection was performed to determine the expression of what genes temporally correlate with carotenoid and chlorophyll biosynthesis and with development of the photosynthetic apparatus. The expression profiles of genes responsible for the synthesis of carotenoids, chlorophylls, quinones and upstream isoprenoid biosynthesis pathways were examined throughout key developmental stages and in response to osmotic stress in *A. thaliana*. This analysis identified several conditions that induced differential expression of individual isoprenoid-related genes and the CBRG. Refer to section *Stimulus specific expression analysis* in Meier et al., 2011 for a detailed description of the results of the microarray stimuli-specific transcription analysis (see Appendix II - Figures 3, 4 and 5).

In general, the observed trends in the measured transcript profiles demonstrated that the induction in the expression of the isoprenoid and CBRG genes following germination was dependent on gibberellic acid and brassinosteroid (BR), respectively, as observed via application of inhibitors or by means of mutant analyses. These findings are consistent with the reported role of these hormones in etioplast development and the requirement for carotenoid and chlorophyll precursor accumulation in developing etioplasts (Rodriguez-Villalon *et al*., 2009).

2.8 Methodology

2.8.1 Co-expression correlation network of the CBRG

1. A web-crawling script written in RUBY (see below), was devised in order to retrieve text files containing the correlation values of the whole genome in relation to the genes that appear in Figure 2.3 (see script below: ACTcrawler.rb).

2. A PERL script was designed to parse the data files and extract the genes most tightly co-expressed with the CBRG (with predefined r-value thresholds) (see script below: grabRowsByRValue.pl)

3. The resulting gene lists were then used with another script (see below: makeCytoscapeNetwork.pl) to create the files needed to build the CBRG co-expression network with Cytoscape.

4. The gene annotations used in this section and throughout my thesis work are based on the TAIR.7 release (latest version updated by the ACT tool in 2009). See Supplementary Table 2.8.1 for the full list of annotations for all probes represented in the ACT files. Even though this annotation list is a bit outdated, for consistency I have used this version of annotations (TAIR 7) through my entire thesis.

ACTcrawler.rb

```
# Script to crawl the ACT web site for downloading
# the co-expression correlation files.
# Usage: Double click on script Icon in the folder the script is saved.
# the output files will be saved on the same folder.

# to use the Watir controller/load package
require "watir"
# starting web site
test_site = "http://www.arabidopsis.leeds.ac.uk/act/coexpanalyser.php#CO2"

# open the IE browser
ie = Watir::IE.new

ie.speed = :fast

# open website
ie.goto test_site

# output file
file = File.new("outputACT.txt", "w")

# gene list
f = File.open("AtProbes.txt", "r")
genes = f.readlines
f.close
c = 0;

genes.each do |gene|
        ie.goto test_site

        ie.text_field(:name, "probe").set gene
        ie.text_field(:value, "50"). set " "
        ie.button(:value, "Submit").click
        text = ie.text

# this matches the ID I am intersted in
r = /[a-zA-z]{2}[0-9]{1,2}[a-zA-z][0-9]{5,7}/
#r = "AT5G02620"
m = text.scan(r)
#print to output:
                file = File.new(m[0]+'.txt', 'w')
                file.puts text
                file.close
end
#file.close
ie.close
```

grabRowsByRValue.pl

```
#!/usr/bin/perl
#use strict;
use Data::Dumper;
# grabRowsByRValue.pl .5 LCYE.filled_in.txt test
my $rval_cutoff = shift;
#my $TF_filename = shift;
my $pathway_coexpression_file = shift;
my $resultDirectory = shift;
#my $TF_filename = "A.t_TF_genes.csv";
#open (TF_FILEHANDLE,$TF_filename);
#my @TFs = <TF_FILEHANDLE>;
#chomp @TFs;
#print Dumper \@TFs;
#my %TFs =
#    map(  ($_, 1) , @TFs);
#print Dumper \%TFs;


open(PATHWAY_COEX, $pathway_coexpression_file);
my $outputfilename = $pathway_coexpression_file . "_" . "_" . $rval_cutoff;
#$outputfilename =~ s/.csv|.txt//g;
$outputfilename .= ".txt";

my $pathway_gene_info = <PATHWAY_COEX>;
my @col_names = split(/\t/,$pathway_gene_info);


<PATHWAY_COEX>;
my $header = <PATHWAY_COEX>;


my @data;
while(<PATHWAY_COEX>) {
   chomp;
   my @row_tokens = split(/\t/);
   my $row_hash  = { probe => $row_tokens[0],
                     r_val => $row_tokens[1],
                     p_val => $row_tokens[2],
                     e_val => $row_tokens[3],
                     gene  => $row_tokens[4],
                     annotation => $row_tokens[5]
   };
   push @data, [ $row_tokens[1] , $row_hash ];
}
```

```
#print Dumper \@data;

 my @sorted_data =
    map $_->[1],
    sort { $a->[0] <=> $b->[0] }
     @data;



#open output filehandle
open(OUTFILE, ">$resultDirectory/$outputfilename"); #open for write, append

#print header
print OUTFILE $pathway_gene_info;
# print OUTFILE "ran this with r-value cutoff of $rval_cutoff\n";
#print OUTFILE "pathway coexpression from file: $pathway_coexpression_file\n";
#print OUTFILE "transcription factors from file: $TF_filename\n\n";
print OUTFILE "Probe\t r-value\t GeneID\t Annotation\n";
 foreach my $gene (@sorted_data) {
#         if ($gene->{r_val} >= $rval_cutoff && ($gene->{'annotation'} =~ m/^expressed/)
|| ($gene->{'annotation'} =~ m/^unknown/) ) {
#         if ($gene->{r_val} >= $rval_cutoff && ($gene->{'gene'} =~ m/^AT/)) {
          if ($gene->{r_val} >= $rval_cutoff ) {
             print OUTFILE $gene->{'probe'};
             print OUTFILE "\t";
             print OUTFILE $gene->{'r_val'};
             print OUTFILE "\t";
             print OUTFILE $gene->{'p_val'};
             print OUTFILE "\t";
             print OUTFILE $gene->{'e_val'};
             print OUTFILE "\t";
             print OUTFILE $gene->{'gene'};
             print OUTFILE "\t";
             print OUTFILE $gene->{'annotation'};
             print OUTFILE "\n";
          }
}

close(PATHWAY_COEX);
#close(TF_FILEHANDLE);
close(OUTFILE);
exit;
```

After obtaining the transcript expression profiles, heat maps were generated using the TM4 (MeV v4.5) microarray software suite.

makeCytoscapeNetwork

```
#!/usr/bin/perl

use strict;

#######usage: perl makeCytoscapeNetwork.pl *.csv

my ($sec,$min,$hour,$mday,$mon,$year,$wday,$yday,$isdst) = localtime(time);

my $timestamp = sprintf ("%4d-%02d-%02d_%02d-%02d-
%02d",$year+1900,$mon+1,$mday,$hour,$min,$sec);

#see if a directory called cytoscapenetworks exists

unless (-e 'cytoscapenetworks') {mkdir 'cytoscapenetworks';}

my $path = "cytoscapenetworks/$timestamp";

mkdir $path;

#opening file handles to write to the files that I will feed into cytoscape

open (SIF, ">$path/ourNetwork.sif");

open (NOA_NODE_TYPE, ">$path/node_type.noa");

open (EDA_RVAL, ">$path/rval.eda");

#defining the delimeter for the cytoscape files as a space

my $delim = ' ';

my $interactionType = 'pd';

# always the attribute name followed by the data type, which in this case is
java.lang.String

print NOA_NODE_TYPE "nodeType" . $delim . "(class=java.lang.String)" . "\n";

#################################### header for edge attribute annotation file

print EDA_RVAL "rval" . $delim . "(class=Double)" . "\n";

####################################

foreach (@ARGV) { ###  ARGV is an array that holds each of the command line
arguments.

   open (PATHWAY, "$_");

### this grabs the first line of the results file
```

```
    my $pathway_info = <PATHWAY>;
    chomp ($pathway_info);
# here we split the first line by tab
    my @pathway_info_tokens = split(/\t/,$pathway_info);
##store into $pathway_gene the second [1] field
    my $pathway_gene = $pathway_info_tokens[1];
    my $cutoff_info = <PATHWAY>;
    chomp($cutoff_info);
### spliting the info in that line by space
   my @cutoff_info_tokens = split(/ /,$cutoff_info);
## so now we can store the cutoff value as the last field [-1]
my $cutoff = $cutoff_info_tokens[-1];
#####################################
    print NOA_NODE_TYPE $pathway_gene . ' = ' . "pathwayGene\n";
#####################################
## so here we use a for loop that will skip the info in the next 3 lines
    for (my $i = 0; $i < 0; $i++) {
          my $line = <PATHWAY>;
    }
    my $i = 0;
    while  (<PATHWAY>) {
          $i++;
          if ($i == 1) {
             print SIF "\n";
             print SIF $pathway_gene  . $delim . $interactionType;
          }
          chomp;
```

```
### split it by tab and store it
        my @tf_info_tokens = split(/\t/);
### so now we would= store each vakue from the array into new variables:
        my ($probe, $r_val, $e_val, $p_val, $gene, $annotation) = @tf_info_tokens;
##########################################printing to the sif file only space and the gene number.
        print SIF $delim . $gene;
##########################################defininning all eXPRESSED PROTEIN nodes as 'CoExpGene'
        print NOA_NODE_TYPE $gene . ' = ' . "CoExpGene\n";
##########################################here we would print to the edge atribute file all the details
        print EDA_RVAL $pathway_gene . $delim . '(' .  $interactionType . ')' . $delim . $gene . $delim . '=' . $delim . $r_val . "\n";
####################################
   }
   close PATHWAY;
}
close SIF;
close NOA_NODE_TYPE;
close EDA_RVAL;
```

2.8.2 CBRG – Microarray stimuli-specific transcription analysis

An *in silico* global expression analysis was performed for the tested gene sets in response to specific stimuli and in selected mutants, to identify conditions that induce differential expression of the genes. The expression profiles were initially screened over all of the available ATH1: 22K array Affymetrix public microarray data in the gene response viewer tool (GRV) in Genevestigator. Subsequently, normalized microarray data were downloaded for experiments that were found to induce differential expression of the genes from the following sites:

NASCArrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl), TAIR-ATGenExpress (http://www.ebi.ac.uk/microarray-as/ae/), GEO (NCBI) (http://www.ncbi.nlm.nih.gov/geo/).

Chapter 3. MORPH: MOdule guided Ranking of candidate PatHway genes in *Arabidopsis thaliana*.

3.1 Introduction

A biological pathway is the set of molecular entities involved in a certain biological process and the interrelations among these entities. Such cellular processes require the participation of multiple genes and their products. Previous studies have shown that genes with common function manifest coordinated expression under different stimuli or during various developmental stages (Stuart et al., 2003; Allocco et al., 2004). Although the current knowledge about some biological pathways is quite substantial and useful for systems-level analysis, it often covers only a fraction of the genes that participate in these pathways. On the other hand, the rapid accumulation of high-throughput biological data, in combination with recent computational advances, allow us to study biological systems on a genome-wide scale, and may help identifying unknown genes in pathways of interest.

While biosynthetic pathways of secondary metabolites have been studied extensively in plants for decades (Saito et al., 2008), little is known about the regulatory mechanisms that control these complicated multi-component processes. Furthermore, we have limited understanding of the nature of interactions between metabolites and gene expression, and have only partial grasp of the relationship between transcriptional regulation and phenotype (Pigliucci, 2009). Even in the best-studied biological pathways there remain ‘information gaps', where some participating genes are still unknown: One common type of information gap in metabolic networks is missing reactions of specific metabolites in the pathway or missing transporters needed to transport pathway intermediates across cellular compartments. Another type of missing knowledge in metabolic pathways are 'orphan reactions', which refers to reactions that are known to take place in the pathway, but the enzymes that catalyze them are still unknown (Orth and Palsson, 2010).

Many whole-genome- computational methods have been developed in an attempt to close gaps in metabolic networks (Yamanishi et al., 2004; Orth and Palsson, 2010). Some of these methods use established knowledge about a pathway as a template for a model that

analyzes microarray expression data**.** Other methods analyze one or several types of high throughput data, e.g. co-expression (Kharchenko et al., 2005; Li et al., 2011), phylogeny similarity profiles (Pellegrini et al., 1999) and spatial clustering of genes on chromosomes (Lee and Sonnhammer, 2003). Distinctively, Li and co-workers designed a computational framework that uses multiple co-expression profiles for mining recurrent profound sub-graphs in a large set of massive weighted networks (Li et al., 2011).

Additional methods integrate different types of data in order to predict a set of genes that fill the gaps in the studied metabolic pathway. One such method is ADOMETA, which combines all the above data types together with a protein-protein interaction network (Kharchenko et al., 2004; Chen and Vitkup, 2006; Kharchenko et al., 2006). In ADOMETA, the genome is sub-divided into two classes, metabolic and non-metabolic genes, and then the genes are compared to their neighbors in the metabolic network using different methods and association scores that are combined using Adaboost. The highest scoring genes are predicted as those encoding a catalyzer of the studied orphan reaction. Other efforts aimed to assign enzymes to known metabolic pathways using solely high throughput gene expression data (Popescu and Yona, 2005). However, all the above methods were designed for organisms such as *E. coli, S. cerevisiae* or *B. subtilis*, for which high throughput information is much more abundant than for plants.

There are numerous systems biology tools designed for studying biological processes in plants. Some of these tools use genomic co-expression data for drawing predicted interactions between genes based on either global co-expression correlation (Arabidopsis Co-expression Tool; Manfield et al., 2006), or on co-expression under specific conditions (Genevestigator; Zimmermann et al., 2004). A recent study described the usage of MAPMAN (Usadel et al., 2006) which is a tool for displaying genomic data sets on diagrams of metabolic pathways, to investigate the coordinated gene networks regulating *Arabidopsis* metabolism in response to various stresses and nutritional stimuli (Less et al., 2011). Although there are several additional approaches and methods designed for studying biological processes in plant model systems, none of these techniques is designed to associate candidate genes from the whole genome to a specific pathway.

In this study, using the available high-throughput plant data, we develop a new computational framework for confidently predicting candidate genes that function in or regulate a given biological pathway. Our prediction method provides robust output despite the limited data available in plants, compared to other model organisms. Our method is not limited to analyzing known enzymes (Popescu and Yona, 2005), or discriminate metabolic from non-metabolic genes as done in ADOMETA. Our method ranks all genes, including both known and unknown proteins, as candidates for membership in the target pathway.

Our method, named "MOdule guided Ranking of candidate PatHway genes" (MORPH), uses resources for the model plant *Arabidopsis thaliana*, which has the most extensive plant data. MORPH receives as input a list of genes known to participate in the target pathway, expression profiles from multiple studies, and possibly additional gene groups with well defined functions. The method tests multiple combinations of expression datasets, clustering algorithms and gene expression pattern similarity scores. Once the best combination for the target pathway is identified, we rank all genes in terms of chance that they belong to the target pathway. This ranking is done by measuring expression similarity of each candidate gene to the target pathway genes that were clustered with it, with appropriate normalization to account for different cluster homogeneities. We validated our method using a cross validation based technique developed by Kharchenko et al. (Kharchenko et al., 2006). We applied our method to 66 plant biological pathways (downloaded from AraCyc; http://www.arabidopsis.org/biocyc/index.jsp ; see Supplementary Table 3.1) and our method was demonstrated to show a statistically significant high prediction quality.

3.2 Results and Discussion

We developed a method to associate candidate genes with a given target pathway in *Arabidopsis*. The method, called MORPH (MOdule guided Ranking of candidate PatHway genes), ranks genes by learning how to choose among various forms of evidence (clustering solutions) to link genes to a specific target pathway. The evidence

sources included gene expression profiles, a network of interacting metabolic genes, a protein interaction network, and gene partitions into functional groups, and could potentially be expanded beyond these.

3.3 Expression data

We collected 216 Arabidopsis gene expression profiles from 24 studies, which included 53 profiles for a seeds developmental time series seeds (the "seeds" data set; Supplementary Table 3.3.1) and collections of treatment-control studies (163 profiles) that were denoted as DS1. DS1 was also sub-divided into a "seedlings" data set (64 experiments; Supplementary Table 3.3.2) and 99 experiments performed on various tissues ("tissues" dataset; Supplementary Table 3.3.4) (for list of experiments used, see Supplementary Table 3.3.3). Background noise was reduced by filtering to remove probes that displayed low detection or low variation. A total of 12,459 probe-sets remained after filtering) (full list of probes survived the filtering step is in Supplementary Table 3.3.2).

3.4 Clustering Solutions.

A key step in MORPH is the partitioning of genes into modules. We used two different strategies to cluster the Arabidopsis genes: gene expression-based clustering and modules defined using external information. In total, we devised 5 different clustering solutions (see Methods) and an additional negative control (NoClustering) to contain all genes:

A. SOM: Clustering by gene co-expression

B. Enzymes: Whether the gene encodes an enzyme or not

C. Orthologs: Whether a gene has a homolog in rice and maize or NOT

D. MD: Whether a gene is co-expressed and shares a metabolite in a metabolic network of 1987 genes (metabolic dependency) (see Supplementary Table 3.4.1)

E. PPI: Whether a gene is co-expressed and linked in a predicted protein-protein interaction (PPI) network containing 8,273 genes (see Supplementary Table 3.4.2)

F. NoClustering: no clustering and contains all genes

Among the clustering solutions, we used two gene networks: The MD and PPI networks of 8,273 genes. A limitation of these two networks was the low coverage of probes (~1200 probes). Therefore, we designed an algorithm to increase coverage of the underlying gene set in order to analyze the networks together with gene expression data. The algorithm called MATISSE*, is a variant of the MATISSE algorithm (Ulitsky and Shamir, 2007) which identifies modules of co-expressed genes that are connected in metabolic or PPI networks (Figure 3.1). MATISSE* expands the modules by adding genes that show a high level of similarity to the average expression pattern of a given module. The expansion step increased module size considerably (over twofold with the MD network and ~1.5 fold when using the PPI network) (data not shown). We added an additional "pseudo-module" containing all genes that were not included in other modules.

3.5 Clustering guided scoring of pathway genes

Given a clustering solution (partitioning the genes in the data sets into modules), and a target pathway, we wished to rank the rest of the genes in terms of how plausible it was that they were associated with a given pathway. A total of 66 pathways (biosynthetic and/or functional groups) derived from ARACYC (see Methods) were used to drive the ranking scheme computed by the MORPH algorithm. For each module generated by the various clustering solutions above, we identified genes from the target pathways, and computed for each other gene in the same module, its average co-expression similarity to the pathway genes in that module. The rationale was that while the modules may reflect various broad functions, some of these functions may be related to the target pathway and

hence on average, the pathway genes would show higher co-expression similarity among themselves than co-expression similarity measured among arbitrary genes in the same module. Since modules varied in size and homogeneity (i.e. average co-expression level), the gene-pathway similarity scores within each module were standardized (see Methods).

The MORPH algorithm (RankerWithoutLOOCV.jar) was applied using all combinations of A) one dataset of four (ds1Data.xls, ds3DataMatrix.xls, SeedlingsMatrix.xls, TissuesStandData.xls; see Appendix III); B) one clustering solution of six, and C) one of two similarity scores (Pearson or Spearman), to compute a total of 48 scores/pathway. We compared different combinations of these three elements on 66 pathways using a statistical technique developed by Kharchenco *et al*. based on leave one out cross validation (LOOCV). The validation procedure repeatedly removes one gene from the target pathway (the test gene), generates the ranking based on the remaining genes (the training set) and calculates the rank of the test gene, denoted as the self-rank of that gene (Supplementary Script 3.5.1; see Appendix IV for instructions on using the script). Then one can plot for every self-rank threshold (in some predefined interval, e.g., 0 to 1000) the fraction of pathway genes that were detected at the threshold when acting as test genes. We extended this approach to obtain a score in a uniform range between 0 and 1, by calculating the relative area under the self-rank curve, where 1 denotes a perfect ranking algorithm (see Methods). We denoted this score as the relative Area Under the Curve of the Self Ranked genes (AUC-SR). For example, when this technique was applied to the CarotenoidCore pathway (Figure 3.2), within 680 scores, all 13 carotenoid pathway genes (Table 3.2) are predicted to be associated with the pathway. From this graph the AUC-SR for the CarotenoidsCore is calculated as 0.92.

MATISSE has been proven to be effective in finding modules of functionally related genes, but it usually categorizes only a small fraction of the genes into modules. We therefore first compared the quality of the rankings produced by MATISSE and MATISSE* using the metabolic dependency network and all gene expression data sets on 66 pathways downloaded from AraCyc containing at least 10 genes in the gene expression data sets. Overall, MATISSE* gave improved results in all four examined data sets, using both Pearson and Spearman correlation as the similarity scores

(Supplementary Figure 3.5.1). On average, the seedlings data set provided the best scores. For the full list of pathway scores refer to Table 3.1. (This list of scores was produced using an older version of our selection algorithm, so AUC scores may vary using the new AUCranker.jar script found in the MORPH folder online). When we used the PPI network, MATISSE* modules received higher AUC-SR scores only in the seeds and tissues data sets (Supplementary Figure 3.5.2). Using the metabolic dependency network provided higher quality predictions than using the PPI network. Since MATISSE* was observed to perform better in terms of both predictability and coverage (i.e. the percentage of genes included in the modules), it was used in all further analyses.

3.6 Customizing the utilization of gene expression data sets

We compared the predictive power of using the tissues and seedlings profiles separately and of using the united dataset DS1. We used MATISSE*, combined with the MD network, to create modules, and Spearman correlation as a similarity function. DS1 was significantly inferior, yielding an average AUC-SR score of 0.47 compared to 0.54 provided by the seedlings data set ($p < 0.01$) and 0.53 provided by the tissues data set ($p=0.032$). Figure 3.3 compares AUC-SR scores for each pathway using either the DS1 or the seedlings data sets. Although some pathways attained better scores using the seedlings data set, other pathways had higher scores in DS1. These results inspired us to refine the MORPH algorithm with a *learning configuration* or *model selection* technique to optimize the analysis process for a specific pathway, as will be explained in the next section.

3.7 Pathway- specific model selection

Optimizing predictive power, given a set of different possibilities to analyze data, is generally referred to as 'model selection' in machine learning, and has received high attention in recent years (Guyon et al., 2010).We define a pathway's *learning configuration* as a triplet of (1) gene expression data set, (2) a clustering method and (3) a

similarity score. Each learning configuration can be used to generate a ranking of candidates for a specific pathway. In order to match the optimal learning configuration to a given biological pathway, we used LOOCV to estimate the predictive power of every configuration and selected the one that produced the highest AUC-SR score (see Table 3.1). We denote this ranking algorithm as 'selection'. It is important to note that for statistical validation of this method, the LOOCV used by the selection algorithm is used internally without taking into account the tested gene, and therefore we avoid over-fitting. The method can potentially accommodate additional datasets and clustering methods to choose the best learning configuration.

We plotted the average AUC-SR scores for all learning configurations and the selection algorithm (Figure 3.4). For every tested pathway, expression data set and a clustering method (except for the Orthologs clustering method which produced inferior results), we show results for the similarity measure that produced the best AUC-SR score. The 'null' clustering (denoted as 'NoClustering'), the SOM clustering solution, grouping by orthologs (data not shown) and MATISSE* using the PPI network, all yielded scores below < 0.4. MATISSE* with the MD network and enzyme modules gave higher scores of 0.44-0.54. These results can be explained by the fact that enzymes and MD sources are derived from previous metabolic information, and therefore we expect them to perform better and reflect more faithfully the signatures of metabolic pathways in expression data. Nevertheless, we observed some exceptional cases where learning schemes resulted in high scores but were not based on enzymes or the MD network. Examples include the aerobic respiration pathway, where the best score was obtained by the SOM clustering algorithm, and the homogalacturonan biosynthesis pathway, where MATISEE* with the PPI network scored best. Importantly, our “model selection” algorithm, which integrates all data used in our framework, yielded the highest average score, 0.66, compared to individual configurations. We therefore use this methodology to rank candidates in the next phase.

3.8 Additional statistical validation

To demonstrate robustness of our method, we evaluated AUC-SR scores obtained for randomly selected gene sets and compared them with AUC-SR generated using known pathways. We tested two ranking algorithms: the selection algorithm, and a clustering guided ranking algorithm that uses (a) the seedlings gene expression data set, (b) MATISSE* with the MD network for creating a clustering solution and (c) Spearman correlation as a similarity function. We denote this latter ranking algorithm as SMDS. We ran each algorithm with target pathways comprised of randomly selected gene sets of different sizes in the same range of the known pathways (10-30), and repeated the process 200 times for each set size. Figure 3.5 shows the results for the selection algorithm on the randomly-generated gene sets, together with the known pathways. Each box-plot summarizes the distribution of AUC-SR scores for random gene sets of a given size. Overall, in this experiment 4000 random gene sets were generated, and no random gene set received an AUC-SR score > 0.5. However, 55 out of the 66 real pathways tested received scores above 0.5 and as high as 0.99**.** Supplementary Figure 3.8 shows the results on the random target networks for the SMDS ranking algorithms, together with the tested pathways of sizes 10-29. Overall, in this experiment 4000 random gene sets were generated, and no random gene set received an AUC-SR score > 0.4, while 48 out of the real pathways tested received a score above 0.4 and as high as 0.98. Comparison of the scores for random gene sets by the SMDS and selection algorithms reveals that the distributions provided by the selection algorithm have higher mean and standard deviation. This result is expected since the selection algorithm chooses internally among more than 40 different ranking combinations (see Methods) and thus has higher ability to capture random patterns in the data. Overall, the statistical validation presented here gives additional support for the robustness of our ranking algorithms. We preferred the selection algorithm since it yields higher AUC-SR scores than the SMDS algorithm, and also showed a larger number of real pathways that received scores higher than the maximal score provided by the random gene sets (55 compared to 48). This validation test further confirms that the selection algorithm avoids over-fitting.

3.9 Ranking candidate genes to be co-regulated with selected biological pathways

3.9.1 'Photosynthesis light reactions' pathway

Out of the examined 66 pathways (Supplementary Table 3.1), 'photosynthesis light reactions', had the best AUC-SR score (0.99) (Table 3.1 lists AUC-SR scores for all 66 pathways) indicating a strong signature of this pathway in the gene expression data. During photosynthesis, absorption of sunlight creates electronic excitations of photosynthetic systems in the chloroplasts and the subsequent transfer of the excitations to a reaction center (Cheng and Fleming, 2009). Plants evolved sophisticated light-harvesting complexes for maximizing efficiency of photosynthesis. Even though these complexes are well studied, the precise molecular principles that enable high efficiency light harvesting remain elusive.

The selected learning configuration for ranking genes for the photosynthesis light reactions was “tissues”/MD network/ Spearman similarity score. Inspecting the top genes ranked by our method (Supplementary Table 3.9.1) revealed genes having a tight biological link with photosynthesis. The top ranked gene (which is annotated in TAIR as ‘expressed protein’; AT5G52220) (see section Supplementary Table 2.8.1 for gene annotation source) was previously predicted to interact with the photosynthetic subunit PSI-D2, by using a chloroplast protein interaction network (Yu et al., 2008). This interaction has been confirmed by a yeast two-hybrid experiment (Yu et al., 2008). Both third and forth ranked genes (AT1G60950 and AT4G03280 respectively) are related to electron transfer as part of phoyosynthesis light reactions. The third ranked gene encodes a major leaf ferredoxin, which was shown to be the preferred electron donor (out of a family of six *Arabidopsis* ferredoxins) for double bond reductions made by phytochromobilin synthase, which participates in the light-sensing machinery (Chiu et al., 2010). The fourth ranked gene encodes a cytochrome B6-F complex iron-sulfur subunit, for which a mutant displayed reduced electron transport at saturating light intensities (Okegawa et al., 2005). Within the top 24 candidate genes of the

'photosynthesis light reactions', there are 16 genes annotated in TAIR as 'expressed proteins'. These candidates provide exciting new avenues for research on photosynthesis.

3.9.2 The carotenoid biosynthetic pathway

Carotenoids are plant pigments that serve as photoprotectors preventing oxidative damage, absorb light in photosynthesis (Niyogi, 2000) and are precursors to cleavage products such the plant hormone *abscisic* acid (ABA) (Nambara and Marion-Poll, 2005), and newly discovered strigolactones (Koltai, 2011a). While the biochemistry of carotenogenesis has been well established over the past decades (Cuttriss et al., 2011), there remain gaps in our knowledge regarding regulation of carotenogenesis and conversion of carotenoids to apocarotenoids. The selection algorithm in MORPH ranked candidate genes linked to the carotenoid biosynthetic pathway (CarotenoidCore gene list – 13 genes) using as the optimal configuration: the MATISSE* module, the seedlings gene expression data set and Spearman correlation similarity. The AUC-SR for this configuration was 0.86.

The top ranked gene for the carotenoid pathway is annotated as ‘squalene monooxygenase' (AT4G37760 / SQE3) (Supplementary Table 3.9.2), an enzyme in sterol biosynthesis. This gene belongs to a family of five genes encoding squalene epoxidase enzymes in *Arabidopsis* that catalyze the conversion of squalene into 2,3-oxidosqualene (Phillips et al., 2006), which is an early precursor of brassinosteroids. We recently showed that expression of genes in the carotenoid pathway are coordinately expressed in response to brassinosteroids (Meier et al., 2011). This ranking result further corroborates the link between brassinosteroids and carotenoids. The second ranked gene is the carotenoid cleavage dioxygenase gene (*CCD1*). *CCD1* is a known carotenoid cleavage enzyme has broad substrate specificity (Vogel et al., 2008).

The next gene ranked with the CarotenoidCore (as number three), is AT4G32770, tocopherol cyclase (*VTE1*) a chloroplast-localized gene that participate in synthesis of vitamin E synthesis (tocopherols) (Mene-Saffrane et al., 2010). Tocopherols, as carotenoids are isoprenoids compounds that are derived from the same precursor-geranylgeranyl diphosphate (GGPP) and also participate in the well-conserved

mechanisms of photoprotection by scavenging reactive oxidant species (Penuelas and Munne-Bosch, 2005)**.** *VTE1*, which is the first enzyme committed to tocopherol biosynthesis (vitamin E), has been shown to be co-localized with several carotenoid biosynthetic genes in pepper (*Capsicum annuum*) chromoplast plastoglobules, which function as a storage and processing site for carotenoids (Ytterberg et al., 2006). Furthermore, the carotenoid cleavage enzyme *CCD4* has been found to be co-localized with *VTE1* in chloroplast plastoglubules suggesting another link between tocopherol biosynthesis and carotenoid degradation (Ytterberg, et al., 2006).

The candidate enzyme ranked by MORPH at number four is solanesyl diphosphathase 2 (*SPS2*) (AT1G17050), which is annotated as solanesyl diphosphate. *SPS2* provides substrates for the plastoquinone (PQ) biosynthesis (Hirooka et al., 2005). PQ are known to be co-localized with carotenoid genes in the chloroplast envelope (Block et al., 2007). In addition, it was shown that PQ plays a role regulating carotenoid biosynthesis in photosynthetic tissue as an intermediate electron carrier between carotenoid desaturases and the photosynthetic electron transport chain (Norris et al., 1995). A recent study has corroborated these findings by showing that the altered carotenoid composition displayed in the *Orr* mutant plants, is probably due to disruption in function of several carotenoid pathway genes (PDS and ZDS) that require oxidized PQ for their proper desaturation activity (Nashilevitz et al., 2010).

The chloroplast synthesized molecule, GGPP is an important metabolic hub that is the immediate precursor for multiple isoprenoid-derived pathways in addition to carotenoids pathway. Overall the additional isoprenoid derived pathways contain ~ 71 genes. Out of the 71 genes, 23 genes were ranked among the top 200 candidates by our method (ranks denoted in green in Supplementary Figure 3.4). Using a hyper-geometric test, this overlap is statistically significant ($p < E-15$), providing additional support for the predictive power of our method.

3.9.3 The homogalacturonan biosynthesis pathway

Pectin is composed of four different polysaccharides that are characterized by a high content of galacturonic acid. These polysaccharides are composed of:

homogalacturonans (HG), xylogalacturonan (XGA), rhamnogalacturonan I and rhamnogalacturonan II (Harholt et al., 2006). According to our configuration learning results, for ranking candidates for the homogalacturonan pathway it is best to use the MATISSE* module (based on the PPI network as a clustering solution), with the seedling gene expression data and Spearman correlation for similarity scoring (AUC-SR of 0.91). All the top three candidates belong to large protein families (Zinc fingers transcription factors, alpha/beta hydrolases and nucleotide-sugar protein family proteins) (see Supplementary Table 3.9.3). The candidate gene that is ranked at number seven (AT5G33290), is annotated as XYLOGALACTURONAN DEFICIENT1 (*XGD1*). In a previous study *XGD1* was described as a xylogalacturonan specific xylosyltransferase , and a mutant of it displayed decreased levels of xylogalacturonan and had decreased cell wall xylose (Jensen et al., 2008). The fact that *XGD1* was ranked among the top candidates of the homogalacturonan pathway highlights once again the robustness of the predictive power of our method.

3.10 Future plans

Biological pathways often have a structure of a grid (rather a linear set of reactions done by enzymes), and as a result branches can be regulated independently and have different expression patterns. Nevertheless collectively the results presented here provide a robust biological support for the ability of MORPH to predict genes that are related to a given biological pathway. MORPH represents a well-needed methodology to discover new components associated with function of biological pathways in plants. Our method could be expanded to incorporate other datasets and clustering solutions for application to pathways in plants and other model systems. The method's great flexibility is due to the fact that the algorithm developed incorporates un-supervised machine learning method, clustering methods and model selection procedure that are not either specie or pathway dependent. For example some pathways might be better studied using microarray data from tissues that reflect the highest-level accumulation of those particular metabolites. Also future RNAseq data will likely improve the abundance and sensitivity of transcripts and expand the number of genes analyzed to include genes that

display lower but variant expression levels. In addition, implementing recent methods for analyzing multiple co-expression networks (Li et al., 2011), and incorporating model networks for predicting links among genes in *A. thaliana* (Lee et al., 2010), will contribute even further to the robustness and biological significance of MORPH predictions regarding candidate genes to a given biological process.

In the next step of the development of our method, we will work on creating a friendly user interface that will allow users to use datasets we collected and also allow the user to add information data sets of their own (for example gene expression arrays, protein-protein interactive networks, metabolomic data, etc). This feature will allow the algorithm to consider also the newly added data by the user in the learning configuration stage, yielding possibly better prediction power of candidate genes that are co-regulated with the biological pathway studied.

3.11 Methods

3.11.1 Microarray data sets and pre-processing

We collected 216 published microarray expression profiles of *Arabidopsis* in response to specific stimuli, including plant hormonal stimuli, biotic/abiotic elicitation, different light regimes, developmental stages, and selected mutants. Normalized matrices were retrieved from: NASCArrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) (Craigon et al., 2004), TAIR-ATGenExpress (http://www.ebi.ac.uk/microarray-as/ae/) and GEO (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2009) **(**For the full list of experiments and their accession numbers see Supplemental Table 3.3.3).

We divided the experiments collected in to three data sets:

1. 64 experiments done on seedling tissues generated by13 different labs.
2. 99 experiments done on different tissues (leaf, roots, seeds and flowers) generated by 10 different labs.

3. 53 experiments derived from the work of one lab (Le et al., 2010), and contain laser dissected seed tissues from different developmental stages.

We analyzed each of the three data sets separately. We also analyzed a combined data set of the seedlings and tissues (1 and 2). We called the combined data set DS1. For DS1 and the seeds data set we removed probes that displayed consistently low detection calls (raw detection call < 100 in at least 80% of the experiments) or low variation (standard deviation < 120). The group of probes that survived this filtering in all data sets (1, 2 and 3) was merged by gene IDs resulting in a group of 12,459 genes that were used for further analysis. Since DS1 profiles were collected from 23 different sources, for each source we averaged replicates, divided treatments by control and standardized each profile. Dividing by the respective controls was applied to each of the seedlings and the tissues data sets but not to the seeds time series data set.

3.11.2 Tested pathways

AraCyc pathways were downloaded from the PMN database (ftp://ftp.plantcyc.org/Pathways/). We filtered out (using an Excel spreadsheet) pathways with less than 10 genes in the GE data sets, leaving 64 pathways. We then manually added the carotenoid biosynthetic pathway (13 genes) and the carotenoid pathway combined with the upstream MEP pathway (23 genes) (Supplementary Table3.1).

3.11.3 Arabidopsis Additional Information Sets used for pathway rankings

3.11.3.1 GE based clustering method

We used the self-organizing map (SOM) clustering algorithm (Kohonen, 1990) with 5X5 grid layout settings, as implemented in the EXPANDER platform (Ulitsky et al., 2010). In addition, we defined one cluster (denoted as 'no-clustering') containing all the genes.

3.11.3.2 Genomic annotation

Enzymatic gene annotations were downloaded from PlantCyc (ftp://ftp.plantcyc.org/Pathways/). For enzymatic annotation, the genes were divided into two sets: a set of 2933 genes annotated as enzymes and a set of the remaining 9526 genes (Supplementary Table 3.11.3.2).

3.11.3.3 Arabidopsis metabolic dependencies map

Arabidopsis metabolic interactions were downloaded from AraCyc (ftp://ftp.plantcyc.org/Pathways/) and were used to construct the 'metabolic dependencies map', as described in (Kharchenko et al., 2004; Kharchenko et al., 2005). Formally this map is an un-weighted and un-directed graph in which nodes denote metabolic genes and edges connect genes whose corresponding enzymes share a common metabolite among their reactants or products. As done in (Kharchenko et al., 2006), in the process of building the network we excluded the most common metabolites. Iteratively we removed the most common metabolite and calculated the percentage of genes that were annotated in the AraCyc reactions covered by the network. We repeated this process until the coverage dropped below 95%. Overall 20 metabolites were removed.

3.11.3.4 Protein- protein interaction network

In a recent study a Predicted *Arabidopsis* Interactome Resource (PAIR; http://www.cls.zju.edu/pair/) was developed (Lin *et al*., 2011). The PAIR database includes 145,404 potential predicted gene interactions and 5990 experimentally reported interactions.

3.12 Module guided ranking algorithm

We developed a new algorithm for prioritizing novel candidate genes in a specific pathway. The algorithm receives as input a set S of genes that are known to participate in the pathway, a set of gene expression profiles, a similarity function □, and a partitioning of all genes in the gene expression data into k modules: □*1,…,* □□ . The modules can be generated using any additional information source available (see below). As a first step we filter out modules that do not contain any pathway genes. For a module □□ that contains a set of pathway genes □*1,…,* □□ and for every gene □ within □□ that is not a part of the pathway, we first calculate its average similarity with □*1,…,* □□ :

(1) □□□ □, □□= *1*□ □=*1*□□*(*□,□□*)*

Where □*(*□,□□*)* is the similarity between the expression pattern of □ and □□. Since there are inherent differences between modules in terms of size and homogeneity, we standardize the similarity scores within each module in order to be able to merge the scores of all candidates. Formally, let □□□*1,…,* □□□□□ be the average similarity scores of all candidate genes within a given module □□. Let μ be the average of □□□*1,…,* □□□□□ and let σ be the standard deviation of these scores. The standardized scores of genes from all clusters are now united into one list, where genes are sorted by the score. For every candidate gene □ we calculate its z-score:

(2) □−□□□□□□□= □□□ □, □□− μ□

The final ranking of candidate genes is set as follows: all genes that were not clustered with a pathway gene are placed at the bottom of the ranking. All the other gene z-scores are sorted in descending order.

3.13 Statistical validation procedure

For each tested pathway S, we ran a leave-one-out cross validation (LOOCV) procedure as follows. One of the genes v in S is removed from the list and the algorithm is applied using the reduced set S\{v} as the pathway genes. v participates in the clustering, scoring and ranking as a non-pathway gene. The performance of the ranking procedure is evaluated using the self-rank (SR) measure of Kharchenko *et al*., 2006. The SR of v is defined as the place of v in the ranking produced by the algorithm. A perfect

prediction would give v an SR of 1 (top candidate), and a completely non-informative method would result in a uniform distribution of ranks. The process is repeated for every gene v in S. The results are summarized by the SR-plot, where for every rank threshold k (k= 1-1000), the percentage of pathway genes with a self rank $\leq$ k is shown (see Figure 3.2 as an example). The value 1000 was chosen empirically so as to cover most pathway genes without including too many irrelevant genes. We calculate the area under the SR-curve and divide it by the area under the line y=1. This ratio is defined as the AUC-SR score, which ranges between 0 and 1.

3.14 Learning pathway specific configuration

The choice of expression data, the specific clustering algorithm used and the similarity measure, all affect the results of the analysis. In order to characterize a specific pathway we tested 40 possible combinations (four gene expression data sets, five clustering methods and two similarity scores) for each specific pathway examined. The selection is done by choosing the combination obtaining the highest AUC-SR score. This stand-alone ranking scheme was statistically evaluated by the same LOOCV procedure. Formally, given a set of learning configurations F, a test set *Te* and a training set *Tr* we calculate the AUC-SR for each configuration in F using LOOCV on the *Tr* alone, and select the algorithm obtaining the maximum score to create the final rankings from *Tr*. We used that configuration to evaluate the rank of *Te*. Finally, we summarize the LOOCV by using the ranks of every *Te* to generate the SR curve.

3.15 Building modules based on co-expression and interaction or dependencies networks

3.15.1 MATISSE modules

MATISSE (Module Analysis via Topology of Interactions and Similarity SEts) is a program for detection of functional modules using interaction networks and expression data (Ulitsky and Shamir, 2007). A MATISSE module is a gene set comprised of highly

co-expressed genes that are connected in the network, possibly through the inclusion of additional genes for which expression data are not available ("back nodes"). Each module was considered as an additional gene set, which was also used by the module guided algorithm.

3.15.2 MATISSE: overcoming the low coverage of networks in plants*

The coverage provided by the MATISSE modules was low (an average of 1204 and 4522 genes were contained in the modules for each gene expression data set using the metabolic network and the PPI network, respectively. We therefore sought an improvement to the coverage of the algorithm: starting with the MATISSE set of modules, we repeatedly inserted the gene with the highest correlation to a module into that module (and updated gene-module correlations) until the correlation dropped below 0.4 (Figure 3.1). Since our modules are derived from PPI and metabolic dependencies networks, we prioritized proteins as candidates for these iterations. This step improved the coverage to 2587 genes for the metabolic network and 6711 in the PPI network. We note that more than 5500 genes out of the 12,459 genes that survived our filtering steps are annotated to date in TAIR as 'expressed protein', therefore our MATISSE* design is fitted to overcome this shortage of knowledge in the plant domain. In addition we did not remove genes that are not predicted to be proteins from the candidate list. Additionally, we added a new module that holds all genes that were not covered by the modules.

3.16 CHAPTER 3 FIGURES:

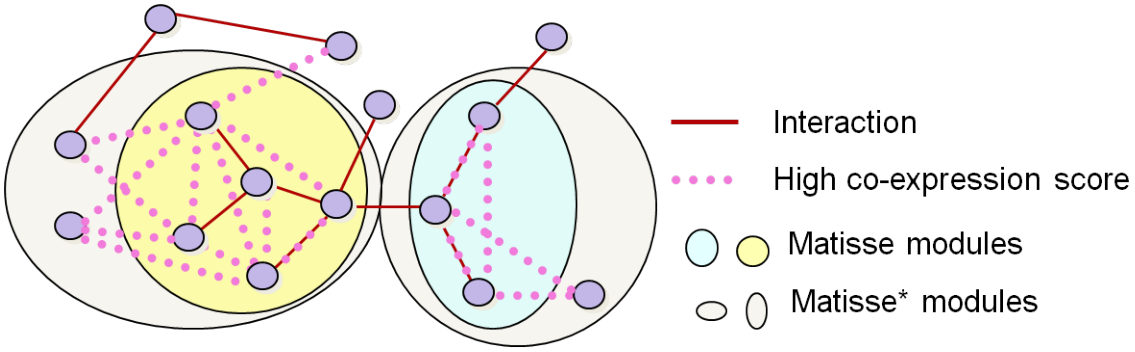


Figure 3.1. MATISSE* modules toy example. Two original MATISSE modules are shown in yellow and blue. The genes in each module are connected in the interaction network (red lines) and each contains many gene pairs with high co-expression score (dotted lines). The two modules are in fact connected by a red edge but not merged by MATISSE to maintain high co-expression within the modules. MATISSE* extends the MATISSE modules to include genes that are not necessarily connected to the module in the network but have high co-expression values with the other genes in that module. MATISSE* modules are colored in gray.

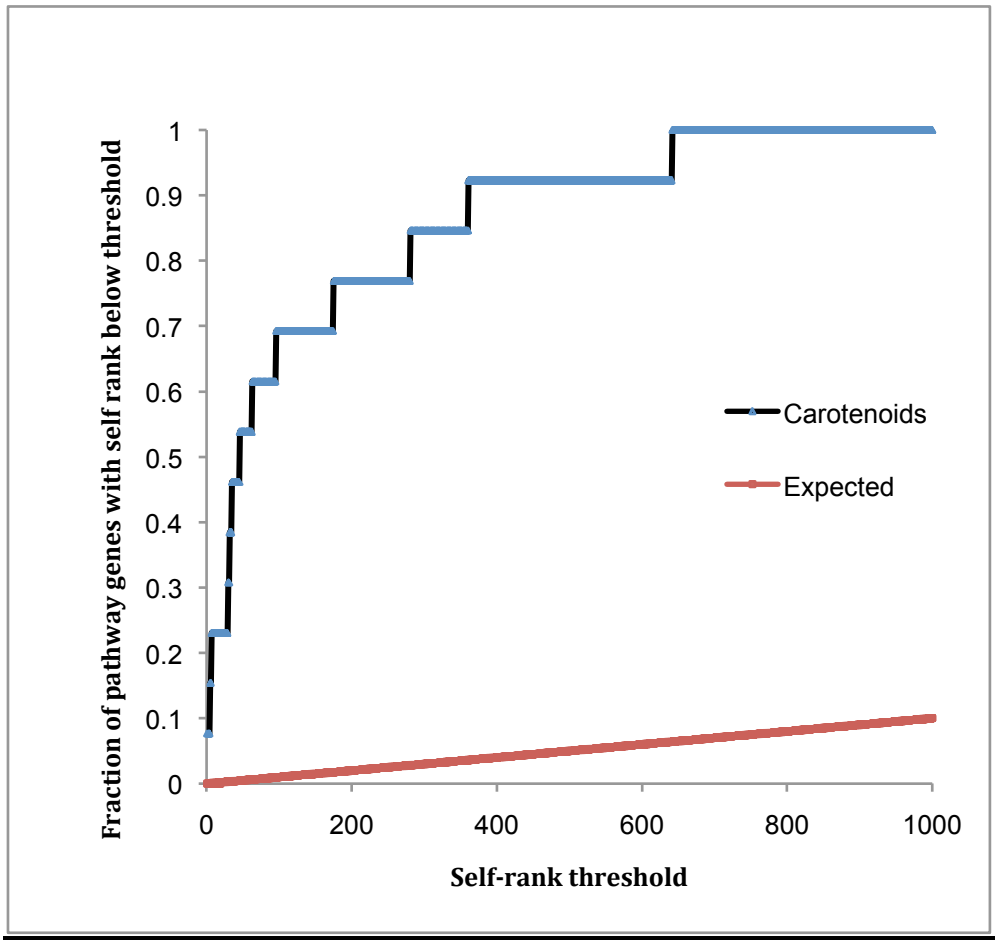


Figure 3.2. The Self Rank (SR) plot of the carotenoid biosynthetic pathway (CarotenoidCore) containing 13 genes (Supplementary Table 3.1). For each value of the SR threshold on the x-axis, the plot shows the fraction of genes in the pathway that were ranked below that threshold (blue line) using the LOOCV method. The red line shows the expected plot for a randomly selected gene set of size 13.

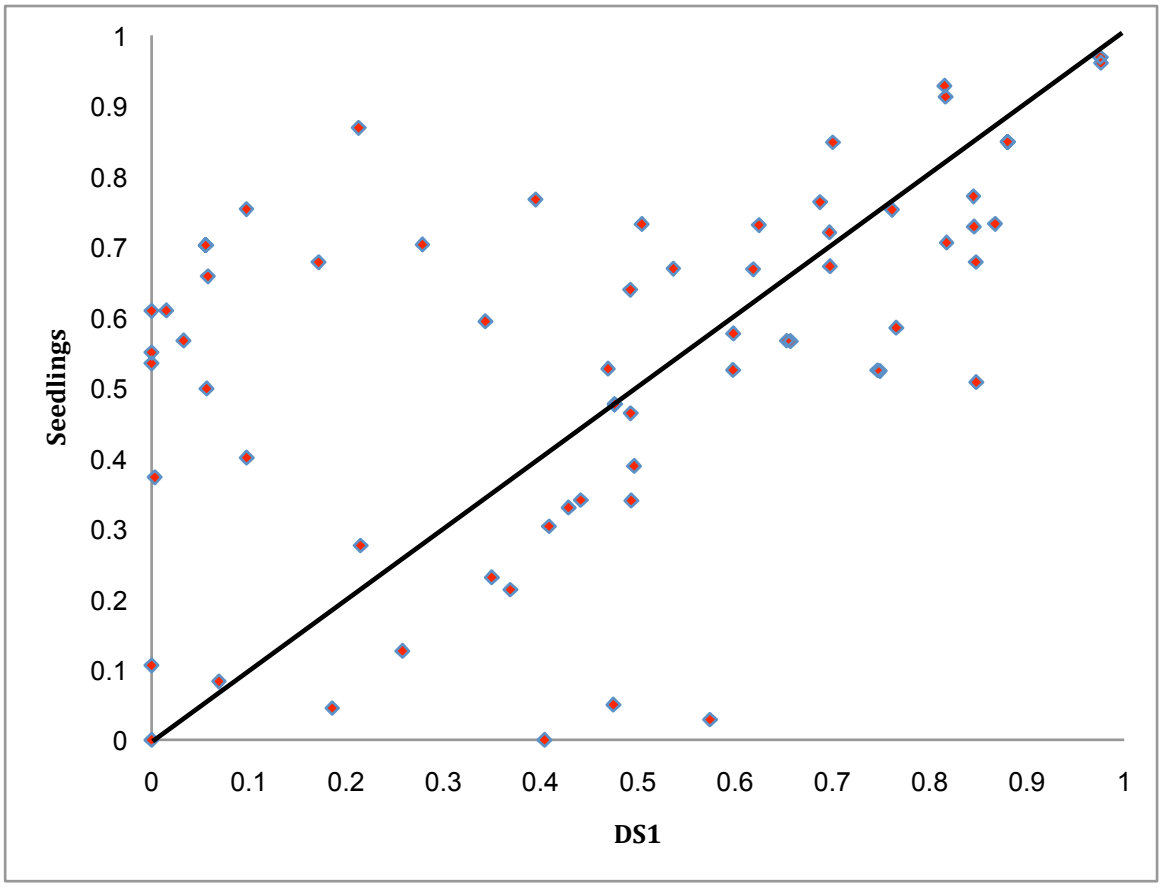


Figure 3.3. The values of AUC-SR scores for data set 1 (DS1) and the seedlings data sets (a sub-data set of DS1). Each red diamond represents an AUC-SR score for one of the 66 tested pathways. The modules used here were created by MATISSE* with the metabolic dependencies network and the Spearman correlation coefficient as the similarity score. The line is x=y divides the graph into two parts: all points above the line have better AUC-SR scores in the seedling data set (y-axis), and all point below the line have better AUC-SR scores in the DS1 set (x-axis).

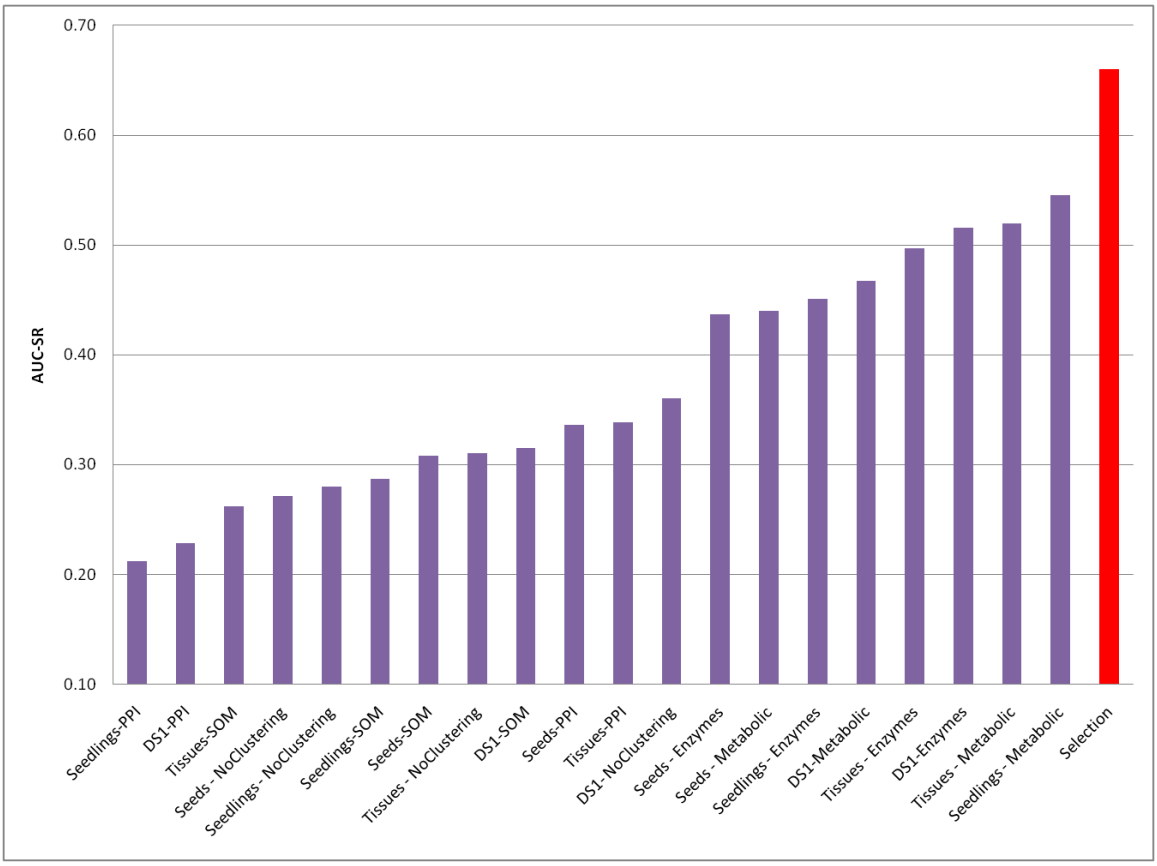


Figure 3.4 The quality of different learning configurations. For each combination of gene expression data set and partitioning algorithm (that can be based on different networks), the average AUC-SR (Area Under the Curve of the Self Ranked genes) over all pathways tested is displayed in a blue column. The value for each individual pathway was taken as the better one among the two possible similarity scores used. The score using the selection algorithm is shown in red. The scores for partitioning using "Orthologs" are not shown because in general this method produced the poorest AUC-SR scores.

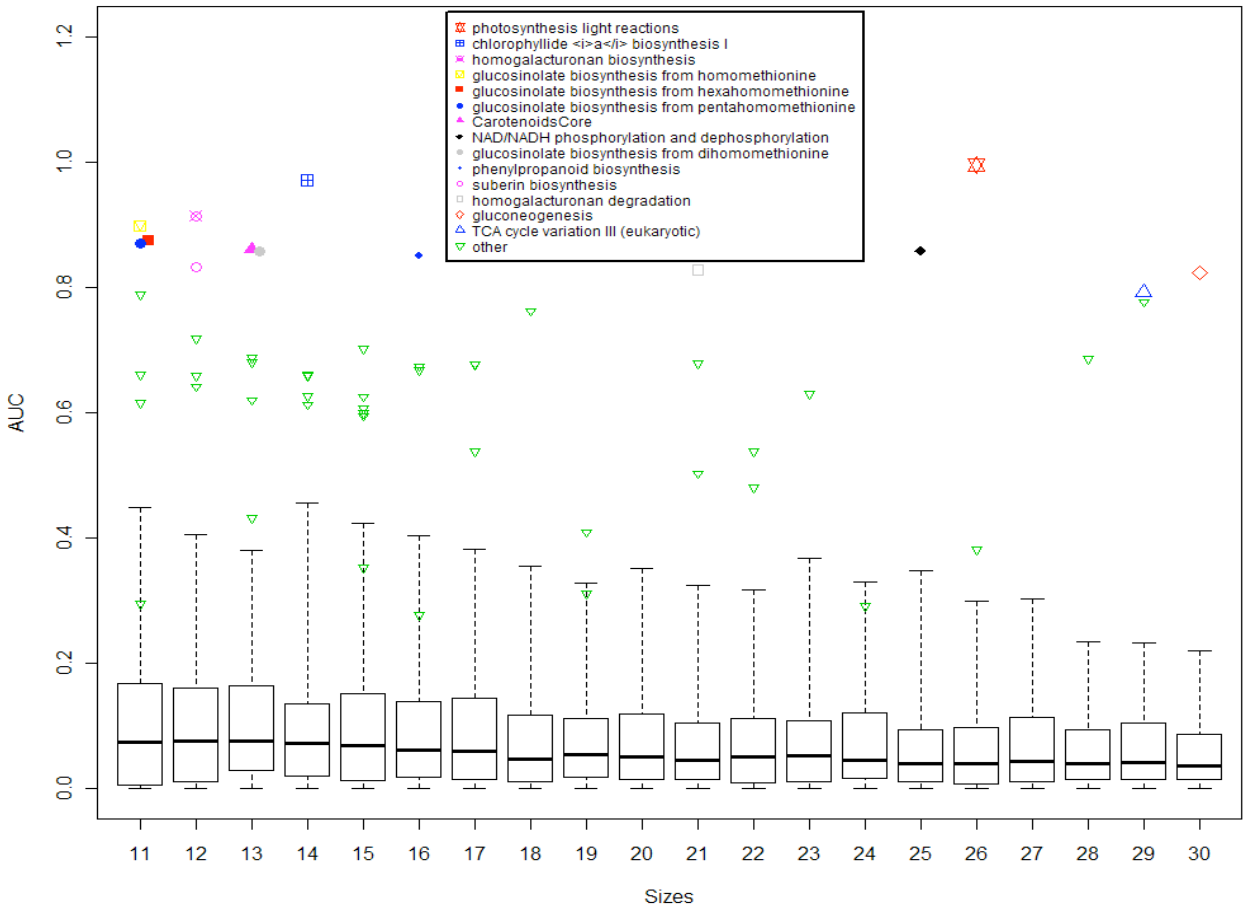
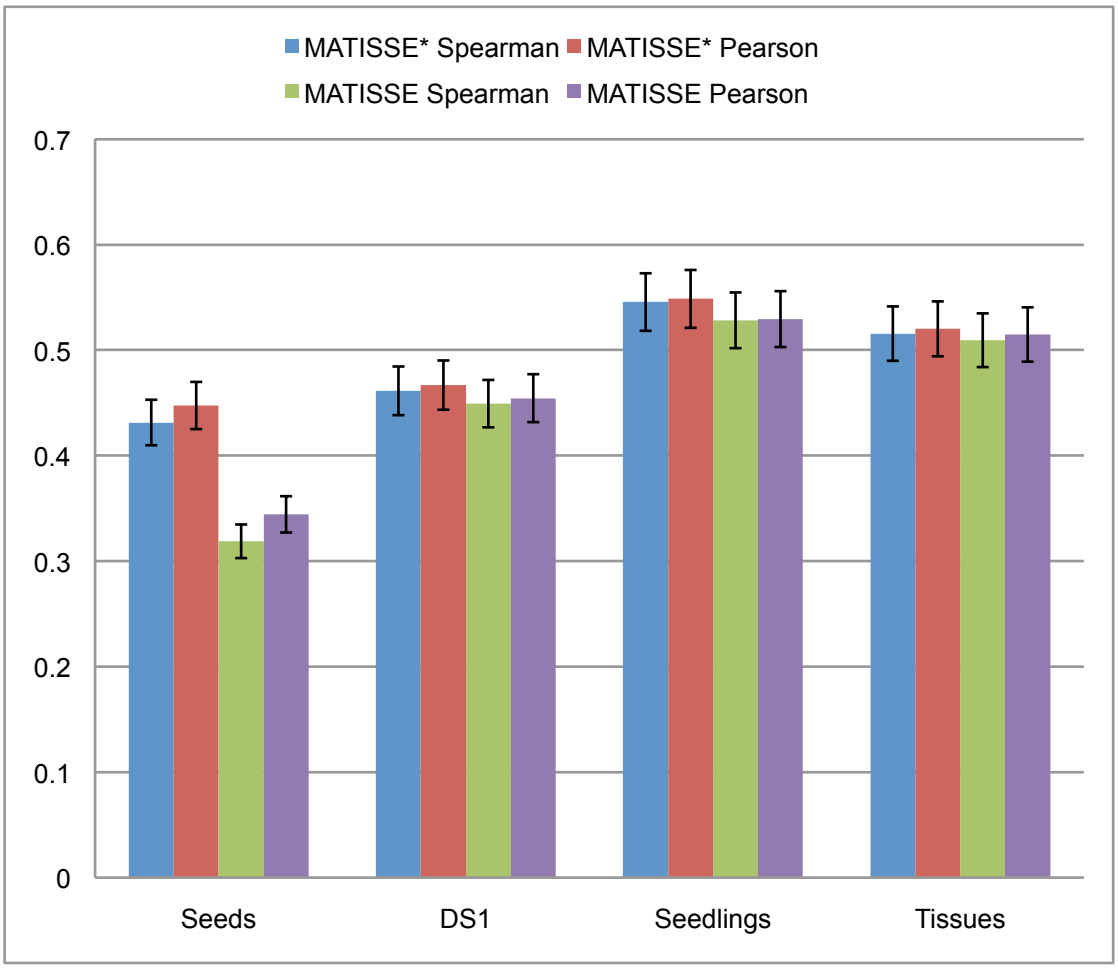


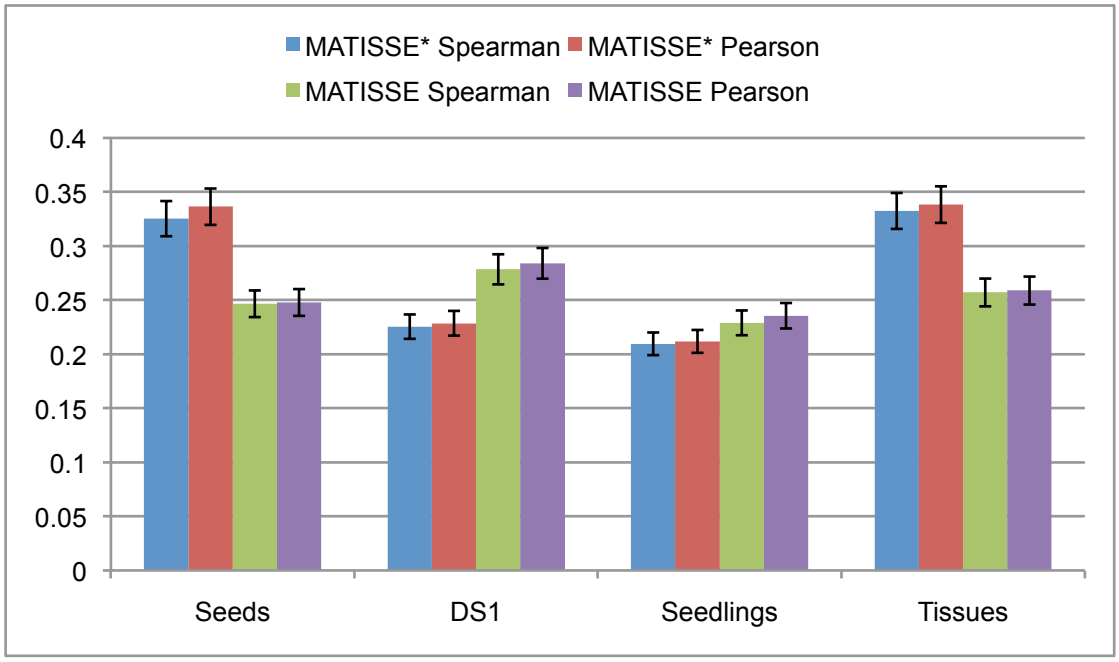
Figure 3.5. The selection algorithm obtains significantly higher scores on real biological pathways than on random pathways of the same size. For each size between 11 and 30 we generated 200 random gene sets (using both metabolic and non metabolic genes) and used the selection-ranking algorithm to get an AUC-SR score. Each box-plot depicts the average and the range of 25% to 75% of the AUC-SR scores provided by the random gene sets. The horizontal bars represent the maximal and minimal scores. The scores of all 'real' biological pathways in the 11-30 size range are plotted; pathways that received a score above 0.79 are named (14 pathways in total).). Box plots represent the median (black line), top 75 percentile (above the black line) and lower 25 percentile (below the black line). The dashed lines represents outliers.

Supplementary Figures:

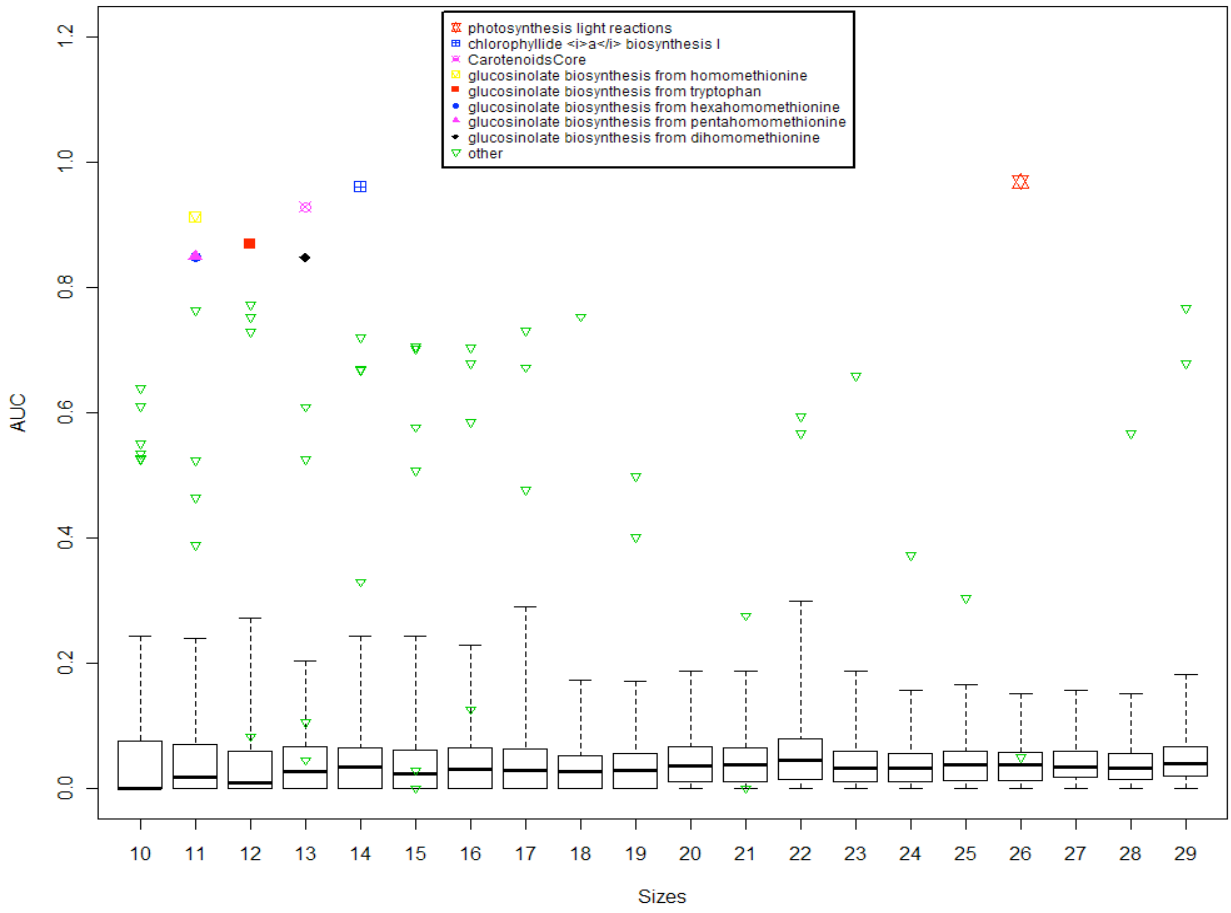


3

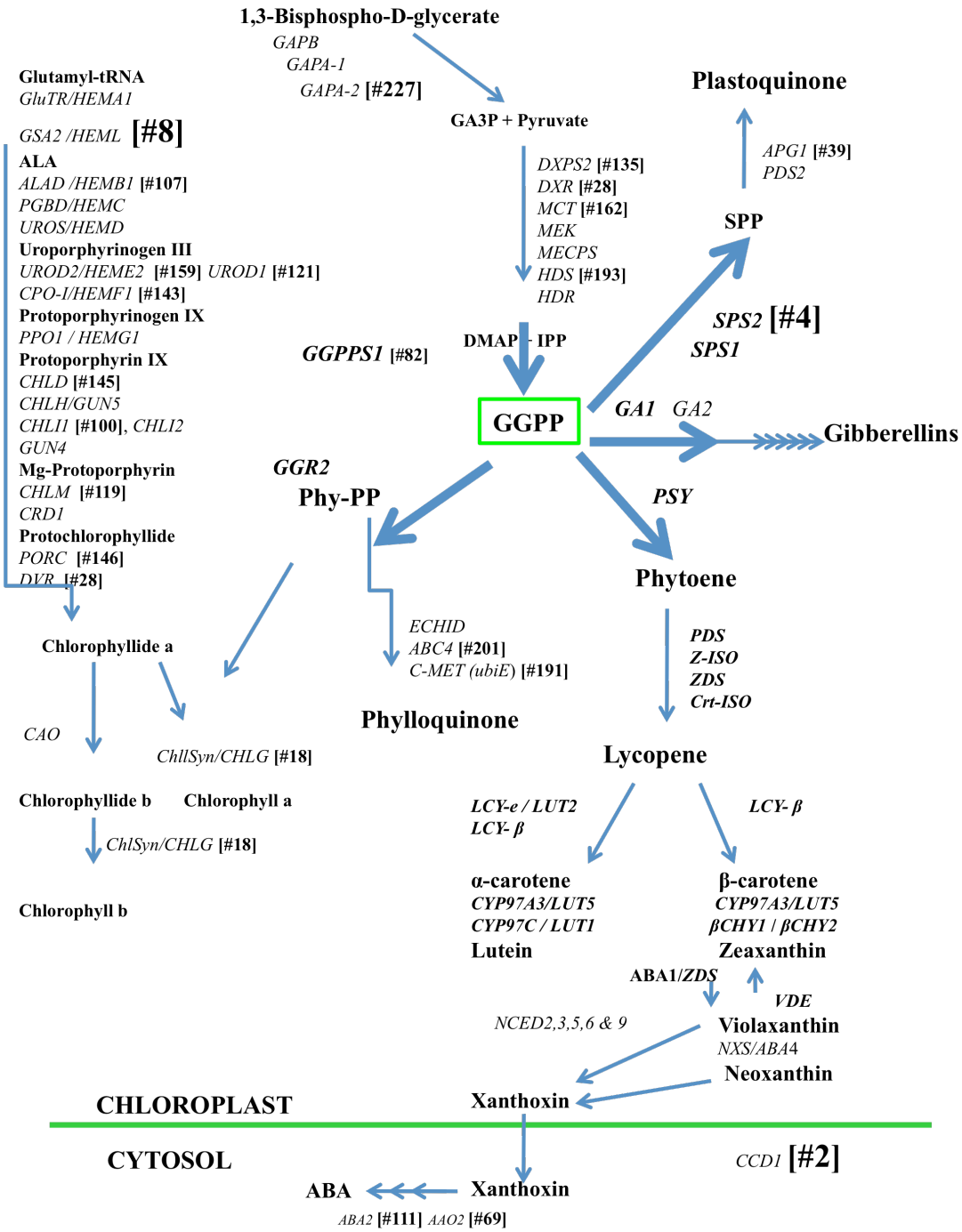
Supplementary Figure 3.5.1. Ranking quality using MATISSE and MATISSE* with the metabolic dependency network. The histogram gives the results using the original MATISSE modules and the extended modules obtained using MATISSE*, in terms of average AUC-SR scores among 66 tested pathways. The metabolic dependencies network was used in the tests as the network input for MATISSE and MATISSE*.



Supplementary Figure 3.5.2. Ranking quality using MATISSE and MATISSE* with the protein-protein interaction (PPI) network. Comparison between the results using the original MATISSE modules and the extended modules obtained using MATISSE*, in terms of average AUC-SR scores among 66 tested pathways.



Supplementary figure 3.8. The SMDS algorithm obtains significantly higher scores on real biological pathways than on random pathways of the same size. For each size between 10 and 29 we generated 200 random gene sets (using both metabolic and non metabolic genes) and used the SMDS ranking algorithm to get an AUC-SR score. Each box-plot depicts the average and the range of 25%-75% of the AUC-SR scores provided by the random gene sets. The horizontal bars represent the maximal and minimal scores. The scores of all 'real' biological pathways in the given size range are plotted and pathways that received a score above 0.79 genes are named (8 pathways in total). Box plots represent the median (black line), top 75 percentile (above the black line) and lower 25 percentile (below the black line). The dashed lines represents outliers.



Supplementary Figure 3.9.1 Diagram of the plastidial isoprenoid biosynthesis pathways. The pathways represented include the Calvin Cycle, MEP, Carotenoid, Chlorophyll, Phylloqinone, Plastoquinone, ABA and Gibberellin biosynthetic pathways. Reaction substrates and products are represented in bold black letters while genes that

encode pathway enzymes are in black italic letters. Numbers in square brackets near the gene names indicate the ranking of these genes by MORPH.

Chapter 3 Table:

Table 3.1. Selection algorithm results summary. For every tested pathway the best learning configuration is presented, the AUC-SR score and the pathway size.

Pathway name	Gene expression	Modules finding algorithm	Similarity score	AUC-SR	Size
beta;-alanine biosynthesis II	Seeds	MD network	Spearman	0.620	13
abscisic acid glucose ester biosynthesis	DS1	IsEnzyme	Spearman	0.784	10
aerobic respiration -- electron donor II	Tissues	SOM	Pearson	0.622	39
aerobic respiration -- electron donor III	DS1	MD network	Pearson	0.848	34
ascorbate biosynthesis I (L-galactose pathway)	Seedlings	MD network	Spearman	0.431	13
ascorbate glutathione cycle	Seedlings	MD network	Spearman	0.539	22
Calvin-Benson-Bassham cycle	Seedlings	MD network	Pearson	0.777	29
CarotenoidsCore	Seedlings	MD network	Spearman	0.861	13
CarotenoidsMEP	DS1	No clustering	Spearman	0.503	21
CDP-diacylglycerol biosynthesis I	DS1	MD network	Spearman	0.538	17
CDP-diacylglycerol biosynthesis II	DS1	MD network	Spearman	0.538	17
cellulose biosynthesis	DS1	MD network	Pearson	0.381	26
chlorophyllide biosynthesis I	DS1	MD network	Pearson	0.971	14
choline biosynthesis III	Seedlings	MD network	Spearman	0.544	10
cysteine biosynthesis I	Tissues	MD network	Pearson	0.788	11
ethylene biosynthesis from methionine	Tissues	MD network	Pearson	0.658	12
fatty acid β-oxidation II (core pathway)	Tissues	MD network	Pearson	0.625	15
flavin biosynthesis I (bacteria)	Seedlings	MD network	Pearson	0.660	14
flavonoid biosynthesis	Seedlings	MD network	Pearson	0.762	18
folate polyglutamylation I	Tissues	MD network	Pearson	0.331	10
folate transformations	Tissues	MD network	Pearson	0.661	11
formylTHF biosynthesis II	Tissues	MD network	Pearson	0.600	15
galactose degradation III	Seedlings	MD network	Pearson	0.291	24
gluconeogenesis	Tissues	MD network	Pearson	0.823	30
glucosinolate biosynthesis from dihomomethionine	Tissues	IsEnzyme	Pearson	0.857	13
glucosinolate biosynthesis from hexahomomethionine	Tissues	IsEnzyme	Pearson	0.874	11
glucosinolate biosynthesis from homomethionine	Tissues	IsEnzyme	Pearson	0.898	11
glucosinolate biosynthesis from pentahomomethionine	Tissues	IsEnzyme	Pearson	0.874	11
glucosinolate biosynthesis from tryptophan	Seedlings	MD network	Spearman	0.718	12
glycolysis I	Tissues	MD network	Pearson	0.748	44
glycolysis IV (plant cytosol)	Tissues	MD network	Pearson	0.747	44
glyoxylate cycle	Seedlings	MD network	Pearson	0.677	17
homogalacturonan biosynthesis	Seedlings	PPI network	Spearman	0.914	12
homogalacturonan degradation	Seeds	PPI network	Spearman	0.828	21
isoleucine degradation I	Seedlings	MD network	Pearson	0.626	14

jasmonic acid biosynthesis	Tissues	SOM	Spearman	0.353	15
leucine degradation I	Seedlings	MD network	Pearson	0.616	11
leucodelphinidin biosynthesis	Seedlings	MD network	Pearson	0.702	15
leucopelargonidin and leucocyanidin biosynthesis	Seedlings	MD network	Pearson	0.702	15
methionine salvage pathway	Tissues	MD network	Pearson	0.648	10
methylerythritol phosphate pathway	DS1	MD network	Pearson	0.661	10
NAD/NADH phosphorylation and dephosphorylation	DS1	IsEnzyme	Pearson	0.858	25
oxidative ethanol degradation I	Tissues	MD network	Spearman	0.294	11
phenylpropanoid biosynthesis	DS1	IsEnzyme	Spearman	0.850	16
phospholipases	Seedlings	MD network	Pearson	0.668	16
photorespiration	DS1	MD network	Pearson	0.594	15
photosynthesis light reactions	Tissues	MD network	Spearman	0.995	26
purine nucleotide metabolism (phosphotransfer and nucleotide modification)	DS1	IsEnzyme	Spearman	0.409	19
pyrimidine ribonucleotides interconversion	Seeds	MD network	Spearman	0.255	10
quercetin glucoside biosynthesis (Arabidopsis)	DS1	IsEnzyme	Spearman	0.687	13
Rubisco shunt	Tissues	MD network	Pearson	0.679	21
salvage pathways of purine nucleosides II (plant)	DS1	MD network	Spearman	0.612	14
starch biosynthesis	Seedlings	MD network	Spearman	0.606	15
starch degradation	Seedlings	MD network	Pearson	0.481	22
suberin biosynthesis	DS1	MD network	Pearson	0.832	12
sucrose biosynthesis	DS1	MD network	Spearman	0.674	16
sucrose degradation III	DS1	MD network	Spearman	0.686	28
TCA cycle variation III (eukaryotic)	DS1	MD network	Pearson	0.792	29
trehalose biosynthesis I	Seeds	MD network	Pearson	0.680	13
triaacylglycerol biosynthesis	Seedlings	MD network	Pearson	0.310	19
triacylglycerol degradation	Tissues	MD network	Spearman	0.641	12
tRNA charging pathway	DS1	IsEnzyme	Pearson	0.880	32
UDP-D-xylose biosynthesis	Seeds	MD network	Pearson	0.658	14
UDP-glucose biosynthesis (from glucose 6-phosphate)	Tissues	PPI network	Pearson	0.277	16
valine degradation I	Seedlings	MD network	Pearson	0.629	23
very long chain fatty acid biosynthesis	DS1	MD network	Pearson	0.678	17

CHAPTER 4. LOOKING TO THE FUTURE

Systems biology is an inter-disciplinary field, which merges molecular biology and biochemical techniques, with computational models and tools. One of the innovative aspects of systems biology is that it can scan a large amount of data for each studied system. Therefore, the post-genome invites the combination of the capacities of different systems biology discipline. The challenge that remains is to analyze the enormous amounts of data, and to put the analysis results in a biologically relevant context. The present work exploited systems biology approaches to perform a wide-scale investigation of the carotenoid biosynthetic pathway in the context of the entire *Arabidopsis* genome.

In the first stages of my work, I queried publicly available data sources and gathered large quantities of high-throughput information. I used open source and web-based tools, to explore carotenoid gene expression patterns collected under various conditions and from behavior of mutant strains (see Aappendix II, Figures 2-5). I then built a CBRG co-expression network, using co-expression data collected from the web (ACT), and visualized the network using Cytoscape, (Figure 2.3). While co-expression networks can uncover unknown relations between genes, and predict functional roles for genes, they often lack information about regulatory factors. One reason for this is the incomplete picture of biological process provided by microarray chips, which do not cover the entire spectrum of processes. Furthermore, small-scale analysis of gene expression under temporal conditions or at various developmental stages is inheritably noisy. For these reasons, caution must be taken when interpreting genome-wide compendia of co-expression data (Atias *et al*., 2009). A previous study showed that statistical significance of co-expression relationships often does not reflect biological relevance (Usadel *et al*., 2009).

Many plant scientists express interest in study of plant pathways. However, while analysis of *Arabidopsis* has yielded an abundance of data, in other plant species, the quantities are still minute, compared to those available for other model systems, such as *Escherichia coli*, *Saccharomyces cerevisiae, Drosophila melanogaster Mus musculus* or *Homo sapiens*. Therefore, there is a need to develop a simple framework that will integrate high-throughput plant data, namely, information generated by

network-based approaches (like co-expression and metabolic networks) with information derived from biological and biochemical investigations (like localization, participation in gene family and function). To date, several systems biology-based tools predict interactions among *Arabidopsis* genes. The Predicted Arabidopsis Interactome Resource (PAIR) (Lin et al., 2009), provides relatively comprehensive and accurate analyses that are inferred from data derived from co-expression and co-localization information, as well as co-evolution, homologous interactions and domain interaction data from other species. This tool includes ~6000 experimentally reported interactions collected from TAIR (Swarbreck et al., 2008), IntAct (Aranda et al., 2010), BioGRID and (Stark et al., 2006), as well as ~145,000 interactions predicted by an evidence integration model (Lin *et al*., 2011). These predicted interactions are expected to cover only ~24% if the complete *Arabidopsis* interactome and its output is estimated to have accuracy of ~40%.

In contrast to the PAIR and other tools based on *Arabidopsis*-derived data, a novel and statistically robust method was devised, in collaboration with David Amar (from Tel Aviv Uinversity in Prof. Ron Shamir's lab). The described method presents a new way of gaining insight into biological pathways via a non-parametric and simple framework. The method (MORPH) features robust prediction power, providing for extrapolation of meaningful biological conclusions. This presented workflow was designed to carefully fit the most suitable subsets of available *Arabidopsis* data, in search of regulatory correlations between any given collections of genes considered a "biological pathway" (see chapter 3).

A recent study presented a comprehensive computational model called AraNet , which has the ability to predict the functions of uncharacterized plant genes (Lee et al., 2010). AraNet, developed for *Arabidopsis thaliana*, contains more than 19,600 genes and the predicted one million functional linkages among them. AraNet is suggesting linkages among neighboring genes based on a “guilt-by-association” model using 24 different data sets. AraNet is based on the idea that genes that physically located in the same neighborhood, or turn on in concert with one another, are probably associated with similar traits. Both MORPH and AraNet combining different data sets and curetted data bases in order to link new genes to known biological processes.

Although the development of the MORPH was initially triggered by the specific needs of our laboratory, namely, for close analysis of the carotenoid pathway, the algorithm is suitable for almost any biological pathway in plants. Moreover the method could be applied to any other model system that has enough available high-throughput data. The method flexibility lays in the incorporation of un-supervised machine learning methods, clustering methods and model selection procedures that are neither species- nor pathway-dependent. Application of the method on several biological pathways in *Arabidopsis* proved the ability of the algorithm to capture experimentally proven gene candidates related to known biological pathways (see Chapter 3).

The robustness of the predictions provided by the computational method presented here creates an exciting research methodology to explore regulation of biological pathways in plants. The next step in the evolution of this newly developed method will be to create a graphical user interface to be distributed for wide usage among plant researchers. The user interface should allow users to add data sets of their own (for example gene expression arrays, protein-protein interaction networks, metabolomic data, etc). This feature will allow the algorithm to also consider the newly added data in the learning configuration stage, possibly yielding candidate co-regulated genes of better prediction power.

APPENDICES:

Appendix I*:

* Permission to publish has been granted by all authors.

The *Phytoene Synthase* gene family in the Grasses: subfunctionalization provides tissue-specific control of carotenogenesis

Faqiang Li[1,2], Oren Tsfadia[1,2], Eleanore T. Wurtzel[1,2]

[1]Department of Biological Sciences, Lehman College, The City University of New York, 250 Bedford Park Blvd. West, Bronx, NY 10468; [2]The Graduate School and University Center-CUNY, 365 Fifth Ave., New York, NY 10016-4309

Corresponding author:

Dr. Eleanore T. Wurtzel

Department of Biological Sciences

Lehman College, The City University of New York

250 Bedford Park Boulevard West

Bronx, NY 10468

Tel.: 718-960-8643; Fax: 718-960-8236

E-mail: wurtzel@lehman.cuny.edu

Addendum to:

Li F, Vallabhaneni R, Wurtzel ET (2008) *PSY3*, a new member of the phytoene synthase gene family conserved in the Poaceae and regulator of abiotic stress-induced root carotenogenesis. Plant Physiol 146: 1333-1345

Li F, Vallabhaneni R, Yu J, Rocheford T, Wurtzel ET (2008) The maize phytoene synthase gene family: overlapping roles for carotenogenesis in endosperm, photomorphogenesis, and thermal stress tolerance. Plant Physiol 147: 1334-1346

KEY WORDS

Carotenoid biosynthesis, phytoene synthase, gene subfunctionalization, ABA, abiotic stress, transcriptional regulation

ABBREVIATIONS

PSY Phytoene synthase

ABA Abscisic acid

RUNNING TITLE: Carotenogenesis in the Grasses

ABSTRACT

Carotenoids represent a diverse group of naturally occurring pigments found in various taxonomies, including plants, fungi and bacteria [1]. In higher plants, carotenoids function as accessory pigments for photosynthesis and prevent photo-oxidative damage [2]. Apocarotenoids, carotenoid cleavage products, include the plant hormones abscisic acid (ABA) and strigolactone [3, 4]. ABA plays an important role in regulating plant abiotic stress responses, differential growth and embryo dormancy; the recently discovered strigolactone inhibits shoot branching. For humans and animals, plant dietary carotenoids serve as precursors of vitamin A and other nutritional factors [5]. In recent years, improvement of pro-vitamin A carotenoid content in seed endosperm of crop staples has been an important goal for alleviating global vitamin A deficiency [6, 7]. However, complexity and poorly understood pathway regulation are barriers to predictive metabolic engineering, especially in the Grass family (Poaceae) containing the major cereal crop staples (e.g. maize, sorghum, millets, rice, wheat, oats, barley, rye).

Predictable manipulation of the carotenoid biosynthetic pathway reviewed in [8] necessitates elucidation of biosynthetic step(s) that control carotenoid

accumulation in various tissues. Previous studies suggested that control of flux to carotenoids is mediated by phytoene synthase (PSY), the first committed enzyme in the plastid-localized pathway [9-13]. PSY catalyzes the condensation of two molecules of geranylgeranyl pyrophosphate into 15-*cis* phytoene, the backbone of all C40 carotenoids and derived apocarotenoids. In Arabidopsis, tomato and other dicot plants, the nuclear-encoded *PSY* genes were shown to be up-regulated during plastid development, leading to carotenoid accumulation in leaf, flower or fruit [14-16]; transgenic overexpression of *PSY* causes elevated carotenoid content [6, 17]. However, additional studies are needed to facilitate predictable manipulation of this pathway in the many agronomically important food crops of the Poaceae. Therefore, we chose maize as a model system and the *PSY* gene as a key potential regulator of pathway flux to study carotenoid regulatory mechanisms in the Grasses. The maize B73 inbred line was the standard for study since carotenoids are found in both leaves and endosperm, and therefore comparisons of gene expression could be made among various carotenogenic tissues.

In Arabidopsis, a single *PSY* gene regulates 15-*cis* phytoene synthesis in all tissues. In contrast, most Grass species contain at least two *PSY* genes; *PSY1* expression was shown to be strongly associated with endosperm carotenoid accumulation in maize and carotenoid absence in rice [10, 13]. More recent examination of the completed rice genome led to identification of a new *PSY* gene, *PSY3*, which was found to be present not only in rice, but also in maize and sorghum [18]. All three genes were shown to encode functional enzymes in a heterologous bacterial platform [13, 18]. Quantitative transcript profiling for the three *PSY* genes revealed that the maize *Y1* locus, *PSY1*, encoded the most abundant transcript in leaf and endosperm, carotenoid-accumulating tissues. In contrast, maize *PSY3* transcripts were found predominately in root and embryo, carotenoid-limited tissues [19]. The tissue-specific transcript patterns suggested that the maize *PSY* genes might be subfunctionlized and not merely redundant copies. We therefore tried to address the following questions: 1) What are internal and external cues that stimulate steady state mRNA levels of each maize *PSY* gene? 2) How do the three *PSY* genes contribute to controlling carotenogenesis for

different physiological purposes? 3) Is *PSY* subfunctionalization unique to maize or is it a more general phenomenon seen in other Grasses? 4) What are the promoter *cis*-elements that may be responsible for the tissue-specific transcript patterns of the *PSY* genes?

Endosperm carotenogenesis. In developing endosperm of the maize B73 inbred, or other genetically diverse inbreds carrying the *Y1* allele, *PSY1* was the only gene family member for which transcript levels significantly increased during endosperm development and correlated with accumulation of endosperm carotenoids [20]; only low levels of *PSY2* and *PSY3* transcripts were detected in endosperm. The contribution of *PSY2* and *PSY3* to endosperm carotenoid accumulation was further evaluated in a line carrying the *PSY1* allele, *y1-602C*. This allele carries a promoter mutation blocking *PSY1* expression in endosperm but not in leaves. In *y1-602C* plants, the presence of *PSY2* and *PSY3* endosperm transcripts which can potentially encode functional proteins, could not compensate for the absence of endosperm *PSY1*. These data indicated that presence of functional *PSY1* in endosperm is critical for carotenoid accumulation and the contribution of *PSY2* and *PSY3* was negligible. It could be that function of PSY2 and PSY3 proteins in endosperm plastids is prevented by an unknown mechanism that interferes with protein translation, protein stability, or metabolon biogenesis, among other possibilities.

Leaf carotenogenesis. The roles of maize *PSY* genes in leaf carotenogenesis were determined by monitoring transcript changes during de-etiolation. In dark-grown plants, *PSY1* represented the major transcript. *PSY2* was the only paralog for which transcript levels increased in response to illumination, suggesting that *PSY2* plays an important role in controlling leaf carotenogenesis during greening. Moreover, the photoinduction of *PSY2* by red or far-red light was repressed in the maize phytochrome deficit mutant *elongated mesocotyl1* (*elm1*) [21], indicating that phytochromes mediate *PSY2* photoinduction by red and far-red light. In contrast, photoinduction of *PSY2* was still observed in blue light illuminated *elm1* seedlings, suggesting that another photoreceptor in addition to phytochrome, might be involved in blue light induction of *PSY2*.

Abiotic stress: thermotolerance. Although maize *PSY1* was unresponsive to light during greening, we could not completely exclude a role in leaf carotenogenesis because of the high abundance of leaf *PSY1* transcripts. Thus, we used a *PSY1* frame-shift mutant, *y1-8549*, to investigate the effect on leaf carotenogenesis when *PSY1* function was eliminated. At high temperature, *y1-8549* mutant leaves were bleached due to photo-oxidative damage and chlorophyll and carotenoid levels decreased dramatically. These results suggested that functional maize *PSY1* is essential for maintaining leaf carotenoid content, particularly under heat stress growth conditions. Therefore, maize *PSY1* in a *Y1* background, has an overlapping role in controlling both endosperm and leaf carotenogenesis.

Abiotic stress: drought and salt. The abundance of *PSY3* transcripts in roots suggested that *PSY3* might have a unique role in root carotenogenesis, perhaps in the role of apocarotenoid formation. Moreover, searching of GenBank revealed a link between rice *PSY3* expression and the ABA pathway or abiotic stresses, indicating that *PSY3* in the Grasses may be involved in abiotic stress-induced carotenogenesis or in regulation of ABA biosynthesis. Maize seedlings were therefore subjected to various abiotic stresses to test this hypothesis. In roots, the levels of maize *PSY3* mRNAs were induced by drought, salt and exogenous application of ABA [18]. Moreover, the elevation in *PSY3* transcripts was accompanied by induced levels of carotenoid intermediates, elevation of other downstream carotenogenic genes, and followed by ABA accumulation. The levels of *PSY2* mRNAs were also affected by abiotic stress in leaves but *PSY1* mRNAs were not elevated in any tissues tested. In particular we showed that root ABA accumulation is limited by *PSY3* expression and *de novo* carotenogenesis in contrast to leaf ABA induction which is known to be regulated by carotenoid cleavage and not by carotenoid synthesis [22].

The prevalence of rice *PSY3* ESTs associated with abiotic stress was the rationale for us to investigate stress-induced regulation of maize *PSY3*. The up-regulation of maize *PSY3* transcript levels in response to abiotic stresses suggested that *PSY3* responses

were not unique to maize. Soon after we characterized the role of maize *PSY3* in root carotenogenesis, the subfunctionlization of *PSY* paralogs in rice was also described by Welsh et al [23]. Both rice *PSY2* and *PSY3* shared similar up-regulation patterns with their maize orthologs; rice *PSY2* was also photoinducible and root *PSY3* mRNA levels increased during ABA formation in response to salt or drought stress. However, maize and rice *PSY1* genes were different in their tissue specificity and in light responsive pattern; transcripts levels of maize *PSY1* showed endosperm developmental responses but lacked photoinduction, whereas the rice paralog showed the reverse, photoregulation without endosperm expression. We were intrigued that orthologs of only two of the three genes shared regulatory responses. We hypothesized that the progenitor allele, *y1*, found in many noncarotenogenic endosperm maize inbreds and the wild ancestor, teosinte, might actually possess photoregulation for *PSY1*; that domestication and cultivation may have led to selection of an allele exhibiting endosperm expression at the expense of light regulation.

Promoter elements and subfunctionalization. Examination of the *PSY1* promoters in the grasses might give some clues as to the basis for the differential gene expression among the *PSY1* alleles and across species. We hypothesized that the different expression pattern of maize *PSY1* might be due to a change in regulatory components. Examination of maize and rice *PSY1* promoter regions revealed that both genes shared a similar *cis*-acting element arrangement except for some transposon insertions in the maize *PSY1* (*Y1*) promoter (**Fig. 1A**). The insertion of transposon *ins2* at 300 bp up-stream of maize *PSY1* start codon has been shown to be statistically associated with endosperm carotenoid accumulation [10]. We suspected this transposon insertion might have pushed away the *cis*-acting elements essential for *PSY1* photoinduction. To verify this hypothesis, we tested for light regulation of *PSY1* in the allele *y1-602C* which lacks the *ins2* element. When *y1-602C* seedlings were illuminated, it was observed that the mRNA levels of both *PSY1* and *PSY2* increased while only *PSY2* transcript levels were photoinduced in B73, which carries the *Y1* allele (**Fig. 1B**). These results suggested that the maize *y1-602C* allele is more similar to rice *PSY1* and the maize ancestor and that the *ins2* insertion in the B73 *Y1* allele is responsible for the loss of light regulation of maize *PSY1* in *Y1* backgrounds. In

summary, maize plants that accumulate endosperm carotenoids, do not have the capacity for *PSY1* photoregulation in green tissue, but must rely on *PSY2* photoregulation alone.

Differential *cis*-acting elements in the *PSY* paralogs could also be responsible for the paralog-specific responses to various environmental cues. Besides the light responsive elements found in *PSY1* and *PSY2* promoters, an ABRE (ABA-responsive element) binding site was found in both rice and maize *PSY3* promoter regions but not in either *PSY1* or *PSY2* [[23] and our unpublished data (Li F, Tsfadia O, Wurtzel ET, unpublished data)].

In summary, the *PSY* gene duplication appears to be a general phenomenon in the Grasses. Subfunctionalization provides for fine control of carotenogenesis that serves numerous physiological purposes. Open questions remain regarding the unknown post-transcriptional mechanisms that control metabolon assembly and membrane association in various plastids found in different tissues.

Acknowledgments

This research was supported by grants (to ETW) from NIH (S06-GM08225, 1SC1GM081160-01, and 5SC1GM081160-02), PSC-CUNY, and NYS.

Literature Cited

1. Britton G, Liaaen-Jensen S, Pfander H, eds. Carotenoids Handbook. Basel: Birkhäuser Verlag, 2004.

2. Niyogi KK. Safety valves for photosynthesis. Current Opinion in Plant Biology 2000; 3:455–60.

3. Gomez-Roldan V, Fermas S, Brewer PB, Puech-Pages V, Dun EA, Pillot JP, et al. Strigolactone inhibition of shoot branching. Nature 2008; 455:189-94.

4. Umehara M, Hanada A, Yoshida S, Akiyama K, Arite T, Takeda-Kamiya N, et al. Inhibition of shoot branching by new terpenoid plant hormones. Nature 2008; 455:195-200.

5. Fraser PD, Bramley PM. The biosynthesis and nutritional uses of carotenoids. Progress in Lipid Research 2004; 43:228-65.

6. Giuliano G, Tavazza R, Diretto G, Beyer P, Taylor MA. Metabolic engineering of carotenoid biosynthesis in plants. Trends in Biotech 2008; 26:139-45.

7. Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, et al. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 2008; 319:330-3.

8. Matthews PD, Wurtzel ET. Biotechnology of food colorant production. In: Socaciu C, ed. Food Colorants: Chemical and Functional Properties. Boca Raton: CRC Press, 2007.

9. Randolph LF, Hand DB. Relation between carotenoid content and number of genes per cell in diploid and tetraploid corn. J Agr Res 1940; 60:51-64.

10. Palaisa KA, Morgante M, Williams M, Rafalski A. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 2003; 15:1795-806.

11. Wong JC, Lambert RJ, Wurtzel ET, Rocheford TR. QTL and candidate genes phytoene synthase and zetacarotene desaturase associated with the accumulation of carotenoids in maize. Theor Appl Genetics 2004; 108:349-59.

12. Pozniak CJ, Knox RE, Clarke FR, Clarke JM. Identification of QTL and association of a phytoene synthase gene with endosperm colour in durum wheat. Theor Appl Genet 2007; 114:525-37.

13. Gallagher CE, Matthews PD, Li F, Wurtzel ET. Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses (Poaceae). Plant Physiol 2004; 135:1776-83.

14. Bartley GE, Scolnik PA. cDNA cloning, expression during development, and genome mapping of Psy2, a second tomato gene encoding phytoene synthase. J Biol Chem 1993; 268:25718-21.

15. von Lintig J, Welsch R, Bonk M, Giuliano G, Batschauer A, Kleinig H. Light-dependent regulation of carotenoid biosynthesis occurs at the level of phytoene synthase expression and is mediated by phytochrome in *Sinapis alba* and *Arabidopsis thaliana* seedlings. The Plant Journal 1997; 12:625-34.

16. Giuliano G, Bartley GE, Scolnik PA. Regulation of carotenoid biosynthesis during tomato development. Plant Cell 1993; 5:379-87.

17. Zhu C, Naqvi S, Breitenbach J, Sandmann G, Christou P, Capell T. Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. Proc Natl Acad Sci U S A 2008; 105:18232-7.

18. Li F, Vallabhaneni R, Wurtzel ET. *PSY3*, a new member of the phytoene synthase gene family conserved in the Poaceae and regulator of abiotic-stress-induced root carotenogenesis. Plant Physiol 2008a; 146:1333-45.

19. Howitt CA, Pogson, Barry J. . Carotenoid accumulation and function in seeds and non-green tissues. Plant, Cell and Environment 2006; 29:435-45.

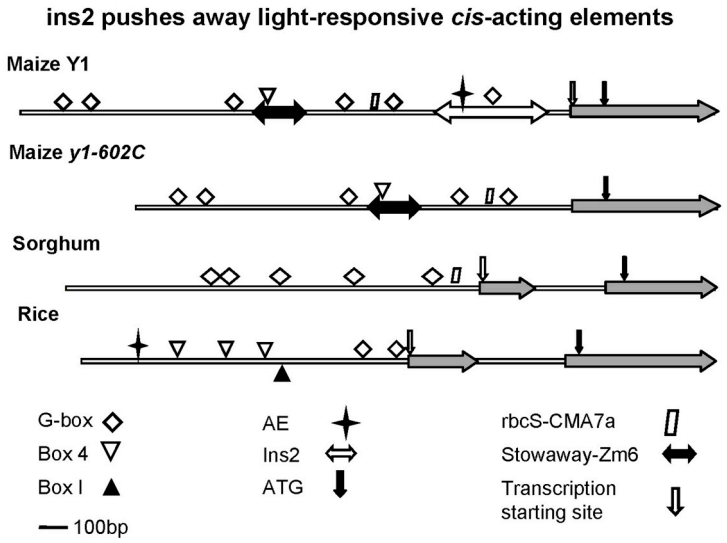
20. Li F, Vallabhaneni R, Yu J, Rocheford T, Wurtzel ET. The maize phytoene synthase gene family: overlapping roles for carotenogenesis in endosperm, photomorphogenesis, and thermal stress-tolerance. Plant Physiol 2008b; 147:1334-46.

21. Sawers RJH, Linley PJ, Farmer PR, Hanley NP, Costich DE, Terry MJ, et al. *elongated mesocotyl1*, a phytochrome-deficient mutant of maize. Plant Physiol 2002; 130:155-63.

22. Nambara E, Marion-Poll A. Abscisic acid biosynthesis and catabolism. Annu Rev Plant Biol 2005; 56:165-85.

23. Welsch R, Wust F, Bar C, Al-Babili S, Beyer P. A third phytoene synthase is devoted to abiotic stress-induced abscisic acid formation in rice and defines functional diversification of phytoene synthase genes. Plant Physiol 2008; 147:367-80.

A


ins2 pushes away light-responsive *cis*-acting elements
Maize Y1
Maize *y1-602C*
Sorghum
Rice
G-box
Box 4
Box I
100bp
AE
Ins2
ATG
rbcS-CMA7a
Stowaway-Zm6
Transcription starting site

B

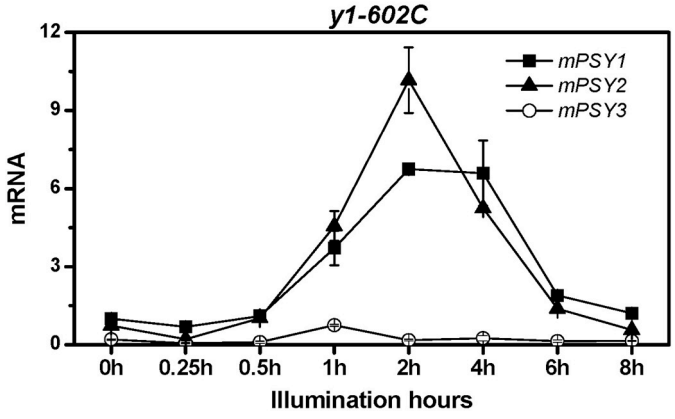

y1-602C
mPSY1
mPSY2
mPSY3
mRNA
12
9
6
3
0
0h
0.25h
0.5h
1h
2h
4h
6h
8h
Illumination hours

Figure 1. *ins2* transposon in maize *Y1* pushes away light-responsive *cis*-acting elements. (A) light-responsive cis-acting elements within maize *Y1* and *y1-602C* allele, sorghum and rice *PSY1* 5' upstream regions were predicted with PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html;). Transposon *ins2* inserts 300bp upstream of transcript starting site in maize *Y1* but not in allele *y1-602C*. (B) light-induced *PSY1* and *PSY2* expresssion during de-etiolation in *y1-602C* allele. Dark-grown nine-day-old dark-grown maize *y1-602C* seedlings were illuminated with white light (50 μmol m^{-2} s^{-1}) for 0 to 8 h. The leaves of illuminated seedlings were harvested for cDNA preparation and used as templates for quantitative RT-PCR as carried out previously [20]. All quantifications were firstly normalized to β-actin amplified using the same conditions and were made relative to *PSY1* transcript levels in unilluminated seedlings. Values represent the mean of three RT-PCR replicates ± SD from five pooled plants.

APPENDIX II*:

* Permission to publish has been granted by all authors.

A transcriptional analysis of carotenoid, chlorophyll and plastidial isoprenoid biosynthesis genes during development and osmotic stress responses in *Arabidopsis thaliana*

Stuart Meier[1*], Oren Tzfadia[2,3], Ratnakar Vallabhaneni[2,3], Chris Gehring[1,4] and Eleanore T. Wurtzel[2,3§]

[1] Division of Chemistry, Life Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

[2] Department of Biological Sciences, Lehman College, The City University of New York, 250 Bedford Park Blvd. West, Bronx, NY 10468

[3] The Graduate School and University Center-CUNY, 365 Fifth Ave., New York, NY 10016-4309

[4] Department of Biotechnology, University of the Western Cape, Private Bag X17, Cape Town - Bellville 7535, South Africa

[§]Corresponding author

Abstract

Background: The carotenoids are pure isoprenoids that are essential components of the photosynthetic apparatus and are coordinately synthesized with chlorophylls in chloroplasts. However, little is known about the mechanisms that regulate carotenoid biosynthesis or the mechanisms that coordinate this synthesis with that of chlorophylls and other plastidial synthesized isoprenoid-derived compounds, including quinones, gibberellic acid and abscisic acid. Here, a comprehensive transcriptional analysis of individual carotenoid and isoprenoid-related biosynthesis pathway genes was performed in order to elucidate the role of transcriptional regulation in the coordinated synthesis of these compounds and to identify regulatory components that may mediate this process in *Arabidopsis thaliana*.

Results: A global microarray expression correlation analysis revealed that the phytoene synthase gene, which encodes the first dedicated and rate-limiting enzyme of carotenogenesis, is highly co-expressed with many photosynthesis-related genes including many isoprenoid-related biosynthesis pathway genes. Chemical and mutant analysis revealed that induction of the co-expressed genes following germination was dependent on gibberellic acid and brassinosteroids (BR) but was inhibited by abscisic acid (ABA). Mutant analyses further revealed that expression of many of the genes is suppressed in dark grown plants by Phytochrome Interacting transcription Factors (PIFs) and activated by photoactivated phytochromes, which in turn degrade PIFs and mediate a coordinated induction of the genes. The promoters of *PSY* and the co-expressed genes were found to contain an enrichment in putative BR-auxin response elements and G-boxes, which bind PIFs, further supporting a role for BRs and PIFs in regulating expression of the genes. In osmotically stressed root tissue, transcription of Calvin cycle, methylerythritol 4-phosphate pathway and carotenoid biosynthesis genes is induced and uncoupled from that of chlorophyll biosynthesis genes in a manner that is consistent with the increased synthesis of carotenoid precursors for ABA biosynthesis. In all tissues examined, induction of the β-carotene hydroxylases are linked to an increased demand for ABA.

Conclusions:

This analysis provides compelling evidence to suggest that coordinated transcriptional regulation of isoprenoid-related biosynthesis pathway genes plays a major role in coordinating the synthesis of functionally related chloroplast localized isoprenoid-derived compounds.

Background

The carotenoids are pure isoprenoids that are synthesized in chloroplasts from geranylgeranyl diphosphate (GGPP) which additionally serves as an immediate precursor for other chloroplastic localized isoprenoid biosynthesis pathways including plastoquinone (PQ), the phytol tail of chlorophylls, phylloquinones (PhQ) and tocopherols as well as the phytohormone gibberellic acid (GA). While the biochemistry of carotenoid biosynthesis (CrtBS) has been extensively studied and most genes encoding enzymes that function in the CrtBS pathway have been identified, little is known about how the synthesis of these enzymes is coordinated and additionally how this synthesis is coordinated with that of other interdependent and interrelated isoprenoid-derived compounds. We have performed a global *in-silico* expression correlation analysis using microarray experimental data to identify genes that share a high level of co-expression and thus may share closely associated functional relationships with phytoene synthase (*PSY*). Comprehensive expression profiling of chloroplastic isoprenoid-related biosynthesis pathway genes was performed over a range of developmental and stress-related conditions in order to identify important regulatory components such as phytohormones and transcriptional regulatory factors that are important in coordinating their collective expression.

A number of chloroplast localized isoprenoid-derived compounds constitute important components of the photosynthetic apparatus. The carotenoids perform a range of functions including the acquisition of light energy and photoprotection [1] and additionally serve as precursors for abcisic acid (ABA) biosynthesis [2]. The chlorophylls are the main light absorbing pigments of the photosynthetic apparatus while PhQ and PQ function in photosynthetic electron transfer reactions. Plastoquinone additionally functions as an essential electron carrier in CrtBS desaturation reactions mediated by phytoene desaturase (PDS) and ζ-carotene desaturase (ZDS) [3]. As the GGPP molecule is an immediate precursor for the biosynthesis of these functionally related molecules, it serves

as an important metabolic hub in the biosynthesis of essential components of the photosynthetic apparatus (see Figure 1, dark text).

The synthesis of GGPP in plastids starts from pyruvate and glyceraldehyde 3-phosphate (GAP), that can be generated directly from the Calvin cycle (photosynthesis) or glycolysis [4], and serve as precursor molecules for the methylerythritol 4-phosphate (MEP) pathway [5]. The MEP pathway consists of a series of seven enzymes that function sequentially to catalyse the synthesis of the prenyl diphosphate precursors, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). GGPP synthase (GGPPS) then catalyses the sequential addition of three molecules of IPP to one molecule of DMAPP resulting in the formation of the poly-isoprenoid, GGPP [4].

The *PSY* gene encodes the first dedicated enzyme of the CrtBS pathway which catalyses the condensation of two molecules GGPP to form 15-*cis*-phytoene (Figure 1) [6-8]. Phytoene is metabolized to lycopene in a linear series of desaturation and isomerisation reactions that involves four enzymes [9,10]. CrtBS then branches into two distinct pathways (Figure 1), the β-cyclase (LCY-β) enzyme converts lycopene into β-carotene while the dual action of epsilon cyclase (LCY-ε) and LCY-β results in the formation of α-carotene. The α- and β-carotenes can then be hydroxylated to form α- and β- branch xanthophylls which are essential components of the photosynthetic apparatus in higher plants where they function in photosystem assembly, light harvesting and photoprotection [11,12]. In addition, violaxanthin and neoxanthin also serve as direct precursors for ABA biosynthesis and are alternative substrates for plastid localized nine *cis*-epoxycarotenoid dioxygenases (NCEDs) [2,13] (Figure 1).

The carotenoids have been shown to have important functional roles in early stages of post-germination development. In etioplasts of dark grown seedlings, lutein and violaxanthin biosynthesis is required for assembly of the prolamellar

body (PLB) [14,15]; a lattice of tubular membranes composed primarily of lipids, carotenoids and a ternary complex of NADPH, protochlorophyllide oxidoreductase (POR) and the chlorophyll precursor, protochlorophyllide (Pchlide) [16]. The synthesis of carotenoids in PLBs is thought to optimize the transition of etiolated plants to photomorphogenic development since it has been shown to enhance chlorophyll accumulation and greening upon light-induced de-etiolation [14,17].

Light activates the differentiation of etioplasts into chloroplasts in a process that is accompanied by a large and coordinated increase in the biosynthesis and accumulation of carotenoids, chlorophylls and pigment-binding proteins; this accumulation supports the development of a functional photosynthetic apparatus [14,15,18]. The light-induced synthesis of carotenoids is characterized by an increase in expression of *PSY* and select MEP pathway genes [18,19] as well as an increase in PSY enzymatic activity [15]. The expression of *PSY* has been shown to be elevated in response to a broad spectrum of continuous (c) light wavelengths including far-red (cFR), red (cR), blue (cB) and white (cW) [18,20,21]. While *PSY* transcript levels have been reported to increase in response to cFR [18], only light wavelengths that activate POR - which catalyzes the light-dependent conversion of Pchlide to chlorophyllide - cause the decay of PLBs, the synthesis of chlorophylls and the transition of etioplasts into chloroplasts [15,22]. These studies demonstrate that the coordinated and co-localized synthesis of carotenoids with chlorophyll precursors and chlorophylls in etiolated and de-etiolated plants respectively is required for normal photomorphogenic development.

The light-induced increase in *PSY* expression has been shown to be mediated by the phytochrome (PHYs) photoreceptors. Mutant studies have shown that the induction of *PSY* expression in response to cFR is dependent on the light-labile PHY-A while the cR-induction is thought to be mediated by light-stable PHYs

other than PHY-B [18]. Upon light-induced activation, the cytoplasmic localized PHYs are translocated to the nucleus where they interact with and mediate the degradation of the Phytochrome Interacting transcription Factors (PIFs); these factors bind to G-boxes in the promoters of light-induced genes and negatively regulate their expression [23]. Recently, the PIFs have been shown to have an important role in regulating the transcription of *PSY* and other carotenoid and chlorophyll biosynthesis genes during light-induced de-etiolation [24]. The PHYs and PIFs are therefore interesting candidate regulatory factors that may function to coordinate the transcription of genes that encode enzymes that function in the interrelated and interdependent chloroplastic isoprenoid biosynthesis pathways during early development.

Transcriptional co-regulation has been shown to play a major role in coordinating cellular responses that involve multiple genes and their products. A number of studies have shown that genes that have been confirmed to be co-expressed in response to a range of conditions have correlated functional relationships, including physical interactions between their encoded proteins [25-28]. These finding also extend to metabolic pathways where it has been shown that many genes encoding metabolic enzymes that function within the same or functionally related pathways form co-expression modules [29,30]. Thus, it is conceivable, that the synthesis of functionally related chloroplast localized isoprenoid molecules is mediated by their transcriptional co-regulation.

The model plant species *Arabidopsis thaliana* is ideal for studying global transcriptional responses since there are thousands of publicly available full-genome microarray experiments that encompass a broad range of experimental conditions including different developmental stages, stress, chemical and hormone treatments and mutants. In addition, analysis tools are available to identify modules of co-expressing genes and genome sequence data allows

analysis of promoter regulatory regions and the identification of putative regulatory elements.

It is pertinent to acknowledge that changes in gene transcription do not necessarily translate to changes in protein abundance and functional activity due to post-transcriptional regulatory mechanisms. However, as these mechanisms rely on a gene being transcribed in the first instance, gene transcription can be considered the primary level of regulation of protein synthesis. While cells can alter the activity of specific proteins/enzymes to fine tune cellular responses, the protein must be synthesized and present at appropriate quantities for this to occur. Changes in gene transcription in response to specific stimuli can be considered a primary regulatory response that reflects a change in requirement for a specific protein(s) at a specific point in time. In addition, in comparison to a single gene, when the expression of a large group of functionally related genes is altered in a uniform manner in response to a specific stimuli, it is a stronger indicator that the transcriptional response is representative of a cells intent to change the associated functional activity in response to the stimuli.

Here we aim to elucidate the role transcriptional regulation plays in coordinating CrtBS and the synthesis of other functionally related isoprenoid-derived compounds during early development and in response to osmotic stress. A global co-expression analysis revealed that *PSY* is highly co-expressed with many photosynthesis-related genes including, those involved in chlorophyll, PQ and PhQ biosynthesis as well as genes that function in the upstream Calvin cycle and MEP pathway that synthesise the commonly required GGPP precursor. Stimuli specific transcription profiling revealed that expression of the isoprenoid biosynthesis genes is almost universally activated following germination, during both etiolated and de-etiolated growth and the induction during early development is positively regulated by BRs and GA and inhibited by ABA. During etiolated growth, the PIFs appear to suppress the expression of the genes while

PHYs mediate their photoactivation. An enrichment in putative BR-auxin response elements and G-boxes (which bind PIFs) in the promoter of *PSY* and the co-expressed genes further supports a role for BRs and PIFs in regulating expression of the genes. In osmotically stressed root tissue, transcription of CrtBS-related genes is induced in a manner that is consistent with the increased synthesis of carotenoid precursors for ABA biosynthesis. In all tissues examined, induction of the β-carotene hydroxylases are linked to increased demand for ABA. We therefore conclude that transcriptional regulation plays a major role in coordinating the synthesis of functionally related isoprenoid-derived compounds in chloroplasts.

Results and Discussion

PSY co-expression analysis

In order to elucidate the role transcriptional regulation plays in coordinating CrtBS and the synthesis of other functionally related isoprenoid-derived compounds; an expression correlation analysis was undertaken using *PSY* as the driver gene in order to determine the level of co-expression that *PSY* shares with all of the other genes represented on the ATH1 microarray (22K) chip. Key to the accuracy of this analysis is that co-expression is measured over a large number of diverse experimental conditions (see methods and Ref [31]). *PSY* was selected as the driver gene for this analysis as it is the first dedicated enzyme of carotenogenesis and its transcription is known to be positively correlated with and a major driving force for carotenoid production [9,14,20,32]. It is thought that genes that are highly co-expressed with *PSY* will have closely associated functional roles.

The expression of *PSY* was shown to be highly correlated with many genes in the genome with the top 50 expression correlated genes (*PSY-ECG50*) having a Pearson correlation coefficient (r-value) ranging from 0.91 to 0.84 (Table 1). In total, approximately 1000 genes (4.3%) had an r-value >0.6 while around 600 (2.6%) had an r-value >0.7 supporting the specificity of the analysis since it indicates that *PSY* is co-expressed with only small percentage of select genes in the *Arabidopsis thaliana* genome. All genes in the *PSY-ECG50* had highly significant p-values ($< 1^{-35}$) and e-values ($<1^{-35}$) supporting the biological significance of the results.

Functional enrichment analysis of the PSY-ECG50

The high expression correlation of the *PSY-ECG50* is a strong indicator that these genes may function in common biological processes. The *PSY-ECG50* was therefore subjected to a functional enrichment analysis using "Fatigoplus" [33] which identified a number of significant enrichments in functional terms associated with the group (Table 1). In the biological process category, significant enrichments are found with genes associated with the terms photosynthesis, plastid organization and biogenesis, PQ biosynthetic process, and carotenoid and tetraterpenoid metabolic processes. In the cellular component category at level nine, genes associated with the terms plastid parts, thylakoid parts and chloroplasts are enriched.

Specifically, a number of genes in the *PSY-ECG50* encode enzymes that directly function in the synthesis of chloroplastic localized isoprenoids. This includes *ZDS* (At3G04870) and *LCY-β* (At3G10230; r = 0.86 for both; Table 1, Figure 1) which function in the CrtBS pathway and the *PHYTOENE DESATURATION* 2 (*PDS2*, At3g11950, r=0.86) [3,34,35] and *ALBINO OR PALE GREEN MUTANT* 1 (*APG1*; AT3G63410, r=0.86) [36] genes that both function in the PQ biosynthesis pathway. In addition to its function as an electron carrier in PSII light-dependent photosynthesis reactions, PQ is also an essential compound in the synthesis of carotenoids where it has a role as a hydrogen acceptor in the desaturation reactions mediated by PDS and ZDS [3]. The *GLCYERALDEHYDE-3-PHOSPHATE DEHYDROGENASE B SUBUNIT (GAPDβ)* gene (At1g42970, r=0.89) functions in the Calvin cycle to synthesise GAP which is a direct substrate for the MEP pathway [5,37].

In general, *PSY* is co-expressed with genes that encode proteins that have critical functional roles in the photosynthetic machinery; these proteins include enzymes that function in the biosynthesis of carotenoids, chlorophylls and components of the photosynthetic electron transport chain including PQ, PhQ,

plastidial NAD(P)H dehydrogenase complex, thioredoxin, ferredoxin, plastocyanin and the cytochrome b6/f complex as well as proteins that form structural components of photosystem I and II. The high co-expression of *PSY* with genes that encode proteins that have important functional roles in the photosynthetic machinery, including a number of isoprenoid biosynthesis genes, illustrates that *PSY* is indeed highly co-expressed with functionally related genes and this gives confidence in the accuracy of the analysis.

PSY co-expression with functionally related isoprenoid biosynthesis genes

The expression correlation values were next extracted for all known Arabidopsis genes that encode enzymes that function in plastidial isoprenoid biosynthesis; this included Calvin cycle and MEP pathway genes as well as carotenoid, chlorophyll, PQ, PhQ, ABA, and GA biosynthesis genes (Figure 1 - see Additional File 1 for full list of genes). These genes will collectively be referred to as the *PSY*-correlated interrelated isoprenoid biosynthesis genes (*PSY-CIIG*).

This analysis revealed that the expression of all nuclear genes that encode enzymes that are known or predicted to function at each of the individual steps in the CrtBS pathway are highly correlated with *PSY* (Figure 1). In addition, *PSY* is also highly co-expressed with many isoprenoid-related biosynthesis pathway genes including, Calvin cycle and MEP pathway genes as well as chlorophyll, PQs and PhQs biosynthesis genes (Figure 1). It is noteworthy that the expression of *PSY* was found to be highly correlated with genes that function in different branches of chlorophyll biosynthesis; this includes *chlorophyll synthetase* (*ChlSyn*, At3g51820, r = 0.82) that functions in phytol side chain biosynthesis, as well as *glutamyl tRNA reductase* (*GluTR/HEMA1,* r = 0.77, e-value $<1^{-35}$) and *glutamate 1-semialdehyde aminotransferase* (*GSA2,* r = 0.72, e-value $<1^{-35}$) that function in the upstream tetrapyrrole branch of chlorophyll biosynthesis. Significantly, the *GluTR and GSA2* genes encode enzymes that catalyse the biosynthesis of 5-aminolevulinic acid (ALA) which is the rate-limiting step for

this pathway [38,39]. The high degree of co-expression of these genes strongly suggests that their transcription is regulated by a common mechanism. In contrast, none of the ABA biosynthesis genes that operate downstream of ABA1/ZEP, or any GA biosynthesis genes are positively expression correlated with *PSY* (Figure 1).

This analysis also revealed that for the carotenoid and chlorophyll biosynthesis-related pathway enzymes that are encoded by multiple genes, only specific family members displayed high co-expression levels; this may imply their functional importance in their respective biosynthesis pathways (Figure 1 and Additional File 1). In the MEP pathway, *1-deoxy-D-xylulose 5-phosphate synthase* (DXPS) is the only enzyme that is encoded for by multiple (three) nuclear genes in Arabidopsis [40] and of these, only the functionally determined *DXPS2* (At4g15560, r = 0.69, e-value $<1^{-35}$) displays a high level of co-expression with *PSY* [41]. The two Arabidopsis *IPP isomerase* (IPPI) genes, show very little correlation with *PSY* (Additional File 1) and this is consistent with, and lends support to a recent study that reported that these enzymes have minor functional roles in plastidial isoprenoid biosynthesis, since IPPI and DMAPP are directly synthesized by the MEP pathway in plastids [42].

Of the family of 12 annotated *GGPPS* genes in Arabidopsis [41], only *GGPPS1* (At4g36810; r = 0.74, e-value $<1^{-35}$), that encodes a functionally active and plastid localized enzyme [43], displays a high level of expression correlation. The GGPPS-like protein, geranylgeranyl reductase (GGR, At4g38460, r= 0.64, e-value = 8.6^{-35}) also shows some expression correlation, however, GGR does not have GGPPS activity *in vitro* and its function remains unknown although it has been suggested to encode a GPPS subunit [43].

Since GGPP is a common substrate, and thus an important metabolic link in the synthesis of multiple isoprenoid-derived compounds, it is significant that in

addition to *PSY*, *GGPPS1* is the most highly co-expressed *GGPPS* in Arabidopsis with a number of other genes that encode chloroplast localized enzymes that directly use GGPP as a substrate. These genes include *GGR2* (At1g74470, *GGR2* to GGPPS1, r = 0.53 (e-value = 2.2^{-20}), data not shown) and two genes that encode solanesyl diphosphate synthase enzymes, *SPS-1* (At1g78510, *SPS1* to *GGPPS1,* r = 0.51 (e-value = 2.7^{-18}), data not shown) and *SPS-2* (At1g17050 , *SPS2* to *GGPPS1*, r = 0.58, (e-value = 6.8^{-26}), data not shown). The GGR2 enzyme reduces GGPP to phytyl pyrophosphate [44] which forms essential phytol side chains for both chlorophyll and PhQ (vitamin K1) biosynthesis while the chloroplast localized SPS-2 enzyme catalyses the synthesis of solanesyl diphosphate (SPP) which is thought to be a precursor of the PQ side-chain in Arabidopsis [45,46]. In addition, *PSY* is also highly co-expressed with all enzymes that function downstream in these pathways including, as mentioned *ChlSyn* for chlorophyll biosynthesis, *C-methyltransferase* for PhQ synthesis (At1g23360, r = 0.74) and, as mentioned above, *APG1* and *PDS2* for PQ synthesis (Figure 1).

A co-correlation scatterplot between *PSY* and *GGR2* (Figure 2) illustrates that both genes have a high level of co-expression with *GGPPS1* and many genes that function in the chlorophyll, PQ and PhQ biosynthesis pathways thus providing strong evidence that the GGPPS1 enzyme plays a major role in generating a common pool of GGPP substrate that is used in the biosynthesis of these compounds. This interpretation is supported by a recent study that shows a reduction in carotenoid and chloroplast levels in a *ggpps1* knock out mutant [47] and suggests that transcriptional regulation of the *GGPPS1* gene serves as an important regulatory node in coordinating carotenoid, chlorophyll, PhQ and PQ biosynthesis.

The scatterplot also shows that expression of carotenoid and chlorophyll biosynthesis genes is not correlated with any of the plastid localized GA biosynthesis genes, and in particular, *GGPPS1* showed no co-expression with

GA1(r = -0.26, e-value = 0.04) or *GA2* (r= - 0.03, e-value > 1), which directly use plastidial GGPP pools as substrates for GA synthesis. This implies that GGPPS1 does not function in the synthesis of GGPP for GA biosynthesis and this in turn is supported by reports that state that *ggpps1* mutants show no signs of GA deficiency [47] and that the generation of GGPP pools for GA biosynthesis is dependent on the action of GPPS (At2g34630) [48].

Of the carotenoid cleavage dioxygenase gene (*CCD*) family, only *CCD1* (At3g63520, r=0.76, e-value $<1^{-35}$), and to a lesser degree *CCD4* (At4g19170, r=0.51, e-value = 1.4^{-18}), showed any degree of expression correlation with *PSY* (Table 2). In Arabidopsis, the CCD1 protein is localized to the cytoplasm and thus inaccessible to plastid localized carotenoid substrates [49]. It has been proposed to function in the metabolism of carotenoids that are localised in the chloroplast envelope [50] or present in dry seeds which lack well defined organelles such as chloroplasts [49]. Significantly, *PSY* is not co-expressed with any of the five plastid localized ABA biosynthesis *NCED* genes (-2, -3, -5, -6) that catalyse the synthesis of xanthoxin from β,β-xanthophylls [50,51], or other downstream enzymes that function in ABA biosynthesis (Figure 2 and 3), indicating that transcription of the CrtBS genes is not directly coupled to ABA biosynthesis. Indeed, previous reports have shown that ABA biosynthesis is correlated with the expression of the NCED genes suggesting that their expression is important in regulating ABA-biosynthesis [2].

Stimulus specific expression analysis

The expression correlation analysis provides a generalized measure of how *PSY* is co-expressed with other genes in the genome since it is performed across multiple tissues and in response to a broad range of experimental conditions. In order to determine if expression of the genes corresponds to the known timing of carotenoid and chlorophyll biosynthesis and development of the photosynthetic apparatus, the expression of individual isoprenoid biosynthesis

genes and genes in the *PSY-ECG50* were examined throughout key developmental stages that are known to involve coordinated changes in the synthesis of carotenoids and chlorophylls.

Many previous studies on CrtBS gene expression have focused primarily on the transcriptional responses of a small subset of genes during de-etiolation since this is when large increases in CrtBS occur concomitant with the development of a functional photosynthetic apparatus [15,18,19,52]. However, carotenoids also have important functional roles during seed development, maturation and germination since they serve as precursors for ABA biosynthesis in developing seeds and CrtBS in dark-grown seedlings has been shown to be essential for PLB formation in etioplasts [14,17]. Thus, expression of the *PSY-CIIG* was examined throughout developmental stages encompassing seed development and maturation, imbibition and germination as well as etiolated and de-etiolated growth. Inhibitor and mutant experiments were also examined in order to determine the role that the early developmental-related phytohormones, ABA, GA and BR have in regulating the expression of these genes. Details of the experimental conditions for the microarray data examined are provided in Additional File 3.

Isoprenoid gene expression during seed and seedling development

The results in Figure 3 are presented as signal values since this provides information regarding the relative expression levels of individual genes at specific developmental stages and can provide insights into genes that may be rate-limiting due to low expression levels. While fold change ratios identify changes in expression in response to different conditions, they do not provide information of the relative abundance of transcripts.

The heat maps generated from the microarray expression analysis revealed that transcription of the *PSY-ECG50* is modulated in a largely uniform manner in response to a range of different experimental conditions (Additional file 2) which is consistent with the high expression correlation of these genes. In general, the expression of *PSY*, other select CrtBS genes, chlorophyll, PQ and PlQ biosynthesis genes decline progressively throughout seed development, remained very low in dry and stratified seeds before being induced during imbibition, germination, skotomorphogenic and photomorphogenic growth (Figure 3). In addition, most genes that function in the upstream biosynthesis pathways to synthesize the commonly required GGPP precursor also follow a very similar expression profile, including the Calvin cycle *GAPD* subunit encoding genes (*GAPD -β*, A-1 and A-2); the *MEP* pathway genes *DXPS2*, *MCT* and *GGPPS1* . The very low expression level of specific genes in mature (stage 10), stratified and dry seeds may translate to very low protein/enzyme levels which would be rate-limiting for the biosynthesis of their respective molecules. While the activity of enzymes may be modulated to fine-tune biosynthesis rates, if there is no enzyme present then transcription of the gene will become the rate-limiting factor.

The observed induction in *PSY* expression in two day dark grown plants and in response to cR (Figure 3) in the absence of increased expression of *PDS* and *ZDS* is consistent with previous reports which additionally show that this response is sufficient to activate carotenoid biosynthesis [14,18,20]. The absence of a requirement for *PDS* and *ZDS* induction at these stages may be explained by the relatively high expression levels of these genes in mature (stage 10) and stratified seeds, when compared to *PSY*, which will presumably maintain relatively high enzyme levels. The activity of the PDS and ZDS enzymes however may be regulated by the abundance of the PDS2 enzyme; PDS2 functions in the biosynthesis of PQ [3,34] which is essential for the desaturation reactions mediated by PDS and ZDS [53]. Since the transcription of *PDS2* closely mirrors that of *PSY* throughout these stages, it may be important in regulating the activities of the PDS and ZDS enzymes.

The collective moderate induction of these genes in dark grown plants and the strong induction in response to light coincides with the timing of carotenoid and chlorophyll biosynthesis during these developmental stages. As previously stated, carotenoids, the chlorophyll precursor Pchlide and the light-activated POR enzyme have been shown to accumulate in PLB of dark grown plants and carotenoid and chlorophyll biosynthesis is strongly activated in response to light [16]. This illustrates a correlation between biosynthesis pathway gene expression and carotenoid and chlorophyll biosynthesis at these developmental stages which implies that the transcriptional coordination of these pathway genes plays a major role in coordinating the synthesis of carotenoids and chlorophylls. Further, since *PSY* expression has been shown to be rate-determining and a major driving force for carotenoid biosynthesis [9,14], the closely coupled expression of other carotenoid and chlorophyll biosynthesis genes during these developmental stages suggests that expression of these genes may also be important in regulating and coordinating the biosynthesis of carotenoids and chlorophylls.

CrtBS gene expression and ABA biosynthesis in developing seeds

The expression profile of a number of CrtBS genes, including, *βCHY1* and *-2* and *ZDS/ABA1*, is in stark contrast with that of *PSY* during seed development in that their transcript levels remain elevated or increase during seed maturation, remain high in dry seeds and sharply decline during imbibition and in dark grown seedlings. These expression profiles are strikingly similar to a number of ABA biosynthesis genes, including NCED5, -6 and -9 that function directly downstream of *ZEP/ABA1*. Indeed, the induction of the ABA-responsive genes *EM6* and *RD29B* [54] at latter stages of seed development (stage 7) strongly supports that an increase in endogenous ABA biosynthesis and accumulation occurs at this stage [55]. The coupled induction of *βCHY1* and *-2*, and *ABA1* with ABA biosynthesis genes during later stages of seed development that coincide

with the accumulation of ABA indicates that *βCHY1* and *-2*, and *ABA1* may function to drive carotenoid intermediates towards β-xanthophyll and ultimately ABA biosynthesis during these stages. This is consistent with the observed reduced expression of *LCY-ε* in dry seeds and reports that *βchy1: βchy2* mutants have a reduced ability to synthesize ABA during drought stress [56]. Interestingly, while the expression of *βCHY1* and *-2* and *ABA1* is reduced during imbibition and in dark grown plants, their expression is rapidly induced by light with *βCHY1* being expressed at levels greater than two fold above that of *βCHY2*. The coupling of expression of these genes with other CrtBS genes in response to light may be indicative of their essential role in synthesizing β-xanthophylls which in turn are required for photoprotection.

Phytohormone regulation of isoprenoid gene expression in early development

Abscisic acid, GA and BRs have been shown to have important roles in regulating germination and post-germinative development and gene expression. Abscisic acid is known to inhibit germination and the expression of photosynthesis-related genes in imbibed seeds [54] while GA acts as essential hormone in promoting germination and etiolated development while negatively regulating ABA levels [57,58] in a process that is dependent on BRs [59]. Given the cross-talk between these important developmental-related phytohormones, we examined their role in regulating transcription of the genes using mutant and chemical treatment studies.

The induction of *PSY* and other photosynthesis-related genes following imbibition in continuous light is negatively regulated by ABA since in the ABA deficient mutant (*aba2*), induction is enhanced, while in the *cyp707a1, -a2 and -a3* triple mutant, which has elevated ABA levels, the induction is almost completely abolished (Figure 3). ABA-mediated suppression is consistent with the known inhibitory role of ABA in germination and reports that ABA inhibits the expression of photosynthesis-related genes at this stage [54]. The observed

ABA-mediated alteration of gene expression only occurs post-germination since gene expression levels in dry mutant seeds are not different from wild type (data not shown).

The presence of exogenous ABA or the GA biosynthesis inhibitor PAC in the growth media of light germinating seeds had very similar effects in that they strongly suppress the induction of *PSY* and other photosynthesis-related genes while maintaining expression of *βCHY2*, *ABA1* and other ABA biosynthesis genes (Figure 3). The high expression level of *EM6* and *RD29B* in PAC treated seeds indicates that these plants maintain high levels of ABA in the absence of GA. This is in line with reports that GA acts as essential hormone in promoting germination and negatively regulating ABA levels [57,58]. These results demonstrate that GA is required to activate the expression of *PSY* and the photosynthesis-related genes during germination in a process that is likely to involve a reduction in endogenous ABA levels. Indeed, ABA and GA are known to antagonistically regulate each others levels and GA levels in developing seeds follow an opposite trend to ABA in that they decrease progressively during seed maturation and increase sharply during germination [55].

The PAC-mediated repression of gene expression in light-germinated seeds observed here is in disagreement with a report that PAC increases expression of *PSY* and CrtBS genes in dark grown seedlings [14]. The discrepancy may be explained by differences in the growth conditions since in the above study [14], seeds were germinated in light for two to six hours before being grown in the dark for three days in the presence of PAC, whereas in the study analysed here, seeds were stratified and germinated in the presence of light and PAC or ABA [60]. Thus, it appears that GA is essential for the early induction of these genes immediately following germination but may inhibit their expression at later developmental stages. This confirms reports that GA is required for germination

and involved in the establishment of etiolated seedling development in darkness while repressing photomorphogenesis in a process that is dependent on BR [59].

The BRs also appear to have a positive role in regulating the expression of the genes. The expression of the *PSY-CIIG* was strongly reduced to non-detectable levels in both root (six day old) and whole shoot tissue (four day old) in the *BREVIS RADIX* (*brx*) loss-of-function mutant; *brx* has an impaired root development phenotype due to a root-specific BR deficiency [61] (Figure 4). The transcription of the genes in *brx* was rapidly restored to control levels following three hour brassinolide (BL) treatment strongly indicating that the miss-regulated expression was due to BL deficiency [61]. The addition of BL to wild type plants, however, failed to alter expression of the genes, indicating that while optimal levels of BL are required for correct expression of the genes, excess BL does not induce further expression. In contrast to the reduced expression of photosynthesis-related genes in the 'minimal' shoot tissue of young *brx* seedlings [61], the shoot system morphology, including leaves of mature *brx* plants have been reported to resemble that of wild-type plants [62]. This suggests that the repression of photosynthesis-related genes in the shoot system of *brx* may only be temporary in young developing seedlings. The effect may result from ABA-mediated inhibition since BRs have been shown to positively regulate germination by reducing ABA sensitivity [63]. In addition, along with GA, BRs have been found to function in the establishment of etiolated development in Arabidopsis seedlings while repressing photomorphogenesis [59]. Thus, like GA, BL appears to be required for the normal expression of carotenoid and chlorophyll biosynthesis genes in young post-germinative tissue in a process that may involve inhibition of ABA sensitivity and biosynthesis.

GA biosynthesis genes

The expression of *GPPS* that functions in GA biosynthesis, increased progressively during seed development, remained high in dry and imbibed seeds

before being reduced in dark- and light-grown plants (Figure 3). This is consistent with the established role of *GPPS* in the synthesis of GGPP pools for GA biosynthesis, but not carotenoid and chlorophyll biosynthesis [48]. Although the *GA1* and *GA2* genes that encode plastid localized enzymes that directly use GGPP as a substrate showed minimal differential expression throughout. The expression profile of *GA3* was similar to *GPPS* which is in line with their functional roles in the early steps in the GA biosynthesis pathway. The high expression level of *GPPS* in maturing, dry and imbibed seeds is likely to contribute to the increased synthesis of GGPP pools that is required for the increase in GA that occurs during germination [55,58]. The expression profile of these GA biosynthesis genes is in marked contrast to those of *GGPPS1*, *PSY* and *GGR2* that decrease during seed maturation and increase in dark- and light-grown seedlings. Thus, these distinct expression profiles are entirely consistent with GPPS functioning in the synthesis of GGPP pools for GA biosynthesis and GGPPS1 catalysing the synthesis of GGPP precursors for carotenoid and chlorophyll biosynthesis. A number of the late GA biosynthesis genes including the GA20 and GA3 oxidases that encode enzymes catalyzing the final steps in the synthesis of bioactive GAs are also strongly induced during imbibition (data not shown) in line with elevated GA levels that occur at this stage [58].

Regulation of isoprenoid gene expression by PHYs and PIFs during early development

The expression of the gene groups was next examined in a time course experiment where dark grown seedlings were exposed to cFR and cR. This experiment was performed on the 8K *Arabidopsis* microarray chip that does not include all genes used in our analysis; it does however include *PSY* and a number of other genes under investigation and thus provides a reasonable representation of the biosynthesis pathways being examined [64,65]. It is not unexpected that exposure to both cFR and cR induced a largely universal increase in expression of the isoprenoid biosynthesis genes including, *PSY*, *GAPDβ*, *GGR2*, *GluTR*, *GSA2* and *ChlSyn* (Figure 4, for 6h time point). These light

responses were additionally examined in *phy-A* and *-B* mutants since the PHYs are considered to be the predominant photoreceptors that mediate light-induced germination [66]. The cFR-induction was largely abolished for most genes in the *phyA* mutant while the cR-induction remained largely unaltered in the *phyB* mutant; these studies illustrate that *phyA* is required for early cFR-induced gene expression while the early cR induction can be mediated by PHYs other than PHYB. The importance of PHYA as an essential signaling component of cFR-regulated gene expression is well documented with a number of studies reporting that *phyA* mutants are disrupted in cFR-induced expression [64,65,67]. In addition, PHYA has been shown to be the dominant PHY in mediating the induction of early-response genes to cR [64,68,69] and to exert an early functional role in inhibiting hypocotyl growth [70]. However, while PHYB does not appear necessary to activate early cR-induced gene expression, *phyB* mutants have been reported to display a distinct morphological phenotype in cR, including long hypocotyls and small cotyledons, pointing to an important functional role for PHYB in plant photomorphogenesis [67].

While it is evident that cFR activates expression of many carotenoid and chlorophyll biosynthesis genes, and can induce de-etiolation (repress hypocotyl elongation) via a PHYA-dependent mechanism [67], it does not activate chlorophyll biosynthesis or chloroplast development which is dependent on light-induced activation of POR that catalyzes the conversion of Pchlide to chlorophyllide [22]. Indeed, it has been demonstrated that dark-grown seedlings exposed to cFR have carotenoid and chlorophyll contents that are around 80% and 20% respectively of the levels present in seedlings exposed to cR [18]. Further, cFR grown plants have a phenotype that is an intermediate between dark and cR grown plants [71] and cFR induces a PHYA-dependent growth pattern essential for soil emerging seeds or seedling survival in conditions of deep canopy shade which are characterized by reduced ratios of R:FR [72]. Hence, while expression of all three Arabidopsis POR genes (-A, -B and -C) is high

in dark grown seedlings, the activation of their enzymatic activity and the induction of chlorophyll biosynthesis is ultimately light-dependent.

The PHYs are known to activate gene expression following their light-induced translocation from the cytoplasm to the nucleus where they specifically interact with PIFs and mediate their degradation [73,74]. The PIFs are a subset of basic helix-loop-helix (bHLH) transcription factors (TFs) that bind to the promoters of light-induced genes and function somewhat redundantly to repress their expression and photomorphogenesis in dark-grown seedlings [23,73]. The PHY-mediated degradation of PIFs allows activation of light-induced genes and *pif* mutants have been shown to have a *constitutive photomorphogenic (cop)-like* phenotype in true dark-grown seedlings [75]. A recent study showed that the PIF1 TF binds specifically to G-box *cis* motifs present in the *PSY* promoter and mutant studies revealed that PIF1 and other members of the PIF family function to inhibit *PSY* expression and carotenoid and chlorophyll biosynthesis in dark-grown seedlings [24]. The expression of the gene sets was thus examined in response to a number of PIF loss-of-function mutants in order to provide a broader systems perspective of the role PIFs have in regulating the synthesis of chloroplast localized isoprenoid derived compounds.

The expression of the gene sets was not substantially altered in dark grown *pif* single and double mutants including *pif1*, *pif3* and *pif4,5* (data not shown) and this is probably a reflection of their redundant functions. In the *pif -1,-3,-4,-5 quadruple mutant (pifq),* however, expression of the CrtBS genes and other genes involved in the synthesis of the photosynthetic apparatus reached quantitatively similar levels to that observed in 2 day cR-exposed wild-type plants (Figure 4). As reported previously for most dark grown *pifq* differentially expressed genes [23], the induction of the CrtBS genes occurs post-germination since their transcript levels are similar in *pifq* and wild type seeds. In this study [23], *PSY* and a number of other genes investigated, including *DXPS2*, *βCHY2*, *ABA1*, *GluTR*,

GUN5, *chlorophyllide a oxygenase* (*CH1*) and *CHL1* were identified as direct target candidates of PIF-mediated repression in the dark based on their expression being, firstly, elevated in dark grown *pifq* mutants compared to wild types, secondly, rapidly elevated after one hour Rc exposure (stimulates rapid PHY-induced PIF degradation) and thirdly, sustained after germination in two days cR. In addition, 84% of genes in the *PSY-ECG50* are induced >1.5 fold in dark-grown *pifq* mutants (Additional File 2).

While the PIFs clearly appear to negatively regulate expression of the genes in dark-grown seedling, it is noted that expression of many genes including *PSY*, Calvin cycle genes, MEP pathway and chlorophyll biosynthesis genes (Figure 3) as well as most genes in the *PSY-ECG50* (Additional File 2) were previously shown to be strongly induced in dark-grown wild-type seedlings when compared to stratified seeds. Although not as great as in response to light, for some genes, including, the three GAPD subunit encoding genes, *PSY, LCY-ε*, *DXPS2, MCT*, *GGR2*, *ChlSyn* and many other chlorophyll biosynthesis genes, the increase was greater than two-fold illustrating that expression of these genes is positively regulated in dark-grown wild-type seedlings. This documents that while the PIFs limit gene expression in dark grown seedlings, the inhibition is not absolute and that increases in expression do occur in the dark in the presence of PIFs. This is consistent with studies which show that increases in the biosynthesis of carotenoids and chlorophyll precursors in the dark is required for optimal greening upon light exposure [14].

In another related mutant experiment, dark-grown *phyAphyB* mutant seedlings expressing the constitutively active $Y^{276}H$ missense allele of Arabidopsis PHYB ($PHYB^{Y276H}$) [76], were similarly able to mimic the cR-induced transcriptional activation of the gene sets as observed in dark-grown *pifq* mutants. In this mutant, $PHYB^{Y276H}$ undergoes light-independent nuclear localization which may mediate degradation of PIFs and allow the induction of light-inducible genes

[77,78]. Thus, while the *phyB* mutant experiment indicates that cR induction of the gene sets can occur independent of PHYB, this experiment clearly shows that active and nuclear localized PHYB can induce expression of the genes in the dark.

In summary, these results strongly support that PHYA and PHYB have important functional roles in coordinating the transcription of the interrelated isoprenoid carotenoid and chlorophyll biosynthesis genes during de-etiolation. This process most likely involves the PHY-mediated degradation of PIFs, thus enabling light-induced gene expression.

Carotenoid gene expression and ABA biosynthesis in response to osmotic stress

The carotenoids are precursors for ABA biosynthesis and we reported in this study that expression of some late CrtBS genes is induced at a time that coincides with increased ABA biosynthesis in maturing seeds (Figure 3). We therefore next examined expression of the *PSY-CIIG* in shoot and root tissue in a time course response to osmotic stress (mannitol) which induces the synthesis of ABA and can thus help resolve how expression of CrtBS genes is coordinated with that of ABA biosynthesis in these tissues. The experimental results reveal some interesting tissue specific expression response patterns (Figure 5). Not surprisingly, responses were more immediate in root tissue where the stress was applied, resulting in an early and sustained increase in expression of a number of the genes including, Calvin cycle genes, MEP pathway genes and dedicated CrtBS genes including, *PSY*, *ZDS*, *βCHY1* and *-2*, *ABA1* and *VDE*. In a similar manner to maturing seeds, this increase was paralleled with a strong increase in expression of a number of ABA biosynthesis genes including *NCED3*, and the ABA-responsive genes, EM6 and RD29B, suggesting an increase in endogenous ABA levels [79]. It is noted that there was little change in the expression of chlorophyll biosynthesis genes here.

The increase in expression of CrtBS pathway genes in root tissue is in contrast to that in the shoot where there is a general reduction in expression of carotenoid and chlorophyll biosynthesis genes from around 3-6 h which progressively decreases up to 24 h. However, there is a strong and transient induction of *βCHY2* and *ABA1* between 3-12 h in shoot tissue while *NCED3* expression is induced early and sustained for the duration. As observed previously, *βCHY2* and *ABA1* expression is also strongly induced independent of other CrtBS genes during seed maturation, a process that also requires increased ABA biosynthesis and illustrates that expression of these genes can be uncoupled from other CrtBS genes, in both non-photosynthetic (seeds) and photosynthetic tissues, under conditions that require increased ABA biosynthesis.

The NCEDs have been proposed to be key regulators of ABA synthesis since their increased expression is correlated with increased endogenous ABA concentrations [80]. Notably here, *NCED2* and *-3* were the predominant *NCED*s induced in root and shoot tissue, which is in contrast to developing seeds where induction of the *NCED-5, -6*, and *-9* genes parallels the increase in ABA production.

The more universal induction of carotenoid-related biosynthesis genes in osmotically stressed roots may reflect the lower concentration of carotenoid precursors that are present in this tissue. Since photosynthetic tissue contains high concentrations of epoxycarotenoids, it appears that transcription, and presumably translation of only the late CrtBS genes, *βCHY2* and *ABA1* are required to increase violaxanthin precursor levels for ABA biosynthesis. In contrast, in root tissue which is a major site of ABA biosynthesis, low concentrations of carotenoids may be rate-limiting for ABA biosynthesis. We therefore propose that an increase in Calvin cycle and MEP pathway genes and a more universal induction of the CrtBS genes is required in root tissue to generate violaxanthin precursors for ABA biosynthesis, a hypothesis that is supported by

studies in maize [81] and rice [82]. The absence of any change in the expression of chlorophyll biosynthesis genes documents that the expression of the carotenoid and chlorophyll genes can be regulated independently, at least in root tissue.

Promoter enrichment analysis

The high expression correlation values of the genes within the *PSY-ECG50* points to the possibility that their expression is coordinately regulated. A promoter content analysis was therefore performed in an attempt to identify the presence of enriched putative regulatory elements that may be causative for their co-expression. The analysis was performed examining regions 2000 base pairs (bp) upstream of the coding region/translation start sites (TlSS) of genes since regulatory elements have previously been identified in 5 prime untranslated regions (5' UTR) of light-induced genes [83,84]. This is particularly relevant to *PSY* which is annotated in TAIR to have a 779 bp sequence upstream of the coding region that includes two 5'UTRs and an intron. A number of elements were found to be significantly enriched in the promoters of the co-expressed *PSY-ECG50* and are thus considered candidate regulatory elements that may coordinate their transcription. In addition, a number of these elements correspond to known plant *cis* regulatory elements including, a slightly degenerate G-box (CACGNG (p-value = 9.8^{-03}) compared to CACGTG) and the auxin-responsive element (AuxRE, TGTCTC (p-value =0.02), Additional File 4).

The G-box is known to be present in the promoters of many light-regulated genes [85-87] and is enriched in the promoters of genes that are rapidly-induced by PHYA [88]. As previously mentioned, G-boxes present in the promoter of *PSY* have been shown to specifically bind the PIF1 TF resulting in inhibition of *PSY* expression [24]. Thus, the identification of an enrichment of G-boxes in the promoters of the *PSY-ECG50* is consistent with the observed PHYA dependency for induction of these genes and the inhibitory effect of PIFs on their expression

in dark grown plants. We noted that one of the two G-boxes identified in the PSY promoter in this analysis is positioned in the 5' UTR in close proximity to the TlSS (-21 to -16 and -919 to -914) and both differ from those identified previously [67] which examined promoters more than 2000 bp upstream of the TlSS.

The enrichment of AuxREs in the promoters of the genes is consistent with the observed BR-dependency for CrtBS gene expression in young tissues. While the AuxRE was initially believed to confer auxin responsiveness to promoters [89,90], more recent studies have indicated that this element is also a target of BR signaling. It has been suggested that the AuxRE should in fact be considered a BR-AuxRE [91] since it has been found to be enriched in auxin- and BR-responsive genes rather than genes specifically regulated by auxin [92]. Thus, the enrichment of BR-AuxRE in the promoters of the genes is in line with the BR-dependent gene expression observed in young tissues in this study and additionally adds strength to studies that have shown BRs have a role in the establishing the etiolated development program in dark-grown Arabidopsis seedlings [59].

In summary, the GA- and BR-dependent induction of the carotenoid and chlorophyll biosynthesis genes following germination is entirely consistent with the role of these hormones in establishing etioplast development [59] and the requirement for carotenoid and chlorophyll precursor accumulation in developing etioplasts [14]. The expression level of the genes appears to be restricted by the PIF TFs in dark grown plants which are subsequently degraded by light activated PHY molecules allowing a strong and coordinated induction of the genes and a subsequent increase in carotenoid and chloroplasts biosynthesis. The identification of an enrichment in putative BR-AuxRE and G-boxes in promoters of the *PSY-ECG50* complements the observed transcriptional regulatory roles of BRs and PIFs respectively.

Conclusions

The tightly coupled expression and induction of PSY and many other isoprenoid biosynthesis genes throughout key developmental stages that correspond to the timing of increased carotenoid and chlorophyll synthesis and development of the photosynthetic apparatus strongly suggests that the coordinated transcription of these biosynthesis genes is critical in regulating and coordinating the biosynthesis of the functionally related carotenoid, chlorophyll, PQ and PhQ molecules. The phytohormones GA, BR and ABA as well as the transcriptional-related PHYs and PIFs appear to have important roles in regulating and coordinating the transcription of these isoprenoid-derived compounds.

Methods

PSY expression correlation analysis

An expression correlation analysis was performed for PSY using the freely available Arabidopsis co-expression tool (ACT) (http://www.arabidopsis.leeds.ac.uk/)[31]. This particular tool uses hybridization signal intensities from microarray experiments to calculate a Pearson correlation coefficient (r-value), which is a scale-invariant measure of expression similarity. The analysis was performed across all of the 322 available Ath1 22K microarrays from the NASC/GARNet dataset which contain probe sets that recognize 21,891 Arabidopsis genes. The arrays included in this analysis are derived from a broad range of experimental samples including specific tissue types, developmental stages, abiotic and biotic treatments, and a range of mutants. Importantly, the ACT tool uses NASC/GARNet data sets that were labeled, hybridised and analysed using a standardised procedure thus providing a homogeneous and readily comparable data set.

The analysis was performed leaving the gene list limit blank resulting in the return of a global correlation analysis of all probe sets relative to PSY ranging from the most positive to the most negatively expression correlated genes (total over 22,500 probe IDs). The top 50 genes that had the highest expression correlation with PSY were extracted from the list as were genes that were included in the *PSY-CIIG*. Both lists were filtered to include only genes that were represented by a unique probe on the microarray chip.

The co-correlation analysis was performed using the 2D scatter plot tool present on the ACT website. Probe IDs for *PSY* and *GGR2* was inserted into the X and Y axis and all other specifically highlighted genes were inserted in the highlight option.

Functional enrichment analysis

A gene ontology (GO) analysis was performed using the "Fatigo plus" (version 3.1) compare tool in the Babelomics suite (http://babelomics.bioinfo.cipf.es/EntryPoint?loadForm=fatigo) [33,93] to determine if there was any statistically enriched terms associated with the *PSY-ECG50* expression correlated genes compared to the expected frequency in the complete genome. The top 50 genes were selected for this analysis since their expression was highly correlated with *PSY* (r-value range 0.91-0.84). All the available functional annotation options for Arabidopsis were selected which include the three GO categories of biological process (BP), cellular component (CC) and molecular function (MF) as well as KEGG pathways. The tool uses a Fisher's exact test and returns adjusted *p*-values (Family Wise Error Rate) to accounting for multiple testing to determine statistical significance.

Microarray stimuli specific transcription analysis

An *in silico* global expression analysis was subsequently performed for both gene sets in response to specific stimuli and in selected mutants to identify conditions that induce differential expression of the genes (Figures 3-5 and Additional File 2). The expression profiles of *PSY* and its positively correlated gene sets were initially screened over all of the available ATH1: 22K array Affymetrix public microarray data in the gene response viewer tool (GRV) in Genevestigator [94]. Normalised microarray data were downloaded for experiments that were found to induce differential expression of the genes from the following sites:

NASCArrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl)[95], TAIR-ATGenExpress (http://www.ebi.ac.uk/microarray-as/ae/), GEO (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) [96].

(see attached file for experiment descriptions).

Promoter enrichment analysis

A number of tools in the POXO (http://ekhidna.biocenter.helsinki.fi/poxo) [97] promoter analysis suite were used to analyze promoter regions 2000 bp upstream of the coding regions of the genes in the *PSY-ECG50*. The POCO tool was used to identify enriched elements and the POBO tool was used to verify the presence of identified elements in the PSY promoter. The identified significantly enriched motifs were filtered to ensure that they were present and enriched in the PSY promoter and were present in greater than 70% of the genes in the *PSY-ECG50*.

Authors' contributions

ETW and CG conceived the initial project. OT and RV performed an initial analysis on the carotenoid biosynthesis pathway genes and promoters. SM expanded the project to include additional isoprenoid biosynthesis pathway genes, generated the results presented, interpreted data and wrote the manuscript with contributions by ETW, OT and CG.

Acknowledgements

This research was supported by grants (to ETW) from NIH (GM081160) and NYS, and grants (to CG) from the South African National Research Fund and the Oppenheimer Memorial Trust (South Africa).

Reference List

FIGURES

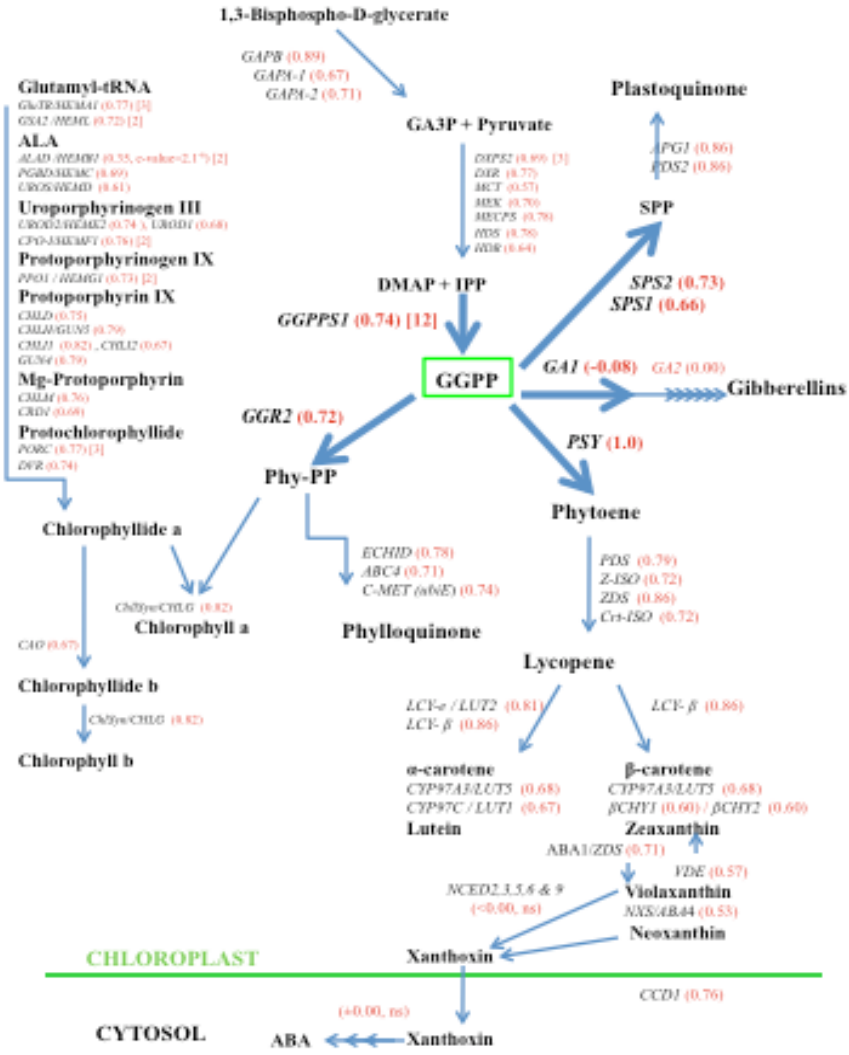


Figure 1. Diagram of the plastidial isoprenoid biosynthesis pathway detailing the level of co-expression that *PSY* shares with individual interrelated isoprenoid biosynthesis pathway genes. The pathways represented include the Calvin Cycle, MEP, Carotenoid, Chlorophyll, Phylloqinone, Plastoquinone, ABA and Gibberellins and are collectively referred to as the *PSY*-correlated isoprenoid interrelated genes (*PSY-CIIG*). Reaction substrates and products are represented in bold black letters while genes that encode pathway enzymes are in black italic letters. Numbers in red parentheses represent expression correlation r-values and numbers in square brackets indicate the number of paralog genes that are annotated to encode the respective enzymes. All r-values > 0.5 had p-values and e-values < 1.0^{-15}. Non-significant r-values are indicated as n.s. Only the highest correlated member of the paralog gene family and those that have a co-expression value > 0.6 are listed. See Additional File 1 for list of corresponding gene IDs, details of statistics for

individual genes and an extended list including additional paralog gene family members.

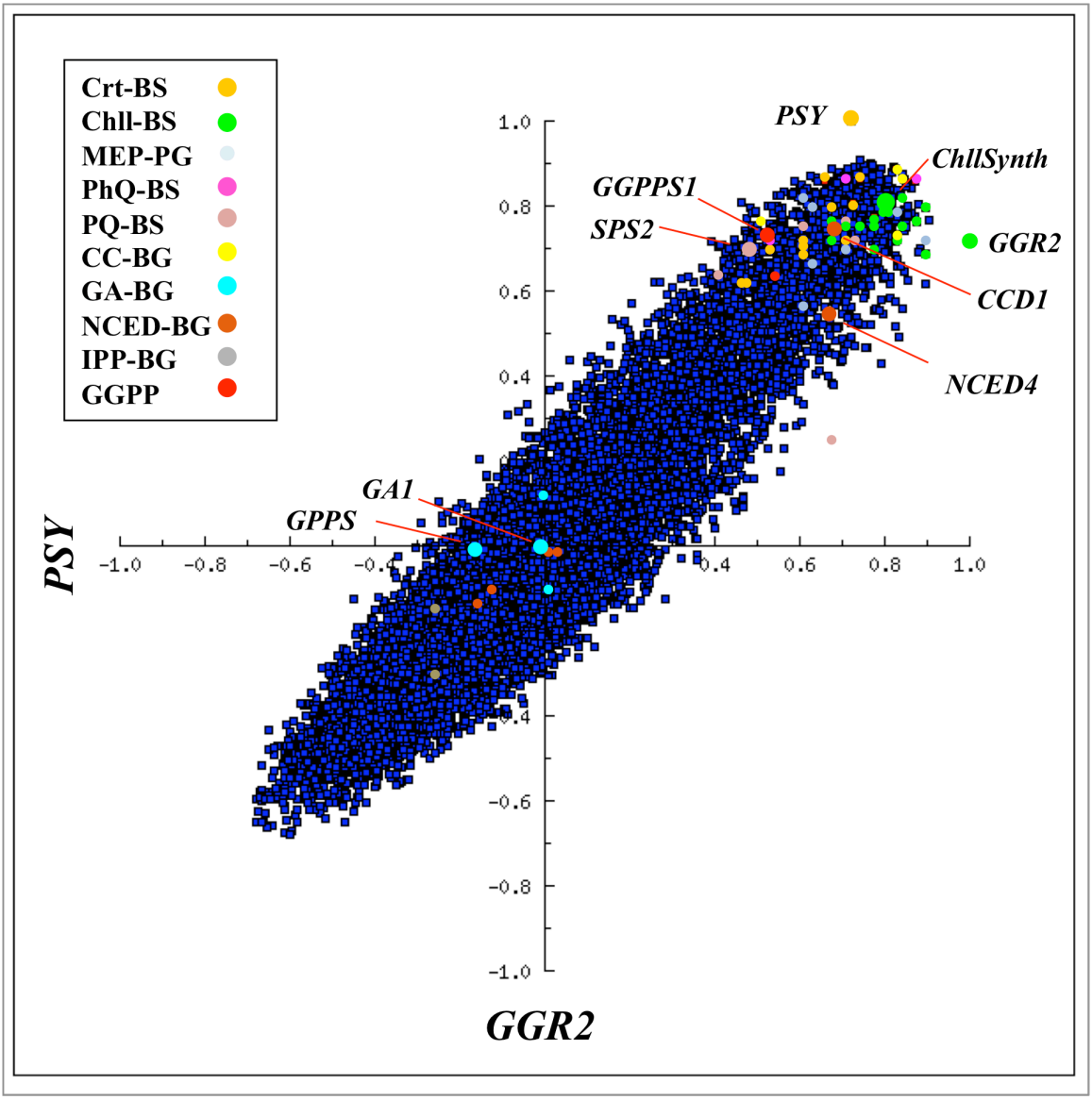


Figure 2. Co-correlation scatter plot illustrating the level of co-expression of all Arabidopsis genes relative to *PSY* and *GGR2*. Genes that function in defined biosynthesis pathways (*PSY-CIIG*) are color highlighted as indicated in the legend. Select individual genes of interest are highlighted. All genes listed in Figure 1 are represented.

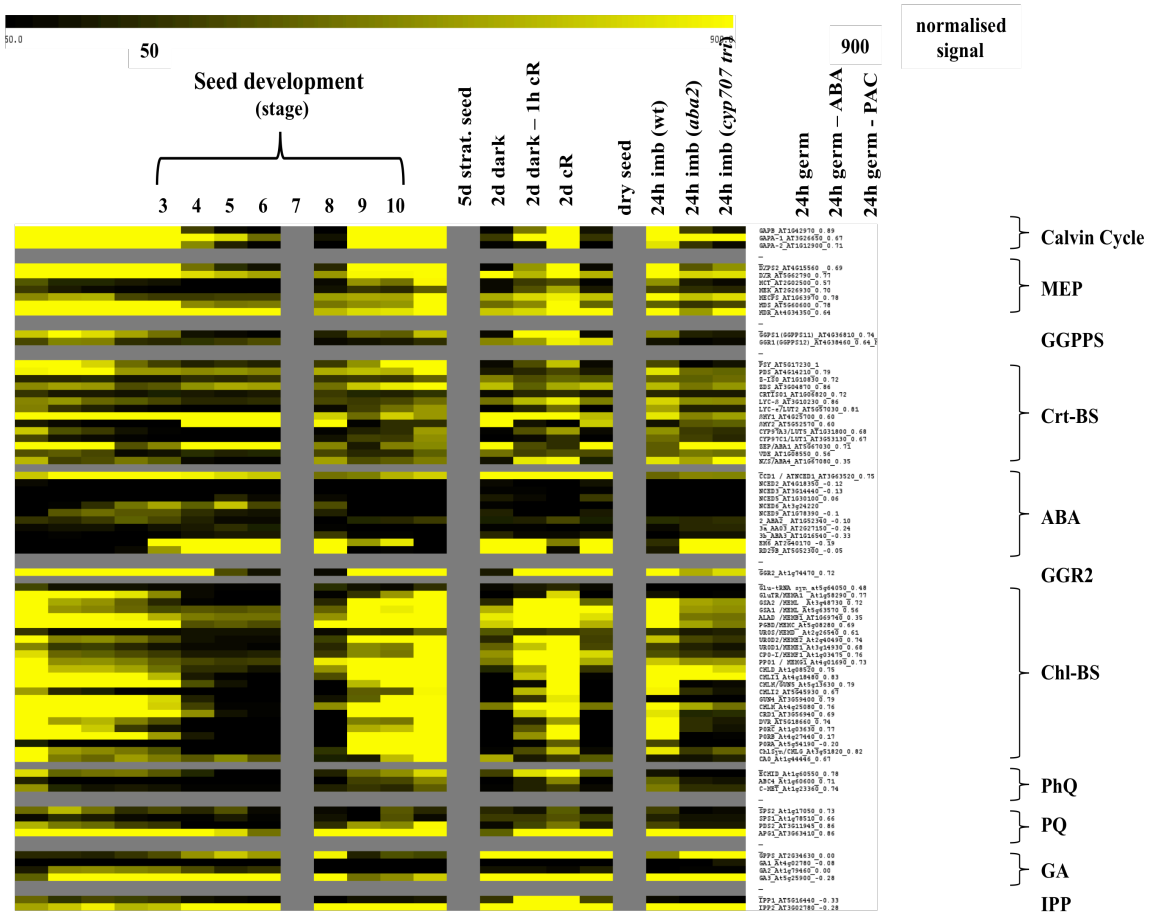


Figure 3. Expression heat map illustrating the relative expression levels of the *PSY-CIIG* during early developmental stages. The experimental conditions examined are listed across the top of the heat map and include, developing seed (GSE5634); stratified seed, etiolated and de-etiolated growth (GSE17159); imbibition in wild-type (wt), *aba2* (ABA-deficient), and the *cyp707a1,a2,a3* triple mutant (elevated ABA levels, GSE15700); post-germinative growth (24h) in the presence and absence of exogenously applied ABA or PAC (GSE5751). Individual genes included in the analysis are listed on the right and are arranged in sequential pathway order. Arrows indicate branch points where reaction products are used in multiple biosynthesis pathways. Results are presented as normalized signal values to reveal the relative expression levels of individual genes at conditions examined. Genes that function in biosynthesis pathway represented in Figure 1 are analysed. Details of the microarray experimental conditions are presented in Text S3 (Supporting Information).

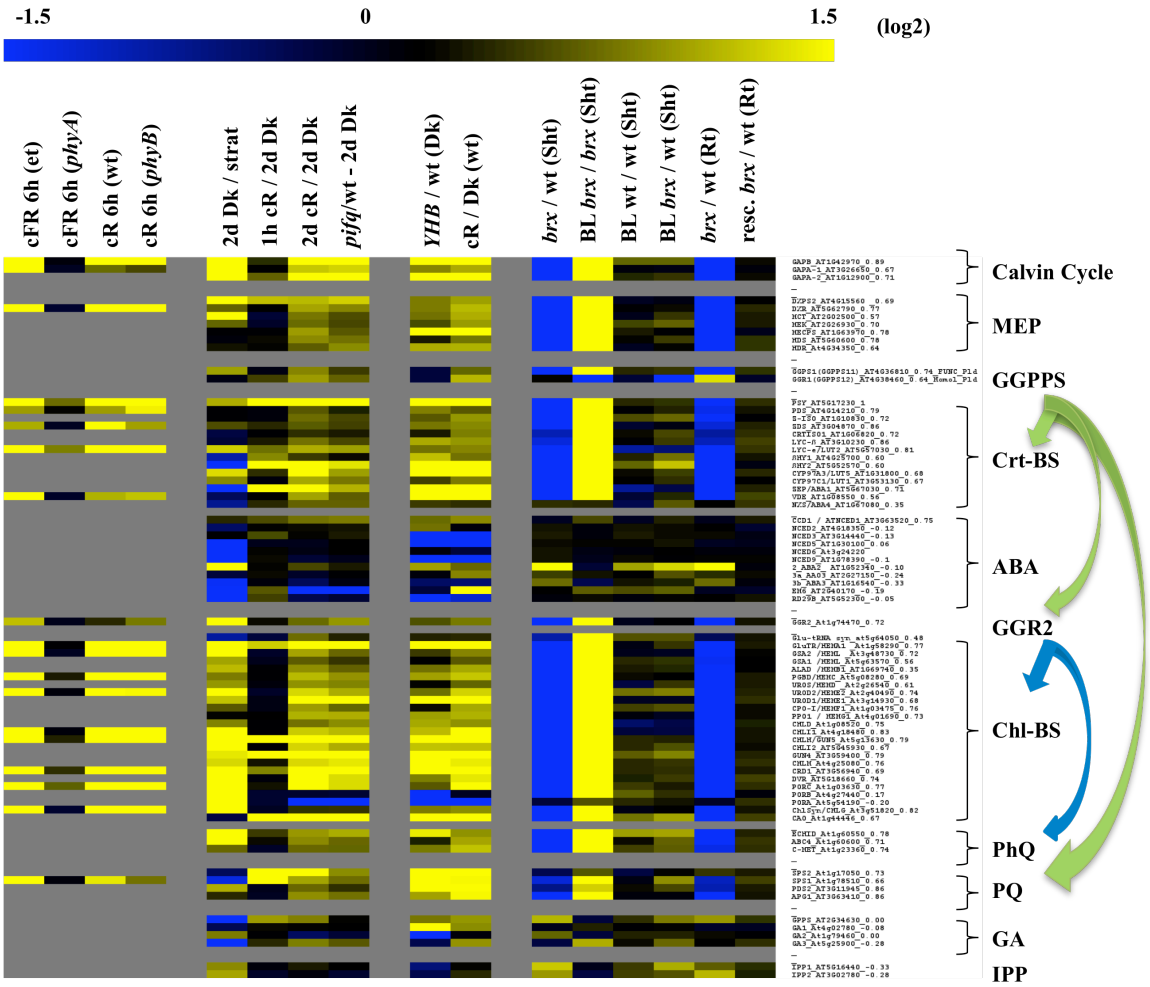


Figure 4. Heat map illustration the fold change in expression of the *PSY-CIIG* in developmental-related mutants. The experimental conditions included, etiolated and de-etiolated growth in *pifq* mutants (GSE17159), effect of constitutively active $PHYB^{Y276H}$ allele in dark-grown *phyAphyB* mutant seedlings (GSE8951), exposure of dark grown *phyA* and *phyB* mutants to cFR and cR light respectively (Tepperman 2001, 2004) and effect of brassinolide (BL) in roots (6 day) and shoots (4 day) of young *brx* loss-of-function plants (E-MEXP-635) Details of the microarray experimental conditions are presented in Text S3 (Supporting Information).

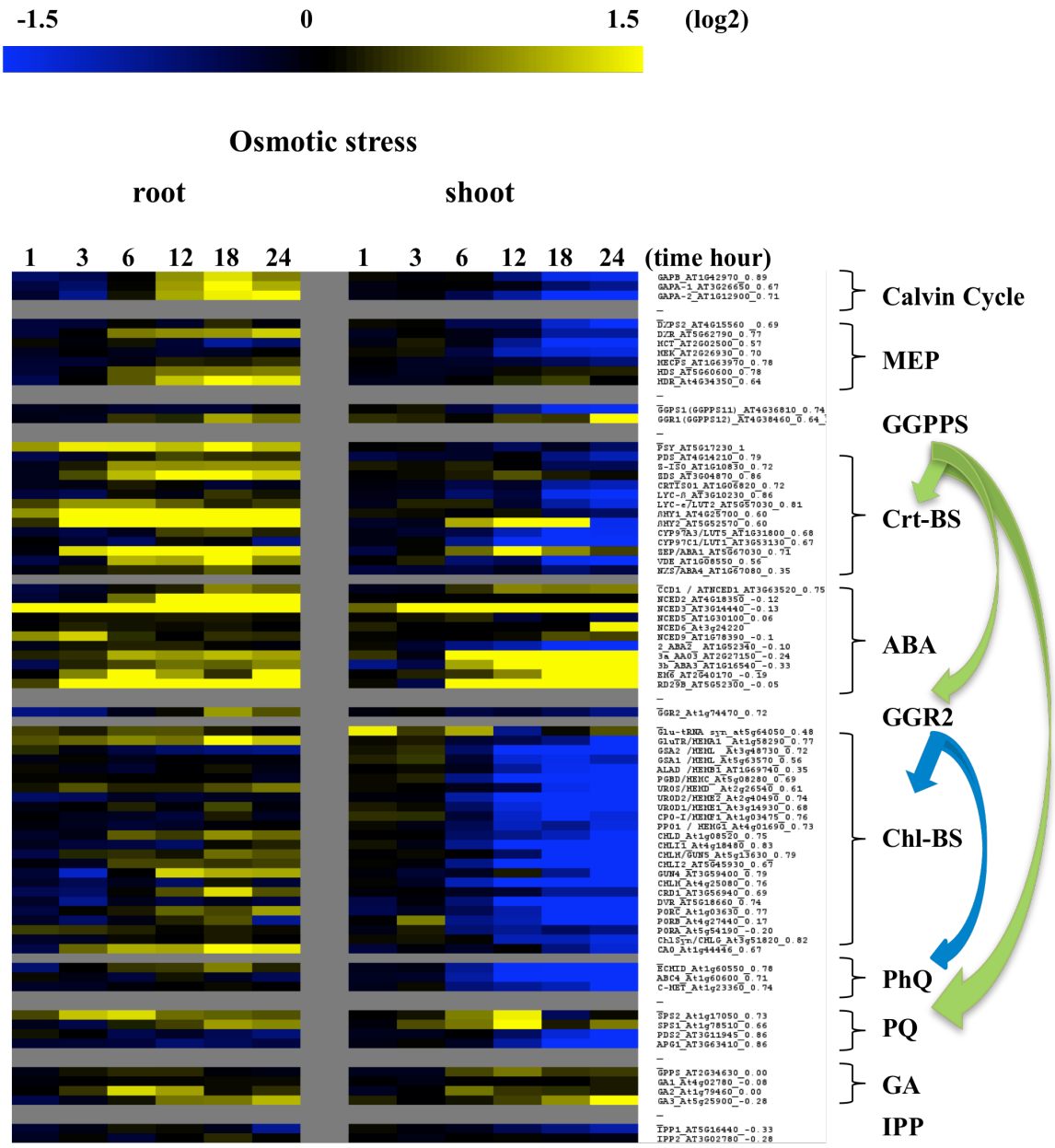


Figure 5. Expression heat map time-course illustrating the effect of osmotic stress on expression of the *PSY-CIIG* in root and shoot tissue. Fold-change (log2) in gene expression was measured in root and shoot tissue at the indicated time points following continuous osmotic stress application (mannitol) to root tissue (ME00327). Details of the microarray experimental conditions are presented in Text S3 (Supporting Information).

TABLES

Table 1. List of the 50 genes that are most highly co-expressed with *PSY*. An expression correlation analysis was performed to identify genes in the Arabidopsis genome that are most highly co-expressed with PSY. Co-expression is measured as expression correlation (r-value). Genes in the top 50 expression correlated genes (*PSY-ECG50*) that were found to belong to a functionally enriched category are indicated.

ID	r	DESCRIPTION	GO
AT5G17230	1	PHYTOENE SYNTHASE (PSY)	CMP, Pd, CPl
AT2G04039	0.910	Expressed protein (ExPr)	Pd, CPl
AT1G62750	0.902	SNOWY COTYLEDON 1 (SCO1), Elongation factor Tu	CO, Pd, CPl, TF
AT1G14345	0.898	Transmembrane domain, oxidoreductase	Pd, PP, TP, CPl
AT1G16880	0.897	Uridylyltransferase-related	AB, TS, Cd, Pd, PP, TP, CPl
AT3G55330	0.896	Photosystem II reaction center PsbP family protein (PPL1)	PS, Pd, PP, TP, CPl
AT1G55480	0.896	similar to LPA1 (Low PSII accum1),	Pd, PP, TP, CPl
AT1G09340	0.890	CHLOROPLAST RNA BINDING (CRB)	AB, CO, TS, Cd, Cd, Pd, PP, CPl
AT1G26220	0.890	GCN5-related N-acetyltransferase (GNAT) family protein	
AT1G42970	0.889	GLYCERALD-3-PHOSPHATE DEHYDROGENASE B SUBUNIT (GAPB)	PS, AB, TS, Cd, PM, PSD, Pd, PP, TP, CPl
AT4G34090	0.888	ExPr // chloroplast stroma	
AT1G50320	0.887	THIOREDOXIN X (ATHX)	Pd, CPl
AT1G54500	0.887	Rubredoxin family protein	PM, Pd, PP, TP, CPl
AT1G17220	0.882	fu-gaeri1 (FUG1), Translation initiation factor IF-2, chloroplast	TF
AT5G44650	0.881	ExPr // chloroplast thylakoid membrane	Pd, PP, TP, CPl
AT3G26570	0.881	PHOSPHATE TRANSPORTER 2;1 (PHT2;1)	Pd, PP, CPl
AT5G04140	0.881	GLUTAMATE SYNTHASE 1 (GLU1) / ferredoxin-dependent	AB, Pd, CPl, OR
AT4G01800	0.877	Preprotein translocase secA subunit, chloroplast [precursor]	
AT1G11860	0.876	Aminomethyltransferase, mitochondrial precursor	
AT1G45474	0.874	PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 5 (LHCA5)	PS, PM, PSL, TP
AT1G73110	0.874	Ribulose bisphosphate carboxylase/oxygenase activase, putative	Pd, PP, TP, CPl
AT2G21330	0.873	FRUCTOSE-BISPHOSPHATE ALDOLASE 1 (FBA1)	PM, Pd, PP, CPl, CF
AT5G58260	0.873	Encodes subunit NDH-N of NAD(P)H:plastoquinone dehydrogenase	Pd, PP, TP, CPl
AT5G43750	0.871	NAD(P)H DEHYDROGENASE 18 (NDH18)	Pd, PP, TP, CPl
AT1G15980	0.870	NDH-DEPENDENT CYCLIC ELECTRON FLOW 1 (NDF1)	Pd, CPl
AT4G10300	0.869	ExPr	Pd, CPl

AT5G17170	0.867	Rubredoxin family protein, enhancer of sos3-1 (ENH1)	PM, Pd, PP, TP, CPl
AT3G04790	0.866	Ribose 5-phosphate isomerase-related	PS, PM, PSD, Pd, PP, TP, CPl, CF
AT1G05140	0.866	Membrane-associated zinc metalloprotease	Pd
AT5G08650	0.866	GTP-binding protein LepA, putative	Pd, CPl, TF
AT5G23120	0.865	HIGH CHLOROPHYLL FLUORESCENCE 136 (HCF136) PS II assembly,	Pd, PP, TP, CPl
AT1G32470	0.865	Glycine cleavage system H protein, mitochondrial precursor	OR
AT1G01320	0.865	Tetratricopeptide repeat (TPR)-containing protein l	
AT1G32080	0.864	Membrane protein, putative contains 12 transmembrane domains	Pd, PP, CPl
AT2G20890	0.863	THYLAKOID FORMATION1 (THF1)	PS, PM, PSL, Pd, PP, TP, CPl
AT3G11950	0.862	PHYTOENE DESATURATION 2 (PDS2), UbiA prenyltransferase	
AT1G18060	0.862	ExPr	
AT3G54050	0.862	Fructose-1,6-bisphosphatase, putative	AB, TS, Cd, PM, Pd, PP, CF
AT3G10230	0.862	LYCOPENE CYCLASE (LYC)	CMP, Pd, CPl,
AT2G34860	0.861	Embryo sac development arrest 3 (EDA3), Heat shock protein 40	Pd, CPl
AT1G27480	0.860	Lecithin:cholesterol acyltransferase family protein (LACT)	
AT3G63410	0.860	ALBINO OR PALE GREEN MUTANT (APGM) , MPBQ methyltransferase	Pd, PP, CPl
AT1G07010	0.860	Calcineurin-like phosphoesterase family protein	Pd, CPl
AT1G76450	0.858	Oxygen-evolving complex-related	Pd, PP, TP, CPl
AT5G42310	0.858	PPR repeat-containing protein	
AT3G04870	0.857	ZETA-CAROTENE DESATURASE (ZDS)	CMP
AT1G77090	0.856	Thylakoid lumenal 29.8 kDa protein i	PS, Pd, PP, TP, CPl
AT1G64680	0.856	ExPr	
AT1G80030	0.855	DNAJ heat shock protein,	Pd, PP, TP, CPl
AT4G17600	0.855	light-harvesting-like protein (Lil3:1)	Pd, PP, TP, CPl
AT5G08050	0.854	ExPr	

Additional files for Meier et al., 2011.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3123201/?tool=pubmed

Additional Table 1. Extended list of genes in the *PSY*-correlated interrelated isoprenoid biosynthesis genes and their expression correlation relative to PSY.

Additional Table 2. Enriched motifs identified in the promoters of genes in the PSY-ECGG50.

Additional Figure 1. Heatmaps illustrating the expression of the PSY-ECGG50 in response to the range of experimental conditions examined.

Additional Text 1. Description of microarray experimental conditions examined.

APPENDIX III:

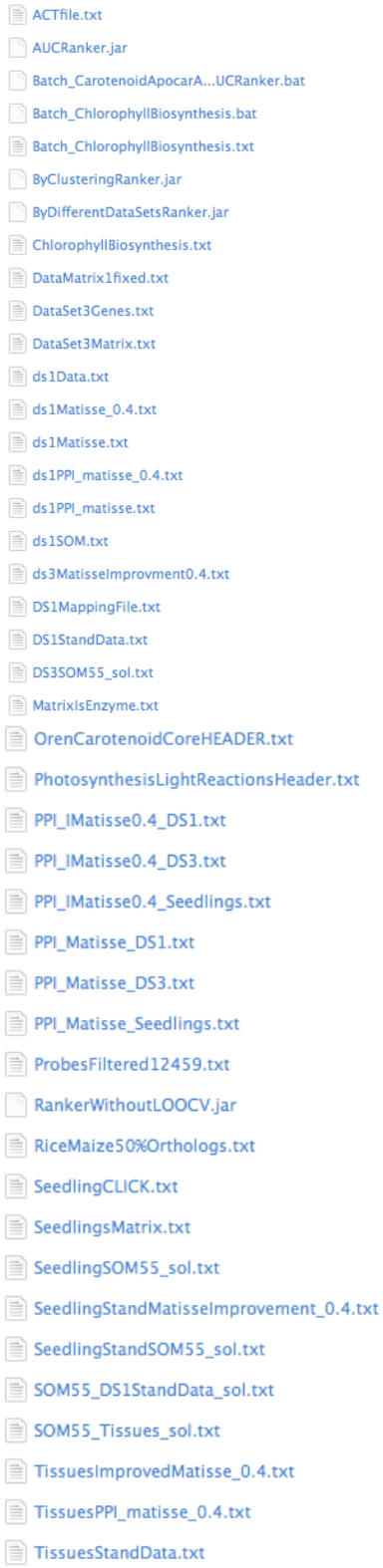
Table listing locations of additional files:

Thesis section the table appears	Supplementary Table name and location	Description
1.5	ThesisFiles\Tables\ SupplementaryTable1.5.xlsx	The lists carotenoid biosynthesis related genes (CBRG) and the upstream MEP pathway genes
2.3	ThesisFiles\Tables\ SupplementaryTable2.3.1xlsx	The top co-expressed genes with *PSY* (r-value >0.6)
2.3	ThesisFiles\Tables\ SupplementaryTable2.3.2xlsx	List of all probes co-expressed with *PSY*
2.4	ThesisFiles\Tables\ SupplementaryTable2.4.xls	Raw data used to generate Figures 2.4 and 2.5
2.5	ThesisFiles\Tables\ SupplementaryTable2.5.xlsx	List of GO terms (biological processes) enriched in the CBRG co-expression network
2.6	ThesisFiles\Tables\ SupplementaryTable2.6.xlsx	The TFs genes that are co-expressed with the CBRG with r-value >0.7
2.6	ThesisFiles\Scripts\ ACTcrawler.rb	A script that crawls the Arabidopsis Co-expression Tool (ACT) to retrieve whole genome co-expression coefficient (Person) values for ~12,500 genes.
2.8.1	ThesisFiles\Tables\SupplementaryTable2.8.1.txt	Arabidopsis genes annotation for all probes represented on the ACT files. This list used throughout the thesis for retrieving gene annotations.
3.1	ThesisFiles\Tables\ SupplementaryTable3.1.xls	Biological pathways donloaded from AraCyc and have at least 10 genes
3.3.1	ThesisFiles\Tables\ SupplementaryTable3.3.1.xlsx	The 'seeds' gene expression data set
3.3.2	ThesisFiles\Tables\ SupplementaryTable3.3.2.xlsx	The 'seedlings' gene expression data set
3.3.3	ThesisFiles\Tables\ SupplementaryTable3.3.3.docx	The list of microarray experiments used in this study

3.3.4	ThesisFiles\Tables\ SupplementaryTable3.3.4.xlsx	The 'seedlings' gene expression data set
3.4.1	ThesisFiles\Tables\ SupplementaryTable3.4.1.sif	Arabidopsis gene network of metabolic dependencies
3.4.2	ThesisFiles\Tables\ SupplementaryTable3.4.2.xls	Arabidopsis protein-protein interaction network (from PAIR)
3.5.1	ThesisFiles\MORPH.2\ DataSetIntegration\ RankerWithoutLOOCV.jar	Script for ranking AUC for input pathway
3.5.1	ThesisFiles\ MORPH.2\ DataSetIntegration\ ds1Data.xls	The ‘DS1’ gene expression data set
3.5.1	ThesisFiles\ MORPH.2\ DataSetIntegration\ ds3DataMatrix.xls	The ‘Seeds’ gene expression data set
3.5.1	ThesisFiles\ MORPH.2\ DataSetIntegration\ SeedlingsMatrix.xls	The ‘Seedlings’ gene expression data set
3.5.1	ThesisFiles\ MORPH.2\ DataSetIntegration\TissuesStandData.xls	The ‘Tissues’ gene expression data set
3.5.2	ThesisFiles\Tables\ SupplementaryTable3.5.2.xls	Top ranked candidate genes for the CarotenoidsCore pathway
3.9.1	ThesisFiles\Tables\ SupplementaryTable3.9.1.xls	Top ranked candidate genes for the 'photosynthesis light reactions' pathway
3.9.2	ThesisFiles\Tables\ SupplementaryTable3.9.2.xls	Top ranked candidate genes for the 'CarotenoidsCore' pathway
3.9.3	ThesisFiles\Tables\ SupplementaryTable3.9.3.xls	Top ranked candidate genes for the 'Homogalacturonan ' pathway
3.11.3.2	ThesisFiles\Tables\ SupplementaryTable3.11.3.2.xls	The list of genes annotated as enzymes or not.

Appendix IV:

MORPH online files (Location: ThesisFiles\MORPH.2\DataSetsIntegration)



MORPH running protocol (Location: ThesisFiles\MORPH2\)**:**

1. Copy MORPH. 2 folder to your C drive on your PC and navigate to this location using the command line. *This directory includes all files to run MORPH including scripts and input data files. This is where new gene lists should be added. See notes below on "To generate the the pathway input file".*
   From START menu, choose "RUN" and type "cmd" *(to open command line)*
   Print on the command line:
   cd c:\ (get to root directory)
   cd MORPH.2
   cd DataSetsIntegration *(note: working area for MORPH containing all needed files)*

2. For learning the best configuration for a given (new) pathway we need to give AUC scores:

   Command line usage>**java -Xmx1000m -jar ByDifferentDataSetsRanker.jar {GOI IDs header} {GE-DataSet} {Clustering solution file} { Pearson/Spearman } {GE-DataSet file} { solution file } { Pearson/Spearman }... > OutputFileName.txt**

   ** This step can be used using a batch file (.bat) for automation (see Batch_example.txt).

Example command >**java -Xmx1000m -jar ByDifferentDataSetsRanker.jar FolateCoreBiosynthesisGenesNoSpaces.txt** SeedlingsMatrix.txt SeedlingSOM55_sol.txt **Pearson >FolateSeedlingSomAUC.txt**

Example > java -Xmx1000m -jar ByDifferentDataSetsRanker.jar FolateCoreBiosynthesisGenesNoSpaces.txt TissuesStandData.txt TissuesImprovedMatisse_0.4.txt Pearson >FolateCoreTissuesMDPearsonAUC.txt

NOTES: To generate the command line, input the following components separated by one space (*best to use a text file to avoid space problems*).

a. JAVA: "**java -Xmx1000m**" can be increased to 1500 or 2000 to decrease running time (depends on computer)

b. SCRIPT to generate AUC: ByDifferentDataSetsRanker.jar**

c. {GOI IDs header} is the Gene input file (e.g. FolateCoreBiosynthesisGenesNoSpaces.txt)

d. {GE-DataSet} is the Gene expression Data Set file (4 available:

SeedlingsMatrix.txt \ ds1Data.txt\ TissuesStandData.txt\ DataSet3Matrix.txt

e. {Clustering solution file} 5 options are available:

	Gene Expression Datasets			
ClusteringSolution:	**Seedlings (SeedlingsMatrix.txt)**	**Ds3 (DataSet3Matrix.txt)**	**Ds1 (ds1Data.txt)**	**Tissues (Tissues**
SOM (Gene Coexpression "Expander"	SeedlingSOM55_sol.txt	DS3SOM55_sol.txt	SOM55_DS1StandData_sol.txt	SOM55_
MD (Metabolic Dependency) "Matisse*"	SeedlingStandMatisseImprovement_0.4.txt	ds3MatisseImprovment0.4.txt	ds1Matisse_0.4.txt	TissuesI
PPI (Prot-Prot Interaction) "Matisse*"	PPI_IMatisse0.4_Seedlings.txt	PPI_IMatisse0.4_DS3.txt	PPI_IMatisse0.4_DS1.txt	TissuesI
Is Enzyme (VLOOKUP in Excel)	MatrixIsEnzyme.txt	MatrixIsEnzyme.txt	MatrixIsEnzyme.txt	MatrixIs
Orthologs+ (Biomart download and clustered with VLOOKUP in Excel)	RiceMaize50%Orthologs.txt	RiceMaize50%Orthologs.txt	RiceMaize50%Orthologs.txt	RiceMa

+"orthologs" clustering solution did not give good results and is therefore not recommended.

f. Similarity score {Pearson/Spearman }: choose one or the other

g. Repeat for each combination (or use the "batch command", see below)

h. Output file: > OutputFileName.txt

** For learning AUCs with minimum output. The following uses the ranker, AUCRanker.jar which gives only AUC values (1,1,1000) and not all the extraneous clustering information.

Example: java -Xmx1500m -jar AUCRanker.jar ChlorophyllBiosynthesis.txt SeedlingsMatrix.txt SeedlingStandSOM55_sol.txt pearson >> ChlorophyllBiosynthesisLearningAUCoren2.txt

4. After the best AUC are determined, rank candidates according to the best data combination:

Command line usage> java -Xmx1000m –jar RankerWithoutLOOCV.jar {GE-DataSet file} {GOI IDs header file} {clustering solution file} {Pearson/Spearman} > OutputFileName.txt

Example: java -Xmx1000m -jar RankerWithoutLOOCV.jar TissuesStandData.txt FolateCoreBiosynthesisGenesNoSpaces.txt TissuesImprovedMatisse_0.4.txt Pearson > FolateCandidateRanking.txt

Note: Transfer results to Excel file and sort by column A, rank # (otherwise ranked genes start at the bottom!).

General Note: To generate the pathway input file, start with an excel file containing a vertical list of genes (accessions in all CAPS-use Find/Replace to correct); transpose all ATG numbers to a horizontal line of continuous columns. Save as a "tab delimited text file". This text file must be *named without any spaces*. Check the file that there are only tabs (and no spaces) between ATGs.

5. To retrieve Ranked candidate annotation:

Use the ACT annotations file (possibly outdated) of all genes: see Supplementary Table 2.8.1.txt

Reference List

Akiyama, K., Matsuzaki, K.-i., and Hayashi, H. (2005). Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. Nature **435,** 824-827.

Allocco, D.J., Kohane, I.S., and Butte, A.J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics **5,** 18.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res **38,** D525-531.

Atias, O., Chor, B., and Chamovitz, D.A. (2009). Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. BMC Syst Biol **3,** 86.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., and Edgar, R. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res **37,** D885-890.

Block, M.A., Douce, R., Joyard, J., and Rolland, N. (2007). Chloroplast envelope membranes: a dynamic interface between plastids and the cytosol. Photosynth Res **92,** 225-244.

Booker, J., Auldridge, M., Wills, S., McCarty, D., Klee, H., and Leyser, O. (2004). MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. Curr Biol **14,** 1232-1238.

Britton, G., Liaaen-Jensen, S., and Pfander, H. (2004). Carotenoids Handbook (Basel: Birkhäuser Verlag).

Burrows, P.A., Sazanov, L.A., Svab, Z., Maliga, P., and Nixon, P.J. (1998). Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes. Embo J **17,** 868-876.

Chander, S., Guo, Y.Q., Yang, X.H., Zhang, J., Lu, X.Q., Yan, J.B., Song, T.M., Rocheford, T.R., and Li, J.S. (2007). Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. Theoret. Appl. Genetics **116,** 223-233.

Chen, L., and Vitkup, D. (2006). Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol **7,** R17.

Chen, Y., Li, F., and Wurtzel, E.T. (2010). Isolation and characterization of the Z-ISO gene encoding a missing component of carotenoid biosynthesis in plants. Plant Physiol. **153,** 66-79.

Cheng, Y.C., and Fleming, G.R. (2009). Dynamics of light harvesting in photosynthesis. Annu Rev Phys Chem **60,** 241-262.

Chiu, F.Y., Chen, Y.R., and Tu, S.L. (2010). Electrostatic interaction of phytochromobilin synthase and ferredoxin for biosynthesis of phytochrome chromophore. J Biol Chem **285,** 5056-5065.

Cordoba, E., Salmi, M., and Leon, P. (2009). Unravelling the regulatory mechanisms that modulate the MEP pathway in higher plants. J. Exp. Bot. **60,** 2933-2943.

Covington, M.F., Maloof, J.N., Straume, M., Kay, S.A., and Harme, S.L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. Genome Biol. **9,** R130.

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 Res **32,** D575-577.

Cuttriss, A., Vallabhaneni, R., and Wurtzel, E.T. . (2011). Learning from maize to enhance cereal crop provitamin A carotenoids. Trends in Biotechnology **submitted**.

Dall'Osto, L., Fiore, A., Cazzaniga, S., Giuliano, G., and Bassi, R. (2007). Different roles of α- and β-branch xanthophylls in photosystem assembly and photoprotection. J. Biol. Chem. **282,** 35056-35068.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95,** 14863-14868.

Endo, T., Ishida, S., Ishikawa, N., and Sato, F. (2008). Chloroplastic NAD(P)H dehydrogenase complex and cyclic electron transport around photosystem I. Mol Cells **25,** 158-162.

Fiore, A., Dall'Osto, L., Fraser, P.D., Bassi, R., and Giuliano, G. (2006). Elucidation of the [beta]-carotene hydroxylation pathway in Arabidopsis thaliana. FEBS Letts. **580,** 4718-4722

Fraser, P.D., and Bramley, P.M. (2004). The biosynthesis and nutritional uses of carotenoids. Progress in Lipid Research **43,** 228-265.

Gantet, P., and Memelink, J. (2002). Transcription factors: tools to engineer the production of pharmacologically active plant metabolites. Trends Pharmacol Sci **23,** 563-569.

Giliberto, L., Perrotta, G., Pallara, P., Weller, J.L., Fraser, P.D., Bramley, P.M., Fiore, A., Tavazza, M., and Giuliano, G. (2005). Manipulation of the blue light photoreceptor Cryptochrome 2 in tomato sffects vegetative development, flowering time, and fruit antioxidant content. Plant Physiol. **137,** 199-208.

Giuliano, G., Tavazza, R., Diretto, G., Beyer, P., and Taylor, M.A. (2008). Metabolic engineering of carotenoid biosynthesis in plants. Trends in Biotech. **26,** 139-145.

Guyon, I., Saffari, A., Dror, G., and Cawley, G. (2010). Model Selection: Beyond the Bayesian/Frequentist Divide. Journal of Machine Learning Research **11,** 61-87.

Harholt, J., Jensen, J.K., Sorensen, S.O., Orfila, C., Pauly, M., and Scheller, H.V. (2006). ARABINAN DEFICIENT 1 is a putative arabinosyltransferase involved in biosynthesis of pectic arabinan in Arabidopsis. Plant Physiol **140,** 49-58.

Harjes, C.E., Rocheford, T.R., Bai, L., Brutnell, T.P., Kandianis, C.B., Sowinski, S.G., Stapleton, A.E., Vallabhaneni, R., Williams, M., Wurtzel, E.T., Yan, J., and Buckler, E.S. (2008). Natural genetic variation in *lycopene epsilon cyclase* tapped for maize biofortification. Science **319,** 330-333.

Hirooka, K., Izumi, Y., An, C.I., Nakazawa, Y., Fukusaki, E., and Kobayashi, A. (2005). Functional analysis of two solanesyl diphosphate synthases from Arabidopsis thaliana. Biosci Biotechnol Biochem **69,** 592-601.

Jensen, J.K., Sorensen, S.O., Harholt, J., Geshi, N., Sakuragi, Y., Moller, I., Zandleven, J., Bernal, A.J., Jensen, N.B., Sorensen, C., Pauly, M., Beldman, G., Willats, W.G., and Scheller, H.V. (2008). Identification of a xylogalacturonan xylosyltransferase involved in pectin biosynthesis in Arabidopsis. Plant Cell **20,** 1289-1302.

Kharchenko, P., Vitkup, D., and Church, G.M. (2004). Filling gaps in a metabolic network using expression information. Bioinformatics **20 Suppl 1,** i178-185.

Kharchenko, P., Church, G.M., and Vitkup, D. (2005). Expression dynamics of a cellular metabolic network. Mol Syst Biol **1,** 2005 0016.

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., and Church, G.M. (2006). Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics **7,** 177.

Kohonen, T. (1990). Cortical maps. Nature **346,** 24.

Koltai, H. (2011a). Strigolactones are regulators of root development. New Phytol **190,** 545-549.

Koltai, H. (2011b). Strigolactones' ability to regulate root development may be executed by induction of the ethylene pathway. Plant Signal Behav **6**.

Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L., Belmonte, M., Kirkbride, R., Horvath, S., Drews, G.N., Fischer, R.L., Okamuro, J.K., Harada, J.J., and Goldberg, R.B. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A **107,** 8063-8070.

Lee, J.M., and Sonnhammer, E.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res **13,** 875-882.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., Rhee, S.Y. (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nat Biotechnol **28,** 149-156.

Less, H., Angelovici, R., Tzin, V., and Galili, G. (2011). Coordinated gene networks regulating Arabidopsis plant metabolism in response to various stresses and nutritional cues. Plant Cell **23,** 1264-1271.

Li, F., Vallabhaneni, R., and Wurtzel, E.T. (2008a). *PSY3,* a new member of the phytoene synthase gene family conserved in the Poaceae and regulator of abiotic-stress-induced root carotenogenesis. Plant Physiol. **146,** 1333-1345.

Li, F., Tzfadia, O., and Wurtzel, E.T. (2009a). The *Phytoene Synthase* gene family in the grasses: Subfunctionalization provides tissue-specific control of carotenogenesis. Plant Signaling & Behavior **4,** 208-211.

Li, F.Q., Tzfadia, O., and Wurtzel, E.T. (2009b). The phytoene synthase gene family in the Grasses. Plant Signaling and Behavior **4,** 208-211.

Lin, M., Hu, B., Chen, L., Sun, P., Fan, Y., Wu, P., and Chen, X. (2009). Computational identification of potential molecular interactions in Arabidopsis. Plant Physiol **151,** 34-46.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics **21,** 3448-3449.

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. Nucleic Acids Res **34,** W504-W509.

Meier, S., Tzfadia, O., Vallabhaneni, R., Gehring, C., and Wurtzel, E. (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,** 77.

Mene-Saffrane, L., Jones, A.D., and DellaPenna, D. (2010). Plastochromanol-8 and tocopherols are essential lipid-soluble antioxidants during seed desiccation and quiescence in Arabidopsis. Proc Natl Acad Sci U S A **107,** 17815-17820.

Moran, N.A., and Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science **328,** 624-627.

Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O., and Persson, S. (2009). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol **152,** 29-43.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell **23,** 895-910.

Nambara, E., and Marion-Poll, A. (2005). Abscisic acid biosynthesis and catabolism. Annu Rev Plant Biol **56,** 165-185.

Nashilevitz, S., Melamed-Bessudo, C., Izkovich, Y., Rogachev, I., Osorio, S., Itkin, M., Adato, A., Pankratov, I., Hirschberg, J., Fernie, A.R., Wolf, S., Usadel, B., Levy, A.A., Rumeau, D., and Aharoni, A. (2010). An orange ripening mutant links plastid NAD(P)H dehydrogenase complex activity to central and specialized metabolism during tomato fruit maturation. Plant Cell **22,** 1977-1997.

Niyogi, K.K. (2000). Safety valves for photosynthesis. Current Opinion in Plant Biology **3,** 455–460.

Norris, S.R., Barrette, T.R., and DellaPenna, D. (1995). Genetic dissection of carotenoid synthesis in *Arabidopsis* defines plastoquinone as an essential component of phytoene desaturation. Plant Cell **7,** 2139-2149.

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res **16,** 249-260.

Okegawa, Y., Tsuyama, M., Kobayashi, Y., and Shikanai, T. (2005). The pgr1 mutation in the Rieske subunit of the cytochrome b6f complex does not affect PGR5-dependent cyclic electron transport around photosystem I. J Biol Chem **280,** 28332-28336.

Orth, J.D., and Palsson, B.O. (2010). Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng **107,** 403-412.

Park, H., Kreunen, S.S., Cuttriss, A.J., DellaPenna, D., and Pogson, B. (2002). Identification of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar body formation, and photomorphogenesis. Plant Cell **14,** 321-332.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96,** 4285-4288.

Penuelas, J., and Munne-Bosch, S. (2005). Isoprenoids: an evolutionary pool for photoprotection. Trends Plant Sci **10,** 166-169.

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

Pigliucci, M. (2009). Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philos Trans R Soc Lond B Biol Sci **365,** 557-566.

Popescu, L., and Yona, G. (2005). Automation of gene assignments to metabolic pathways using high-throughput expression data. BMC Bioinformatics **6,** 217.

Quinlan, R., Jaradat, T., and Wurtzel, E.T. (2007). *Escherichia coli* as a platform for functional expression of plant P450 carotene hydroxylases. Arch. Biochem. Biophysics **458,** 146-157.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., and Yu, G. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290,** 2105-2110.

Rodriguez-Villalon, A., Gas, E., and Rodriguez-Concepcion, M. (2009). Phytoene synthase activity controls the biosynthesis of carotenoids and the supply of their metabolic precursors in dark-grown Arabidopsis seedlings. Plant J.

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

Semba, R.D., Muhilal, Scott, A.L., Natadisastra, G., West, K.P., Jr., and Sommer, A. (1994). Effect of vitamin A supplementation on immunoglobulin G subclass responses to tetanus toxoid in children. Clin Diagn Lab Immunol **1,** 172-175.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res **13,** 2498-2504.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., and Tyers, M. (2006). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res **39,** D698-704.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science **302,** 249-255.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A.,

Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36,** D1009-1014.

Sweetlove, L.J., Fell, D., and Fernie, A.R. (2008). Getting to grips with the plant metabolic network. Biochem J **409,** 27-41.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A **96,** 2907-2912.

Tan, B.C., Joseph, L.M., Deng, W.T., Liu, L., Li, Q.B., Cline, K., and McCarty, D.R. (2003). Molecular characterization of the Arabidopsis 9-cis epoxycarotenoid dioxygenase gene family. Plant J **35,** 44-56.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nat Genet **22,** 281-285.

Toledo-Ortiz, G., Huq, E., and Rodriguez-Concepcion, M. (2010). Direct regulation of phytoene synthase gene expression and carotenoid biosynthesis by phytochrome-interacting factors. Proc Natl Acad Sci U S A **107,** 11626-11631.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23,** 137-144.

Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. BMC Syst Biol **1,** 8.

Ulitsky, I., and Shamir, R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics **25,** 1158-1164.

Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D., Linhart, C., Elkon, R., Tanay, A., Sharan, R., Shiloh, Y., and Shamir, R. (2010). Expander: from expression microarrays to networks and functions. Nat Protoc **5,** 303-322.

Underwood, B.A., and Arthur, P. (1996). The contribution of vitamin A to public health. The FASEB Journal **10,** 1040-1049.

Usadel, B., Nagel, A., Steinhauser, D., Gibon, Y., Blasing, O.E., Redestig, H., Sreenivasulu, N., Krall, L., Hannah, M.A., Poree, F., Fernie, A.R., and Stitt, M. (2006). PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics **7,** 535.

Vallabhaneni, R., and Wurtzel, E.T. (2009). Timing and biosynthetic potential for carotenoid accumulation in genetically diverse germplasm of maize. Plant Physiol. **150,** 562-572.

Vallabhaneni, R., Gallagher, C.E., Licciardello, N., Cuttriss, A.J., Quinlan, R.F., and Wurtzel, E.T. (2009). Metabolite sorting of a germplasm collection reveals the *Hydroxylase3* locus as a new target for maize provitamin A biofortification. Plant Physiol. **151,** 1635-1645.

van den Berg, H., Faulks, R., Granado, H.F., Hirschberg, J., Olmedilla, B., Sandmann, G., Southon, S., and Stahl, W. (2000). The potential for the improvement of carotenoid levels in foods and the likely systemic effects. Journal of the Science of Food and Agriculture **80,** 880-912.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol **150,** 535-546.

von Lintig, J., Welsch, R., Bonk, M., Giuliano, G., Batschauer, A., and Kleinig, H. (1997). Light-dependent regulation of carotenoid biosynthesis occurs at the level of phytoene synthase expression and is mediated by phytochrome in *Sinapis alba* and *Arabidopsis thaliana* seedlings. The Plant Journal **12,** 625-634.

Wang, D., Guo, Y., Wu, C., Yang, G., Li, Y., and Zheng, C. (2008). Genome-wide analysis of CCCH zinc finger family in Arabidopsis and rice. BMC Genomics **9,** 44.

Welsch, R., Medina, J., Giuliano, G., Beyer, P., and Von Lintig, J. (2003). Structural and functional characterization of the phytoene synthase promoter from Arabidopsis thaliana. Planta **216,** 523-534.

Welsch, R., Maass, D., Voegel, T., Dellapenna, D., and Beyer, P. (2007). Transcription factor RAP2.2 and its interacting partner SINAT2: stable elements in the carotenogenesis of Arabidopsis leaves. Plant Physiol **145,** 1073-1085.

West, K.P., Jr. (2002). Extent of Vitamin A deficiency among preschool children and women of reproductive age. J. Nutr. **132,** 2857S-2866.

Wong, J.C., Lambert, R.J., Wurtzel, E.T., and Rocheford, T.R. (2004). QTL and candidate genes phytoene synthase and zetacarotene desaturase associated with the accumulation of carotenoids in maize. Theor. Appl. Genetics **108,** 349-359.

Wurtzel, E.T. (2004). Genomics, genetics, and biochemistry of maize carotenoid biosynthesis. In Recent Advances in Phytochemistry, J. Romeo, ed (Elsevier Ltd.), pp. 85-110.

Wurtzel, E.T., and Grotewold, E. (2006). Plant Metabolic Engineering. In The Encyclopedia of Chemical Processing S.K.B. Lee, ed (New York, USA: Marcel Dekker, Inc.), pp. 2191-2200.

Yamanishi, Y., Vert, J.P., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. Bioinformatics **20 Suppl 1,** i363-370.

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., Zhiqiang, L., Yunfei, Z., Xiaoxiao, W., Xiaoming, Q., Yunping, S., Li, Z., Xiaohui, D., Jingchu, L., Xing-Wang, D., Zhangliang, C., Hongya, G., and Li-Jia, Q. (2006). The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. Plant Mol Biol **60,** 107-124.

Ytterberg, A.J., Peltier, J.-B., and van Wijk, K.J. (2006). Protein profiling of plastoglobules in chloroplasts and chromoplasts. A surprising site for differential accumulation of metabolic enzymes. Plant Physiol. **140,** 984-997.

Yu, Q.B., Li, G., Wang, G., Sun, J.C., Wang, P.C., Wang, C., Mi, H.L., Ma, W.M., Cui, J., Cui, Y.L., Chong, K., Li, Y.X., Li, Y.H., Zhao, Z., Shi, T.L., and Yang, Z.N. (2008). Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in Arabidopsis thaliana. Cell Res **18,** 1007-1019.

Zhou, X., Van Eck, J., and Li, L. (2008). Use of the cauliflower Or gene for improving crop nutritional quality. Biotechnol Annu Rev **14,** 171-190.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiology **136,** 2621-2632.

Reference list for Appendix II:

1. Demmig-Adams B, Gilmore AM, Adams WW, III: **Carotenoids 3: in vivo function of carotenoids in higher plants.** *FASEB J* 1996, **10:**403-412.

2. Nambara E, Marion-Poll A: **Abscisic acid biosynthesis and catabolism.** *Annu Rev Plant Biol* 2005, **56:**165-185.

3. Norris SR, Barrette TR, Dellapenna D: **Genetic dissection of carotenoid synthesis in arabidopsis defines plastoquinone as an essential component of phytoene desaturation.** *Plant Cell* 1995, **7:**2139-2149.

4. Cordoba E, Salmi M, Leon P: **Unravelling the regulatory mechanisms that modulate the MEP pathway in higher plants.** *J Exp Bot* 2009, **60:**2933-2943.

5. Rodriguez-Concepcion M: **Supply of precursors for carotenoid biosynthesis in plants.** *Arch Biochem Biophys* 2010, **504:**118-122.

6. Beyer P, Weiss G, Kleinig H: **Solubilization and reconstitution of the membrane-bound carotenogenic enzymes from daffodil chromoplasts.** *Eur J Biochem* 1985, **153:**341-346.

7. Dogbo O, Laferriere A, d'Harlingue A, Camara B: **Carotenoid biosynthesis: Isolation and characterization of a bifunctional enzyme catalyzing the synthesis of phytoene.** *Proc Natl Acad Sci U S A* 1988, **85:**7054-7058.

8. Matthews PD, Wurtzel ET: **Biotechnology of food colorant production.** In *Food Colorants: Chemical and Functional Properties*. CRC Press Boca Raton; 2007:347-398.

9. Sandmann G, Romer S, Fraser PD: **Understanding carotenoid metabolism as a necessity for genetic engineering of crop plants.** *Metab Eng* 2006, **8:**291-302.

10. Chen Y, Li F, Wurtzel ET: **Isolation and characterization of the Z-ISO gene encoding a missing component of carotenoid biosynthesis in plants.** *Plant Physiol* 2010, **153:**66-79.

11. Dall'Osto L, Fiore A, Cazzaniga S, Giuliano G, Bassi R: **Different roles of alpha- and beta-branch xanthophylls in photosystem assembly and photoprotection.** *J Biol Chem* 2007, **282:**35056-35068.

12. Fiore A, Dall'Osto L, Fraser PD, Bassi R, Giuliano G: **Elucidation of the beta-carotene hydroxylation pathway in Arabidopsis thaliana.** *FEBS Lett* 2006, **580:**4718-4722.

13. Wasilewska A, Vlad F, Sirichandra C, Redko Y, Jammes F, Valon C, Frei dit FN, Leung J: **An update on abscisic acid signaling in plants and more..** *Mol Plant* 2008, **1:**198-217.

14. Rodriguez-Villalon A, Gas E, Rodriguez-Concepcion M: **Phytoene synthase activity controls the biosynthesis of carotenoids and the supply of their metabolic precursors in dark-grown Arabidopsis seedlings.** *Plant J* 2009.

15. Welsch R, Beyer P, Hugueney P, Kleinig H, Von Lintig J: **Regulation and activation of phytoene synthase, a key enzyme in carotenoid biosynthesis, during photomorphogenesis.** *Planta* 2000, **211:**846-854.

16. Paddock TN, Mason ME, Lima DF, Armstrong GA: **Arabidopsis protochlorophyllide oxidoreductase A (PORA) restores bulk**

chlorophyll synthesis and normal development to a porB porC double mutant. *Plant Mol Biol* 2010, **72:**445-457.

17. Park H, Kreunen SS, Cuttriss AJ, Dellapenna D, Pogson BJ: **Identification of the carotenoid isomerase provides insight into carotenoid biosynthesis, prolamellar body formation, and photomorphogenesis.** *Plant Cell* 2002, **14:**321-332.

18. Von Lintig J, Welsch R, Bonk M, Giuliano G, Batschauer A, Kleinig H: **Light-dependent regulation of carotenoid biosynthesis occurs at the level of phytoene synthase expression and is mediated by phytochrome in *Sinapis alba* and *Arabidopsis thaliana* seedlings.** *Plant J* 1997, **12:**625-634.

19. Botella-Pavia P, Besumbes O, Phillips MA, Carretero-Paulet L, Boronat A, Rodriguez-Concepcion M: **Regulation of carotenoid biosynthesis in plants: evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors.** *Plant J* 2004, **40:**188-199.

20. Li F, Vallabhaneni R, Yu J, Rocheford T, Wurtzel ET: **The maize phytoene synthase gene family: overlapping roles for carotenogenesis in endosperm, photomorphogenesis, and thermal stress tolerance.** *Plant Physiol* 2008, **147:**1334-1346.

21. Li F, Tsfadia O, Wurtzel ET: **The phytoene synthase gene family in the Grasses: subfunctionalization provides tissue-specific control of carotenogenesis.** *Plant Signal Behav* 2009, **4:**208-211.

22. Philippar K, Geis T, Ilkavets I, Oster U, Schwenkert S, Meurer J, Soll J: **Chloroplast biogenesis: the use of mutants to study the etioplast-chloroplast transition.** *Proc Natl Acad Sci U S A* 2007, **104:**678-683.

23. Leivar P, Tepperman JM, Monte E, Calderon RH, Liu TL, Quail PH: **Definition of early transcriptional circuitry involved in light-induced**

reversal of PIF-imposed repression of photomorphogenesis in young Arabidopsis seedlings. *Plant Cell* 2009, **21:**3535-3553.

24. Toledo-Ortiz G, Huq E, Rodriguez-Concepcion M: **Direct regulation of phytoene synthase gene expression and carotenoid biosynthesis by phytochrome-interacting factors.** *Proc Natl Acad Sci U S A* 2010, **107:**11626-11631.

25. Meier S, Bastian R, Donaldson L, Murray S, Bajic V, Gehring C: **Co-expression and promoter content analyses assign a role in biotic and abiotic stress responses to plant natriuretic peptides.** *BMC Plant Biol* 2008, **8:**24.

26. Meier S, Ruzvidzo O, Morse M, Donaldson L, Kwezi L, Gehring C: **The Arabidopsis wall associated kinase-like 10 gene encodes a functional guanylyl cyclase and is co-expressed with pathogen defense related genes.** *PLoS ONE* 2010, **5:**e8904.

27. Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** *BMC Bioinformatics* 2004, **5:**18.

28. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302:**249-255.

29. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A: **Transcriptional coordination of the metabolic network in Arabidopsis.** *Plant Physiol* 2006, **142:**762-774.

30. Ihmels J, Bergmann S, Barkai N: **Defining transcription modules using large-scale gene expression data.** *Bioinformatics* 2004, **20:**1993-2003.

31. Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR: **Arabidopsis Co-expression Tool (ACT): web server**

tools for microarray-based gene expression analysis. *Nucleic Acids Res* 2006, **34:**W504-W509.

32. Gallagher CE, Matthews PD, Li F, Wurtzel ET: **Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses.** *Plant Physiol* 2004, **135:**1776-1783.

33. Al Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J: **FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.** *Nucleic Acids Res* 2007, **35:**W91-W96.

34. Tian L, Dellapenna D, Dixon RA: **The pds2 mutation is a lesion in the Arabidopsis homogentisate solanesyltransferase gene involved in plastoquinone biosynthesis.** *Planta* 2007, **226:**1067-1073.

35. Sadre R, Gruber J, Frentzen M: **Characterization of homogentisate prenyltransferases involved in plastoquinone-9 and tocochromanol biosynthesis.** *FEBS Lett* 2006, **580:**5357-5362.

36. Motohashi R, Ito T, Kobayashi M, Taji T, Nagata N, Asami T, Yoshida S, Yamaguchi-Shinozaki K, Shinozaki K: **Functional analysis of the 37 kDa inner envelope membrane polypeptide in chloroplast biogenesis using a Ds-tagged Arabidopsis pale-green mutant.** *Plant J* 2003, **34:**719-731.

37. Munoz-Bertomeu J, Cascales-Minana B, Mulet JM, Baroja-Fernandez E, Pozueta-Romero J, Kuhn JM, Segura J, Ros R: **Plastidial glyceraldehyde-3-phosphate dehydrogenase deficiency leads to altered root development and affects the sugar and amino acid balance in Arabidopsis.** *Plant Physiol* 2009, **151:**541-558.

38. Tanaka R, Tanaka A: **Tetrapyrrole biosynthesis in higher plants.** *Annu Rev Plant Biol* 2007, **58:**321-346.

39. Masuda T, Fujita Y: **Regulation and evolution of chlorophyll metabolism.** *Photochem Photobiol Sci* 2008, **7:**1131-1149.

40. Floss DS, Hause B, Lange PR, Kuster H, Strack D, Walter MH: **Knock-down of the MEP pathway isogene 1-deoxy-D-xylulose 5-phosphate synthase 2 inhibits formation of arbuscular mycorrhiza-induced apocarotenoids, and abolishes normal expression of mycorrhiza-specific plant marker genes.** *Plant J* 2008, **56:**86-100.

41. Lange BM, Ghassemian M: **Genome organization in Arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism.** *Plant Mol Biol* 2003, **51:**925-948.

42. Okada K, Kasahara H, Yamaguchi S, Kawaide H, Kamiya Y, Nojiri H, Yamane H: **Genetic evidence for the role of isopentenyl diphosphate isomerases in the mevalonate pathway and plant development in Arabidopsis.** *Plant Cell Physiol* 2008, **49:**604-616.

43. Okada K, Saito T, Nakagawa T, Kawamukai M, Kamiya Y: **Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in Arabidopsis.** *Plant Physiol* 2000, **122:**1045-1056.

44. Keller Y, Bouvier F, d'Harlingue A, Camara B: **Metabolic compartmentation of plastid prenyllipid biosynthesis--evidence for the involvement of a multifunctional geranylgeranyl reductase.** *Eur J Biochem* 1998, **251:**413-417.

45. Hirooka K, Bamba T, Fukusaki E, Kobayashi A: **Cloning and kinetic characterization of Arabidopsis thaliana solanesyl diphosphate synthase.** *Biochem J* 2003, **370:**679-686.

46. Hirooka K, Izumi Y, An CI, Nakazawa Y, Fukusaki E, Kobayashi A: **Functional analysis of two solanesyl diphosphate synthases from Arabidopsis thaliana.** *Biosci Biotechnol Biochem* 2005, **69:**592-601.

47. Ruppel N, Hangarter R: **A mutant in geranylgeranyl diphosphate synthase 1 (GGPS1) of Arabidopsis thaliana that affects chloroplast development in adult leaves.**

48. van Schie CC, Ament K, Schmidt A, Lange T, Haring MA, Schuurink RC: **Geranyl diphosphate synthase is required for biosynthesis of gibberellins.** *Plant J* 2007, **52:**752-762.

49. Auldridge ME, Block A, Vogel JT, Dabney-Smith C, Mila I, Bouzayen M, Magallanes-Lundback M, Dellapenna D, McCarty DR, Klee HJ: **Characterization of three members of the Arabidopsis carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family.** *Plant J* 2006, **45:**982-993.

50. Tan BC, Joseph LM, Deng WT, Liu L, Li QB, Cline K, McCarty DR: **Molecular characterization of the Arabidopsis 9-cis epoxycarotenoid dioxygenase gene family.** *Plant J* 2003, **35:**44-56.

51. Iuchi S, Kobayashi M, Taji T, Naramoto M, Seki M, Kato T, Tabata S, Kakubari Y, Yamaguchi-Shinozaki K, Shinozaki K: **Regulation of drought tolerance by gene manipulation of 9-cis-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in Arabidopsis.** *Plant J* 2001, **27:**325-333.

52. Ghassemian M, Lutes J, Tepperman JM, Chang HS, Zhu T, Wang X, Quail PH, Markus Lange B: **Integrative analysis of transcript and metabolite profiling data sets to evaluate the regulation of biochemical pathways during photomorphogenesis.** *Arch Biochem Biophys* 2006, **448:**45-59.

53. Mayer MP, Beyer P, Kleinig H: **Quinone compounds are able to replace molecular oxygen as terminal electron acceptor in phytoene desaturation in chromoplasts of Narcissus pseudonarcissus L.** *Eur J Biochem* 1990, **191:**359-363.

54. Okamoto M, Tatematsu K, Matsui A, Morosawa T, Ishida J, Tanaka M, Endo TA, Mochizuki Y, Toyoda T, Kamiya Y et al.: **Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays.** *Plant J* 2010, **62:**39-51.

55. Braybrook SA, Harada JJ: **LECs go crazy in embryo development.** *Trends Plant Sci* 2008, **13:**624-630.

56. Tian L, Dellapenna D, Zeevaart JAD: **Effect of hydroxylated carotenoid defi ciency on ABA accumulation in *Arabidopsis*.** *Physiology of Plants* 2004, **122:**314-320.

57. Seo M, Hanada A, Kuwahara A, Endo A, Okamoto M, Yamauchi Y, North H, Marion-Poll A, Sun TP, Koshiba T et al.: **Regulation of hormone metabolism in Arabidopsis seeds: phytochrome regulation of abscisic acid metabolism and abscisic acid regulation of gibberellin metabolism.** *Plant J* 2006, **48:**354-366.

58. Seo M, Nambara E, Choi G, Yamaguchi S: **Interaction of light and hormone signals in germinating seeds.** *Plant Mol Biol* 2009, **69:**463-472.

59. Alabadi D, Gil J, Blazquez MA, Garcia-Martinez JL: **Gibberellins repress photomorphogenesis in darkness.** *Plant Physiol* 2004, **134:**1050-1057.

60. Penfield S, Li Y, Gilday AD, Graham S, Graham IA: **Arabidopsis ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm.** *Plant Cell* 2006, **18:**1887-1899.

61. Mouchel CF, Osmont KS, Hardtke CS: **BRX mediates feedback between brassinosteroid levels and auxin signalling in root growth.** *Nature* 2006, **443:**458-461.

62. Mouchel CF, Briggs GC, Hardtke CS: **Natural genetic variation in Arabidopsis identifies BREVIS RADIX, a novel regulator of cell proliferation and elongation in the root.** *Genes Dev* 2004, **18:**700-714.

63. Steber CM, McCourt P: **A role for brassinosteroids in germination in Arabidopsis.** *Plant Physiol* 2001, **125:**763-769.

64. Tepperman JM, Zhu T, Chang HS, Wang X, Quail PH: **Multiple transcription-factor genes are early targets of phytochrome A signaling.** *Proc Natl Acad Sci U S A* 2001, **98:**9437-9442.

65. Tepperman JM, Hudson ME, Khanna R, Zhu T, Chang SH, Wang X, Quail PH: **Expression profiling of phyB mutant demonstrates substantial contribution of other phytochromes to red-light-regulated gene expression during seedling de-etiolation.** *Plant J* 2004, **38:**725-739.

66. Mathews S: **Phytochrome-mediated development in land plants: red light sensing evolves to meet the challenges of changing light environments.** *Mol Ecol* 2006, **15:**3483-3503.

67. Ma L, Li J, Qu L, Hager J, Chen Z, Zhao H, Deng XW: **Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways.** *Plant Cell* 2001, **13:**2589-2607.

68. Tepperman JM, Hwang YS, Quail PH: **phyA dominates in transduction of red-light signals to rapidly responding genes at the initiation of Arabidopsis seedling de-etiolation.** *Plant J* 2006, **48:**728-742.

69. Quail PH: **Phytochrome-regulated Gene Expression.** *Journal of Integrative Plant Biology* 2007, **49:**11-20.

70. Parks BM, Spalding EP: **Sequential and coordinated action of phytochromes A and B during Arabidopsis stem growth revealed by kinetic analysis.** *Proc Natl Acad Sci U S A* 1999, **96:**14142-14146.

71. Jiao Y, Ma L, Strickland E, Deng XW: **Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis.** *Plant Cell* 2005, **17:**3239-3256.

72. Yanovsky MJ, Casal JJ, Whitelam GC: **Phytochrome A, phytochrome B and HY4 are involved in hypocotyl growth responses to natural radiation in Arabidopsis: weak de-etiolation of the phyA mutant under dense canopies.** *Plant, Celt and Environment* 1995, **18:**794.

73. Castillon A, Shen H, Huq E: **Phytochrome Interacting Factors: central players in phytochrome-mediated light signaling networks.** *Trends Plant Sci* 2007, **12:**514-521.

74. Nagatani A: **Light-regulated nuclear localization of phytochromes.** *Curr Opin Plant Biol* 2004, **7:**708-711.

75. Leivar P, Monte E, Oka Y, Liu T, Carle C, Castillon A, Huq E, Quail PH: **Multiple phytochrome-interacting bHLH transcription factors repress premature seedling photomorphogenesis in darkness.** *Curr Biol* 2008, **18:**1815-1823.

76. Hu W, Su YS, Lagarias JC: **A Light-Independent Allele of Phytochrome B Faithfully Recapitulates Photomorphogenic Transcriptional Networks.** *Mol Plant* 2009, **2:**166-182.

77. Oh E, Kim J, Park E, Kim JI, Kang C, Choi G: **PIL5, a phytochrome-interacting basic helix-loop-helix protein, is a key negative regulator of seed germination in *Arabidopsis thaliana*.** *Plant Cell* 2004, **16:**3045-3058.

78. Al Sady B, Ni W, Kircher S, Schafer E, Quail PH: **Photoactivated phytochrome induces rapid PIF3 phosphorylation prior to proteasome-mediated degradation.** *Mol Cell* 2006, **23:**439-446.

79. Chaves MM, Flexas J, Pinheiro C: **Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell.** *Ann Bot* 2009, **103:**551-560.

80. Schwartz SH, Qin X, Zeevaart JA: **Elucidation of the indirect pathway of abscisic acid biosynthesis by mutants, genes, and enzymes.** *Plant Physiol* 2003, **131:**1591-1601.

81. Li F, Vallabhaneni R, Wurtzel ET: **PSY3, a new member of the phytoene synthase gene family conserved in the Poaceae and regulator of abiotic stress-induced root carotenogenesis.** *Plant Physiol* 2008, **146:**1333-1345.

82. Welsch R, Wust F, Bar C, Al Babili S, Beyer P: **A third phytoene synthase is devoted to abiotic stress-induced abscisic acid formation in rice and defines functional diversification of phytoene synthase genes.** *Plant Physiol* 2008, **147:**367-380.

83. Ovadia A, Tabibian-Keissar H, Cohen Y, Kenigsbuch D: **The 5'UTR of CCA1 includes an autoregulatory cis element that segregates between light and circadian regulation of CCA1 and LHY.** *Plant Mol Biol* 2010, **72:**659-671.

84. Tompa M: **Identifying functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1143-1144.

85. Martinez-Garcia JF, Huq E, Quail PH: **Direct targeting of light signals to a promoter element-bound transcription factor.** *Science* 2000, **288:**859-863.

86. Huq E, Quail PH: **PIF4, a phytochrome-interacting bHLH factor, functions as a negative regulator of phytochrome B signaling in Arabidopsis.** *EMBO J* 2002, **21:**2441-2450.

87. Moon J, Zhu L, Shen H, Huq E: **PIF1 directly and indirectly regulates chlorophyll biosynthesis to optimize the greening process in Arabidopsis.** *Proc Natl Acad Sci U S A* 2008, **105:**9433-9438.

88. Hudson ME, Quail PH: **Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data.** *Plant Physiol* 2003, **133:**1605-1616.

89. Ulmasov T, Ohmiya A, Hagen G, Guilfoyle T: **The Soybean *GH2/4* Gene That Encodes a Glutathione S-Transferase Has a Promoter That Is Activated by a Wide Range of Chemical Agents.** *Plant Physiol* 1995, **108:**919-927.

90. Ulmasov T, Hagen G, Guilfoyle TJ: **ARF1, a transcription factor that binds to auxin response elements.** *Science* 1997, **276:**1865-1868.

91. Nemhauser JL, Mockler TC, Chory J: **Interdependency of brassinosteroid and auxin signaling in Arabidopsis.** *PLoS Biol* 2004, **2:**E258.

92. Goda H, Shimada Y, Asami T, Fujioka S, Yoshida S: **Microarray analysis of brassinosteroid-regulated genes in Arabidopsis.** *Plant Physiol* 2002, **130:**1319-1334.

93. Al Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.

94. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W: **GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox.** *Plant Physiol* 2004, **136:**2621-2632.

95. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S: **NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.** *Nucleic Acids Res* 2004, **32:**D575-D577.

96. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al.: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37:**D885-D890.

97. Kankainen M, Pehkonen P, Rosenstom P, Toronen P, Wong G, Holm L: **POXO: a web-enabled tool series to discover transcription factor binding sites.** *Nucleic Acids Res* 2006, **34:**W534-W540.