ABSTRACT

Title: **TIME-SERIES TRANSCRIPTOMIC ANALYSIS OF A SYSTEMATICALLY PERTURBED *Arabidopsis thaliana* LIQUID CULTURE SYSTEM: A SYSTEMS BIOLOGY PERSPECTIVE**

Bhaskar Dutta, Ph.D., 2007

Directed By: **Professor Maria I. Klapa,**
Department of Chemical and Biomolecular Engineering

Revealing the gene regulation network has been one of the main objectives of biological research. Studying such a complex, multi-scale and multi-parametric problem requires educated fingerprinting of cellular physiology at different molecular levels under systematically designed perturbations. Conventional biology lacked the means for holistic analysis of biological systems. In the post-genomic era, advances in robotics and biology lead to the development of high-throughput molecular fingerprinting technologies. Transcriptional profiling analysis using DNA microarrays has been the most widely used among them.

My Ph.D. thesis concerns the dynamic, transcriptional profiling analysis of a systematically perturbed plant system. Specifically, *Arabidopsis thaliana* liquid cultures were subjected to three different stresses, i.e. elevated CO_2 stress, salt (NaCl) stress and sugar (trehalose) applied individually, while the latter two stresses were also applied in combination with the CO_2 stress. The transcriptional profiling of these conditions involved carrying out 320 microarray hybridizations, generating thus a vast amount of transcriptomic data for *Arabidopsis thaliana* liquid culture system. To upgrade the dynamic information content in the data, I developed a statistical analysis strategy that enables at each time point of a time-series the identification of genes whose expression changes in statistically significant amount due to the applied stress. Additional algorithms allow for further exploration of the dynamic significance analysis results to extract biologically relevant conclusions. All algorithms have been incorporated in a software suite called *MiTimeS*, written in C++. MiTimeS can be applied accordingly to analyze time-series data from any other high-throughput molecular fingerprint.

The experimental design combined with existing multivariate statistical analysis techniques and MiTimeS revealed a wealth of biologically relevant dynamic information that had been unobserved before. Due to the high-throughput nature of the analysis, the study enabled the simultaneous identification and correlation of parallel-occurring phenomena induced by the applied stress. Stress responses comparisons indicated that transcriptional response of the biological system to combined stresses is usually not the cumulative effect of individual responses. In addition to the significance of the study for the analysis of the particular plant system, the experimental and analytical strategies used provide a systems biology methodological framework for any biological system, in general.

TIME-SERIES TRANSCRIPTOMIC ANALYSIS OF A SYSTEMATICALLY PERTURBED *Arabidopsis thaliana* LIQUID CULTURE SYSTEM: A SYSTEMS BIOLOGY PERSPECTIVE

Bhaskar Dutta

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Maria I. Klapa, Chair
Professor John Quackenbush
Professor Nam Sun Wang
Professor Evanghelos Zafiriou
Professor George H. Lorimer
Professor Srinivasa R. Raghavan

Dedicated to my parents and Indrani

Acknowledgements

Like many others, Ph.D. has been a long and eventful expedition for me. Those who go through the odyssey start from the same mark and also share the same destination, but fortunately or unfortunately traverse different paths. So I believe it's the journey that makes them different and it's the journey that matters when we all reach the three letter destination. As I see the shimmering lights from the lighthouse somewhere close I look back to recount the voyage I took and the voyagers who accompanied me.

I joined the graduate program at University of Maryland in fall of 2002. I made another literal journey half across the globe to this place to explore a new world, at least the most happening part of it. The transition to this seemingly new world was made smoother due to effort and advice of countless people from our Chemical Engineering Department and Student Council of India. I thank them for giving me warm welcome and helping me to settle down.

After couple of hasty months with coursework it was time to head blues of Atlantic. On a bright morning and I boarded the ship that I felt is most exciting, equipped and can meet my career aspirations. I joined Prof. Maria Klapa's group as Ph.D. student. I thank Prof. Klapa for selecting me as one of her sailors and for constant encouragement, support and guidance that she has provided over the years. We went on sometimes with or against the wind. But we never let our morals down and often kept going on till quite late night. I also thank Harin Kanani for all the wonderful time we spent together and synergistic work that we continued. Things were going fine till the winter of 2003 when we hit the iceberg and the Titanic was about to sink in an ice cold winter. Prof. Klapa informed us she has to avail a leave of absence for a year for personal reasons. It was quite an unusual situation and was difficult to foresee the future implications. A ship without a captain can wander away in deep blues. I thank those who have given me advice and courage in those days and appreciate Dr. Klapa's effort to remain virtually close to us always from that time in spite of all the work she has to do in Greece. I am also in debt to numerous 0s and 1s for carrying our ideas, discussions and emotions across the Atlantic without which this intercontinental research work couldn't have flourished.

I am thankful to Dr. Quackenbush for providing the facility, nearly 400 microarray slides and researchers in his lab like Dr. Fenglong Liu, Linda Moy in his lab for patiently guiding me through the delicate microarray experiments. During my seven month stint in TIGR I enjoyed the company of all these interesting people. Though it was the toughest time of my Ph.D. career but will remain as one of my sweetest memories for years. I am grateful to Dr. Nirmal Bhagabati for our personal communications. I enjoyed working with undergraduate students Robert Snyder and Quan Pham while coding for MiTimeS software. I would also like to extend my gratitude to all the professors who taught me at University of Maryland, especially Dr. Wang, Dr. Bentley and some senior students for helping me to understand the maps and read compass. I would extent my thanks to Dr. Zafiriou, Dr. Lorimer and Dr. Raghavan for their participation and advice as committee members.

Journey would have been much boring without all the fun I had with my roommates like Basu, Mainak, Kaushik and other fiends like Kousik, Amardeep, Ranjan, Swarat, Subhra, Angela, Rinku, Yechun, Ye and many more. They have accompanied me at different phases of journey sharing equally the fun and the burden. Possibly the only soul who has always been with me from crest to trough in this journey is Indrani. She had been like an aromatic balm soothing the days wound and preparing me for the next.

Last but not the least, I would like to thank my parents who have helped me to develop the sense of reasoning, love for science and thrill in adventures, in spite of knowing mere thanks do not commensurate their effort. I hope I could make them proud of me and I will miss them at my commencement.

As I move forward in life with the values that I learnt, the experiences that I garnered through my journey in last five years, now it is the destination that matters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 Background and Motivation

Better understanding of the biological systems requires them to be studied in their entirety. Rather than investigating small and isolated sections, systems engineering approaches should be applied to biological systems to reveal the functionality and interaction between different parts of the system [Klapa and Quackenbush, 2003]. For the implementation of systems engineering approaches in complex biological systems, high resolution maps of cellular fingerprints are required. Indeed, in the post genomic era, analysis of biological systems has moved from measuring a small set of biological markers, to the measurement of entire cellular fingerprints. This was made possible due to the development of high-throughput technologies, like DNA microarrays [Schena et al., 1995; Brown et al., 1999; Fodor 1997], mass spectral analysis of proteins [Gevaert et al., 2003; Lopez et al., 2003; Manabe et al., 2003; Wang et al., 2003] and metabolites [Roessner et al., 2000; Fiehn, 2000; Taylor, 2002].

Gene expression analysis using DNA microarrays being the most widely used “omics”, it is becoming increasingly clear that comprehensive picture of a biological system requires the integrated high-throughput analysis of multiple levels of cellular function [Ideker et al., 2001, Hwang et al 2005 a and b, Hirai et al. 2004, Dutta et al. 2007a]. In a systematically perturbed system, integrated analyses can provide insight about the function of unknown genes, the relationship between gene and metabolic regulation, and even the reconstruction of gene regulatory networks [Klapa et al., 2003].

Selection of model system is one of the most important parameters of the systems biology experimental design. *Arabidopsis thaliana* liquid cultures were selected as the model system in the proposed research. Beyond the traditional role of plant in providing food, neutraceuticals and natural polymers, for commercial, environmental and (bio)ethical reasons, plants are now taking central stage in bio-fuel [Ragauskas et al., 2006], engineered bio-polymer [Slater et al., 1999] and chemical [Oksman-Caldentey and Inze, 2004] industry. Research efforts to engineer plants to produce desired products are increasing considerably. Plants are complex eukaryotic systems, so it is also expected that the developed/ applied experimental and computational techniques and the conclusions about transcriptional regulation might be easily extended to other higher organisms. *A. thaliana,* which is a model system for plant research, was chosen as it has a small genome constituting of 5 chromosomes which is fully sequenced and best annotated. It has a small growth cycle of 3-6 weeks making it an ideal for conducting experiments. Though most of the plant physiology research was carried out in soil grown plants, liquid culture was chosen for this experiment as it provides a more controllable growth environment. Liquid culture also allows the perturbations applied to the growth media to get homogeneously and immediately distributed.

In order to study the regulation mechanisms those are active in any biological system, it needs to be systematically perturbed from different perspectives and its responses should be analyzed and compared. In case of the transcriptional response of a systematically perturbed biological system, this can provide information about function of unknown genes and structure of gene regulation network. Comparing the change in transcriptional profile over a large set of perturbations, it might be possible to differentiate between stress specific and common stress activated genes. Specifically in *A. thaliana* from multiple stress studies [Kreps et al., 2002; Cheong et al., 2002] it was observed that common stress induced gene expression is predominant as acute effect, whereas the stress specific changes in gene expression is more delayed effect. Studying the response of combination of stresses is even more interesting because comparing the combined stress response with respect to its constitutive ones can reveal part of the regulatory network that remains conserved for a particular stress. Multiple stress response studies can help to identify a core gene regulatory network involved in general stress response of the system under investigation.

To compare the transcriptomic data sets obtained from two different physiological conditions statistical methods like Fold Change (FC), t-test [Wang et al., 2004], SAM [Tusher et al., 2001; Saidi et al., 2004; Tian et al., 2004], one-way and multi-way ANOVA [Orlando et al., 2004; Zhao et al., 2002] are commonly used. FC is a crude measure to identify the differentially expressed genes. SAM is a modified t-test [Tusher et al., 2001] which also provides false detection rate (FDR) which is a measure of how reliable the test is. Yang et al.

[2004] provides a comprehensive comparison of different statistical methods used for testing differentially expressed genes. A statistic, derived from different statistics, was shown to give a better measure of the difference in gene expressions.

For systems biology experiments time-series is preferable compared to snapshot experimets because a snapshot experiment can not reveal the causality among the variables and usually insufficient to capture the full picture of the change initiated from this perturbation. This is true, because gene expression inherently is a temporal process. Different proteins that are required for performing different functions are produced from gene expression [Lewin et al., 2000]. Even under normal condition, due to degradation of proteins, mRNA is continuously transcribed and new proteins are generated. One of the major ways this process is regulated is by using a feedback loop. Genes are transcribed to lead to the production of proteins, some of which, like transcription factors (TF) can in turn regulate the transcription of other genes [Alberts et al., 2003]. So the causality in gene expression can lead to a time lag or dependencies in temporal expression profiles [Bar-Joseph et al., 2004]. When cells are exposed to a new condition (treatment or stress) they respond to the situation; thus by changing their gene expression. In most of the cases, the gene expression process starts by activating few transcription factors, which in turn activate the other genes that will respond to the new condition. If a snapshot of the new condition is compared with old condition, a set of genes in transient state at this time point of change can be found. Therefore, in order to determine the complete set of genes that are undergoing change and to explain the interaction between the genes that were involved in the process, it is necessary to study the change in expression profile over time. This allows us to determine not only the gene expression at the new state but also the pathways and networks that were involved to arrive at this new state [Bar-Joseph et al., 2004]. Few experiments have been conducted that capture the temporal profile of the response of biological systems to perturbations, especially in the case of higher eukaryotic organisms.

An important problem in the design of the time-series experiments is the selection of the proper sampling time. If the experiment is under-sampled (large time difference between the samples) then events might be missed on a shorter time scale. On the other hand over-sampled experiments are more expensive and difficult to carry out for higher eukaryotes as plants or mammalians. Therefore, shorter sampling periods usually leads to shorter duration of the experiment, missing important events that are occurring at a later stage. Most of the experiments conducted for the study of the expression change over time refer to a sampling time of 7 to 15 min [Spellman et al., 1998; Chu et al., 1998; Zhu et al., 2000]. Experiments were conducted to detect the genes that are periodically expressed [Spellman et al., 1998]. Identifying such genes is challenging, because different genes may have different phase, amplitude and periodicity [Joseph et al., 2004] of expression. Another notable effort was made by Holter et al. [2001] to build a linear time-delayed model of gene expression for different data sets [Spellman et al., 1998; Chu et al., 1998]. A more sophisticated AutoRegressive with eXogenous (ARX) model was proposed by Schmitt et al. [2003] for modeling time series gene expression data obtained by subjecting cultures of the photosynthetic bacterium *Synechocystis* to consecutive light-and-dark transitions. Akike's information criterion (AIC) was used for model selecting optimal model which gives best prediction without over fitting the data.

The methods that have been employed to analyze the difference between two transcriptional snapshots can not be used as such to provide conclusions about the change between two time profiles. Therefore suitable application of time series analysis algorithms to identify differentially expressed genes is required.

In this context of current biological research, my Ph.D thesis addressed challenges in the experimental and analytical techniques for high-throughput time-series transcriptional profiling analysis of a systematically perturbed *A. thaliana* liquid culture system.

1.2 Main Objective and Specific Aims

The main objective of my PhD work was the high-throughput, quantitative, time-series transcriptional profiling analysis of a systematically perturbed *Arabidopsis thaliana* liquid culture system. For this objective to be achieved, the following specific aims were pursued:

Specific Aim 1:

To develop a methodology for significance analysis of time-series transcriptomic data.

Specific Aim 2:

To determine systematic perturbations, to design and carry out the relevant experiments.

Specific Aim 3:

To apply multivariate statistical analysis of transcriptomic data and interprete the results in the context of the known plant (*A. thaliana*) physiology.

1.3 Description of the thesis

The thesis is organized into 7 chapters.

Chapter 1: The background and motivation behind the present work is presented and an overview of the main and specific objectives of the research is provided. A short description of each chapter of the thesis is also provided.

Chapter 2: It provides a brief introduction to DNA microarray technology and a detailed description of different normalization and clustering techniques used for microarray data processing and analysis. The techniques described were used to analyze the data of the present study.

Chapter 3: It describes the algorithms developed for significance analysis of time-series transcriptomic data. Methods were also proposed for better exploration of information contain in any high-throughput time-series molecular fingerprint data.

Chapter 4: Experimental design was explained. In materials and methods section common experimental and computational steps followed for DNA microarray analysis is discussed in detail.

Chapter 5: It contains the results from individual and combined stress response studies. Different multivariate statistical analysis was used to compare the stress response. Results were discussed in the context of *Arabidopsis thaliana* physiology. Multiple stress responses were compared to identify if (a) some of the stress responses are conserved, (b) there exist a common pool of stress response genes (c) the stress responses are additive.

Chapter 6: The focus of this chapter is to analyze the multiple stress responses simultaneously.. Genes were clustered based on all the experiment to reveal insight about their regulation. Clustering results were compared in the context of metabolic pathways, chromosomal locations and sequence alignment.

Chapter 7: Based on the conclusion and experience derived from the current experiment this chapter provides suggestion for better experimental design and methodologies that can integrate gene expression, metabolic and chromosomal location data.

2 TRANSCRIPTIONAL PROFILING

2.1 DNA microarray technology

Two different technologies are used for microarray slide preparation [Vivian et al., 1999]. Commercially it is manufactured by Affymetix [http://www.affymetrix.com]. It is produced by adding nucleotides sequentially using photolithographic technique to obtain desired sequence of oligo-nucleotides attached to the plate. The other technology cDNAs are printed onto chemically modified glass slides with the help of an arraying robot [Brown et al., 1999] and called spotted arrays. For this experiment spotted arrays printed in TIGR were used. In the rest of the document microarray refers to spotted arrays.

The first step in the preparation of microarray slides is proper probe (the sequence that are arranged on the microarray) selection. Then the probes are spotted. The arrayed genes are probes that can be used to query pooled, differentially labeled targets derived from RNA samples from different cellular phenotypes to determine the relative expression levels of each gene.

Two mRNA samples, one for control and another for query, from the tissues of interest are labeled with two different fluorescent dyes Cy3 and Cy5. Then they are purified and hybridized on the arrays. After hybridization, slides are scanned and independent images for control and query channels are generated. The relative fluorescence intensities give us a measure of relative amount of mRNA in control and query. After image processing data are normalized.

Normalization adjusts for differences in labeling and detection efficiencies for fluorescent labels and for difference in the quantity of initial mRNA from the two samples [Quackenbush, 2001].

The normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control is called "expression ratio" [Quackenbush, 2001]. The logarithm of the expression ratio is used because it is easy to understand. Genes that are up-regulated by a factor of two have a expression ratio of 2, hence $\log_2$(expression ratio) will have a value of 1. Similarly the genes that are down-regulated by the same factor will have a expression ratio 0.5 and $\log_2$(expression ratio) as -1. If the logarithm of expression level ranges between 1 to -1 then the expression level varies within 2 fold. So taking the logarithm of the expression makes the expression profile symmetric for a certain factor of up and down regulation.

There are a number of data analysis steps followed in sequence after the microarray slides are hybridized and scanned. TIGR TM4 software was used for microarray data analysis and the steps will be discussed in this context.

2.2 Image Processing

TIGR TM4 software spotfinder was used for image processing. The TIFF image files generated from the scanning of hybridized files is used for image processing. Image processing software takes the scanned image of both the dyes corresponding to each slide. Spotfinder generates TAV file which contains the information like position of the spot on the slide, intensity of the two dyes for each spot and whether the spot should be rejected or not.

2.3 Data Normalization

In many field comparisons are needed to extract conclusions, for an effective comparison appropriate normalization of the data is needed. In the context of DNA microarray analsis there is need for comparison among

i. Two different dyes

ii. Gene spots on the same slide

iii. Gene spots on different slides

In this process the source of systematic error that introduces difference between comparable data should be taken into consideration, so that data are compared only with respect to experimental perturbation. In the case of cDNA microarray analysis, such sources of systematic error arise in the experimental process of cDNA microarray development and hybridization. Following are sources of systematic error:

- Unequal quantities of starting RNA: in cDNA microarray RNA concentration of sample set is measured with respect to a reference. Equal amount of sample and reference RNA is taken so that they can be compared get relative expression of the sample with respect to reference.
- Difference in labeling efficiencies: fluorescent dye is attached to a mRNA sample through a biochemical reaction. Some dye can have preferential binding to one of the mRNA samples. Hence that mRNA sample will

always be shown at higher abundance compared to the other mRNA sample.

- Difference in scanning efficiencies: sample and reference are attached with two different dyes and after hybridization the slide is scanned for two different dye intensities in two different channels. Difference in sensitivity of the scanner for the two dyes can cause one of the dyes to be detected more effectively.
- Variation of the intensity across the slide: cDNA microarray is printed by a pen assembly and different parts (metablocks) of a microarray are printed by different pens. If there is variation among pens, this will translate into variation in the spots printed by different pens.

To account for the systematic errors various normalization methods have been proposed. In the rest of the text only those used in the present analysis in the context of MIDAS (TIGR TM4 software for normalization) are explained in greater detail.

2.3.1 Total intensity normalization

Total intensity normalization can eliminate the biases caused by difference in labeling and scanning efficiencies of the two dyes. It can also compensate for the unequal quantities of starting mRNA of the two sets. The total intensity normalization is based on the following hypothesis [Quackenbush 2002]. If the two samples to be compared have equal weight of mRNA, if the average mass of each molecule is approximately the same then each sample will have equal number of mRNA. It is also assumed that arrayed genes on the microarray slide equally interrogate the two mRNA samples. Hence the total number of mRNA molecules attached to the microarray slide is same for the two samples. Intensity of a spot is proportional to the amount of mRNA bound to the spot. As the total amounts of mRNA with two different dies are equal, the total fluorescent intensity for each die will also be equal. This can be checked by calculating the ratio of sum of intensities of two dyes, called normalization factor and is given by

$$N_{total} = \frac{\sum_{i=}^{N_{array}} R_i}{\sum_{j=1}^{N_{array}} G_j} \tag{2.1}$$

where R_i and G_i corresponds to the intensity of the red and green dye (two dyes used for two samples) for ith gene and N_{array} is total number of genes in the slide. In absence of any systematic error N_{total} value should be 1. When the value is not 1, then one of the samples (depending on which one is taken as reference) is scaled up or down depending on the value of N_{total}, so that, after the scaling the sum of the intensities of both the dyes are same. This process is equivalent to subtracting a constant from the logarithm of expression ratio.

$$\log_2(t_i) = \log_2(T_i) - \log_2(N_{total}) \tag{2.2}$$

where, t_i is normalized expression ratio and is given by

$$t_i = \frac{R_i}{\mathrm{N}_{\mathrm{total}} G_i} \tag{2.3}$$

T_i is expression ratio before normalization and is given by

$$T_i = \frac{R_i}{G_i} \tag{2.4}$$

N_{array} can be the number of genes on a section of the slide, a whole slide or number of slides. In the same way as above, in stead of comparing mean intensities, median intensities of the two samples can also be equated.

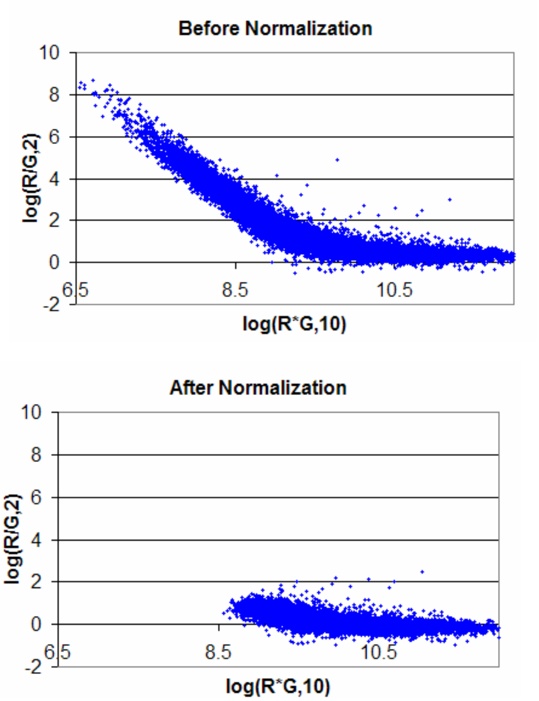


Figure 2.1: RI plot before after total intensity normalization. (R-I plot obtained from the data of one of the time points of the experiment, displaying the ratio of the intensities ($\log_2(R_i/G_i)$) as a function of the product of the intensities ($\log_{10}(R_i*G_i)$) before and after total intensity normalization.) Figures adopted from a microarray data used for my Masters thesis in 2004.

2.3.2 Lowess

It is observed very often that $\log_2(R_i/G_i)$ values can have a systematic dependence on intensity [Yang Y. et al., 2002 and Yang I. et al., 2002], which most commonly appears as a deviation from zero for low or high intensity spots. This leads to a long tail in R-I plot (plot of ratio of the intensities ($\log_2(R_i/G_i)$) as a function of the product of the intensities, $\log_{10}(R_i*G_i)$). Locally weighted regression (Lowess)

[Cleveland et al 1979] can take care of this systematic error in microarray data. It carries out a locally weighted regression between $\log_{10}$(Ri*Gi)and $\log_2$(Ri/Gi) and gets the best fit curve which predicts $\log_2$(Ri/Gi) as a function of $\log_{10}$(Ri*Gi). Best fit curve, which captures the systematic error in the data, is subtracted from each data ($\log_2$(Ri/Gi)) point to remove the systematic error in the data. The weights assigned in this locally weighted regression are function of the distance of the data points from the fitted curve. If a point is far from the curve then it has very low weight, as the point has more chance of being an outlier. Lowess carries out the regression for each block of the microarray slide separately. Lowess can also be applied globally by considering whole data set (all the spots of the microarray slide).

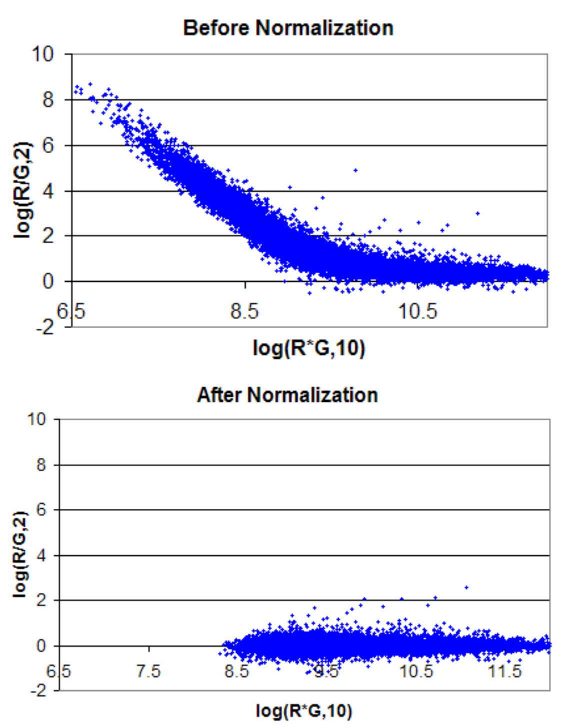


Figure 2.2: RI plot before and after lowess normalization. Notation and data used were same as that of figure 2.1.

The data after total intensity normalization in Fig 2.2 shows a systematic bias in RI plot. The plot is showing a small tail at low intensity values due to systematic error. This error is eliminated in the data after lowess normalization (Fig 2.2).

2.3.3 Standard Deviation Regularization

In the above normalization methods mean intensity of the two sets are equated. How the points are scattered around the mean is also an important criterion to study. In a spotted array different meta-blocks are printed by different pens, so the spots may vary slightly from meta-block to meta-block due to difference in pen. Standard deviation regularization scales the data so that there is same variation for all the meta-blocks. [Yang Y. et al., 2002],

It is assumed that the mean of $\log_2$(ratio) is already zero for each meta block, by applying the normalization methods discussed above. So the variance of the nth meta-block will be given by

$$\sigma^2{}_n = \sum_i^N \left(\log_2(T_i)\right)^2 \tag{2.5}$$

where T_i is ratio of the dye intensity for ith gene and is given by

$$T_i = \frac{R_i}{G_i} \tag{2.6}$$

N is the number of spots in a meta-block. Appropriate scaling factor for the jth meta-block is given by

$$a_j = \frac{\sigma_j{}^2}{\left[\prod_{k=1}^{N_{metablock}} \sigma_k{}^2\right]^{1/N_{metablock}}} \tag{2.7}$$

where $N_{\text{metablock}}$ is the number of meta-blocks in a slide. All the elements of the jth meta-block is scaled by dividing them with the scaling factor. Hence

$$\log_2(T_i) = \frac{\log_2(T_i)}{a_j}$$

(2.8)

Where, T_i is the ratio of red to green dye intensity for the ith gene in the jth meta-block. This is same as taking the a_j th root of all the intensities of the jth meta block. So the transformed intensities after the normalization become:

Or $G_i' = [G_i]^{1/a_j}$ and $R_i' = [R_i]^{1/a_j}$

(2.9)

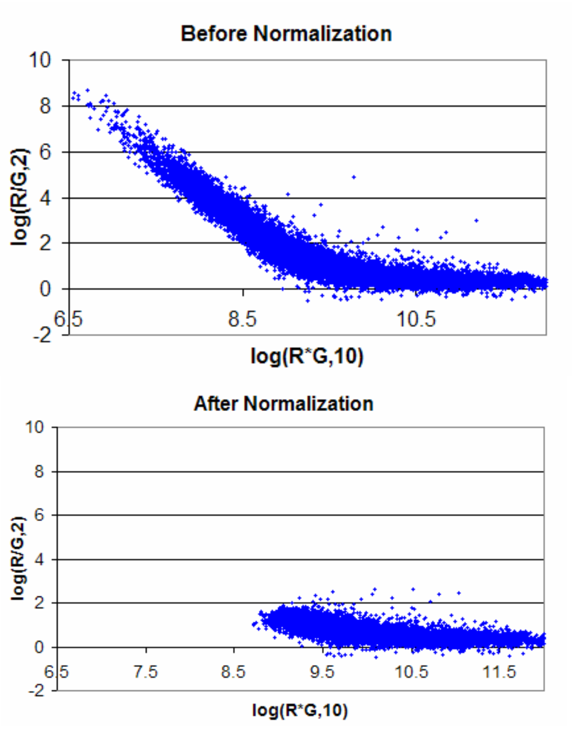


Figure 2.3: RI plot before and after standard deviation normalization. Notation and data used were same as that of figure 2.1.

2.3.4 Flip dye analysis

By performing a flip dye analysis biases that may occur during labeling and scanning, for example, some die may preferentially bind to mRNAs, can be eliminated [Quackenbush, 2002]. If one of the dyes has higher average intensity over the other, then the sample tagged with that dye will show higher expression, which is misleading. So the same experiment is carried out by swapping the dyes among the samples. If there are two samples A and B, then they can be tagged by two possible combinations, red and green or green and red dye respectively. In the first case when A and B are attached with red and green dye respectively, the ratio will be given by

$$T_{1i} = \frac{R_{1i}}{G_{1i}} = \frac{A_{1i}}{B_{1i}} \tag{2.10}$$

After the dyes are reversed the ratio will become

$$T_{2i} = \frac{R_{2i}}{G_{2i}} = \frac{B_{2i}}{A_{2i}} \tag{2.11}$$

As the same experiment is being performed and only the dyes are reversed, $\frac{A_{1i}}{B_{1i}}$ and $\frac{A_{2i}}{B_{2i}}$ are expected to be same. Hence

$$\frac{A_{1i}}{B_{1i}}\frac{B_{2i}}{A_{2i}} = (T_{1i} * T_{2i}) = 1 \tag{2.12}$$

$$\log_2\left(\frac{A_{1i}}{B_{1i}}\frac{B_{2i}}{A_{2i}}\right) = \log_2(T_{1i} * T_{2i}) = 0 \tag{2.13}$$

If the measurements are consistent then the value of $\log_2(T_{i1}*T_{i2})$ is expected to be zero, if it is not zero then close to zero. But if the value is far from zero, then the measurements are inconsistent. Either one of the measurements or both could be erroneous. The user can decide how stringent the rejection criteria of the erroneous data would be. Stringent criteria means only a small range of values around zero is acceptable.

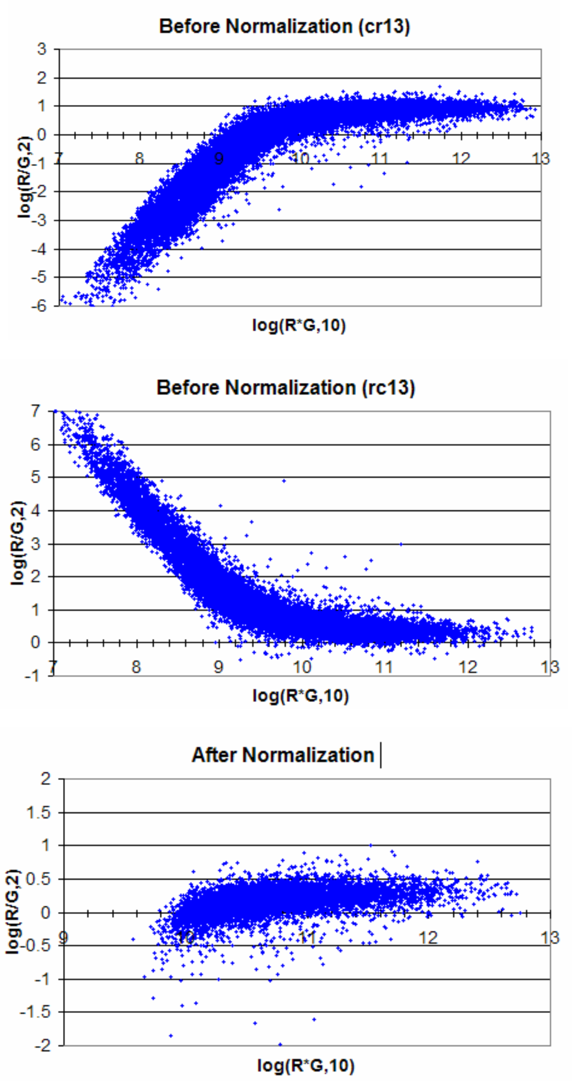


Figure 2.4: RI plot before and after flip dye normalization. Before normalization the RI plots has long tails and look like mirror image with respect to the line y =0 line. After normalization the dye based bias is gone. Notations and data used were same as that of figure 2.1.

2.4 Clustering Methods/ Statistical Analysis of DNA Microarray Data

Several clustering algorithms are used for the identification of the patterns in the gene-expression data. Clustering techniques can be classified as decisive or agglomerative [Quackenbush 2001]. A decisive method begins with all elements in one cluster that is gradually broken down into smaller and smaller clusters.

Agglomerative techniques start with single member clusters and gradually fuse them together. There are two types of clustering algorithms supervised or unsupervised [Quackenbush 2001]. Supervised methods use existing biological information about specific genes that are functionally related to 'guide' the clustering algorithm. Most of the algorithms described in this chapter are unsupervised.

2.4.1 Distance Metrics

Suppose N number of experiments are conducted to study the expression profiles of M genes. Then the expression of a particular gene in N experiments can be represented by a single point in N dimensional space. This is called expression space, as it has the same number of dimension as the number of experiments. Clustering algorithms group the genes together based on their "distance" from each other in the expression space. Distance gives a measure of similarity between the genes. There are various methods for calculating distances.

1. **Euclidean distance** is the most commonly used distance. It is a metric distance. Following are the characteristic of metric distances [Quackenbush 2001]. If d_{ij} is the distance between two vectors i and j,

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \qquad (2.14)$$

where, x_{ik} and x_{jk} are expression level of i^th and j^th genes respectively and n is the number of experiments

- Distance must be positive and definite, d_{ij} >o
- Distance must be symmetric, $d_{ij} = d_{ji}$

- An object is zero distance from itself, $d_{ii} = 0$
- It follows triangular inequality

2. **Manhattan distance** is given by:

$$d_{ij} = \sum_{k=1}^{n} | x_{ik} - x_{jk} | \qquad (2.15)$$

where n is the dimension of the expression space [Heyer et al., 1999].

3. **Pearson correlation** is given by [Eisen et al., 1998]

$$S(G_i, G_j) = \frac{1}{n} \sum_{k=1}^{n} \left(\frac{G_{i,k} - G_{i,offset}}{\Phi_i} \right) \left(\frac{G_{j,k} - G_{j,offset}}{\Phi_j} \right) \qquad (2.16)$$

Where,

$$\Phi_i = \sqrt{\sum_{k=1}^{n} \frac{(G_{i,k} - G_{i,offset})^2}{n}} \qquad (2.17)$$

$G_{i,\ offset}$ is the mean and Φ_i is the standard deviation of observation of the ith gene.

4. **Cosine correlation** is given by the following expression [Eisen et al., 1998]

$$C(x_{ik}, x_{jk}) = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\|x_{ik}\| \|x_{jk}\|} \qquad (2.18)$$

Distance between two clusters can be calculated in different ways:

Average linkage clustering: This is most frequently used. The distance between two clusters i and j is calculated by calculating the average of the distance between each gene of ith cluster with all other genes in the jth cluster. Two clusters with lowest average distance is joined together to form a new cluster.

Complete linkage clustering: Complete linkage clustering is known as the maximum or furthest-neighborhood method. The distance between two clusters is calculated as the greatest distance between the members of relevant clusters. This method often produces clusters that are often similar in size.

Single linkage clustering: The distance between two clusters is calculated as the smallest distance between the members of the relevant clusters. In this method there is a sequential addition of single samples in to an existing cluster. This produces trees with many long, single addition branches representing clusters that have grown by accretion.

If the expression level of a gene at each time point is viewed as a coordinate, then the standardized expression level of each gene at all n time points describes a point in *n* dimensional space, and the Euclidean distance between any two points in this space can be computed. It can be shown that the two points for which the distance is minimized are precisely the points that have the highest correlation. In other words, genes pairs with highly correlated expression pairs are close in expression space. It should be noted that simply using Euclidean distance without standardizing the data is ineffective, because gene pairs whose expression patterns have the same shape but different magnitudes will not score well.

To gauge the measure of a performance, one might consider taking gene pairs those are known to be co-regulated or functionally related, and computing the score (distance or correlation) of each pair. These scores could then be compared with the scores of unrelated gene pairs. The measure that gives high scores only to related genes would be chosen. Unfortunately neither Euclidean distance nor Pearson Correlation consistently gives high scores only to related gene pairs. In fact, not all related genes are coexpressed, and some unrelated genes have similar expression patterns. Because there is a connection between coexpression and functional relation, coexpressed genes provide excellent candidates for further study. However, the connection is complex, and it cannot be derived so easily [Heyer et al.,1999].

Two genes may be close according to one distance definition but may be far apart according to other. So the way we define distance between two expression vectors has a profound effect on the cluster they produce.

To study gene expression patterns statistical and clustering techniques have been proposed. In the rest of the text only the techniques that were used for the resent analysis will be discussed in detail.

2.4.2 Hierarchical Clustering

Hierarchical clustering is one of the first and widely used clustering techniques for expression data. The reason being, it is simple and the results can be visualized easily. Hierarchical clustering is an agglomerative approach in which expression profiles are joined in groups, which are further joined and this continues till completion, so that finally it forms a single tree. The algorithm of Hierarchical clustering is as follows. Initially each cluster contains a single gene. Then the pair-wise distance is calculated for all of the genes to be clustered. If they are formulated in a matrix form it forms a square matrix which is symmetric. This matrix is called distance matrix or similarity matrix. This matrix is scanned to figure out smallest value (if Euclidean distance is used, because it selects the genes that are closest in the expression space) or highest value (if Pearson correlation distance is used, because it finds the genes that have most similar expression profile). These two genes are most similar or closest, hence they are clustered together. If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives [Quackenbush, 2001]. A node is created joining these two genes, and gene expression profile is computed for the node by averaging observations for the joined elements [Eisen et al., 1998]. The similarity matrix is updated with this new node replacing the two joined element and the process for any set of *n* genes the process repeated n-1 times until only a single cluster remains.

There are several variations in Hierarchical clustering that differs in the rule governing how distances should be calculated among the clusters as they are constructed. There are three ways of calculating distances between two clusters, they are average linkage, complete linkage and single linkage. They are explained in detail in section 2.5.1.

There are several limitations of hierarchical clustering. Decisions to join two elements are based only on the distance between the two elements, and once the elements are joined they can not be separated [Tamayo et al., 1999]. This is a local decision making scheme that doesn't consider the data as a whole, and it may lead to mistakes in the overall clustering.

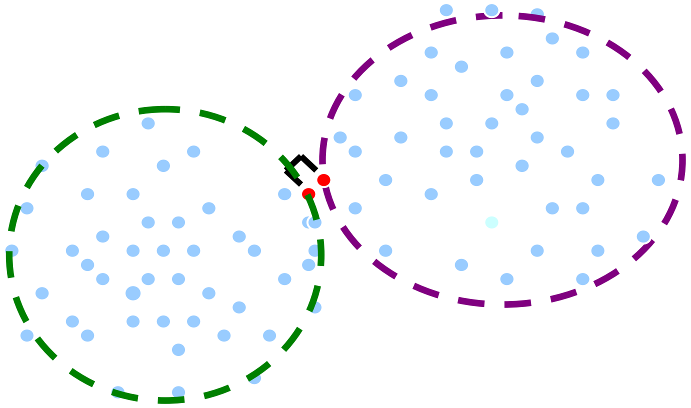


Figure 2.5: Limitation of hierarchical clustering. Hierarchical cluster start growing from the genes closest to each other, but they may belong to different cluster if overall picture is considered.

The Fig 2.5 shows there are two distinct clusters and the red points belong to different clusters but close to each other in expression space. Hierarchical clustering will join the points which are closest to each other in expression space. So the red points will be clustered together. But these points belong to two different clusters. So two points might have minimum distance but that doesn't necessarily mean that they have to belong to the same cluster. Hierarchical clustering has a shortcoming of suffering from lack of robustness and non-uniqueness problems [Tamayo et al., 1999]. An alternative approach to avoid some of the shortcomings are to use decisive clustering approach, such as k-means or self organizing maps, to partition data into groups which has similar expression pattern.

2.4.3 k- means clustering

This is a statistical algorithm [Velculescu et al., 1995] by which objects are partitioned into a fixed number (k) of clusters, such that the clusters are internally similar but externally dissimilar. If the advance knowledge of the number of clusters is known then k-means can separate the objects effectively. K-means clustering uses a supervised clustering algorithm that is conceptually simple but computationally intensive [Quackenbush 2001]. First all initial objects are randomly assigned to one of the k clusters. Then an average expression vector is calculated for each cluster which is eventually used to compute the distance between the clusters. Using an iterative method, objects are moved between clusters and intra and inter cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster. After each move, the expression vectors for each cluster are recalculated. The shuffling proceeds until moving any more objects will increase the intra-cluster distances and decrease inter-cluster dissimilarity.

Tavazoie (1999) used data gathered by Cho (1998) and applied k-means clustering algorithm and found the members of each cluster to be significantly enriched for genes with similar functions. They used k means algorithm to cluster 3000 genes into different regulation classes. Algorithm was repeated for 200-400 iterations and partitioned the data into 10, 30 and 60 clusters. It was observed that by 200 iterations the algorithm was converged. They finally chose 30-cluster partitioning because it provided the best compromise between number of clusters and separation between them.

2.4.4 Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a statistical technique that allows the key variables (or combination of variables) in a multidimensional data set to be identified. PCA determines those key variables in the data set that best explains the difference in the observations [Raychaudhuri et al., 2000].

PCA is very effective when some of the data might contain redundant information. For example if a group of experiments are more closely related than we had expected, we could ignore some of the redundant experiments or can take some average vale of the data without losing any information[Qucakenbush 2001]. PCA projects a high dimensional data into a lower dimensional space so that we can find the view, that gives the best separation of the data. Given a matrix of expression data, A, where each row corresponds to a different gene and each column corresponds to one of several different experimental conditions. The a_{it} entry of the matrix corresponds to i^{th} gene's relative expression ratio with respect to a control population under condition *t*. Using PCA each of the n components can be calculated for a given gene. To compute the principal components, the n (smallest of the number of experiments or number of genes) eigenvalues and their corresponding eigenvectors are calculated from the *n x n* covariance matrix of experimental conditions or time points. Each eigenvector defines a principal component.

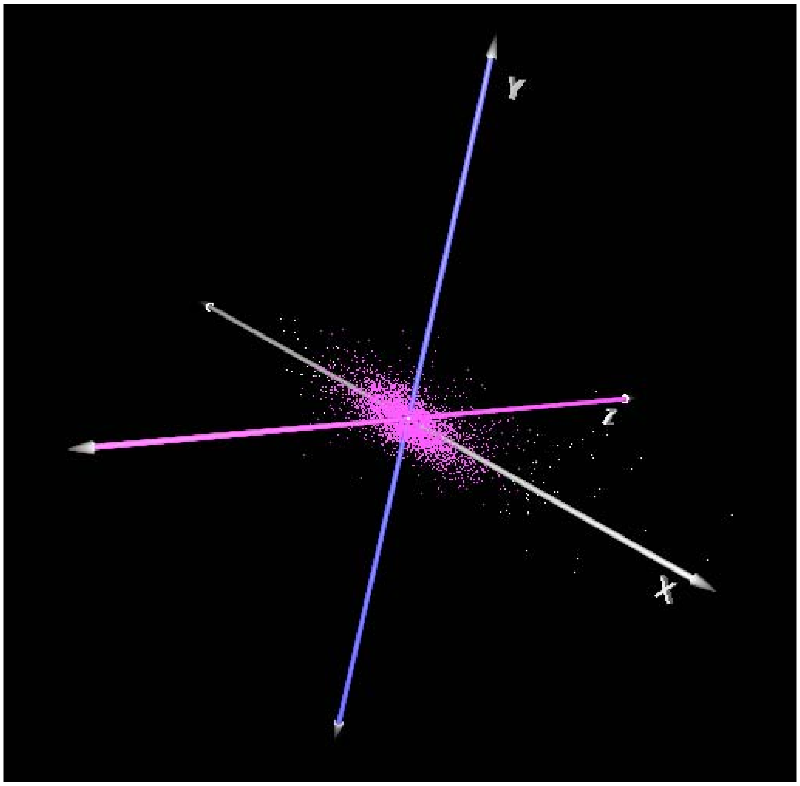


Figure 2.6: PCA of data (generated using TIGR TM4 software).

A component can be viewed as a weighted sum of the conditions (or time points) where the coefficients of the eigenvectors are the weights. Consequently, the eigenvectors with large eigenvalues are the once that contain most of the information; eigenvectors with small eigenvalues are uninformative [Raychaudhuri et al., 2000]. Data can be converted in terms of principal components from the following relation

$$a_{ij}' = \sum_{t=1}^{n} a_{it} v_{tj} \tag{2.19}$$

where v_{tj} is the t^th^ coefficient of the j^th^ principal component. a_{it} is the expression measurement for gene i under t^th^ condition. A' is the data in terms of principal components and V is the set of ortho-normal eigenvectors.

2.4.5 Statistical analysis using Significance Analysis of Microarrays (SAM)

SAM is a statistical method to identify the genes that are undergoing considerable change in expression between two sets of microarray data [Tusher et al., 2001]. SAM is a hypothesis testing based on student t test. Suppose n_1 observations of x_i and n_2 observations of y_i are given. It is assumed that x_i and y_i are normally distributed. Then a hypothesis is created that the population means are equal. Then it can be found out if the observations are consistent with the hypothesis [Meyer, 1975]. For unpaired SAM, a statistic is defined [Tusher et al., 2001] based on the ratio of change in gene expression to standard deviation in data for that gene.

$$d(i) = \frac{r(i)}{s(i) + s_o}$$

(2.20)

$$r(i) = \overline{x}(i) - \overline{y}(i) \qquad (2.21)$$

where $\overline{x}(i)$ and $\overline{y}(i)$ are defined as the average levels of expression for gene i in two different sets. *s(i)* is the standard deviation of repeated expression measurements.

$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \overline{x}(i)]^2 + \sum_n [y_n(i) - \overline{y}(i)]^2\right\}}$$

(2.22)

where,

$$a = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) * \frac{1}{n_1 + n_2 - 2} \qquad (2.23)$$

s_o is a positive constant which ensures the variance of *d(i)* is independent of gene expression.

Genes are ranked according to the magnitude of their *d(i)* values, therefore *d(1)* has the largest relative difference, *d(2)* has the second largest and *d(i)* has ith largest difference.

A large number of surrogate data is generated by permutation of the data used for analysis. For each of the permutations relative differences $d_p(i)$ were also calculated and the genes were ranked in the same way, so that $d_p(i)$ has the ith largest relative difference for pth permutation. Expected relative difference was calculated by

$$d_E(i) = \frac{\sum_p d_p(i)}{N} \tag{2.24}$$

Where N is the total number of permutations. To identify the significant changes in expressions, observed relative difference *d(i)* is plotted against the expected relative difference *d_E(i)*. For vast majority of the genes *d(i)* and *d_E(i)* values are expected to be same, hence they should be close to d(i) = d_E(i) line. Some genes can also be far from the line. If the distance of a gene from the line is greater than a threshold value, say delta (Δ), that gene can be called significant [Tusher et al., 2001].

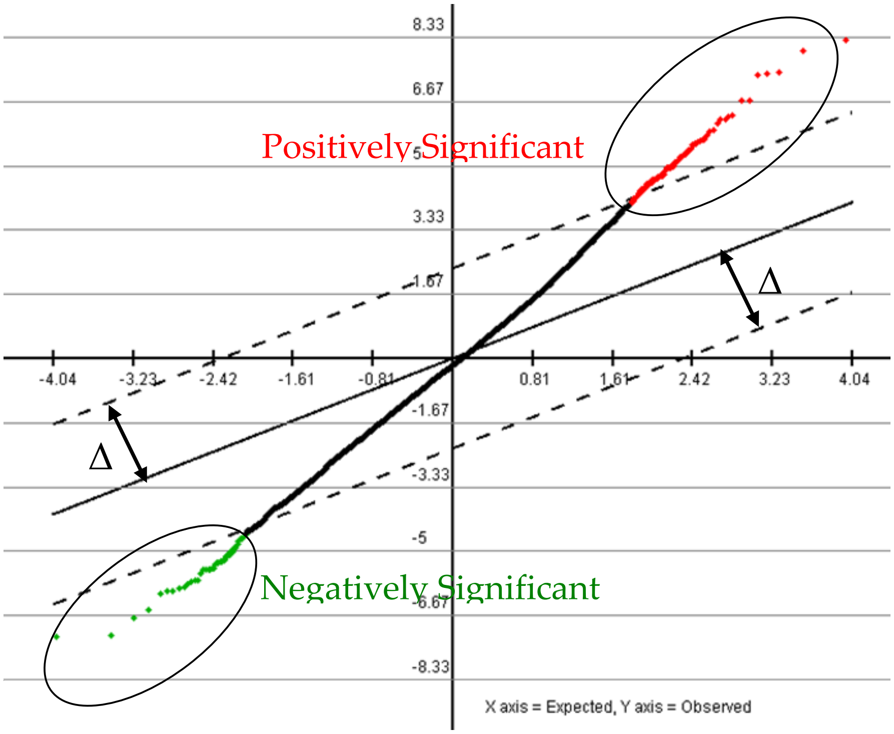


Figure 2.7: SAM graph showing the genes identified as positively and negatively significant marked with red and green respectively.

SAM can also give a measure of false discovery rate (FDR). It's a measure of percentage of genes identified as significant by chance. To determine the number of falsely significant genes generated by SAM, two parallel cutoffs were defined.

Cutoffs are lines on both sides of $d(i) = d_E(i)$ and parallel to it. The distance of the parallel lines from the line $d(i) = d_E(i)$ is given by the threshold value. The genes that are above the upper line can be called significantly induced and the genes which are lying below the lower line are called significantly repressed. The number of falsely significant genes corresponding to each permutation was computed by counting the number of genes that exceeded the horizontal cutoffs for that permutation. The estimated number of falsely significant genes is the average of the significant genes found in all the permutations.

2.4.5 Paired SAM

In control and perturbed experiments plants were harvested at same time points. So the difference in expression level of the perturbed and control samples should be compared for each time points separately. If unpaired SAM (explained in 2.4.5) is used, then it calculates the average expression level of the control and perturbed sets separately and finds the genes that are differentially expressed based on the averages calculated. Here we lose the information of individual time points by taking the average. Paired SAM computes the difference in expression of a gene between controlled and perturbed at each time point and calculates the statistic based on that. If there are K time points [1, 2, 3,… k] and x_{ij} of control pairs with y_{ij} of perturbed, r_i and s_i are calculated from the following equations [Stanford SAM manual]:

$$z_{ij} = x_{ij} - y_{ij} \tag{2.25}$$

$$r_i = \sum_j z_{ij} / K \tag{2.26}$$

$$s_i = \left[\sum_j (z_{ij} - r_i)^2 / \{K(K-1)\} \right]^{1/2} \tag{2.27}$$

Paired SAM can only be used if there is equal number of observations (time points) in the two sets to be compared and the samples are collected at the same time points.

3 ANALYSIS OF TIME-SERIES TRANSCRIPTOMIC DATA

3.1 Introduction:

In our effort to identify the mechanisms and networks underlying cellular function, biological studies have traditionally involved the perturbation of a cellular system in multiple ways and the monitoring of its response through various markers. Prior to the genomic revolution, these markers were mainly macroscopic. The high-throughput post-genomic era provided the tools, DNA microarrays (Brown and Botstsein, 1999; Schena et al., 1995) being the most often utilized among them, to also monitor simultaneously a great number of molecular markers (Klapa and Quackenbush, 2003). The computational problems, therefore, that biology has often to solve concern the identification of differentially expressed markers due to the applied perturbation(s). In the case of transcriptional profiling, in particular, these problems refer to the identification of differentially expressed genes between transcriptional profile populations representing different sets of physiological conditions. There are two types of experiments: “snapshot” and “time-series”. In the first type, each population comprises the same with respect to time transcriptomic snapshot of the cellular function under the particular set of conditions, but in different biological and/or experimental (i.e. injections of the same sample) replicates. In the time-series experiments, however, the transcriptional profiles that are acquired under each of the examined sets of conditions correspond to different, sequential in time snapshots of the biological process/system under investigation. In this case, it is of interest to identify the genes whose expression profile over time changes drastically due to the applied perturbation. Moreover, it would be of interest to compare the various time points with respect to the change in their transcriptional profile due to the applied perturbation, taking, however, into consideration that they are components of the same time-series.

The identification of differentially expressed genes in "snapshot" experiments is achieved using hypothesis testing methods like *t-test* (Pan, 2002; Korn et al., 2001; Baldi et al., 2001; Wang et al., 2004), F-test (Chen et al., 2005; Cui et al., 2005), *ANOVA* (Zar, 1999; Draghici et al., 2003; Orlando et al., 2004; Zhao et al., 2002), non-parametric t-test and Wilcoxon rank sum test (Troyanskaya et al., 2002), and the *Significance Analysis of Microarrays* (SAM) (Tusher et al., 2001; Larsson et al., 2005; Wu, 2005), a recent permutation estimation method tailored for the analysis of transcriptional profiling data. Permutation-based (non-parametric) compared to parametric hypothesis testing methods have the advantage of not requiring the data to follow a particular distribution. They also provide an estimation of the "False Discovery Rate (FDR)", i.e. the probability that a given gene identified as differentially expressed is a false positive. SAM provides an additional benefit: the flexibility for the user to adjust the threshold of significance and observe the sensitivity of FDR and number of significant genes to the threshold change.

The application, however, of these hypothesis testing techniques for the significance analysis of time-series data is not straightforward. They cannot directly take into consideration the specific order of the transcriptional profiles in time. For example, based on them, the change in the expression of the gene shown in Figure 3.1 from the physiological state 1 to states 2, 3 and 4 would be considered identical. While this is true for the gene's average change in time, it does not reflect its dynamic expression change. In this context, to upgrade the time-related information content of the measurements requires particular handling.

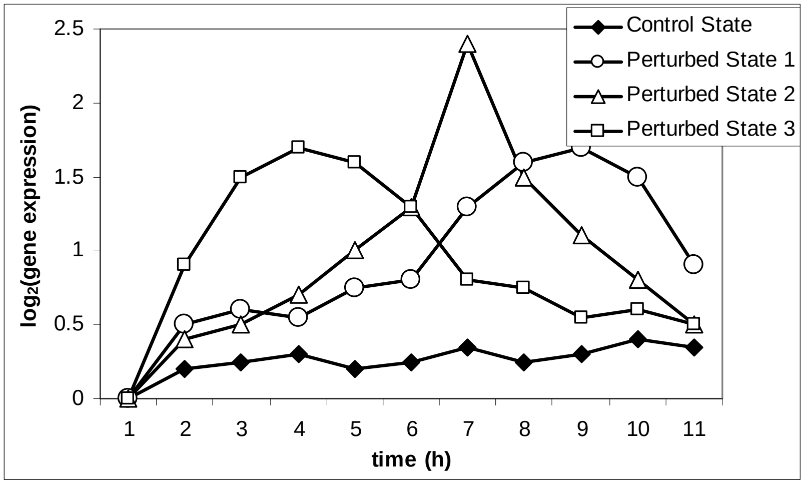


Figure 3.1 Paired-SAM bases conclusions on the average and not the dynamic gene expression profile. Paired-SAM allocates the same significance score to all three depicted changes in the expression profile of a gene over time due to three different perturbations.

Classical statistical methods for the modelling of time-series data that have been successfully applied to other fields, e.g. Moving Average, Auto Regressive or Auto Regressive Moving Average Modeling (Chatfield, 2003), are not usually expected to be equally effective within the context of transcriptional profiling data in particular or any other high-throughput biological dataset in general. This is true, because the number of time-points in biological experiments is most of the times – due to current experimental limitations - much smaller than the number of variables, the latter being equal to the number of monitored gene expressions in the case of transcriptional profiling analysis. A recent publication (Ernst *et al.*, 2005) pointed out that more than 80% of the reported time-series gene expression datasets, referring to thousands of genes, involves fewer than 9 time points. In these cases, the derived statistical models are expected to be rudimentary and simplistic. In this context, there are currently few reported algorithms for the significance analysis of dynamic gene expression data. Specifically, Bar-Joseph et al. (2003a) proposed a method for the analysis of time-series transcriptional datasets, based on fitting a continuous curve to the discrete data to describe the time profile of a gene's expression. Then, the two curve sets, each representing the time profiles of all genes' expressions under each of the examined experimental conditions, are compared to conclude whether they are independent or a noisy realization of each other (2003b). On a similar basis, Storey et al. (2005) proposed a model that describes each gene's expression as function of time; subsequently they test for which genes the model parameters are significantly different between the two investigated experimental conditions. In the time-series feature that is incorporated in SAM software (Chu et al.), the area under the time profile of each gene's expression is calculated for each biological sample in any of the two examined physiological conditions. Then, the area datasets referring to the two compared physiological conditions are analysed using SAM. These methods are quite significant as they allow for the identification of differentially expressed genes based on their expression profile over time. They do not enable, however, the comparison between the various timepoints. In a different approach, Park et al. (2003) used 2-way ANOVA to study how stress, but also time, affect the transcriptional profile, individually and in combination. Kamimura et al. (2000), in the context of fermentation process data, used mean hypothesis testing to identify the most discriminatory variables and time windows. Liu et al. (2005), on the other hand, compared the time-points of a plant growth process directly through the SAM-identified differentially expressed genes at each time point, considering, however, each time-point as an independent "snapshot" experiment. Consequently, each SAM analysis was conducted independently, without using a common reference for normalization among the time points.

A SAM-based algorithm is presented here [Dutta et al., 2007] that enables the identification of the differentially expressed genes at each timepoint of a time sequence, taking, however, into consideration that they correspond to sequential snapshots of the same biological process. This is achieved by comparing the gene expression profile of all timepoints with a common reference distribution and by identifying the differentially expressed genes at each timepoint based on a common threshold of significance. The extracted information is further explored to obtain insight about the regulation of gene networks. No similar type of time-series data analysis exists currently in the literature. Specifically, I present a systematic methodology that allows for (a) deducing and appropriately storing the individual gene and gene class variability in significance level with time, and (b) comparing genes, gene classes and time points based on (a). The derived information is expected to unravel significant characteristics of a biological system's dynamic response to particular perturbation(s). This is demonstrated in chapter 5 and 6 of this thesis. The applicability of the proposed algorithm and subsequent data analysis methodology is not limited to transcriptomic data, but they could be accordingly applied to time-series high-throughput biological data of any other type (e.g. proteomic or metabolomic), as it has been demonstrated in (Dutta et al., 2007).

3.2 Proposed Algorithms

3.2.1 SAM-based algorithm for the identification of differentially expressed genes at each timepoint

SAM identifies the genes that are differentially expressed between two experimental groups based on whether the difference between a gene's *observed* (d(i)) and *expected* ($d_e(i)$) "relative differences" is greater than a significance threshold 'delta' (Tusher et al., 2001). Paired-SAM, in particular, deals with the analyses in which the samples of the two experimental groups can be paired according to the experimental design, time-series analyses being a characteristic example. In these cases, the "per pair" information is used in the estimation of the relative differences d(i) and $d_e(i)$. Specifically, d(i) is defined as follows:

$$d(i) = \frac{(\overline{X}_1(i) - \overline{X}_2(i))}{S(i) + S_o} \tag{3.1}$$

where: $\overline{X}_k(i)$ represents the mean expression of gene i in experimental group k (k = 1 or 2); S(i) represents the standard deviation of the per pair differences in expression of gene i between the two experimental groups; and S_o depicts a positive fudge factor used to eliminate numerical biases at low values of S(i). The observed scores are ranked in decreasing order and d(i) corresponds to the i-th ranked gene of the distribution. $d_e(i)$ is also estimated from Equation 3.1, but in this case the samples are multiple times divided into two groups of the same size as the original by random sampling permutations. For each permutation, the calculated based on Equation 3.1 gene scores are also ranked in decreasing order. The average of the scores in the i-th position among all permutations is considered as $d_e(i)$. Finally, the two distributions are plotted in a quantile-quantile plot and the genes whose absolute difference between the observed and the expected scores is larger than delta are identified as differentially expressed. FDR is estimated based on two cuttoffs defined by the minimum and maximum (least negative) d(i) values from the cluster of positively and negatively significant genes, respectively (Tusher *et al.*, 2001). For each permutation of the expected distribution, the number of genes "laying outside" the cutoff region is determined; the median of this number over all permutations is multiplied by a correction factor to estimate FDR.

Hence, in the case of time-series analysis, SAM identifies as differentially expressed the genes whose average over time expression has changed due to the applied perturbation to a greater than delta extent than what it would have been anticipated due to random differences among samples, the latter being quantified by the relative expected difference distribution, d_e. In the context of this analysis, the expression of a gene at any time point is represented by its average expression over all sampled biological and experimental replicates at this time point. Following the same concept, I define as differentially expressed at a particular time point the genes whose expression at this time point has changed due to the applied perturbation to a greater than delta extent than what it would have been expected based on the null distribution d_e. In this way, I use the same reference distribution of expected gene expression differences and the same significance threshold delta for all time points, taking inherently into consideration that they are part of the same time sequence. Specifically, the "time-dependent" statistic that is proposed for the new algorithm is the observed score of gene i at a particular time point t, which is defined based on Equation 3.1 as follows:

$$d_t(i) = \frac{(X_1{}^t(i) - X_2{}^t(i))}{S(i) + S_0} \qquad (3.2)$$

where: $(X_1{}^t(i) - X_2{}^t(i))$ represents the difference in the expression of gene i between the two experimental groups at timepoint t; the rest of the symbols represent the same quantities as in Equation 3.1. For each timepoint, the distribution of observed scores is separately ranked in the decreasing order and $d_t(i)$ represents the observed score of the i-th ranked gene at the t-th timepoint. At a particular time point t, gene i is identified as differentially expressed, if the absolute difference between its observed score at this time point and the i-th expected score is larger than delta (see schematic diagram in Figure 3.2).

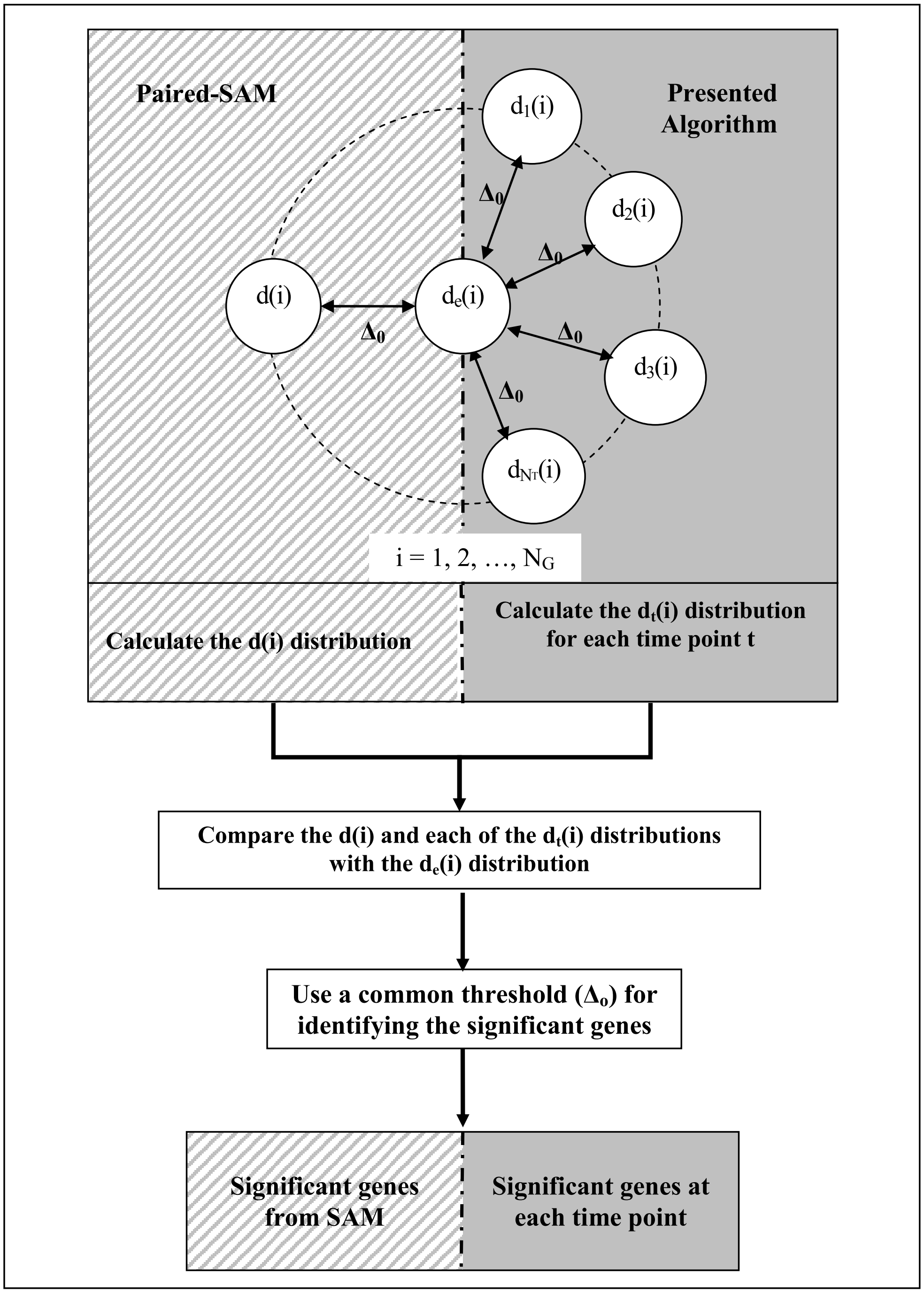

Paired-SAM
Presented Algorithm
$d_1(i)$
$d_2(i)$
$d_3(i)$
$d_{N_T}(i)$
$d(i)$
$d_e(i)$
Δ_0
Δ_0
Δ_0
Δ_0
Δ_0
$i = 1, 2, \ldots, N_G$
Calculate the d(i) distribution
Calculate the $d_t(i)$ distribution for each time point t
Compare the d(i) and each of the $d_t(i)$ distributions with the $d_e(i)$ distribution
Use a common threshold (Δ_o) for identifying the significant genes
Significant genes from SAM
Significant genes at each time point

Figure 3.2 - Schematic representation of the presented time-dependent modified SAM algorithm. d(i), $d_e(i)$, $d_t(i)$, N_T, N_G depict, respectively, the observed relative difference of gene i based on the SAM definition as described in Equation 3.1, the expected relative difference of gene i based on the SAM definition as described in the text, the observed relative difference of gene i at time point t according to the proposed algorithm as described in Equation 3.2, the total number of time points and the total number of gene expressions included in the significance analysis.

There have been concerns regarding the use of the same expected distribution for each time-point and paired-SAM and to what extent this could lead to high number of false positives (personal communication). Regarding the first concern, the I support that application of the presented methodology enables the comparison (a) between time-points, since they are members of the same time-series, and (b) of each time-point with paired-SAM. Specifically, it enables the identification of the genes whose observed expression at a particular time-point is larger in absolute value than its expected to an extent higher than the threshold delta, the expected value being estimated from permutations of all time-point samples. If the average over time observed expression of a particular gene has this characteristic, this gene's expression is considered as changing significantly between the two time-series groups of samples. In this way, application of the presented algorithm enables the comparison between the time-profile of- with respect to the average over time- significance level of a particular gene. Ability to carry out this comparison is important in identifying biologically relevant conclusions both with respect to the experimental design and the selected time intervals, but mainly regarding the dynamic behaviour of biological processes.

To increase the confidence in the time-point significant genes, a more stringent Bonferroni-like (Bland et al., 2003) corrective algorithm could also be applied (provided as option to the user in the accompanying software). Specifically, if F, F^1_0, F^2_0, …, F^n_0 are the %FDR(median) of paired-SAM and of the significance analyses at time point 1, 2, …, n, respectively, for a particular delta value, Δ_0, as presented earlier, this is the starting point of the iterative corrective algorithm. Let us define F'$_j$ at the j-th iteration of the corrective algorithm, j being a nonnegative integer, as follows:

$$F'_j = 1 - \prod_{i=1}^{n} (1 - F^i{}_j) \tag{3.3}$$

The criterion for the termination of the iterative corrective algorithm is for F'$_j$ to become equal to F (Figure 3.3). Thus, the corrective algorithm involves carrying out the significance analyses at each timepoint iteratively, based on increasing delta value at each iteration, until the termination criterion is satisfied (Figure 3.3). The relative increment in delta at the jth iteration is proportional to the difference between F'$_{j-1}$ and F as follows:

$$\frac{\Delta_j}{\Delta_{j-1}} = 1 + c(F'_j - F) \tag{3.4}$$

where c is the proportionality constant (0.1. has been identified as optimized default value). The delta value used at the last iteration being larger than Δ_0, this process will certainly result in smaller number of significant genes and lower %FDR(median) at each time point [see results in chapter 5 and 6]. The need or not of the Bonferroni-type correction can be evaluated each time in the context of the particular biological dataset and the biologically relevant conclusions that the inclusion of this correction could provide. Finally, after applying the new “time-dependent” significance analysis algorithm, with or without correction, at each time point each gene will belong to a particular significance level. The latter might be different from the significance level in which the gene is classified based on paired-SAM. In subsequent section, we will discuss some characteristic cases of this difference. The results of the new algorithm could be stored in a matrix, which we accordingly called “time-dependent significance matrix” (TDSM). TDSM has as many rows as the number of genes (N_G) and as many columns as the number of time points (N_T). The (i,j)-th element of TDSM is equal to +1, 0, -1, depending on whether the i-th gene’s change in expression between the two experimental groups at time point j has been, respectively, identified as positively, non or negatively significant. “Augmented” TDSM (A-TDSM) comprises one additional column that corresponds to the significance level of the genes based on paired-SAM.

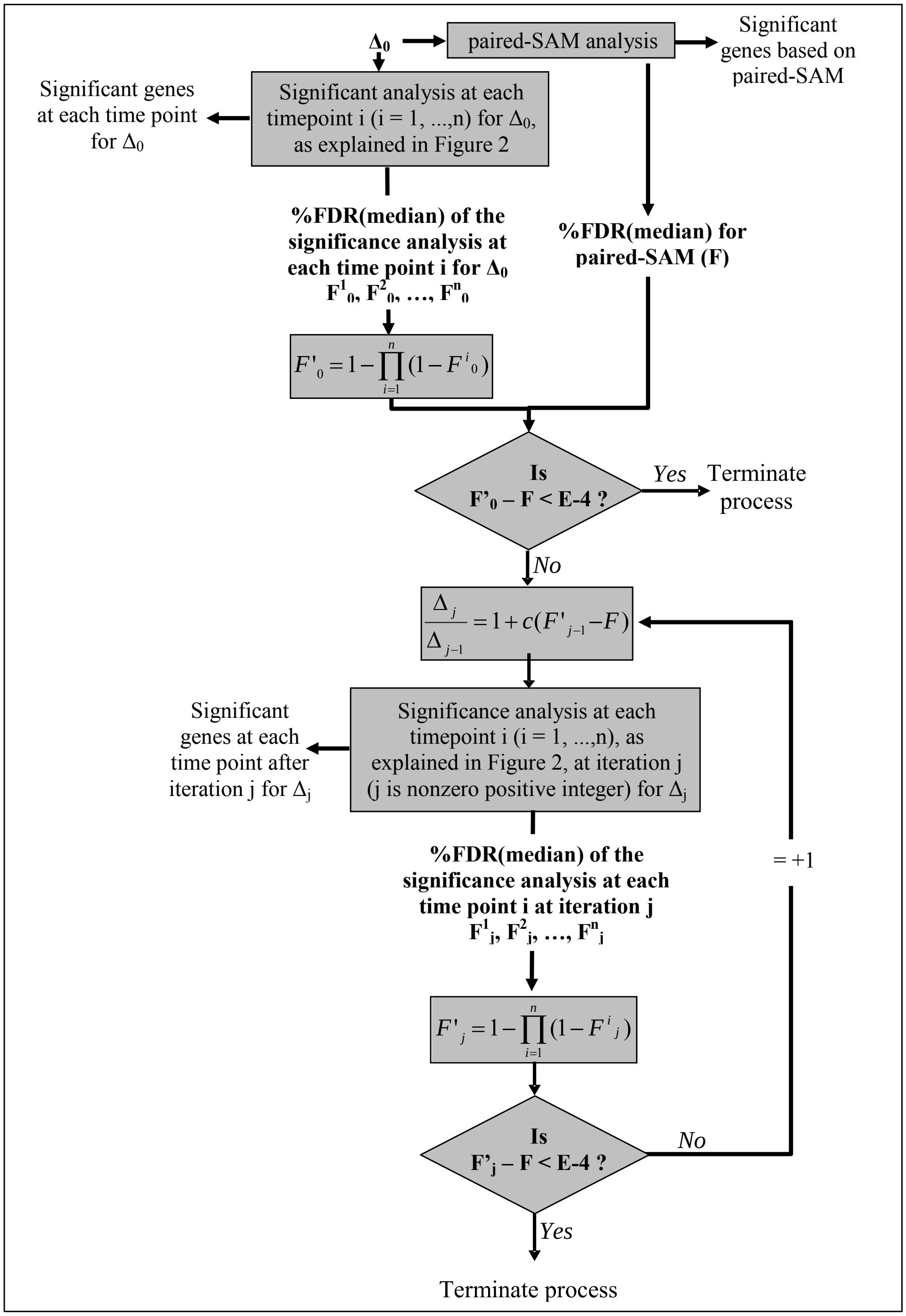

Δ_0
paired-SAM analysis
Significant genes based on paired-SAM
Significant genes at each time point for Δ_0
Significant analysis at each timepoint i (i = 1, ...,n) for Δ_0, as explained in Figure 2
%FDR(median) of the significance analysis at each time point i for Δ_0 $F^1{}_0, F^2{}_0, \ldots, F^n{}_0$
%FDR(median) for paired-SAM (F)
$F'_0 = 1 - \prod_{i=1}^{n} (1 - F^i{}_0)$
Is $F'_0 - F <$ E-4 ?
Yes
Terminate process
No
$\frac{\Delta_j}{\Delta_{j-1}} = 1 + c(F'_{j-1} - F)$
Significant genes at each time point after iteration j for Δ_j
Significance analysis at each timepoint i (i = 1, ...,n), as explained in Figure 2, at iteration j (j is nonzero positive integer) for Δ_j
%FDR(median) of the significance analysis at each time point i at iteration j $F^1{}_j, F^2{}_j, \ldots, F^n{}_j$
= +1
$F'_j = 1 - \prod_{i=1}^{n} (1 - F^i{}_j)$
Is $F'_j - F <$ E-4 ?
No
Yes
Terminate process

Figure 3.3 Schematic representation of the iterative corrective algorithm for the significance analyses at each time point. Δ_0, Δ_j, $F^1{}_0$, $F^2{}_0$,…, $F^n{}_0$, $F^1{}_j$, $F^2{}_j$,…, F^n depict the initial delta value, the delta value at the j-th iteration (where j a nonzero positive integer), the %FDR(median) of the significance analysis at time point 1, 2, …, n (where n the total number of time points) based on Δ_0, the %FDR(median) of the significance analysis at time point 1, 2, …, n (where n the total number of time points) based on Δ_j, respectively. The proportionality constant ‘c’ determines the rate of convergence; 0.1. has been identified as optimized default value.

Clearly, if statistical significance is related to biological significance, the information in TDSM or A-TDSM could be the basis for data mining to extract time-dependent biologically relevant conclusions, which would have been otherwise missed. Obviously, this is true, independent of the algorithm by which the information in TDSM, i.e. the significance level of each gene at each time point, might have been derived. An obvious data mining exploration of TDSM data would be the clustering of the genes based on their significance level profile over time. In the next sections, we present a series of methods that allow for further use of the information in TDSM towards the extraction of significant biological conclusions.

3.2.2 Algorithms for exploring gene variability in significance level over time to extract biologically relevant conclusions

A. Significance Variability Score

The information in TDSM could be used to rank the genes based on their variability in significance level over time. For this to become possible, the genes' "significance variability" (SV) score needs to be estimated; we propose the following algorithm:

1. Use TDSM to construct the "Significance Variability Matrix" (SVM). SVM should have as many rows as the number of genes (N_G) and columns by one fewer than the number of time points (N_T-1). The elements of SVM are estimated to reflect the number of "significance levels" that a gene ascends or

descends from one time point to the next. Specifically, for i = 1, 2, …, N_G and j = 2, …, N_T:

$$SVM[i, (j-1)] = \left| TDSM[i, j] - TDSM[i, (j-1)] \right| \tag{3.5}$$

Clearly, the genes could be also clustered based on their SVM profile. The genes clustering together would have similar dynamic significance profile. In this case, genes remaining in the same significance level over time would be clustered together independent of the significance level; the same for genes, which follow the opposite significance level profile with time, if an absolute distance metric is used. An easy way to determine the number of the genes in these clusters and focus on a particular cluster of interest is the estimation of their "Significance Variability" (SV) score as indicated below.

2. Estimate the SV score of the i-th gene as the average of the i-th SVM row's elements:

$$SV_i = \frac{\sum_{j=1}^{N_T - 1} SVM[i, j]}{N_T - 1} \tag{3.6}$$

Based on its definition, the SV score could range from 0 to 2. The distribution of the genes with respect to their SV score might reveal significant information about the biological problem under investigation. For example, the genes whose SV score is equal to 2 "fluctuate" between the positively and negatively significant levels from one time-point to the next throughout the entire time period. Determination of the number and type of these genes could prove significant for understanding the response of the biological system to the investigated perturbation, but also for correctly selecting the time points in future experiments to capture subtler changes in gene expression. On the other hand, the genes whose SV score is equal to zero belong to the same significance level at all examined time points. These are the genes whose expression was significantly affected (either positively or negatively) or remained (statistically) unaffected by the investigated biological perturbation. Obviously, the distribution of all genes around these two numbers (0, 2) will give simple, but strong, indications regarding the transcriptional response of the system to the examined perturbation over time. Paired-SAM results are expected to have stronger similarity to the results of the presented algorithm the more the genes with SV score closer to zero are.

B. New metric for time point correlation

The change in the physiology of a biological system due to a particular perturbation at two different time points could be initially compared with respect to the number of genes in each significance level. However, two time points could correspond to the same number of genes in all three significance categories, but still not be biologically correlated, because each category comprises different genes at each time point. Therefore, another metric that takes into consideration the number of *common* genes in each of the significance categories should be defined. Of note, the same correlation metric could also be used if, instead of time points, two experimental conditions or two biological perturbations are to be compared.

We defined "Significance correlation matrix" (SCM) with respect to positively, negatively or non-significant genes the N_T x N_T symmetric matrix, whose elements are estimated as follows:

$$SCM_k\,[\,i\,,j\,] \;=\; \begin{cases} \dfrac{G_k{}^{i} \cap\; G_k{}^{j}}{\sqrt{G_k{}^{i} \cdot G_k{}^{j}}} & \textit{for } i \neq j \\[2ex] \dfrac{\overline{G_k{}^{i}}}{G_k{}^{i}} & \textit{for } i = j \end{cases} \tag{3.7}$$

where, k depicts the significance level with respect to which the time point comparison is performed (for example, k = P, N, O or P∩N, if the comparison is made with respect to the positively, negatively, non or the union of positively and negatively significant, respectively, genes); $G_k{}^{\ell}$ depicts the number of genes in the k-th significance level at the $\ell -$th timepoint, $\ell = 1,2,\ldots,$$N_T$; $\overline{G_k{}^{\ell}}$ depicts the number of genes in the k-th significance level *only* at the $\ell -$th timepoint (i.e $\overline{G_k{}^{\ell}} \cap\; G_k{}^{q} = \mathbf{0} \;\; \forall\; q \neq \ell\,,$ q = 1, 2, …, N_T). By definition, the elements of a SCM may take values between 0 and 1. Two time points might be considered strongly correlated if the corresponding SCM element(s) is(are) larger than a certain value-threshold, usually larger than 0.5. In addition, large diagonal element implies that at this time point the response of the system to the particular perturbation(s) is largely different than at the rest.

C. Gene Class Comparison

If (a) particular gene class(es) is(are) of interest, then the matrices described in the above sections should be constructed to contain only the gene set associated with this(these) gene class(es); the same analytical methodologies described above could be used to extract biologically relevant conclusions focused only on this(these) gene class(es).

In order to identify the gene class(es) that are highly enriched in significant genes hypergeometric distribution could be used as follows: let us suppose that the total number of genes used in the analysis is N and among these n genes are significant at a particular timepoint t. In addition, let us assume that among the y genes that have been associated with a particular gene class (among all N genes), x have been identified as significant at timepoint t. For the null hypothesis $\mathbf{H_0}$: gene class *i* is not significantly enriched and alternate hypothesis $\mathbf{H_1}$: gene class *i* is significantly enriched, the p-value can be computed in the following way

$$p = \frac{\sum_{i=x}^{y} {}^{y}C_{i}\,{}^{(N-y)}C_{(n-i)}}{{}^{N}C_{n}} \tag{3.8}$$

Where ${}^{a}C_{b}$ represents number of ways we can select “b” elements out of “a” without replacement. If $p < 0.05$, then gene class *i* is significantly enriched.

Specifically, matrices corresponding to each (or to the union of more than one) of the significance levels could be formed; each of the matrices will have as many columns as the number of the sampled time points and as many rows as the number of gene class(es) that are to be investigated (in a high-throughput unsupervised way, the latter could be all the gene class(es) that are associated with the gene list under investigation). The [i,j]-th element of a particular significance level’s matrix will be equal to the *p value* of the i-th gene class corresponding to j-th time timepoint. Studying the information in these matrices, it would be possible to answer a variety of questions regarding the response of the various gene class(es) to the applied perturbation based on their significance level profile over time.

Another way to identify the gene class(es) highly enriched in significant genes would be to construct matrices of number of rows and columns equal to number of gene class(es) and sampled timepoints respectively, whose [i,j]-th element of a particular significance level's matrix will be equal to the percentage of the i-th gene class that has been classified in the particular significance level at the j-th time point. Analyzing these matrices, for example, it would be possible to identify all gene class(es) whose more than 50% of the genes have been consistently classified as (positively or negatively) significant at each time point.

MiTimeS software suite:

Algorithms proposed here were implemented in a software suite called MiTimeS written in C language and compatible to both windows and Macintosh computers. This software is free for all the academic users and the executable files can be obtained by requesting me or Prof. Klapa.

The software has 4 modules corresponding to four main features of the algorithm. Following are the modules and their description:

1. <u>DEGenes:</u> Calculates the list of significant genes and FDR for each timepoint. Tree files continuing the list of positively, negatively and non-significant genes are created corresponding to each timepoint.

2. ExpressionChanges: from the output files of DEGenes program TDSM and SVM matrices are computed.

3. TimeCorr: creates the SCM matrices for all the three significant categories

4. GOComp: creates GO comparison table based on algorithm explained above.

3.3 Conclusions

In light of the importance of time-series transcriptional profiling analysis to derive conclusions regarding a biological system's regulation, we developed an algorithm based on SAM principles that enables the identification of differentially expressed genes at each time point of a sampled time sequence using a common reference distribution and significance threshold for all timepoints. This algorithm enables the direct comparison between the different phases of a time-dependent process. In this chapter, I also presented three additional algorithms that assist in further exploring the results of the initial method regarding the gene variability in significance level with time. All four proposed algorithms, programmed in the form of executable files under the overall name MiTimeS, provides a platform for the significance analysis of time series transcriptomic (or any other high-throughput biological) data that could lead to biologically relevant conclusions, which would have not been easily reachable otherwise.

The software suite was used for analysis of time-series transcriptional profiling data obtained from this project. It was used for each pair-wise comparison in concert with paired SAM analysis. The results obtained were analyzed and explained in detail in the contest of plant physiology [please see chapters 5 and 6]. Results revealed wealth of biological information unobserved before due to non-existence of the presented algorithm.

4 EXPERIMENTAL DESIGN AND SETUP

4.1 Experimental Design

Arabidopsis thaliana liquid culture system was subjected to various environmental stresses applied individually or in combination. This will not only reveal the response of the plant to the specific stresses applied but also provides us a framework for comparing the different stresses that are applied. Application of multiple stresses is believed to reveal the group of genes that are differentially expressed under all the stresses. These genes are believed to play an important role in gene regulation network. If it is found that some unknown gene is always clustering with a group of known genes in different perturbations, then this information might help us to assign the functionality of that unknown gene. Based on our previous findings [Dutta B, 2004] and other studies following stresses were applied:

4.1.1 Elevated CO_2

Elevated CO_2 stress was found affect the carbon fixation, central carbon metabolism and amino acid bio-synthesis at within short period of time which can be observed from transcriptional profiling. The results obtained can be compared with literature as the effect of elevated CO_2 stress is well studied at genomic and metabolomic level for most of the metabolic pathways. Elevated CO_2 stress was applied individually to plants grown in sucrose and glucose media and in combination with trehalose and NaCl stress for the plants grown in sucrose media.

4.1.2 Trehalose Stress

Previous studies have shown that trehalose plays an important role in carbohydrate utilization and plant growth [Moore el al., 2003; Wingler el al., 2000]. Though the exact role of trehalose pathway is not yet elucidated, but previous studies have shown that expression of genes encoding trehalase and trehalose-6-phosphate phosphatase gets affected at elevated CO_2. Trehalose stress of 12mM will be chosen as it is believed [Moore el al., 2003] to create an observable response at the transcriptional level.

4.1.3 Salt Stress

Most of the organisms respond to the osmotic stress at genomic and metabolic level [Verala et al., 2003; Taiz and Zeiger, 2002]. It will be interesting to study how the gene expression of *A. thaliana* gets affected in short term by salt stress. 50mM salt stress was applied as it is believed [Essah et al., 2003] to create enough stress that can be observed at genomic level but plants would be able to sustain it.

4.1.4 Combined Stresses

Two combined stress experiments were carried out where CO_2 stress was applied in conjunction with NaCl and trehalose stress. In the combined stress, then strengths of the stresses applied were same as that of individual ones. The objective of carrying out the combined stress experiment is to compare then with respect to individual stresses and to study if the combined response is constitute of the individual responses.

Figure 4.1 shows the experimental design in detail. Each of the rectangles represents an experiment that was conducted. Each rectangle is divided into two parts; the upper part corresponds to air composition and the lower to growth media condition. If a compound is added to the media it is represented as a separate rectangle.

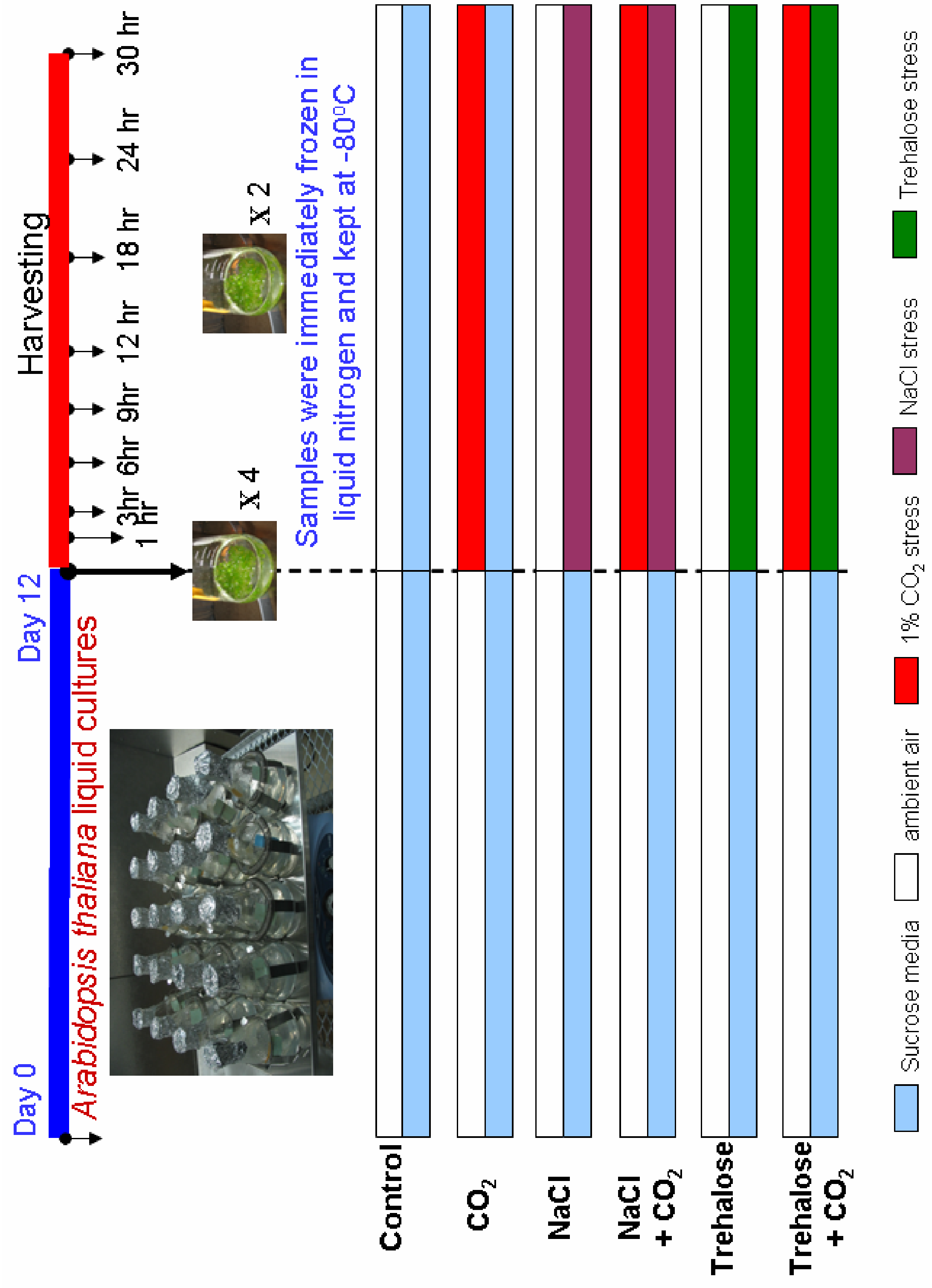


Figure 4.1: The figure shows the experimental design and setup. The stresses applied at 6 different experiments are shown by colors of two rectangles. The top rectangle shows the air composition (white – ambient air, red – elevated CO_2) and the bottom rectangle represent the stress applied in the media (blue – no stress, purple – NaCl stress and green – trehalose stress).

Specifically, *A. thaliana* (Columbia Strain) plants were grown in shake flasks in a growth chamber (model M-40, EGC Inc., Chagrin Falls, OH) for 12 days under constant light intensity (80 - 100 μE m^{-2} s^{-2}) and $23^{O}C$ at ambient air condition. At the beginning of 13th day following stresses was applied:

- In experiments 2, 4, 6 and 8 CO_2 concentrations were increased to 1% from 0.035%.
- In experiment 3 and 4 10ml of 240mM trehalose solution was added to the media.
- In experiment 5 and 6 10ml of 1M NaCl solution was added.

Short term dynamic response of the *A. thaliana* system was studied to see how the gene expression changes with time in first 30 hours of the applied stress. From the previous studies [Dutta B, 2004] it was observed that plants respond to environmental stresses at the transcriptional level in this time scale. 4 plant cultures were harvested at the beginning of the experiment (0hr) and 2 plants at each of the time points 1hr, 3hr, 6hr, 9hr, 12hr, 18hr, 24hr and 30hr (a total of 20 plants). Time points are selected such that they are mostly equally spaced, but it is also possible to observe the initial response of the plant to the applied stresses.

Following the protocols discussed in the “Acquiring DNA Microarray Data” section of “General Methodologies” the time series transcriptional profiles for all 8 experiments will be obtained. A pool consisting of equal amount of mRNA from all the samples of experiment 1 and 2 will be used as common reference for all the hybridizations. This reference will provide us a common platform for the comparison of individual and combined stresses.

To understand the size of the experimental dataset to be provided from this experiment and the effort invested in it, below is a summary of all steps.

- 8 liquid cultures
- 20x8 = 160 total RNA extraction cycles
- 20x8 = 160 mRNA amplification
- 4x(20x8) = 640 cDNA synthesis
- 2x(20x8) = 320 microarray hybridizations.
- Image processing of 4x(20x8) files using TIGR Spotfinder
- Normalization of 2x(20x8) = 320 TAV files using TIGR MIDAS
- 8 full genome profiles of 8 time points each; each time point corresponds to the average profile of 2 replicates, while time 0 corresponds to 4 replicates.

4.2 Materials and Methods

4.2.1 Selection of plant liquid cultures as model system

The plant cultures grew in 500 ml shake flasks, each containing 200 ml B5 Gamborg media [Gamborg et al., 1976] with minimal organics (Sigma, St. Louis), 2% (w/v) sucrose (or glucose) and 0.1% agar. Agar is also added to increase the viscosity of the liquid media and consequently the support of the plants, permitting there by the growth of the seeds in the liquid media. Liquid cultures were grown for 12 days on an orbital shaker platform (Barnstead, IL) at 150 rpm, in the ambient air (350ppm CO_2).

4.2.3 Seed preparation and inoculation

A. thaliana Columbia strain seeds were washed and stored at 4°C for 24 hours covered with aluminum foils. Seeds were added in agar solution and inoculated in 200ml of autoclaved media. Each flask contained around 80-100 seeds.

4.2.4 Experimental setup

At the end of 12th day 4 liquid cultures were harvested from each of the experimental set. These liquid cultures will serve as pretreatment control. Immediately after the plants were harvested, one set continued to grow under same conditions (experiment 1) which will be compared as "control experiment" in the rest of the thesis. However, environmental perturbations were applied to the remaining set of experiments in the following way:

2. CO_2 concentration in ambient air composition was increased to 10,000ppm. The CO_2 concentration increase in the perturbed system's growth chamber was achieved in 5min [WMA-4 CO_2 Analyzer, PP Systems, Amesbury, MA].
3. Trehalose concentration in the growth media was increased by adding 10 ml of 240 mM trehalose solution to the media, so that the trehalose concentration in each flask reaches 12mM. The solution was added in each flask separately and to minimize the contamination injection syringe was used.
4. 10ml of 240mM trehalose solution was added along with elevated CO_2 concentration of 10,000 ppm.
5. Salt stress was applied in the growth media by adding 10 ml of 1 M NaCl solution in each flask, which makes the NaCl concentration in each flask

50mM. NaCl solution was also added with an injection syringe to minimize contamination.

6. 10ml of 1M NaCl solution was added along with elevated CO_2 concentration of 10,000 ppm.

4.2.5 DNA microarray hybridization and data acquisition

Slide Preparation: *Arabidopsis thaliana* genomic DNA amplicon microarrays were constructed as described previously (Kim et al., 2003). Briefly, using the TIGR *Arabidopsis* genome release 2.0, genomic DNA fragments representing the predicted 3'-ends of the 26,777 protein-encoding nuclear, plastid or mitochondrial genes were amplified by PCR. The PCR amplicons were purified and resuspended in 50% DMSO and printed onto UltraGAPS aminosaline-coated slides (Corning Inc, Corning, NY) using an Intelligent Automation System (IAS) arrayer (Cambridge, MA). After printing, the spotted DNA was cross-linked to the slide surface by UV irradiation at an integrated intensity of 120 mJ cm^{-2} using a Stratalinker UV Crosslinker (Stratagene, La Jolla, CA) and slides were stored in a desiccated chamber until used. Functional annotations for the arrayed elements are from TIGR *Arabidopsis* genome release 5.0 (http://www.tigr.org/tdb/e2k1/ath1/) and the current annotation can be downloaded from the Plant RESOURCERER database at (http://www.tigr.org/tigr-scripts/magic/p1.pl).

Total RNA extraction and mRNA amplification: Total RNA was extracted from the ground plant samples using trizol [see detailed protocol http://atarrays.tigr.org/arabprotocols.shtml]. From the total RNA extracted, mRNA was selectively amplified through cDNA synthesis using reverse transcriptase [http://atarrays.tigr.org/arabprotocols.shtml].

Hybridization and Scanning: Probe labeling and hybridization protocols were described previously (Kim et al., 2003), and are available in detail at http://atarrays.tigr.org . Starting with 1 µg of poly(A)-enriched mRNA, single-stranded cDNAs were synthesized during reverse transcription reaction using random hexamer primers (Invitrogen, Carlsbad, CA) in the presence of aminoallyl-dUTP (aa-dUTP; Sigma, St. Louis, MO). Following the removal of unincorporated aa-dUTP and dNTPs using Microcon YM-30 columns (Millipore, Bedford, MA), the reaction products were conjugated to either Cy3 or Cy5 NHS-ester fluorescent dye (Amersham-Pharmacia, Piscataway, NJ). The Cy3- and Cy5-labeled probes were further purified using Qiaquick PCR Purification Kit (Qiagen, Valencia, CA), combined as an appropriate pair, and lyophilized.

Gene expression levels were measured using a reference design for microarray analysis in which each test sample was hybridized to a common reference created by pooling equal quantities of poly(A) RNA from every experimental and control sample. All experimental and control mRNA samples were labeled and compared to the labeled reference pool RNA in co-hybridization assays and all hybridizations were repeated using a dye-reversal replication (in which the use of Cy3 and Cy5 dyes were switched between experimental/control and reference RNAs) approach to compensate for any potential dye-specific biases.

Slides were pre-hybridized in 1% bovine serum albumin (BSA) in 5×SSC, 0.1% SDS for 45 minutes at 42ºC, followed by several washes in water and isopropanol, and then dried by centrifugation. The labeled probes were resuspended in hybridization buffer containing 50% formamide, 5× SSC, 0.1% SDS and 0.2 µg/µl salmon sperm DNA and hybridized to the microarray slide at 42ºC for 16 – 20 hours in a sealed, humidified chamber. Following hybridization, slides were sequentially washed once in 2×SSC and 0.1% SDS for 4 minutes at 42ºC, once in 0.1×SSC and 0.1% SDS for 4 minutes at room temperature, and twice in 0.1×SSC for 4 minutes at room temperature, and then dried by centrifugation. Slides were scanned using an Axon 4000B microarray scanner (Axon Instruments, Union City, CA), and data were saved as two independent 16-bit TIFF files corresponding to the Cy3 and Cy5 channels, respectively. Therefore the relative intensity of the same spot between the two scanned images provides a measure of the relative amount of mRNA between the query and reference samples.

The protocols of total RNA extraction, RNA amplification, dye coupling and hybridization are described at *Arabidopsis* functional genomics webpage [http://atarrays.tigr.org/].

4.2.6 Gene and metabolic pathway databases

The TIGR *A. thaliana* annotation database, regularly updated with new annotation (please check *http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml*) will be used for assigning function to the observed genes. Other public databases, i.e. metabolic pathway database KEGG (*www.kegg.com*) and ExPasy (*www.expasy.org*) will be consulted to study the gene expression profiles in the context of metabolic application. The regularly application categorization of gene functions based on gene ontologies (www.geneontology.org) will also be used to cluster genes based on a particular property.

4.2.7 Data normalization and multivariate statistical analysis

TIGR TM4: This is an open source software package (www.tm4.org) [Saeed et al., 2003] for DNA microarray data processing and analysis. The different software(s) included in TM4 are to be used in our analysis are listed below:

1. MicroArray Data Manager [MADAM] (data storage)
2. TIGR Spotfinder (image processing)
3. TIGR Microarray Data Analysis System (MIDAS) (normalization)
4. TIGR MultiExperiment Viewer [MeV] (clustering)

Image Processing: The raw intensity data were extracted from the two TIFF images using TIGR Spotfinder [V2.2.1_NoDB] [Saeed et al., 2003], and data points were not considered for further analysis if a spot was flagged during data acquisition as saturated or non-detectable at either channel, or if greater than 50% of the pixels within the spot were less than the median plus one standard deviation of background intensity. For the remaining spots, the raw signal intensity was reported as the mean spot intensity minus the median background intensity.

Normalization: Microarray data normalization is necessary, because it eliminates the systematic biases of the DNA microarray data acquiring process [Quackenbush, 2000]. Errors in DNA microarray data could be originated from:

- Unequal quantities of starting mRNA among the query and the reference samples.
- Difference between the labeling efficiencies of the Cy3 and Cy5 dyes because of preferential binding of one of the dyes to the samples.
- Difference in the scanner sensitivity between the two dyes lead to more effective detection of one of the dyes.
- Variation of spot intensity across the slide due to variation among the pins used for slide printing.

To eliminate the above mentioned errors various normalization methods have been proposed, the most eminent ones are listed below:

- *Total intensity normalization:* compares the sum of the intensities of all the spots of one channel with that of the other. If they are not equal then the intensity of all the spots in one of the channels are scaled up or down accordingly [Quackenbush, 2002].
- *Lowess:* The ratio of the spot intensity values in the two channels has been observed to have a systematic dependence on intensity [Yang Y. et al., 2002 and Yang I. et al., 2002]. This dependence is most commonly exhibited as a deviation from 1 for the low or high intensity spots. Lowess normalization accounts for this bias and scales the data accordingly.
- *Standard Deviation Regularization:* in a spotted array, the variation among the sets of pins can result into variation in intensities among the meta-blocks. Standard deviation regularization scales the data in such a

way that they have the same variation across the slide [Yang Y. et al., 2002].

- *Flip-dye normalization:* Comparison between the flip-dyes slides can provide information about the biases due to the difference between the two dyes [Quackenbush, 2002]. Flip-dye normalization scales data accordingly to eliminate such biases.

The normalization involved locally weighted scatterplot smoothing regression (LOWESS) (smooth parameter: 0.33; reference: Cy3), variance regularization (reference: Cy3) and “flip-dye” data consistency trim (data trim option: SD cut; cross log ratio data keep range: +/- 2SD).

Outlier detection and data preparation:

Gene expression profiles of the samples (corresponding to each flask) we clustered using hierarchical clustering. Samples that have distinctly different physiological state will also appear as outlier from clustering. Weight of the plants harvested is also a measure of their physiological and growth state. Samples that are outliers based on gene expression, metabolomic data as well as from weight were excluded from further analysis. Subsequently, the control and perturbed expression of each gene at each time point was estimated as the geometric mean of its expression in all control and perturbed, respectively, biological replicates harvested at this timepoint. Finally, the control and perturbed timepoint expressions of each gene were divided with the control and perturbed, respectively, 0h expression of this gene.

Significance analysis using paired SAM: Normalized data was used for clustering and multivariate statistical analysis. Paired SAM implemented in TIGR MeV software was used for pair-wise comparisons of any two experimental data sets. Paired SAM is a non-parametric hypothesis testing methodology which allows user to change the level of significance conveniently. It also provides a measure of false discovery rate (FDR), i.e. number of genes found significant by chance. For an effective "comparison of the comparisons" from different stress responses a common level of significance should be used. Environmental stress levels are very different and so are the data sets. For all the individual pair-wise comparisons a significance level was chosen such that it has maximum number of significant genes with minimum FDR. FDR is a non-negative number; hence the threshold value selected had maximum number of differentially expressed genes with FDR 0.

Significance Analysis at individual timepoints:

MiTimeS software developed in our lab for significance analysis of microarray time-series data was used for the extraction of time-dependent information for the data. Expected distributions obtained from the permutation were saved from TIGR MeV software and was used for MiTimeS analysis. In the first module of the software where list of significant genes at each timepoints is calculated, delta value provided was same as that of paired SAM. The program calculates the combined FDR based on all the timepoints and if its significantly greater than the FDR obtained from SAM (0 in this case), internally a new delta is calculated until the program converges. Subsequent module of the MiTimeS software was also used for identification of TDSM, SVM and SCM matrices, which serves as basis for studying significance analysis results in the context of metabolic pathways. Results obtained form these analyses are explained in detail in the following section.

5 ANALYSIS OF THE PLANT TRANSCRIPTIONAL RESPONSE TO EACH APPLIED STRESS

This chapter includes the results from all the experiments explained before. Figure 5.1 shows the schematic diagram of the experimental design, where each colored circle represents an experiment and an arrow connecting two circles signifies a comparison between them. The unique color of each circle provides the color convention to be used to represent a particular experiment, in this and the following chapter. The metric representation of the circles signifies the stress(es) that is(are) applied to that experiment. Each row and column shows the environmental perturbation applied to the media and the ambient air composition respectively. Following are the notation used in the figure 5.1 and in subsequent part of the thesis:

SC: Control Experiment (no stress)
NC: NaCl stress experiment
TC: trehalose stress experiment
SP: CO_2 stress experiment
NP: NaCl and CO_2 stress experiment
TP: trehalose and CO_2 stress experiment

Different experiments were compared using hypothesis testing techniques. When experiment Y is compared with respect to experiment X, the results are shown as X_Y. The arrows with the same color imply the same stress effect. Continuous arrows signify the stress response with respect to the control state, whereas the dotted arrows imply the comparison between single and multiple stress conditions. There are 5 possible stress response analysis with respect to the control sate and 4 between stresses.

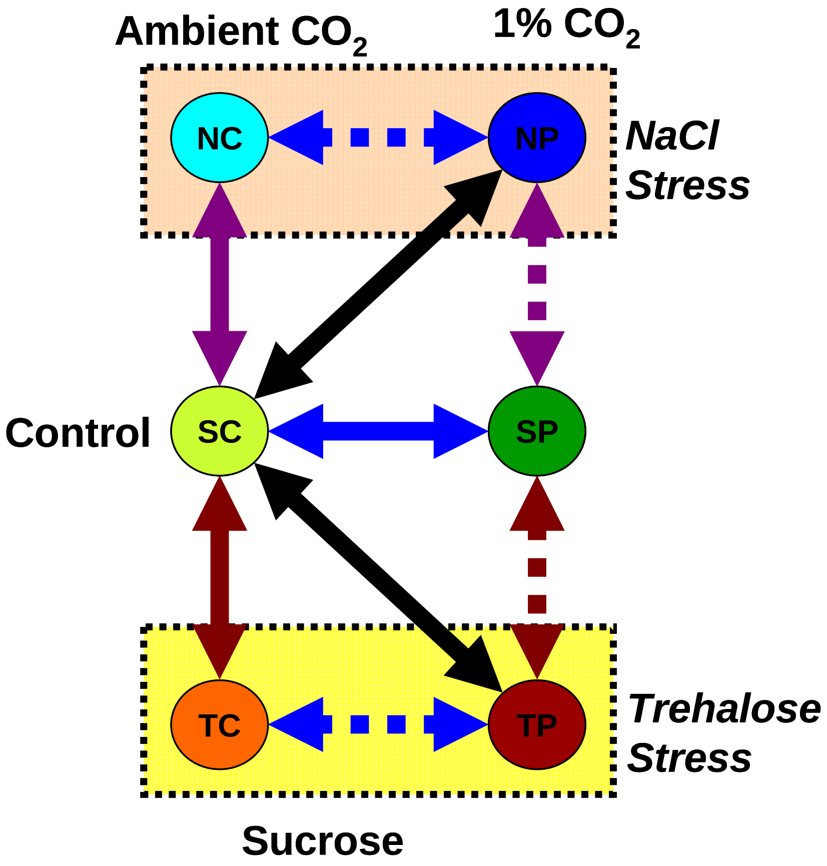


Figure 5.1: The figure shows the overall experimental design. Each of the 6 circles represents a stress response experiment. Each of these experiments has 9 timepoints including pretreatment control. Arrows connecting two circles signifies a stress response comparison. Arrows with same color shows similar stress, i.e. blue arrows represent CO_2 stress response. Continuous line arrows imply comparison with respect to control state SC, while the dashed arrows imply comparison between two stress conditions.

Each pair-wise comparison (shown by arrows) was done independently. One of the underlying objectives of carrying out this multiple stress experiment was to compare the stress responses. For an effective comparison of the significant genes from each stress response the significance levels of analysis should be comparable. One way to achieve this is to use the same threshold value (delta) for hypothesis testing using paired-SAM. However the reference distribution of expected scores changes for every comparison, hence using the same delta doesn't ensure the same significance level. Having the same FDR from all the stress comparisons will be an effective way of making the significance levels comparable. Hence, for each comparison delta value was chosen such that it corresponds to maximum number of genes with 0 FDR. FDR value 0 was chosen to keep the false positives minimum or to have maximum confidence in the results.

Results from the statistical significance analysis are explained and assuming that the statistical significance implies biological significance they are discussed in the context of *A. thaliana* physiology.

5.1. Study of individual stress responses

5.1.1 Transcriptional response of Arabidopsis thaliana liquid cultures subjected to elevated CO_2 stress

Elevated CO_2 stress was the first stress response studied in the course of this experiment. In the later experiments elevated CO_2 stress was coupled with other stresses like salt stress and trehalose stress. Following measurements of plant weight and media pH were obtained from both the experimental set [table 5.1]. Sample 5 of the control experiment (SC) was found to cluster separately from rest of the samples of control experiment possibly due to its different physiological condition. This is supported by the finding that sample 5 also has exceptionally low weight [table 5.1]. This sample was thus removed from further analysis and the timepoint 6h was represented by the only sample 6.

Table 5.1 Weight and media pH of the samples from control and CO_2 stress experiments

		Sucrose (SC)		Sucrose perturbed (SP)	
Time Pt.	Sample No	Weight	pH	Weight	pH
0	20	14	6.16	16.7	6.15
0	19	16.9	6.21	19.2	6.15
0	18	19	6.18	18.3	5.98
0	17	17.8	6.45	15.7	6.03
1	1	13.7	6.13	11.6	6.1
1	2	12.8	6.43	8.4	6.03
3	3	14.7	6.09	20.2	6.03
3	4	16.3	6.23	25	6.1
6	5	9.2	6.32	20.7	6.12
6	6	15.1	6.3	25.1	6.25
9	7	18	6.35	18.3	6.13
9	8	21	6.24	19.7	6.21
12	9	12	6.28	13.5	6.14
12	10	14.5	6.4	12.6	6.19
18	11	22.9	6.3	22.3	6.39
18	12	21.9	6.36	14.2	6.26
24	13	21.8	6.27	31.2	6.21
24	14	20.1	6.45	30.6	6.34
30	15	28.3	6.51	22.7	6.18
30	16	30.6	6.4	27.3	6.36

5.1.1.2 Multivariate statistical analysis

With 75% cutoff a repository of genes were selected that are present in at least 12 out of 16 timepoints. A total of 11231 genes were selected and this gene pool was used for all further analysis. Principal component analysis (PCA), as implemented in TIGR TM4 software suite [Saeed et al., 2003] was used for clustering of the experiments. PCA shows a clear separation of timepoints from control and perturbed group in 3-D reduced gene space. First 3 principal components captured 39, 20 and 14% of the variance respectively, i.e. 73% in total [Figure 5.2].

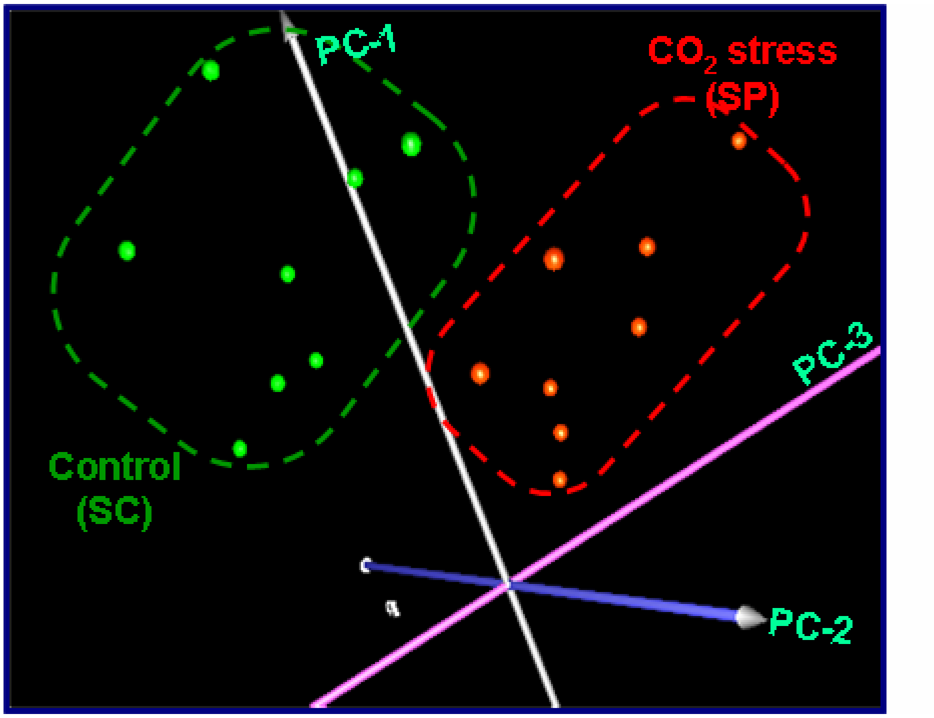


Figure 5.2 PCA analysis shows the timepoints of control and CO_2 stress experimental timepoints on reduced gene space. The timepoints are clearly separated implying elevated CO_2 stress is producing a significant change in *A. thaliana* physiology.

Experimental timepoints were also clustered using hierarchical clustering (HCL) with Pearson's correlation distance. HCL also effectively separates the two sets.

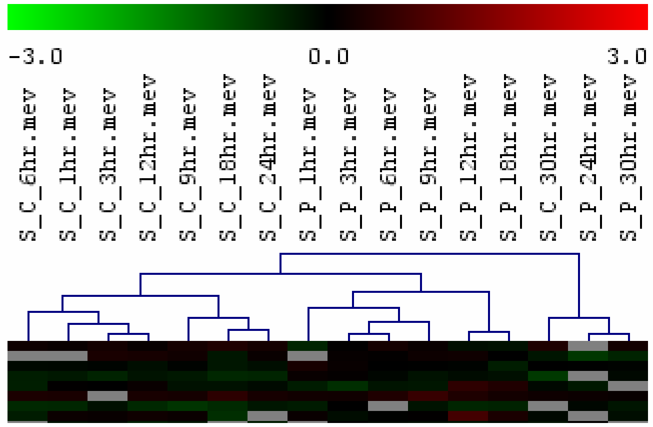


Figure 5.3: Hierarchical clustering of the experimental timepoints show they are broadly producing two clusters corresponding to two different stress conditions. 30h timepoints from both the groups and 24h timepoint of perturbed set are clustering separately.

Timepoints 24 and 30h of perturbed and 30h of control clustered separately from the group of other control and perturbed timepoints showing how the response changes at the later part of the stress.

Paired SAM as implemented in TIGR TM4 [Saeed et al., 2003] was used for overall significance analysis while MiTimeS [Dutta et al., 2007] helped reveal the significance analysis results at individual timepoints. Delta value 1.16 was used for paired SAM, as this significance level provides maximum number of differentially expressed genes with 0 FDR. Paired SAM identified 313 and 143 genes as positively and negatively significant which is only 3 and 1% respectively of the genes used for analysis. Multiple test correction was used in MiTimeS for significance analysis; hence significance threshold used for individual timepoints (1.56) was higher than that of paired SAM. Use of multiple test correction is another way to ensure that genes identified as significant are truly significant. Percentages of genes that are positively and negatively significant at individual timepoints and also from paired SAM are shown in figure 5.4. It is evident from the figure that, numbers of significant genes of both types are increasing for first 9h of the applied stress. From 12 to 30h period significant gene numbers are remaining almost at the same level. For all timepoints and also from paired SAM, number of positively significant genes was higher than negatively significant genes. Significant gene numbers from paired SAM is much smaller compared to that of individual timepoints. As most of the genes are changing their significance level between timepoints, paired SAM, which is based on average of all the timepoints couldn't identify these genes as significant [explained in more detail in chapter 6].

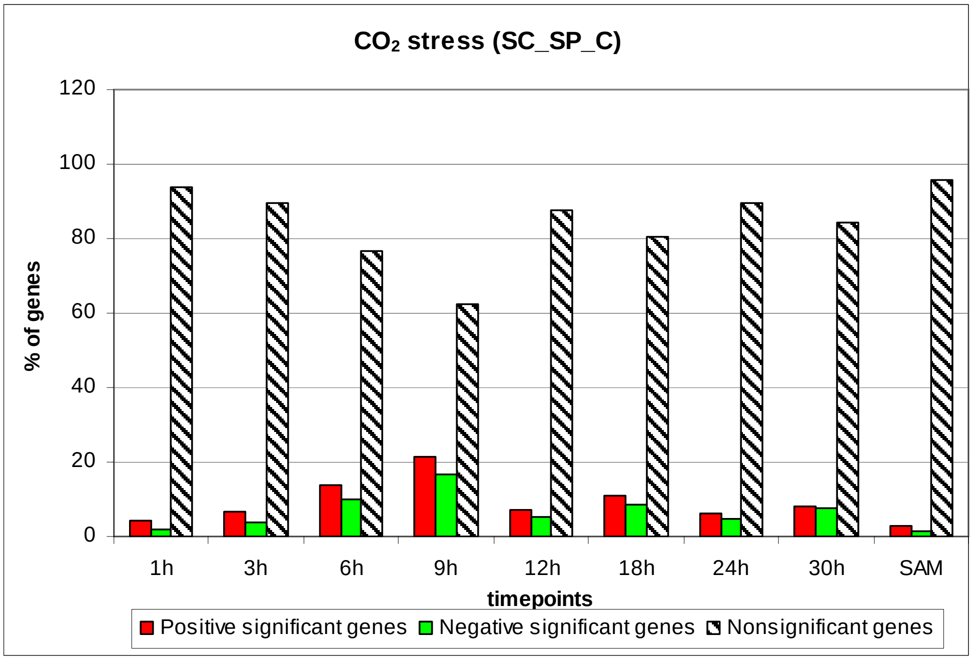


Figure 5.4: The bar diagram shows the percentage of genes that are identified as positively, negatively and non-significant at individual timepoints and also from paired SAM. Gene number of both positively and negatively significant type gradually increases for first 9 hours.

5.1.1.2 Data validation and interpretation in the context of plant physiology

Calvin cycle, sucrose and starch biosynthesis

CO_2 fixation in Calvin cycle is catalyzed by Rubisco [Nelson et al., 2002]. Rubisco comprises two subunits, small (*rbcS*) and large (*rbcL*), which are encoded by nuclear and chloroplast genes respectively. The *rbcL* gene is positively significant at 3 and 18h. The *Arabidopsis rbcS* gene family consists of four members, namely 1A, 1B, 2B and 3B [Krebbers et al., 1988]. In the present study all four subunits were identified as negatively significant at 9, 24 and 30h of perturbation. Subunit 1A was identified as negatively significant also at 3h timepoint. The gene encoding phosphoglycerate kinase, the enzyme catalyzing the conversion of 3PG to 1,3-bis-phosphoglycerate, is also negatively significant at 3, 6 and 24h.

Triose-phosphates transported from the chloroplasts to the cytoplasm are converted to hexose-phosphates. The gene encoding UDP-glucose pyrophosphorylase, which catalyzes the conversion of glucose-1-phosphate to UDP-glucose, is significantly under-transcribed at 6-18h of perturbation (Figure 5.5). Sucrose is synthesized from UDP-glucose through two sequential reactions (Figure 5.5) catalyzed by the enzymes sucrose phosphate synthase (SPS) and sucrose-phosphatase [Denis et al., 2001]. SPS is potentially the main regulatory enzyme [Stitt et al., 1991], activated by glucose-6-P and inhibited by inorganic phosphate [Denis et al., 2001]. In the present study, the SPS expression was observed to be significantly decreasing due to the applied perturbation in an average over all the timepoints and at 6, 9 and 18h specifically, after the initiation of the perturbation. Sucrose could be also directly produced from UDP-glucose through a reversible reaction catalyzed by sucrose synthase (SS) [Smith et al., 1993]. However, SS is considered to be mainly used in the breakdown of sucrose [Denis et al., 2001]. The gene encoding SS was identified as positively significant from 3-24h timepoints and also from paired-SAM, possibly increasing sucrose dissociation.

Starch, produced in chloroplast, serves as a transient sink to accommodate excess photosynthate that cannot be converted to sucrose and exported (Figure 5.4) [Smith et al., 1993]. Hence, when sucrose synthesis is restricted, starch synthesis is promoted. ADP-glucose pyrophosphorylase (AGPase) is a key enzyme catalyzing ADP-glucose formation and regulated by triose-phosphate/Pi ratio and fructose-1P (Figure 5.5) [Smith et al., 1993]. One gene encoding for AGPase family protein was identified to be positively significant at 6h of perturbation. Genes encoding starch phosphorylase and beta-amylase, enzymes involved in starch dissociation, were identified negatively significant at most of the timepoints, indicating possible decrease in starch degradation.

To check if it is possible to extract more information, with reduction in significance threshold, delta values was reduced to 0.9. At this delta value FDR was 0.47%, which is still quite low. Using MiTimeS with this delta value reveals lot more genes as significant, but the confidence involved in these results is much low. Rubisco activase, involved in removing the inhibition of rubisco activity by ribulose-1,5-bisphosphate [Nelson et al., 2002], was identified positively and negatively significant at 1h and 9h, respectively. Rubisco subunits 2B and 3B were also identified as positively significant at 6h.

In conclusion, the rate of carbon fixation is possibly increasing significantly for the first 9h of perturbation, as this is supported by the expression change of both Rubisco activase and both subunits of Rubisco. This coincides with the observation that number of significant genes increase for first 9 hours of the experiment. Without being adequately conclusive, measurements related to sucrose and starch production indicate that the latter is favored over the former. This agrees with previous studies that connect elevated CO_2 physiology with starch accumulation [Paul et al., 2001]. More organelle-specific studies are required to validate these indications.

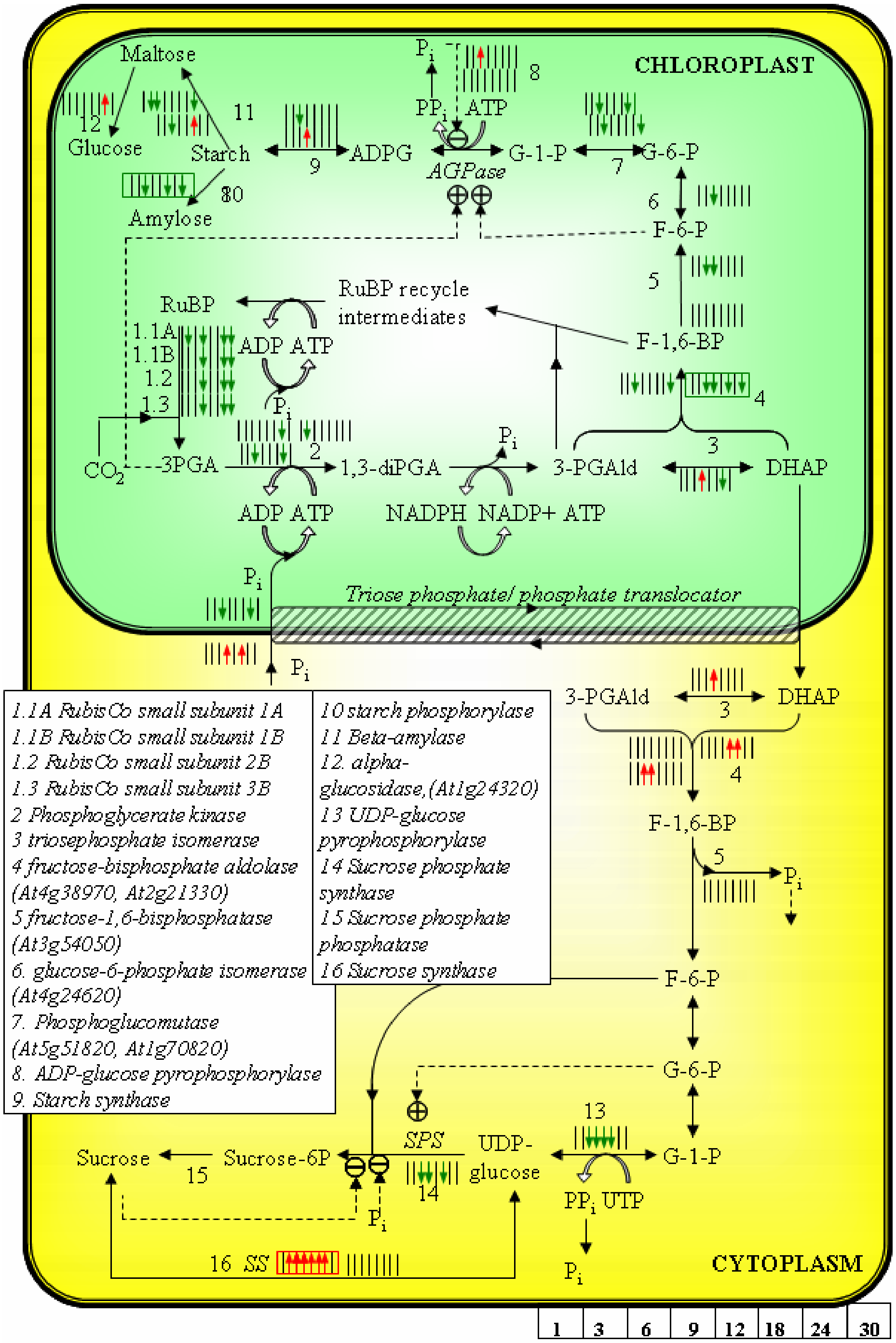

CHLOROPLAST
Maltose
12
Glucose
11
Starch
10
Amylose
9
ADPG
P$_i$
8
PP$_i$
ATP
AGPase
G-1-P
7
G-6-P
6
F-6-P
5
F-1,6-BP
4
RuBP recycle intermediates
RuBP
1.1A
1.1B
1.2
1.3
ADP ATP
P$_i$
2
CO$_2$
3PGA
1,3-diPGA
P$_i$
3-PGAld
3
DHAP
ADP ATP
NADPH NADP+ ATP
P$_i$
Triose phosphate/ phosphate translocator
P$_i$
1.1A RubisCo small subunit 1A
1.1B RubisCo small subunit 1B
1.2 RubisCo small subunit 2B
1.3 RubisCo small subunit 3B
2 Phosphoglycerate kinase
3 triosephosphate isomerase
4 fructose-bisphosphate aldolase (At4g38970, At2g21330)
5 fructose-1,6-bisphosphatase (At3g54050)
6. glucose-6-phosphate isomerase (At4g24620)
7. Phosphoglucomutase (At5g51820, At1g70820)
8. ADP-glucose pyrophosphorylase
9. Starch synthase
10 starch phosphorylase
11 Beta-amylase
12. alpha-glucosidase,(At1g24320)
13 UDP-glucose pyrophosphorylase
14 Sucrose phosphate synthase
15 Sucrose phosphate phosphatase
16 Sucrose synthase
3-PGAld
3
DHAP
4
F-1,6-BP
5
P$_i$
F-6-P
G-6-P
13
G-1-P
SPS
UDP-glucose
14
PP$_i$ UTP
P$_i$
Sucrose
15
Sucrose-6P
P$_i$
16 SS
CYTOPLASM
1
3
6
9
12
18
24
30

Figure 5.5: Observed effect of the applied perturbation on the physiology of Calvin cycle, starch and sucrose biosynthesis pathways at the transcriptional level. At individual timepoints significance level of a gene encoding a reaction's enzyme is represented by an arrow. Corresponding to 8 timepoints there are 8 arrows/lines. Red facing up, green facing down and black with no arrow head implies that the gene is positively, negatively and non-significant at timepoint. A red or green box around the timepoint arrows signifies that the gene was identified as positively or negatively, respectively, significant by paired-SAM. The arrows with continuous lines imply reactions whereas those with dotted lines imply positive or negative regulation depending on the sign.

Photorespiration

Carbon fixation and photorespiration “compete” for Rubisco activity (Figure 5.6). Changes in the CO_2/O_2 ratio have been shown to affect the flux distribution between the two pathways [Lehninger et al., 2002]. Transcriptomic measurements indicated that this was indeed the case in the present study, in which the liquid cultures experienced a 25-fold increase in the CO_2/O_2 ratio in their growth environment. Three of the enzymes involved in photorespiration, serine hydroxylmethyl transferase, NAD+ hydroxypyruvate reductase and 2-phosphoglycolate phosphatase, were also identified as negatively significant at most of the time-points and by paired-SAM (Figure 5.6). Thus, the most commonly observed effect of elevated CO_2 stress in soil grown plants, i.e. inhibition of photorespiration, is also observed conclusively in the liquid culture system as well, at both the metabolomic and the transcriptomic levels.

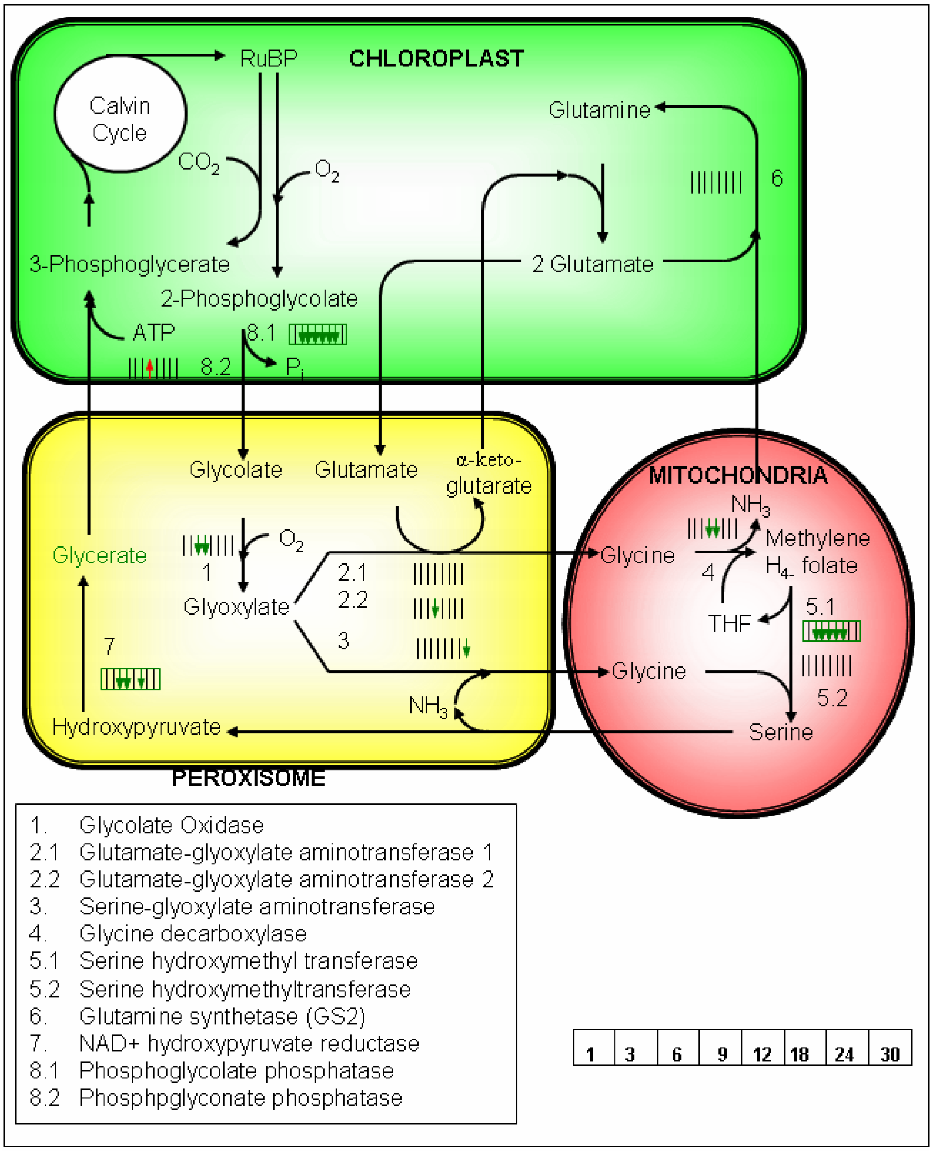


Figure 5.6 Observed effect of the applied perturbation on the physiology of photorespiration, at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Nitrogen assimilation and amino-acid biosynthesis

Amino acid synthesis requires source of carbon and nitrogen, hence it is dependent on central carbon metabolism and nitrogen fixation. Based on their precursor and bio-synthetic pathway amino-acids can be divided into four classes [Lee and Leagood, 1993]. Though this division is to some extent arbitrary, however studying them based on their precursor will make it easier to analyze them. Following are the four main precursors:

1. *Aspartate:* Asparagine, lysine, threonine, methionine, isoleucine
2. *Glutamate:* Glutamine, arginine and praline.
3. *Pyruvate:* alanine, serine, cysteine, glycine. Pyruvate also donates carbon to lysine, isoleucine and valine.
4. *Erythrose 4-phosphate:* Aromatic amino acids phenylalanine, tyrosine and tryptophan.
5. *Ribose 5-phosphate:* Histidine.

Each of the above classes will be analyzed separately in the following sections. Nitrate is assimilated in the leaves, and also in the roots. In most of the full grown plants, nitrate assimilation occurs primarily in leaves [Heldt, 2005]. The transport of nitrate into the root cells proceeds as symport with two protons. Root cells contain several nitrate transporters in their plasma membrane; among them are transporters with low affinity and transporter with very high affinity. The latter one is induced only when required by metabolism, so that capacity of nitrate uptake is adjusted to the environmental conditions.

Nitrate taken up by roots are stored temporarily in vacuoles. Nitrate is reduced to NH_4^+ in the leucoplast and is used for production of glutamine and asparagine. When capacity of nitrate assimilation in the roots is exhausted, nitrate is released from the roots to the xylem vessel and is carried to the leaves. It is taken up into mesophyll cells and is reduced subsequently. First step is the reduction of nitrate to nitrite by nitrate reductase (NR) present in cytosol and then to NH_4^+ by nitrite reductase in the chloroplast. Nitrate reductase mostly uses NADH as reductant.

In *Arabidopsis* there are two isoenzymes of nitrate reductase NR1 and NR2. Both of these genes are expressed in root and leaves and are induced by nitrate and show differential response [Cheng et al., 1991]. It was observed that NR1 (not NR2) mRNA maintains a higher basal level in *Arabidopsis* plants grown in the absence of nitrate than in the presence of nitrate. By maintaining a higher basal level of one gene, the plant could scavenge nitrate at levels below that required for induction [Cheng et al., 1991].

Reduction of nitrite to ammonia requires the uptake of six electrons. Nitrite reductase is located exclusively in plastids and utilizes reduced ferredoxin as electron donor which is supplied by photosystem I. Glutamine synthetase in chloroplast transfers the newly formed NH_4^+ at the expense of ATP to glutamate, forming glutamine. The same reaction fixes NH_4^+ released during photorespiration. Because of high-rate of photorespiration, the amount of NH_4^+ produced by the oxidation of glycine is about 5 to 10 times higher than amount of NH_4^+ generate by nitrate assimilation. Thus only a minor proportion of glutamine synthesis in the leaves is actually involved in nitrate assimilation.

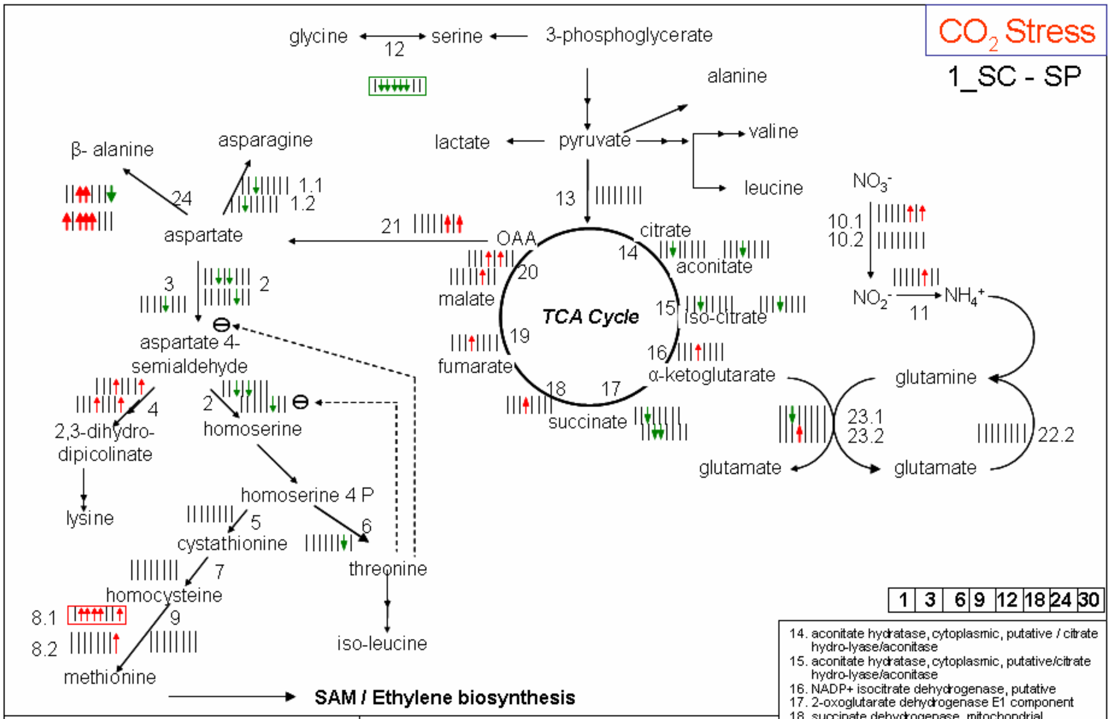
Glutamine synthetase (GS) is the key enzyme in this nitrogen assimilatory process, as it catalyzes the first step in the conversion of inorganic nitrogen into its organic form. Distinct isoenzymes of GS exist in the chloroplast (GS2) and cytosol (GS1) in *Arabidopsis*. These distinct GS isoenzymes are encoded by distinct nuclear genes in all higher plants studied. Expression studies showing that the distinct GS genes display organ-specific, cell-specific, developmental, and temporal patterns of gene expression. The levels of mRNA for the chloroplastic GS2 or the cytosolic GS1 are each induced by light or by carbon metabolites [Oliveira and Coruzzi, 1999]. The dramatic light induction of mRNA for GS2 is mediated in part by phytochrome and in part by light-induced changes in levels of Sucrose. In contrast, light induction of cytosolic GS1 mRNA can be accounted for by metabolic induction by sucrose alone. Interestingly, the non-hexose carbon source 2-oxoglutarate also induced accumulation of mRNA for cytosolic GS1, but had negligible effects on the levels of mRNA for chloroplastic GS2 [Oliveira and Coruzzi, 1999].

The glutamine formed in chloroplast is converted via glutamate synthase (also called glutamine oxoglutarate amino transferase, abbreviated as GOGAT) by reaction with α-ketoglutarate to two molecules of glutamate with ferrodoxin as reductant. *Arabidopsis* in fact contains two expressed genes encoding Fd-GOGAT isoforms (GLU1 and GLU2). These genes show contrasting patterns of gene expression [Coschigano et al., 1998]. *GLU1*gene plays a major role in photorespiration as well as a role in primary nitrogen assimilation in leaves. The Fd-GOGAT isoenzyme encoded by *GLU2* is proposed to be involved mainly in primary nitrogen assimilation in roots [Coschigano et al., 1998]. It was also observed that *GLU1* gene product functions in concert with chloroplastic GS2 in leaves [Coschigano et al., 1998].

α-ketoglutarate, which is required for the glutamate synthase reaction, is transported into the chloroplast by oxoglutarate/malate translocator in the counter exchange for malate. Glutamate formed is also transported out of the chloroplast into cytosol by plastidic glutamate/malate translocator, also in exchange for malate [Heldt, 2005].

During photosynthesis CO_2 assimilation and nitrate assimilation have to be matched to each other. Nitrate assimilation can progress only when CO_2 assimilation provides carbon skeletons for the amino acids. Moreover, nitrate assimilation must be regulated such that the production of amino acids does not exceed demand. Finally it is important that nitrate reduction doesn't proceed faster than nitrite reduction, since otherwise toxic levels of nitrite would accumulate in cells [Heldt, 2005].

Under elevated CO_2 stress nitrate reductase 1 (NR1) gene was positively significant at 18 and 30h timepoints, although NR2 was non-significant at all timepoints [Figure 5.7]. Nitrite reductase, the next enzyme of the same pathway was also positively significant at 18h timepoint, showing similar significance profile (for details please check chapter 6). Possibly there is an increase in rate or nitrogen assimilation at the later stage of elevated CO_2 stress when enough carbon has been assimilated. Most of the genes coding for enzymes catalyzing TCA cycle reactions like NADP+ isocitrate dehydrogenase, succinate dehydrogenase, fumarate hydratase, malate dehydrogenase [NAD] shows similar significance profiles and becomes positively significant at 9h timepoint, again showing the importance of this timepoint during the course of the experiment. GLU2 gene involved in nitrogen assimilation is also positively significant at this time-point. Homocysteine S-methyltransferase(HMT-1) gene in methionine synthesis is strongly up-regulated leading to the increased methionine production under elevated CO_2 stress. Glutamate decarboxylase and dihydrodipicolinate synthase genes of beta-alanine and lysine production pathway are also positively significant at some of the timepoints and concentration of these amino-acids has increased in first 24hs of the experiment. None of the other genes was found strongly negatively significant, which implies most of the amino-acids are over-produced under elevated CO_2 stress especially at the later stage when sucrose production and starch biosynthesis is stopped.



1.1 Asparagine synthetase 2 (ASN2)
1.2 Asparagine synthetase 3 (ASN3)
2 Bifunctional aspartate kinase/ homoserine dehydrogenase / AK-HSDH
3. Aspartate kinase, lysine-sensitive
4. Dihydrodipicolinate synthase
5. Cystathionine γ-synthase
6. Threonine synthase

7. cystathionine beta-lyase, chloroplast
8.1 homocysteine S-methyltransferase(HMT-1)
8.2. homocysteine S-methyltransferase(HMT-2)
9. methionine synthase, putative
10.1 Nitrate reductase1 (NR1)
10.2 Nitrate reductase 2 (NR2)
11. ferredoxin--nitrite reductase, putative
12. Serine hydroxyl-methyl transferase
13. citrate synthase

14. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
15. aconitate hydratase, cytoplasmic, putative/citrate hydro-lyase/aconitase
16. NADP+ isocitrate dehydrogenase, putative
17. 2-oxoglutarate dehydrogenase E1 component
18. succinate dehydrogenase, mitochondrial
19. fumarate hydratase, putative / fumarase, putative
20. malate dehydrogenase [NAD], mitochondrial
21. aspartate aminotransferase, cytoplasmic isozyme 1 / transaminase A (ASP2)
22.1 Glutamine Synthetase 1(GS1)
22.2 Glutamine Synthetase 2(GS2)
23.1 Glutamate Synthase 1 (GLU1)
23.2 Glutamate Synthase 2 (GLU2)
24. glutamate decarboxylase

Figure 5.7 Observed effect of the applied perturbation on the physiology of the nitrogen assimilation and amino acid biosynthesis at the transcriptional level. Positively and negatively significant genes and metabolites are color-coded as described in the caption of Figure 5.5

Under aerobic condition ATP and NADPH is produced from TCA cycle provides energy for other cellular function. However, under anaerobic condition when supply of oxygen is limited, anaerobic fermentation pathway could be used and both of the pathways use pyruvate as precursor. In fermentation, pyruvate is first reduced to acetaldehyde and then to ethanol. Pyruvate decarboxylase, the enzyme that catalyses the reduction of pyruvate to acetaldehyde was found overproduced as two genes (At4g33070, At5g01320) coding that enzyme are significantly up regulated. The following reduction reaction for ethanol is catalyzed by alcohol dehydrogenase and this gene is positively significant at 6 out of 8 timepoints and also from paired SAM. This provides a strong indication that fermentation pathway flux is possibly increasing to provide energy to cells. Pyruvate can also be converted to lactate an-aerobically by lactate dehydrogenase and this gene is positively significant at 3 timepoints.

Ethylene Synthesis and Signaling

Ethylene is a potent modulator of plant growth and development [Ecker, 1995]. The plant hormone ethylene is involved in many aspects of the plant life cycle, including seed germination, root hair development, root nodulation, flower senescence, abscission, and fruit ripening [reviewed in Johnson and Ecker, 1998]. The production of ethylene is tightly regulated by internal signals during development and in response to environmental stimuli from biotic (e.g., pathogen attack) and abiotic stresses, such as wounding, hypoxia, ozone, chilling, or freezing. To understand the roles of ethylene in plant functions, it is important to know how this hormone is synthesized, how its production is regulated, and how the signal is transduced.

Ethylene response has been shown to be regulated at the level of ethylene synthesis. S-adenosylmethionine (SAdoMet) is the precursor for ethylene biosynthesis [reviewed in Yang and Hoffman, 1984; Kende, 1993]. In addition to being an essential building block of protein synthesis, nearly 80% of cellular methionine is converted to *S*AdoMet by *S*AdoMet synthetase (SAM synthetase, EC 2.5.1.6) at the expense of ATP utilization [Ravanel et al., 1998]. *S*AdoMet is the major methyl donor in plants and is used as a substrate for many biochemical pathways, including polyamines and ethylene biosynthesis. On the basis of the Yang cycle, the first committed step of ethylene biosynthesis is the conversion of *S*AdoMet to ACC by ACC synthase (*S*adenosyl-L-methionine methylthioadenosine-lyase, EC4.4.14) [reviewed in Yang and Hoffman, 1984; Kende,1993]. The rate-limiting step of ethylene synthesis is the conversion of *S*AdoMet to ACC by ACC synthase [reviewed in Kende, 1993]. The observations that expression of the ACS genes is highly regulated by a variety of signals and that active ACC synthase is labile and present at low levels suggest that ethylene biosynthesis is tightly controlled. ACC is further oxidized by ACC oxidase to produce ethylene and is activated by CO_2 [Thrower et al., 2001]. ACC Synthase and ACC oxidase enzymes belong to a multigene family and are regulated by a complex network of developmental and environmental signals responding to both internal and external stimuli.

Five ethylene receptors exist in Arabiodpsis: ETR1, ETR2, ETHYLENE RESPONSE SENSOR 1 (ERS1), ERS2, and EIN4 [Chang et al., 1993; Hua et al., 1995; Hua and Meyerowitz, 1998; Sakai et al., 1998]. Among these receptors, only ETR1, ETR2, and EIN4 contain a receiver domain that shows similarity to bacterial response regulators Since homodimerization of ETR1 and ERS1 has been observed in plants [Schaller et al., 1995; Hall et al., 2000], receptors that do not have receiver domain, ERS1 and ERS2, have been postulated to use the receiver domains of other proteins by forming heterodimers with them

Although various factors have been demonstrated to regulate ethylene levels in the plant [Abeles et al., 1992], only limited information is available on the regulation of ethylene receptor levels. Interestingly, one factor that affects the expression of ethylene receptors is ethylene itself, which induces the expression of ETR2, ERS1, and ERS2, but not of ETR1 and EIN4. Expression of the ethylene receptor ETR1 is downregulated by salt and osmotic stress at the transcript and protein levels. This decrease in receptor levels should cause increased sensitivity of the plant to ethylene. Thus, abiotic stresses, in addition to regulating ethylene signal transduction by modulating hormone levels, may also do so by modulating receptor levels.

Genetic studies have predicted that hormone binding results in the inactivation of receptor function [Schaller et al., 2002]. In absence of ethylene, therefore, the receptors are hypothesized to be in functionally active form that activates CTR1, which is a negative regulator of the pathway [Schaller et al., 1995]. The receptors ETR1 and ERS1 have high affinity for CTR1, whereas

ETR2 posses a low binding affinity for CTR1 [Guo and Ecker, 2004] and they were found to show similar expression profiles [Please see chapter 6 for more details]. EIN2, EIN3, EIN5 and EIN6 are positive regulators of ethylene responses, acting downstream of CTR1. The nuclear protein EIN3 is a transcription factor that regulates the expression of its immediate target genes such as ETHYLENE RESPONSIVE FACTOR 1 (ERF1) [Hall et al., 2000; Chen et al., 2002,] It has been shown recently that ERF1 also regulates other hormone responses particularly the jasmonate (JA) mediated defense response. Like ethylene, JA is a volatile signal that rapidly induces the expression of ERF1, and its expression is activated synergistically by both the hormones. Both signaling pathways are required concurrently for induction of ERF1 expression and the activation of its target gene PDF1.2. Hence EFR1 functions as transcription factor that integrates signal from ethylene and JA pathways. The mechanism of simultaneous requirement for both pathways to activate ERF1 expression is unclear.

Ethylene modulates the responses to other plant hormones, such as JA, salicylic acid, auxin, abscisic acis (ABA) and cytokinin, but the mechanism that control each of these critical hormone-hormone interactions are largely unknown. *A. thaliana* genome comprises of multiple copies of ACC synthase (regulatory step of ethylene formation) and ACC oxidase (final step); two copies of each were considered in the present analysis after normalization and filtering. One gene corresponding to each of the enzymes was positively significant at most of the timepoints and also from paired SAM. Though the other genes encoding the same enzymes were significant at some of the timepoints, but didn't pass the significance test from paired SAM analysis. These observations indicate that even at the first hour of perturbation, the ethylene synthesis was significantly induced at the transcriptional level. This is in agreement with previous studies of the short- and long-term effect of elevated CO_2 stress, which have reported a sustained ethylene release in the photosynthetic leaves of higher plants [Bassi et al., 1982, Dhawan et al., 1981, Grodzinski et al., 1996]]. This is, however, the first time that the same is reported for the *A. thaliana* physiology.

ETR2 and ERS were identified as positively significant at 5 and 6, respectively, out of the 8 time points. ETR2 was also positively significant from paired-SAM. EIN4 was identified as positively significant at 30h, whereas ETR1 was identified positively significant at 24h. ETR1 gene, as explained before, was not affected significantly to increased ethylene production. The present transcriptional observations are in complete agreement with previous reports regarding soil-grown *A. thaliana* plants [Chen et al., 2005, Hua et al., 1998], providing additional support to our hypothesis that the molecular responses of both plant systems to a particular perturbation are similar.

Quite limited experimental data exits about the transcriptional regulation of the ethylene signaling cascade(s) [Chen et al., 2005]. The transcription of CTR1, the second protein of the ethylene cascade, has been previously observed being induced in response to ethylene in tomatoes [Adams-Phillips et al., 2004], but this had not been previously confirmed in *A. thaliana* [Keiber et al., 1993, Gao et al., 2003]. In the present study, the gene encoding for CTR1 synthesis was identified as positively significant at 3 out of the 8 time points (3, 9 & 24h). Regarding three genes encoding for subsequent proteins in the ethylene signaling cascade, i.e. EIN2, EIN3, EIL1, these were identified as significant at only 1,1 and 0, respectively, out of the 8 time points (see Figure 5.8). It has been previously reported [Chen et al., 2005] that indeed these genes are not regulated by ethylene at the transcriptional level. In the absence of ethylene, EIN3 is degraded by the ubiquitin-ligases EBF1 & EBF2. In the presence of ethylene, EIN3 degradation is halted. The mechanism by which this regulation is achieved is not well-understood to-date, but its elucidation would be of great importance for understanding ethylene signaling [Chen et al., 2005].

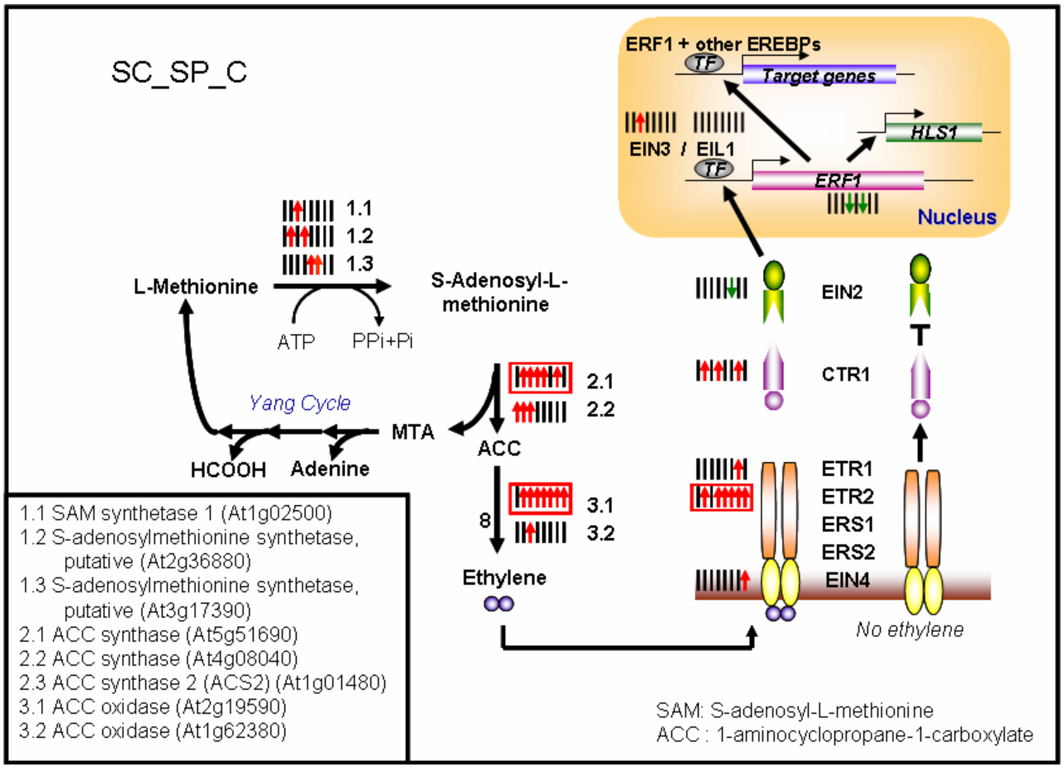

SC_SP_C
ERF1 + other EREBPs
TF
Target genes
EIN3 / EIL1
TF
ERF1
HLS1
Nucleus
1.1
1.2
1.3
L-Methionine
S-Adenosyl-L-methionine
ATP
PPi+Pi
Yang Cycle
MTA
HCOOH
Adenine
2.1
2.2
ACC
8
3.1
3.2
Ethylene
EIN2
CTR1
ETR1
ETR2
ERS1
ERS2
EIN4
No ethylene
1.1 SAM synthetase 1 (At1g02500)
1.2 S-adenosylmethionine synthetase, putative (At2g36880)
1.3 S-adenosylmethionine synthetase, putative (At3g17390)
2.1 ACC synthase (At5g51690)
2.2 ACC synthase (At4g08040)
2.3 ACC synthase 2 (ACS2) (At1g01480)
3.1 ACC oxidase (At2g19590)
3.2 ACC oxidase (At1g62380)
SAM: S-adenosyl-L-methionine
ACC : 1-aminocyclopropane-1-carboxylate

Figure 5.8 Observed effect of the applied perturbation on ethylene biosynthesis and signaling at the transcriptional level. Positively and negatively significant genes and metabolites are color-coded as described in the caption of Figure 5.5.

Moreover, the gene encoding for the transcription factor ERF1 is activated by EIN3 [Chen et al., 2005], was identified as negatively significant at 9 and 18h of perturbation (Figure 5.8). These observations complement data from an ethylene stress study in soil-grown *A. thaliana* plants, in which it had been observed that the expression of ERF1 was not up-regulated under conditions of high ethylene concentration [Zhong et al., 2003]. However, among the 7 analyzed Ethylene Responsive Element Binding (EREBP) transcription factor family genes, which are regulated by EIN3 / ERF1, three were identified as strongly negatively significant and from paired-SAM (for details on the significance profile over time of all 7 genes, see Table 5.2). Most of other EREBP genes, which are not significant from paired SAM were also negatively significant at 9h timepoint implying an overall under-regulation of these family of genes at this timepoint. Some of these apparent contradictions are one additional indication of the plethora of open questions that need to be pursued towards the elucidation of the ethylene signaling pathway, the present study contributing important information to the currently available relevant database.

Table 5.2 Significance level of ethylene-responsive element-binding family genes at individual timepoints and also from paired SAM. The red box with 1, green box with -1, white box with 0 imply that gene is positively negatively and non-significant respectively.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
ethylene-responsive element-binding family protein	-1	-1	-1	-1	0	-1	0	0	-1
ethylene-responsive element-binding family protein	-1	0	-1	-1	0	-1	0	0	-1
ethylene-responsive element-binding protein, putative	0	0	-1	-1	0	-1	0	0	-1
ethylene-responsive element-binding factor 4 (ERF4)	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	-1	0	0	0

ethylene-responsive element-binding protein, putative	0	-1	-1	-1	0	0	-1	0	0
ethylene-responsive element-binding protein, putative	0	0	1	0	0	0	0	-1	0

Conclusion: Elevated CO_2 stress affected the carbon fixation reactions which in turn affects the sucrose and starch biosynthesis reactions. Increased carbon pool available can be used for increased amino-acid biosynthesis and plant growth. Amino-acid biosynthesis also requires nitrogen supply and up-regulation of nitrate reductase genes are possibly providing the increased nitrogen supply. Increased CO_2/O_2 ratio is inhibiting the photorespiration reactions as expected. Sulfur metabolism was down regulated and the elevated CO_2 stress was found to affect the ethylene biosynthesis and signaling cascade.

5.1.2 Transcriptional response of Arabidopsis thaliana liquid cultures subjected to NaCl stress

As an part of our multiple stress response studies of *A thaliana* liquid cultures, I applied NaCl stress of 50mM by applying NaCl solution in the media [please see the experimental design chapter for detail]. Salt stress was found to create physiological change to the culture within first 30 hours of its application. Table 5.3 shows the weight of the plant samples and the corresponding media pH measurements. Sample 9 of the control experiment (NC) was found to cluster separately from rest of the samples of control experiment possibly due to its different physiological condition. This is supported by the finding that sample 9 also has exceptionally low weight [table 5.3]. This sample was removed from further analysis and the timepoint 6h was represented by the only sample 6.

Table 5.3 Weight and media pH of the samples from control and NaCl stress experiments

		Sucrose control (SC)		NaCl stress (NC)	
Time Pt.	Sample No	Weight	pH	Weight	pH
0	20	14	6.16	26.3	6.95
0	19	16.9	6.21	19.2	6.79
0	18	19	6.18	N.A.	N.A.
0	17	17.8	6.45	N.A.	N.A.
1	1	13.7	6.13	21.1	6.59
1	2	12.8	6.43	N.A.	N.A.
3	3	14.7	6.09	23.8	6.58
3	4	16.3	6.23	22.4	6.56
6	5	9.2	6.32	26.9	6.84
6	6	15.1	6.3	23.0	6.51
9	7	18	6.35	20.6	6.76
9	8	21	6.24	23.6	6.46
12	9	12	6.28	12.8	6.34
12	10	14.5	6.4	24.1	6.51
18	11	22.9	6.3	22.5	6.63
18	12	21.9	6.36	22.3	6.64
24	13	21.8	6.27	26.5	6.55
24	14	20.1	6.45	23.9	6.56
30	15	28.3	6.51	25.1	6.56
30	16	30.6	6.4	24.3	6.67

Plants were immediately frozen in liquid nitrogen and kept at -80^{o}C until they were ground in liquid nitrogen. During hand grinding of the frozen plants in liquid nitrogen it was experienced that these plants are much easier to grind compared to frozen plants from other experiments. This is only a qualitative observation and can not be quantified; nevertheless it is an important observation as it implies there have been some physiological changes.

5.1.2.1 Multivariate statistical analysis

As there are many genes that have missing expression values at one or more timepoints, before the analysis is started, a common repository of genes is selected. The selected 12049 genes have non-zero expression values for at least 12 out of 16 timepoints. According to TIGR MeV Principal Component Analysis (PCA), the control transcriptomic profiles can be clearly differentiated from their perturbed counterparts (Figure 5.9). This implies that the physiology of the plant liquid cultures is affected by the applied perturbation at transcriptional level, even during the first 30h of treatment. First 3 principal components were found to capture 60, 16 and 6% of the information. Hence, when the experiments are viewed at 3-D space it can account for most of the variance (82%). It can also be seen due to the application of NaCl stress timepoints have moved along principal component 1, which accounts for maximum variability.

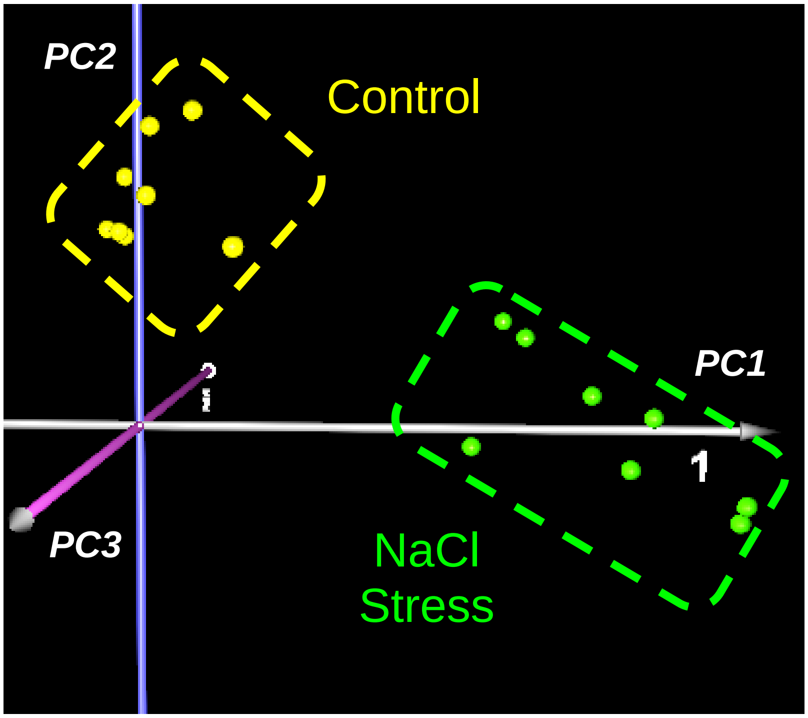


Figure 5.9: PCA analysis shows the timepoints of control and NaCl stress experimental timepoints on reduced gene space. The timepoints are clearly separated implying NaCl stress is producing a significant change in *A. thaliana* physiology.

Experimental timepoints were also clustered using hierarchical clustering and it also shows a clear separation between them (Figure 5.10).

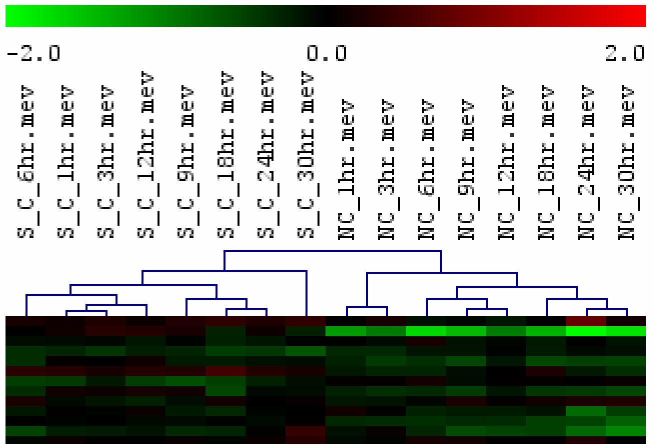


Figure 5.10: PCA analysis shows the timepoints of control and CO2 stress experimental timepoints on reduced gene space. The timepoints are clearly separated implying elevated CO2 stress is producing a significant change in *A. thaliana* physiology.

Both paired SAM and MiTimeS were used for significance analysis based on overall and individual timepoints as explained earlier. Delta value of 2.677 was selected for paired SAM, as this delta value has highest number of significant genes with minimum (0 in this case) FDR. There were 1643 and 1653 genes found positively and negatively significant from paired SAM with this delta value, which constitutes around 14% (in both significant types) of genes used for analysis (12049). The whole list of positively and negatively significant genes can be found in supplementary table S1. The delta value used for MiTimS analysis was same as that of paired SAM. Number of genes positively, negatively and non-significant at individual timepoints obtained from MiTimeS were plotted with that of paired SAM results in figure 5.11. All the time-points including paired SAM shows almost equal number of significant genes of both types. It is clear 6 and 12h timepoints have maximum and minimum number of significant genes respectively.

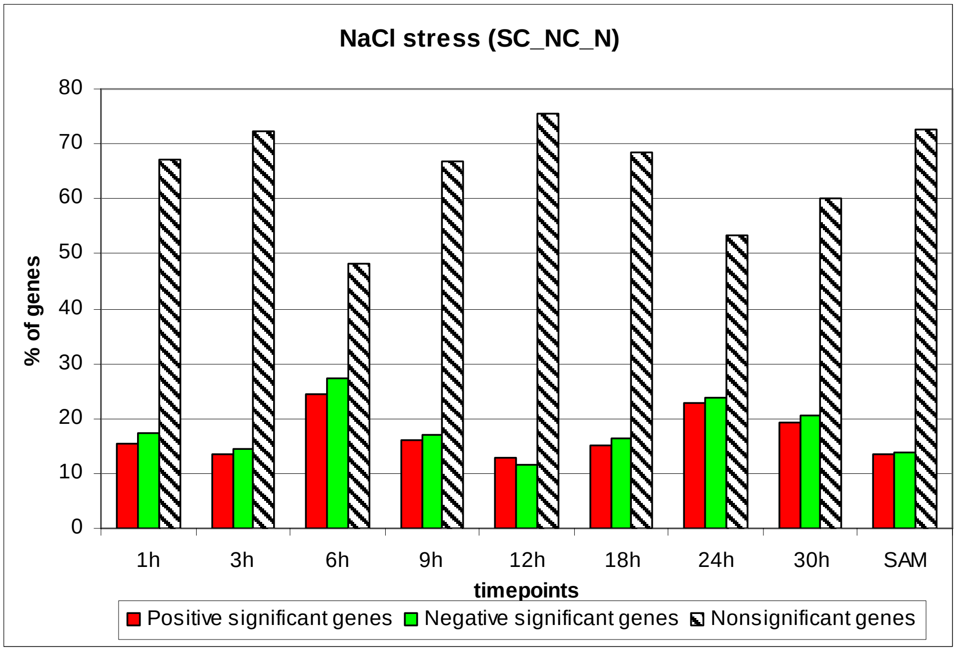


Figure 5.11 Percentage of positively, negatively and non-significant genes at individual timepoints and from paired SAM. Significant gene numbers from paired SAM is comparable with that of individual timepoints.

5.1.2.2 Data validation and interpretation in the context of plant physiology

As a result of genetic, molecular and biochemical analysis salt stress response pathway is well studied [Zhu JK et al., 2000]. Figure 5.12 shows a schematic diagram of the salt stress response pathway, also called SOS pathway. Calcium signal is induced by salt stress which is further sensed by calcium binding protein SOS3. SOS3 interacts and activated SOS2 which is a serine threonine protein kinase. SOS1 is a salt tolerance effector gene encoding a plasma membrane Na+/H+ antiporter is regulated by combined activity of SOS2 and SOS3 [Zhu JK, 2002] . SOS1 gene regulates transcriptionally and post-transcriptionally expression of other genes related to salt tolerance [Shi et al., 2000]. SOS pathway may also be involved in stimulation or suppression the activities of other transporters involved in ion homeostasis under salt stress, such as vacuolar H+-

ATPases and pyrophosphatases (PPase), vacuolar Na+/H+ exchanger (NHX), and plasma membrane K+ and Na+ transporters [Zhu JK, 2002] .

From the hypothesis testing results of this experiment it was observed SOS1 gene (sodium proton exchanger, putative (NHX7)) was positively significant at all the timepoints and also from paired SAM, which is exactly what is expected based on previous literature. SOS3 gene was also significantly up-regulated at 1, 18 and 24h timepoints. Different response of SOS1 and SOS3 gene is interesting and creates a scope for future study. Information about SOS2 gene was missing in this analysis. Responses of several other sodium proton exchanger genes (NHX) were also studied. Among them NHX2 and NHX3 genes were positively significant at 7 and 6 timepoints respectively and also from paired SAM. However other NHX genes were non-significant at most of them timepoints and also from paired SAM. Three genes (At4g19960, At2g30070, At4g13420) coding for potassium transporter were found to be positively significant and paired SAM analysis, again confirming the results from the literature. It was also observed Na+/Ca2+ antiporter gene (At2g47600) is positively significant at first 4 timepoints (1-9h period) and also from paired SAM analysis.

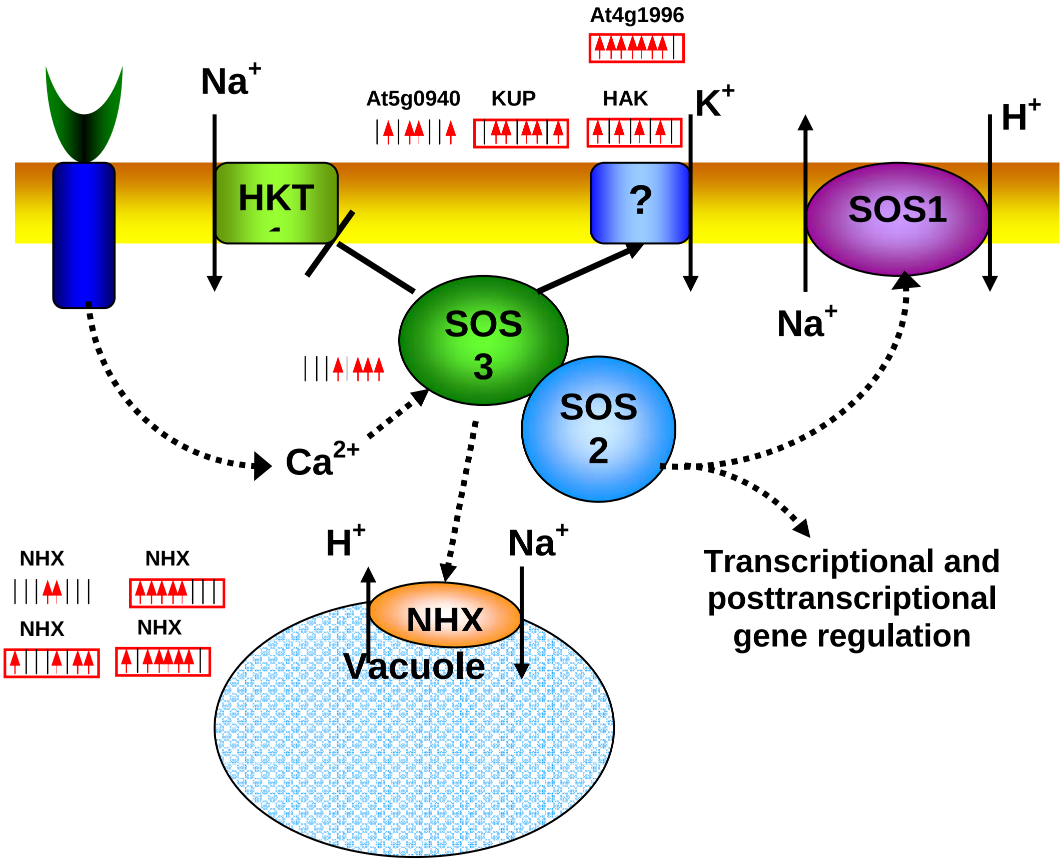


Figure 5.12: Pathway for salt stress signaling and response [redrawn from Zhu et al., 2002].

From the data I observed, 5 (At1g27770, At4g37640, At3g63380, 60608.m00041, At3g57330) out of 6 genes coding "calcium-transporting ATPase – plasma membrane-type/ Ca(2+)-ATPase" are positively significant at all the timepoint and also from paired SAM analysis. Though it is a strong indication, but association of this gene class with salt stress response was not found in literature. High-throughput approach used allowed me to come up with this speculation without knowing much from literature, which would not have been possible with a conventional hypothesis-driven approach. The only gene (At5g57110) that was not positively significant was coding for "Ca(2+)-ATPase isoform 8 (ACA8)". The reason for anomalous behavior of this isoform couldn't be explained and requires further investigation. Interestingly, the genes “calcium-transporting ATPase – endoplasmic reticulum-type” were found to be non-significant. Another gene (At5g38710) coding “proline oxidase, putative / osmotic stress-responsive proline dehydrogenase”, believed to be involved in salt stress was also found positively significant at all the timepoints. The gene coding pyrroline-5-carboxylate reductase, the enzyme that controls the proline biosynthesis from glutamate was found positively significant at only 30h timepoint. None of the other genes involved in proline was found to be significant from paired SAM analysis. In some of the plants it was observed that under salt or water stress, proline, glycine betaine or manitol accumulates in cytosol, chloroplast and mitochondria [Heldt, 2005] which minimizes the damaging effect under these stresses. These substances also participate as anti-oxidant in elimination of reactive oxygen species (ROS) [Heldt, 2005]. Water shortage or high salinity of soil causes and inhibition of CO_2 assimilation, resulting in an over-reduction of photosynthetic electron transport carriers, which in turn leads to an increased formation of ROS [Heldt, 2005].

Response to salt stress might not always translate to up-regulation of genes. Here I found two genes At2g41720 and At3g05890 coding for “putative salt-inducible protein” and “hydrophobic protein (RCI2B) / low temperature and salt responsive protein (LTI6B)” respectively negatively significant from paired SAM.

Calvin cycle, starch and sucrose production

All the four genes encoding small subunits (1A, 1B, 2B and 3B) of RuBisCO were found negatively significant at 24-30h timepoints (figure 5.13). These genes are part of nuclear genome and have important regulatory activity. The large subunit of rubisco was also negatively significant at 24h timepoint. Alpha and beta subunits of RuBisCO subunit binding-protein (At1g55490 and At2g28000) were also negatively significant at 5 out of 8 timepoints and also from paired SAM analysis (figure 5.13). The rate of carbon fixation is possibly decreasing at the last 6 hour of the experiment. Two phosphoglycerate kinase genes were also negatively significant at some of the timepoints especially at the later stage, implying a possible overall reduction of calvin cycle flux. Most of the genes catalyzing the starch synthesis reactions in cytosol are moderately negatively significant especially at last two timepoints. Possibly a reduction in Calvin cycle is translated into reduction in starch biosynthesis. It was also found from literature that starch production decreases under salt stress and starch degradation increases providing energy for carrying out cellular mechanism. ADP-glucose pyrophosphorylase plays an important regulatory role in starch biosynthesis and corresponding 2 genes were found negatively significant at 1 and 6h signifying a possible decrease in starch biosynthesis.

Triose phosphate is transported out of chloroplast for sucrose synthesis. Triose phosphate/ phosphate translocator gene, responsible for transferring the triose phosphates produced in chloroplast to cytosol is negatively significant from 3-24h timepoints and also from overall analysis. Fructose-bisphosphate aldolase gene was found positively significant at all the timepoints and also from paired SAM analysis. Sucrose phosphate phosphatase and sucrose synthase genes were also found positively significant at most of the timepoints and also from paired SAM analysis. However, UDP-glucose pyrophosphorylase gene was negatively significant which is inconsistent with other genes of the pathway (figure 5.13). The gene corresponding to the cytosolic copy of this enzyme was not found and this gene is believed to be active in endomembrane system. Sucrose synthase gene catalyzes reversible reaction and mainly used for degradation of sucrose. It is not clear from the results why sucrose production and degradation are simultaneously increasing. It could be possible under this condition both the pathways are contributing to sucrose synthesis.

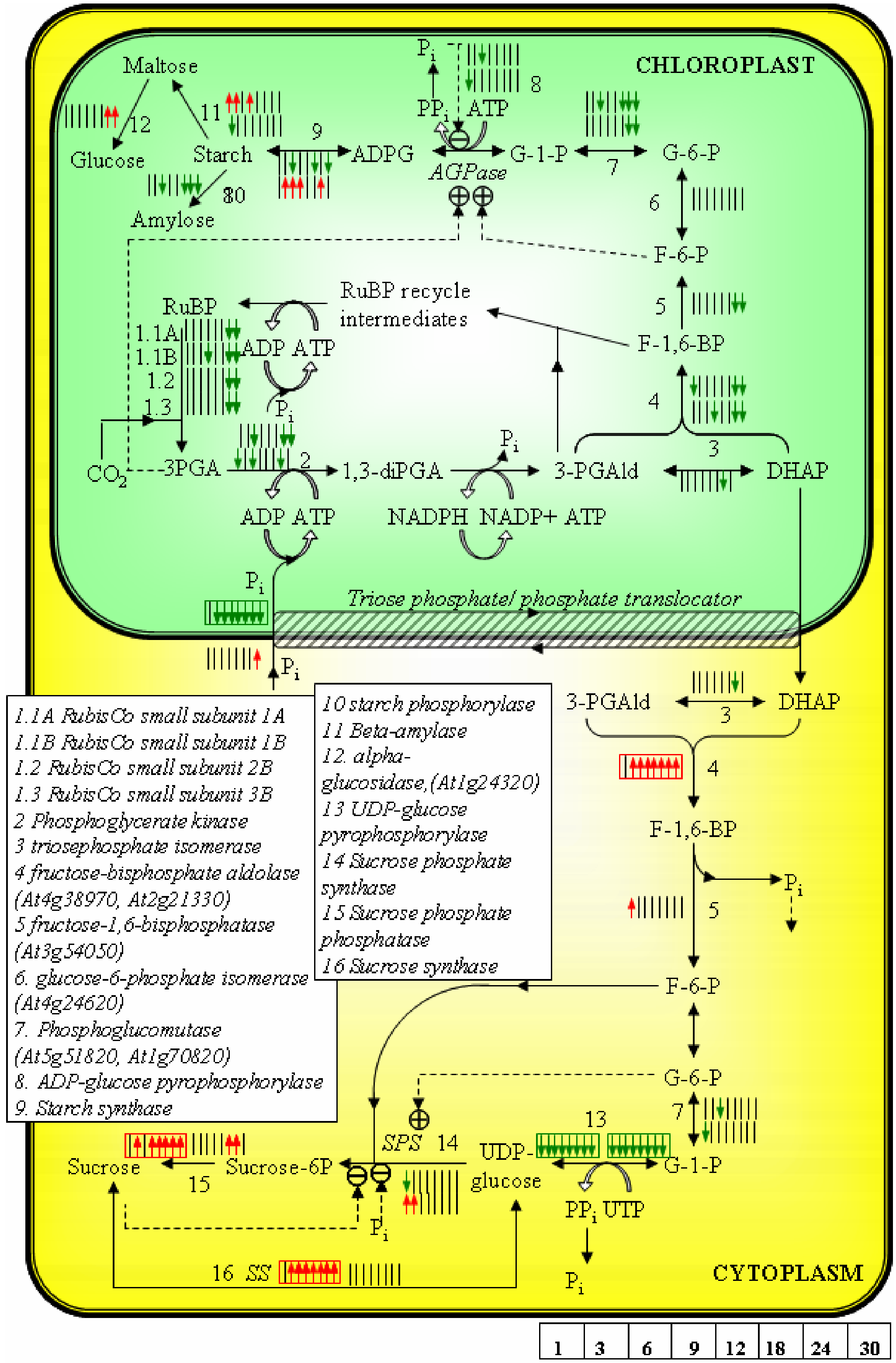

CHLOROPLAST
Maltose
12
11
Glucose
Starch
9
ADPG
10
Amylose
P_i
8
PP_i
ATP
AGPase
G-1-P
7
G-6-P
6
F-6-P
RuBP
RuBP recycle intermediates
1.1A
1.1B
1.2
1.3
ADP ATP
P_i
5
F-1,6-BP
4
CO_2
3PGA
2
1,3-diPGA
3-PGAld
3
DHAP
ADP ATP
NADPH NADP+ ATP
P_i
Triose phosphate/ phosphate translocator
P_i
1.1A RubisCo small subunit 1A
1.1B RubisCo small subunit 1B
1.2 RubisCo small subunit 2B
1.3 RubisCo small subunit 3B
2 Phosphoglycerate kinase
3 triosephosphate isomerase
4 fructose-bisphosphate aldolase (At4g38970, At2g21330)
5 fructose-1,6-bisphosphatase (At3g54050)
6. glucose-6-phosphate isomerase (At4g24620)
7. Phosphoglucomutase (At5g51820, At1g70820)
8. ADP-glucose pyrophosphorylase
9. Starch synthase
10 starch phosphorylase
11 Beta-amylase
12. alpha-glucosidase,(At1g24320)
13 UDP-glucose pyrophosphorylase
14 Sucrose phosphate synthase
15 Sucrose phosphate phosphatase
16 Sucrose synthase
3-PGAld
3
DHAP
4
F-1,6-BP
P_i
5
F-6-P
G-6-P
7
13
SPS 14
UDP-glucose
G-1-P
Sucrose
15
Sucrose-6P
P_i
PP_i UTP
P_i
16 SS
CYTOPLASM
1 3 6 9 12 18 24 30

Figure 5.13: Observed effect of the applied perturbation on the physiology of Calvin cycle, starch and sucrose biosynthesis pathways at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Photorespiration

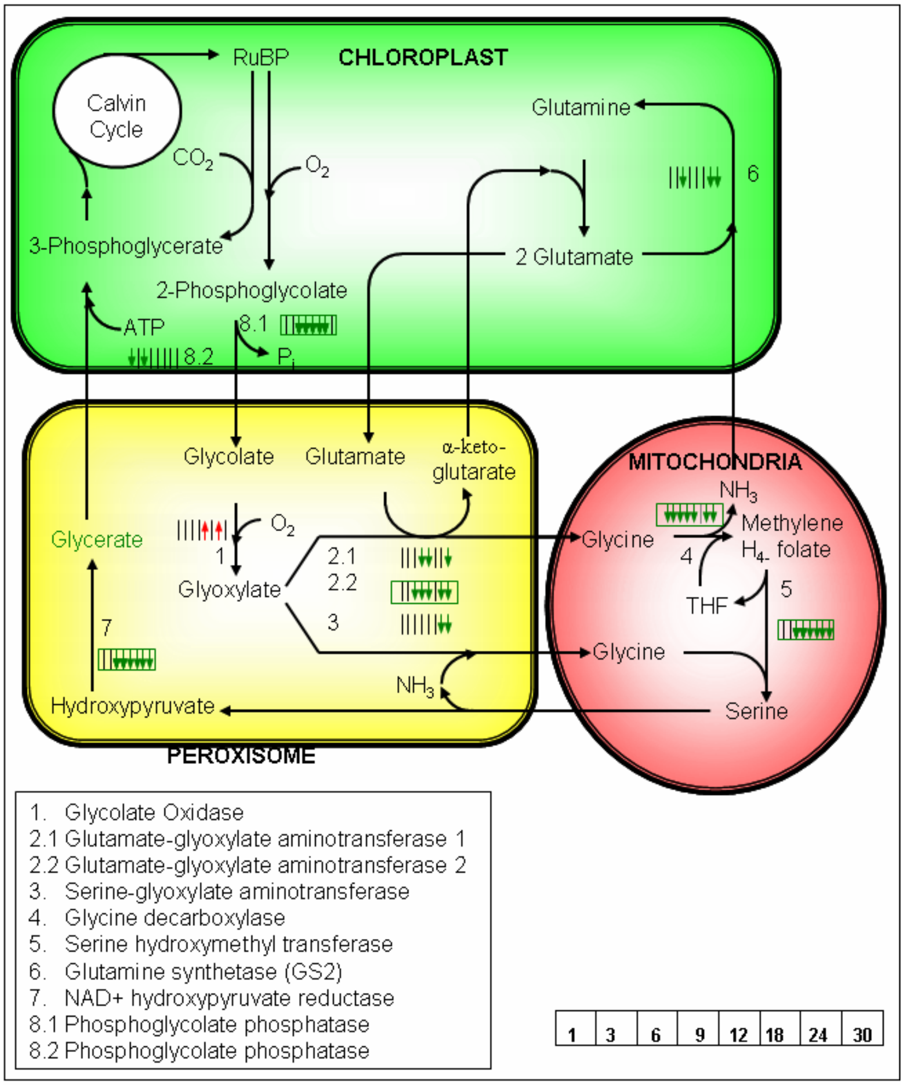


Figure 5.14: Observed effect of the applied perturbation on the physiology of photorespiration, at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Though CO_2/O_2 ratio remains unchanged during this stress, but photorespiratory pathway is significantly affected by the applied perturbations (Figure 5.14). Most of the genes of this pathway like glutamate-glyoxylate aminotransferase 2, glycine decarboxylase, serine hydroxymethyl transferase, NAD+ hydroxyl pyruvate reductase, phosphoglycolate phosphatase are all negatively significant at most of the timepoints and also from paired SAM analysis, implying a net reduction in photorespiratory flux. Reduction in photorespiration under NaCl stress was not reported anywhere before and need to be verified from independent analysis.

Amino-acid biosynthesis pathways

Under NaCl stress it was observed that two nitrate reductase genes NR1 and NR2 are showing differential response. NR1 is negatively significant at 1 and 6h timepoints while NR2 is positively significant at 5 out of 8 timepoints and also from paired SAM analysis (Figure 5.15). Nitrite reductase was also found negatively significant at 1 and 6h timepoints, again showing similarity with NR1 response. As explained above, to stop toxic nitrite from accumulating, the rate of nitrate reduction should not exceed that of nitrite reduction. However our finding about NR2 contradicts this observation. From this observation I speculate that NR1 is possible regulatory enzyme involved in nitrate reduction. When expression profile of NR1, NR2 and nitrite reductase was compared, it was observed that NR1 and nitrite reductase has high degree of expression correlation while NR2 shown distinctly different expression profiles (please see section 6.2.3 in chapter 6). NR2 response is turn quite similar to GS1 stress response. The regulation involved in NR1 and NR2, especially under different environmental stress conditions could be a subject of future research.

After nitrate is reduced to NH_4^+ it is assimilated by glutamine synthetase, as explained above. Distinct isoenzymes of GS exist in the chloroplast (GS2) and cytosol (GS1) [Oliveira and Coruzzi, 1999]. GS1 was positively significant at all timepoints and also from paired SAM analysis. However, GS2, the gene active in chloroplast was negatively significant at 6, 24 and 30h timepoints (figure 5.15). Ferredoxin-dependent glutamate synthase, the enzymes that catalyses the following reaction has two isoenzymes Fd-GOGAT 1 and Fd-GOGAT 2. The gene encoding Fd-GOGAT 1 is negatively significant at 6 and 18-30h, hence shows significance profile similar to GS2 which is in accord with the previous finding [Coschigano et al., 1998].

Significance profiles of NR1 and nitrite reductase genes involved in nitrate reduction do not match with that of GS2 and GLU1 genes involved in NH4+ assimilation. A plausible reason could be NH_4^+ assimilated in chloroplast is coming from both photorespiration and nitrate reduction; hence the overall GOGAT reaction is the cumulative effect of both.

NaCl stress was also found to affect the amino-acid biosynthesis pathway to a great extent. Figure 5.15 depicts the transcriptional response of the nitrogen assimilation, TCA cycle and amino-acid biosynthesis pathway. Citrate synthase, the enzyme that catalyses conversion of pyruvate to citrate, the first reaction of TCA cycle was found negatively significant at all timepoints and also from paired SAM. The influx to TCA cycle is possibly decreasing due to the down regulation of this reaction. Aconitate hydratase (cytoplasmic) is the enzyme that catalyses two consecutive reactions of citrate to aconitate and again aconitate to isocitrate, is overproduced. Aconitate concentration was also significantly increasing from metabolomic analysis. 2-oxoglutarate dehydrogenase E1 component gene producing α-ketoglutarate was also up-regulated at most of the timepoints.

Following four reactions of the TCA cycle from α-ketoglutarate to oxaloacetate is found mostly non-significant. Hence possibly there is an accumulation of α-ketoglutarate concentration, which is responsible for up-regulation of NR2 gene.

Oxaloacetate (OAA) is taken out of TCA cycle for production of arpartate and the amino acids that are produced from aspartate. Aspartate aminotransferase, cytoplasmic isozyme 1/ transaminase A (ASP2) gene involved in OAA to aspartate reaction up-regulated at 18-30 h timepoints (figure 5.15). The increase in flux to aspartate is possibly used for beta- alanine production as the 2 glutamate decarboxylase genes encoding this reaction was positively significant from all timepoints. Beta-alanine is also significantly over-produced from metabolomic analysis. However there is a strong indication from transcriptomic analysis that the flux towards production of other amino acids like asparagine, homoserine, threonine, homosynteine, methionine, lysine are decreasing. This observation was apparently contradictory to some of the metabolomic observations where homoserine and methionie concentrations were significantly increased.

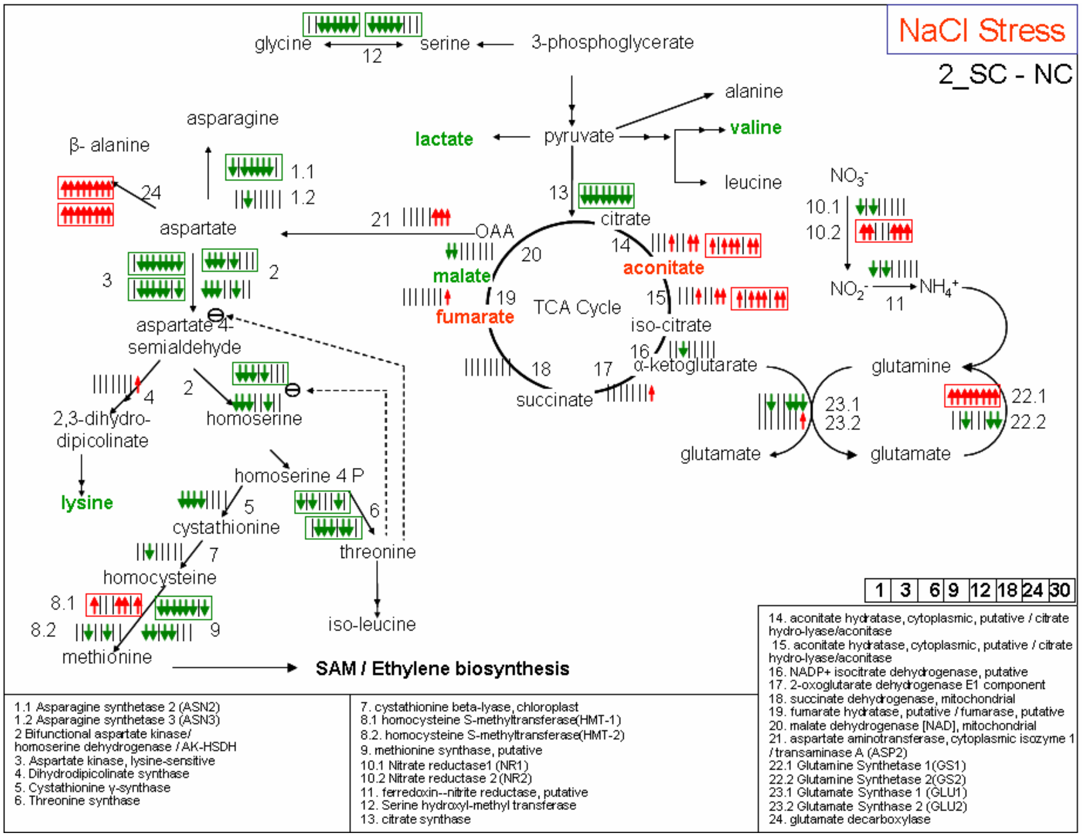

NaCl Stress
2_SC - NC
glycine
12
serine
3-phosphoglycerate
alanine
asparagine
lactate
pyruvate
valine
β- alanine
leucine
1.1
1.2
24
13
citrate
NO$_3^-$
10.1
10.2
aspartate
21
OAA
20
14
aconitate
malate
3
2
19
TCA Cycle
15
fumarate
iso-citrate
NO$_2^-$
NH$_4^+$
11
aspartate 4-semialdehyde
16
α-ketoglutarate
glutamine
4
2
18
17
succinate
23.1
23.2
22.1
22.2
2,3-dihydro-dipicolinate
homoserine
glutamate
glutamate
homoserine 4 P
lysine
5
6
cystathionine
7
threonine
homocysteine
8.1
8.2
9
iso-leucine
methionine
SAM / Ethylene biosynthesis
1
3
6
9
12
18
24
30
1.1 Asparagine synthetase 2 (ASN2)
1.2 Asparagine synthetase 3 (ASN3)
2 Bifunctional aspartate kinase/ homoserine dehydrogenase / AK-HSDH
3. Aspartate kinase, lysine-sensitive
4. Dihydrodipicolinate synthase
5. Cystathionine γ-synthase
6. Threonine synthase
7. cystathionine beta-lyase, chloroplast
8.1 homocysteine S-methyltransferase(HMT-1)
8.2. homocysteine S-methyltransferase(HMT-2)
9. methionine synthase, putative
10.1 Nitrate reductase1 (NR1)
10.2 Nitrate reductase 2 (NR2)
11. ferredoxin--nitrite reductase, putative
12. Serine hydroxyl-methyl transferase
13. citrate synthase
14. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
15. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
16. NADP+ isocitrate dehydrogenase, putative
17. 2-oxoglutarate dehydrogenase E1 component
18. succinate dehydrogenase, mitochondrial
19. fumarate hydratase, putative / fumarase, putative
20. malate dehydrogenase [NAD], mitochondrial
21. aspartate aminotransferase, cytoplasmic isozyme 1 / transaminase A (ASP2)
22.1 Glutamine Synthetase 1(GS1)
22.2 Glutamine Synthetase 2(GS2)
23.1 Glutamate Synthase 1 (GLU1)
23.2 Glutamate Synthase 2 (GLU2)
24. glutamate decarboxylase

Figure 5.15: Observed effect of the applied NaCl stress on Nitrogen assimilation and amino-acid biosynthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

It is already explained that TCA cycle flux is possibly decreasing, so fermentation pathway was investigated. In fermentation pyruvate is first reduced to acetaldehyde and then to ethanol. Pyruvate decarboxylase, the enzyme that catalyses the reduction to acetaldehyde was found overproduced as three genes (At4g33070, At5g01320, At5g17380) coding that enzyme are significantly up regulated. The following reduction reaction is catalyzed by alcohol dehydrogenase and this gene is also positively significant at 6 out of 8 timepoints and also from paired SAM. This provides a strong indication that fermentation pathway flux is possibly increasing to provide energy to cells. Pyruvate can also be converted to lactate an-aerobically by lactate dehydrogenase but the expression of this gene didn’t change significantly over the course of the experiment.

Number of genes related to tryptophan biosynthesis pathway was found positively significant from paired SAM and MiTimeS analysis (figure 5.16). Up-regulation of genes related to this pathway subjected to NaCl stress was not observed before. MiTimeS and paired SAM results are plotted in the context of tryptophan biosynthesis pathway in figure 5.16. Figure 5.16 shows, starting from chorismate all the genes except two involved in tryptophan biosynthesis pathway are significantly up-regulated, while the competitive reactions producing Phenylalanine and Tyrosine are decreasing.

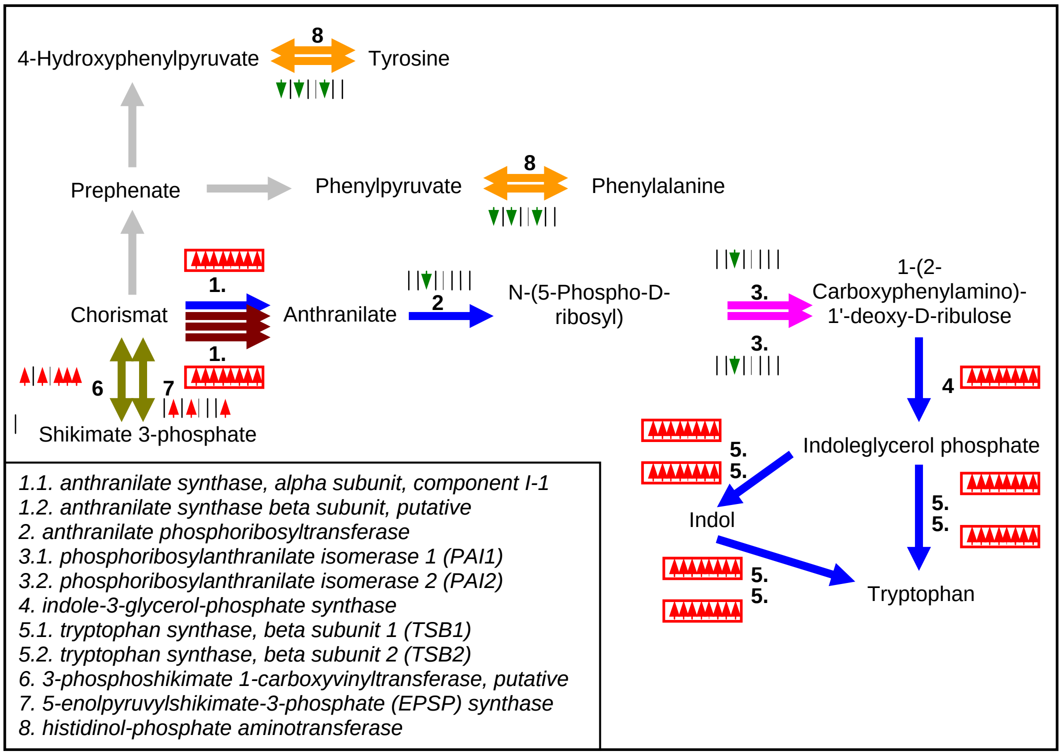

4-Hydroxyphenylpyruvate
8
Tyrosine
Prephenate
Phenylpyruvate
8
Phenylalanine
1.
Chorismat
Anthranilate
2
N-(5-Phospho-D-ribosyl)
3.
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose
1.
6
7
3.
4
Shikimate 3-phosphate
5.
5.
Indoleglycerol phosphate
Indol
5.
5.
5.
5.
Tryptophan
1.1. anthranilate synthase, alpha subunit, component I-1
1.2. anthranilate synthase beta subunit, putative
2. anthranilate phosphoribosyltransferase
3.1. phosphoribosylanthranilate isomerase 1 (PAI1)
3.2. phosphoribosylanthranilate isomerase 2 (PAI2)
4. indole-3-glycerol-phosphate synthase
5.1. tryptophan synthase, beta subunit 1 (TSB1)
5.2. tryptophan synthase, beta subunit 2 (TSB2)
6. 3-phosphoshikimate 1-carboxyvinyltransferase, putative
7. 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
8. histidinol-phosphate aminotransferase

Figure 5.16: Observed transcriptional response of the NaCl stress on tryptophan synthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

The decrease in TCA cycle flux is possibly responsible for down regulation of most of the amino-acid production reaction from aspartate, except beta-alanine. Production of serine and glycine from 3-phosphoglycerate is also decreasing. So the general conclusion that can be drawn is rate of most of the amino-acid biosynthesis is decreasing, while tryptophan and beta-alanine biosynthesis is increasing. The reason for increase in flux towards tryptophan biosynthesis is to produce different secondary metabolites for which tryptophan acts as precursor.

Ethylene Biosynthesis and Signaling

It is already explained before how ethylene plays an important regulatory role in plant. Pathway for ethylene bio-synthesis and its response was also explained in detail. Most of the genes' expressions in this pathway were significantly affected the applied salt stress.

Similar to elevated CO_2 stress, ACC oxidase was positively significant implying a possible increase in ethylene synthesis, especially after first 9 hours (figure 5.17). Both the genes were positively significant from 12 to 30h period while one of them was also significant from paired SAM analysis. However, ACC synthase (ACS), shows difference from CO_2 stress response. One of the ACS genes was found significant from 6-12h period, while the other ACS gene was missing. ACC synthase 2 (ACS2) gene was found positively significant at all timepoints and also from paired SAM (figure 5.17). It is not clear which of the ACS gene plays dominant role in ethylene biosynthesis. Both of them are up-regulated from NaCl stress, but the time-period over which they are significant varies.

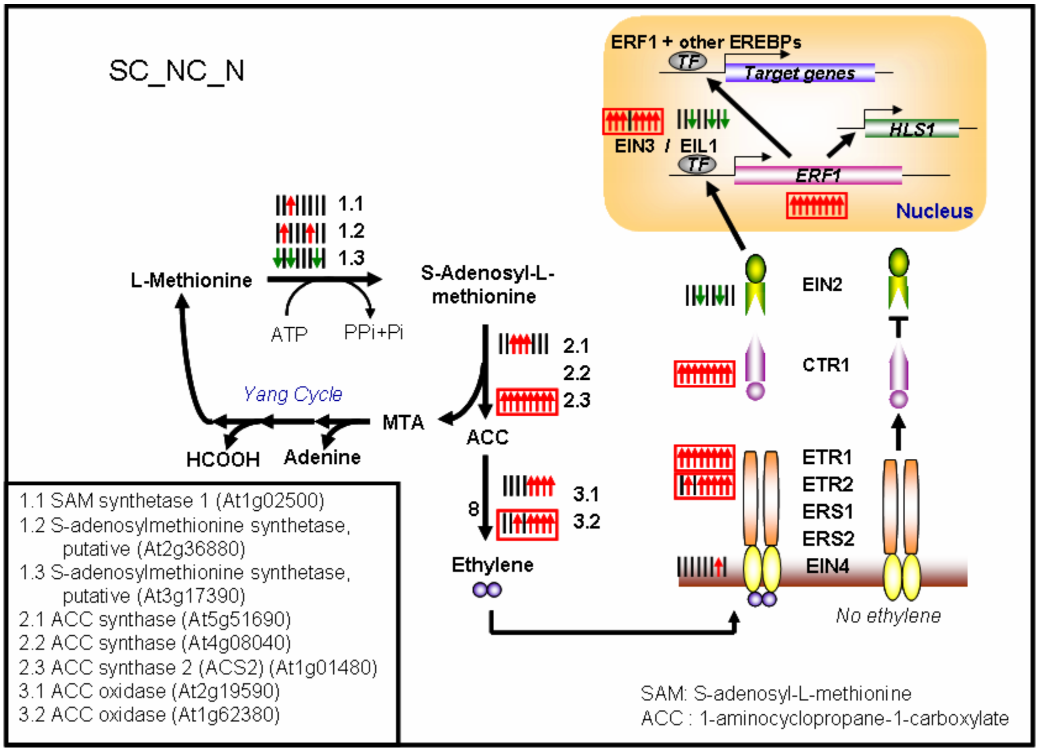

SC_NC_N
ERF1 + other EREBPs
TF
Target genes
HLS1
EIN3 / EIL1
TF
ERF1
Nucleus
1.1
1.2
1.3
L-Methionine
S-Adenosyl-L-methionine
ATP
PPi+Pi
EIN2
CTR1
2.1
2.2
2.3
Yang Cycle
MTA
ACC
HCOOH
Adenine
ETR1
ETR2
ERS1
ERS2
EIN4
8
3.1
3.2
Ethylene
No ethylene
1.1 SAM synthetase 1 (At1g02500)
1.2 S-adenosylmethionine synthetase, putative (At2g36880)
1.3 S-adenosylmethionine synthetase, putative (At3g17390)
2.1 ACC synthase (At5g51690)
2.2 ACC synthase (At4g08040)
2.3 ACC synthase 2 (ACS2) (At1g01480)
3.1 ACC oxidase (At2g19590)
3.2 ACC oxidase (At1g62380)
SAM: S-adenosyl-L-methionine
ACC : 1-aminocyclopropane-1-carboxylate

Figure 5.17 Observed effect of the NaCl stress on ethylene biosynthesis and signaling at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

With chnage in rate of ethylene production, ethylene signaling cascade is also affected which could be seen from the over-expression of genes coding proteins ETR1, ETR2, CTR1. ETR1 gene was not found to be significantly affected by ethylene, was strongly up-regulated here. The reason for its up-regulation under salt stress was an apparent contradiction with [Chen et al., 2005] where it was found down-regulated. CTR1 inhibits EIN2, from the expression analysis we also see EIN2 being negatively significant at some of the timepoints with over-expression of CTR1. This cascade leads to over-expression of other genes like

EIN3. EIN3 acts as transcription factor for the gene ERF1, from the figure it is clear that over-expression of EIN3 is up-regulating the expression of ERF1.

Table 5.4 Significance levels of ethylene-responsive element-binding family genes under NaCl stress. Notation used is same as table 5.2.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
ethylene-responsive element-binding family protein	-1	-1	-1	-1	0	0	-1	-1	-1
ethylene-responsive element-binding protein, putative	-1	-1	-1	-1	0	0	-1	0	-1
ethylene-responsive element-binding family protein	0	-1	-1	-1	-1	-1	0	-1	-1
ethylene-responsive element-binding factor 4 (ERF4)	0	0	0	-1	0	0	-1	-1	0
ethylene-responsive element-binding family protein	0	0	0	0	0	0	0	-1	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	0	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	0	0	0	0

Fatty acid biosynthesis and metabolism

Vast majority of fatty acid biosynthesis in plants occurs in plastids. Fatty acid is also produced in mitochondria but rates of this activity are very small in comparison to those found in plastid [Lea and Leegood]. Fatty acid biosynthesis is a multistep process involving number of enzymes. It was also observed that under the salt stress several genes involved in fatty acid biosynthesis from acetyl CoA are negatively significant. Genes related to fatty acid elongation were also found negatively significant. Following table 5.5 shows the list of genes involved in fatty acid biosynthesis and their significance level at individual timepoints and also from paired SAM.

While there is a possible indication of reduction in fatty acid biosynthesis, it was also observed that genes related to fatty acid metabolism are positively significant. List of the genes involved in fatty acid metabolism and their significance level from paired SAM and MiTimeS are shown in table 5.6.

Table 5.5: Significance level of genes at individual timepoints and also from paired SAM related to fatty acid biosynthesis under salt stress. Notation used is same as that of table 5.2

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
3-oxoacyl-[acyl-carrier-protein] synthase I	-1	-1	-1	-1	-1	-1	-1	-1	-1
3-oxoacyl-[acyl-carrier-protein] synthase III, chloroplast / beta-ketoacyl-ACP synthase III / 3-ketoacyl-acyl carrier protein synthase III (KAS III)	-1	-1	-1	-1	-1	-1	-1	-1	-1
3-ketoacyl-CoA thiolase	-1	0	0	0	0	0	0	0	0
3-oxoacyl-[acyl-carrier protein] reductase, chloroplast / 3-ketoacyl-acyl carrier protein reductase	-1	0	-1	0	0	0	0	0	0
3-oxoacyl-[acyl-carrier-protein] synthase II, putative	0	0	0	0	0	0	1	0	0
acetyl-CoA C-acyltransferase 1 / 3-ketoacyl-CoA thiolase 1 (PKT1)	-1	-1	-1	0	-1	-1	-1	-1	-1
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	-1	-1	-1	0	-1	0	0	-1
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	0	0	0	0	0	0	0	0
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	0	0	0	0	0	0	0	0
acetyl-CoA carboxylase 1 (ACC1)	0	0	-1	0	0	0	0	0	0
acetyl-CoA carboxylase 2 (ACC2)	0	0	-1	0	0	0	0	0	0
acetyl-CoA carboxylase, biotin carboxylase subunit (CAC2)	0	0	-1	0	-1	-1	-1	0	0
acyl carrier family protein / ACP family protein	0	0	-1	-1	-1	-1	-1	-1	-1
acyl carrier family protein / ACP family protein	0	0	-1	-1	-1	-1	-1	-1	-1
acyl carrier protein 3, chloroplast (ACP-3)	-1	0	0	0	0	0	0	0	0
acyl carrier protein, chloroplast, putative / ACP, putative	-1	-1	-1	-1	-1	-1	-1	-1	-1
acyl carrier protein, mitochondrial / ACP / NADH-ubiquinone oxidoreductase 9.6 kDa subunit	0	0	-1	-1	-1	-1	0	0	-1
beta-hydroxyacyl-ACP dehydratase, putative	-1	-1	-1	0	0	0	-1	0	-1
beta-hydroxyacyl-ACP dehydratase, putative	-1	0	-1	0	0	0	-1	0	0

beta-ketoacyl-CoA synthase family (FIDDLEHEAD) (FDH)	0	0	-1	0	0	0	-1	-1	0
beta-ketoacyl-CoA synthase family protein	0	0	0	0	0	0	-1	-1	0
beta-ketoacyl-CoA synthase, putative	-1	-1	-1	0	-1	-1	-1	-1	-1
beta-ketoacyl-CoA synthase, putative	0	0	0	0	0	0	-1	-1	0
enoyl-[acyl-carrier protein] reductase [NADH], chloroplast, putative / NADH-dependent enoyl-ACP reductase, putative	-1	-1	-1	-1	-1	-1	-1	-1	-1
very-long-chain fatty acid condensing enzyme (CUT1)	0	0	0	-1	0	0	-1	-1	0
very-long-chain fatty acid condensing enzyme, putative	0	-1	-1	-1	-1	0	-1	-1	-1
very-long-chain fatty acid condensing enzyme, putative	-1	-1	-1	-1	-1	-1	-1	-1	-1

Table 5.6: Significance level of genes at individual timepoints and also from paired SAM related to fatty acid metabolism under salt stress. Notation used is same as that of table 5.2

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
3-ketoacyl-ACP synthase, putative	0	0	1	1	1	1	1	1	1
acyl-CoA dehydrogenase-related	0	0	0	0	0	0	1	1	0
acyl-CoA oxidase (ACX1)	1	1	1	1	1	1	1	1	1
acyl-CoA oxidase (ACX2)	1	1	1	1	0	0	1	1	1

A general conclusion can be drawn that under the stressed condition, when photosynthesis, the source of energy, is decreasing in plants, fatty acid degradation is possibly increasing to release energy which can be used for plant’s survival and to make possible changes in physiology for stress acclimation. There is no excess photosynthetic product which needs to be stored, hence fatty acid biosynthesis reactions are negatively significant.

Universal stress protein (USP)

Stress doesn’t necessarily create up-regulation of genes’ expression. In this analysis there were 2 genes encoding universal stress protein (USP) were found to be differentially expressed. One of them was positively significant all timepoints while the other was significantly down-regulated at all timepoints and also from paired SAM.

Table 5.7: Significance levels of Universal Stress Protein (USP) genes at individual timepoints and also from paired SAM under salt stress. Notation used is same as that of table 5.2

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	0	0	-1	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	0	0	1	1	0	0	1	0	0
universal stress protein (USP) family protein	-1	0	-1	0	0	-1	0	0	0
universal stress protein (USP) family protein	0	0	-1	0	0	0	-1	0	0
universal stress protein (USP) family protein	-1	-1	-1	0	-1	-1	0	0	0
universal stress protein (USP) family protein	0	0	0	0	0	0	0	1	0
universal stress protein (USP) family protein	1	0	1	0	1	0	0	0	0
universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
universal stress protein (USP) family protein	1	1	1	1	1	1	1	1	1

5.1.3 Clustering results of Arabidopsis thaliana liquid cultures subjected to trehalose stress

To study the response of trehalose, a non-reducing disaccharide involved in sugar partitioning and stress response in several organisms, on plant growth media experiments were conducted as explained in the experimental design section. For an effective comparison in this combined stress experiment with the other experiment of this project stresses were applied in the same fashion and samples were also harvested at the same timepoints. Table 5.8 shows the weight of the plant samples and the corresponding media pH measurements for control experiment (SC) and trehalose stress experiment (TC). Sample 17 and 2 were removed from analysis as I noticed bacterial contamination in these liquid cultures. From the table it is clear sample 18 or TC experiment is weight substantially lower than weights of the rest of the samples. The plants in this flask possibly didn't grow completely for some unknown reasons. When the samples were clustered based on gene expressions this sample was found to cluster separately implying significant difference in the physiology of this sample compared to the others. I believe this sample is an outlier and hence removed for further analysis. The timepoint 0h for TC experiment is represented by geometric mean of the expressions of sample 19 and 20.

Table 5.8: Weights and media pH measurements for harvested samples of control and trehalose stress experiments.

		Control (SC)		trehalose stress(TC)	
Time Pt.	Sample No	Weight	pH	Weight	pH
0	20	14	6.16	17.9	6.83
0	19	16.9	6.21	17.6	6.65
0	18	19	6.18	3.5	5.80
0	17	17.8	6.45	N.A.	N.A.
1	1	13.7	6.13	19.4	6.75
1	2	12.8	6.43	N.A.	N.A.
3	3	14.7	6.09	22.5	
3	4	16.3	6.23	20.9	6.79
6	5	9.2	6.32	21.5	6.70
6	6	15.1	6.3	23.2	7.06
9	7	18	6.35	22.9	6.78
9	8	21	6.24	23.3	6.63
12	9	12	6.28	24.3	6.83
12	10	14.5	6.4	23.4	6.73
18	11	22.9	6.3	21.2	7.06
18	12	21.9	6.36	19.3	6.76
24	13	21.8	6.27	29.2	6.83
24	14	20.1	6.45	20.7	6.91
30	15	28.3	6.51	28.7	6.70

30	16	30.6	6.4	19.9	6.75

Plants were immediately frozen in liquid nitrogen and kept at -80^{o}C until they were ground in liquid nitrogen. Whole plans were ground in liquid nitrogen and 2 grams of this sample was used for transcriptional profiling analysis. Please see the materials and methods section for details of the steps followed in the experiment.

5.1.3.1 Multivariate statistical analysis

Similar to the previous comparison analysis, before the analysis is started, a common repository of genes is selected. The selected 11416 genes have non-zero expression values for at least 12 out of 16 timepoints. PCA of the timepoints for control (SC) and trehalose stress (TC) experiments show that they are separated in reduced gene space (figure 5.18), implying that applied trehalose stress is causing a significant change in the transcriptional level. First 3 principal components contain 35, 28 and 11% of the information. When the three components were combined, around 75% of the variance is retained. Control timepoints were more spread out while the trehalose stress experimental timepoints are not, except for one timepoint.

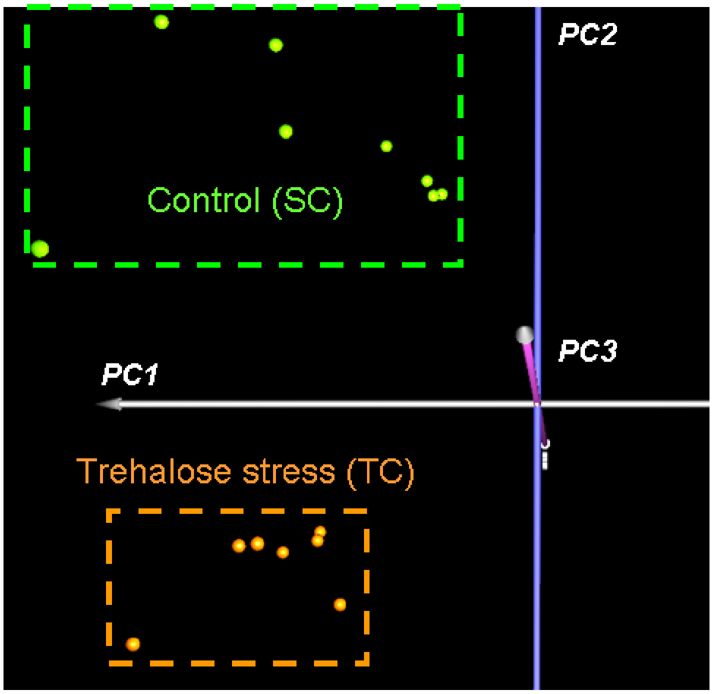


Figure 5.18 PCA analysis of the experimental timepoints of control and perturbed experiments. The experiments are separated in reduced gene space as trehalose stress is moving the timepoints along PC2.

Hierarchical clustering also shows a clear separation of the control and trehalose stress timepoints [Figure 5.19].

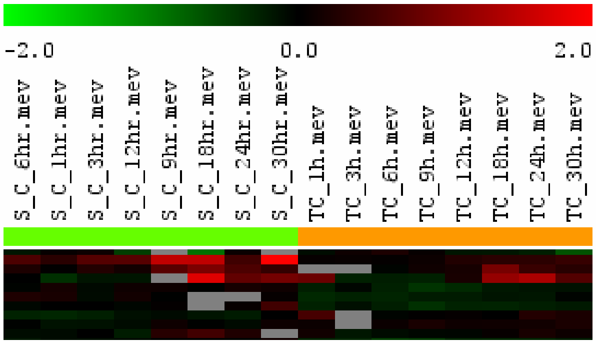


Figure 5.19 Hierarchical clustering of the samples shows two experimental groups form two distinct clusters.

Both paired SAM and MiTimeS were used for significance analysis based on overall and individual timepoints as explained earlier. Delta value of 2.14 was used for paired SAM, because at this significance level there was maximum number of significant genes with minimum possible false discovery rate. There were 709 and 885 genes found positively and negatively significant from paired SAM analysis with the delta value mentioned above, which are 6 and 8% of the genes used for analysis. These numbers are greater than that CO_2 stress response and smaller than NaCl stress response. In MiTimeS analysis when multiple test correction was used, 2.34 was selected as the delta value. At this delta, the combined FDR from all the timepoints was same as the FDR of the paired SAM analysis. Corrected delta value and the significant gene numbers at each timepoint were automatically calculated from the iterative algorithm of MiTimeS. The number of significant genes at each timepoint and also from paired SAM is shown in figure 5.20. The figure shows the significant gene numbers for individual timepoints is at the same range as that of paired SAM. At 1h timepoint, numbers of differentially expressed genes at both the significance categories are maximum, which implies trehalose is possibly creating a strong initial stress response, unlike previous stresses. It is also interesting to note, number of negatively significant genes are higher at most of the timepoints and also from paired SAM analysis.

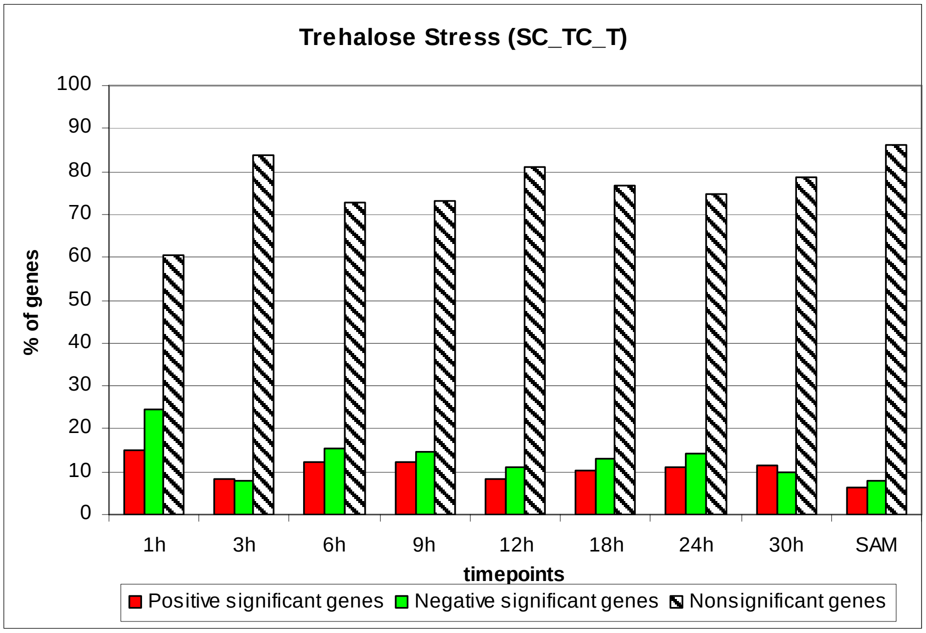


Figure 5.20 The bar diagram show the number of positively, negatively and non-significant genes at individual timepoints and also from paired SAM. Timepoint 1h has maximum number of positive and negative significant genes.

5.1.3.2 Data validation and interpretation in the context of plant physiology

Trehalose is a non-reducing disaccharide that occurs in a large range of organisms, such as bacteria, fungi, nematodes and crustaceans. In addition to its function as a storage carbohydrate and transport sugar, trehalose plays an important role in stress protection, especially during heat stress and dehydration [Wiemken, 1990; Crowe et al., 1998]. Trehalose has been shown to stabilise proteins and membranes under stress conditions, especially during desiccation. Furthermore, trehalose remains stable at elevated temperatures and at low pH and does not undergo Maillard browning with proteins. These protective properties of trehalose are clearly superior to those of other sugars, such as sucrose, making trehalose an ideal stress protectant [Wingler A., 2000].

The observation that most of the trehalose formed in Arabidopsis is simultaneously being degraded by trehalase raises the questions of the function of trehalose biosynthesis. The precursor of trehalose, trehalose-6-phosphate (T6P), prevents an uncontrolled influx of glucose into glycolysis. The synthesis of T6P may also play a role in the regulation of photosynthetic carbon metabolism. Similar to sucrose, trehalose induces enzymes involved in the accumulation of storage carbohydrates in photosynthetic tissues. In Arabidopsis, trehalose strongly induces the expression of ApL3, a gene encoding a large subunit of ADP-glucose pyrophosphorylase, which is an important enzyme in starch biosynthesis. This induction of ApL3 expression leads to increased ADP-glucose pyrophosphorylase activity, an overaccumulation of starch in the shoots and decreased root growth [Wingler et al., 2000; Fritzius et al., 2001].

Trehalose is produced from UDP-glucose in two steps. Trehalose-6P produced from UDP glucose is catalyzed by trehalose-6-phosphate synthase (TPS). Trehalose-6-phosphate phosphatase (TPP) catalyses the subsequent reaction of trehalose production. Functional genes encoding enzymes of trehalose synthesis, i.e. TPS and TPP, have been identified in Arabidopsis [Blazquez et al., 1998; Vogel et al., 1998].

Trehalase activity normally keeps cellular trehalose concentrations low in order to prevent detrimental effects of trehalose accumulation on the regulation of carbon metabolism. Such a role of trehalase may be of particular importance in interactions of plants with trehalose-producing micro-organisms. In support of this hypothesis, expression of the Arabidopsis trehalase gene and trehalase activity were found to be strongly induced by infection of Arabidopsis plants with the trehalose-producing pathogen Plasmodiophora brassicae [Brodmann et al., 2002].

In accordance with the previous observations, it was found the trehalase gene is positively significant at all the timepoints and also from paired SAM analysis under trehalose stress. Hence, external trehalose added to the media follows the same response as that of trehalose-producing pathogen. The increased trehalose concentration might be affecting the sugar partitioning, hence need to be degraded in order maintain homeostasis. Among the other genes of this pathway one of the trehalose-6-phosphate phosphatase genes (At4g22590) was also positively significant from paired SAM analysis. There were couples of other trehalose-6-phosphate phosphatase genes differentially expressed at only few timepoints. Significance level of these genes at individual timepoints and from paired SAM are shown in table 5.9

Table 5.9: Table shows the significance level of genes encoding enzymes in trehalose synthesis and degradation pathway under trehalose stress. Notation used is same as that of table 5.2. Gene coding for trehalose, the enzyme that catalyses trehalose degradation was positively significant at all timepoints.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
glycosyl hydrolase family protein 37 / trehalase, putative	1	1	1	1	1	1	1	1	1
trehalose-6-phosphate phosphatase (TPPA)	0	0	0	0	0	0	0	1	0
trehalose-6-phosphate phosphatase (TPPB)	0	1	0	0	0	0	0	1	0
trehalose-6-phosphate phosphatase, putative	-1	-1	0	0	-1	0	0	0	0
trehalose-6-phosphate phosphatase, putative	0	0	0	0	0	0	0	0	0
trehalose-6-phosphate phosphatase, putative	1	0	1	1	0	1	1	1	1
trehalose-6-phosphate synthase, putative	0	0	-1	-1	0	-1	0	0	0
ADP-glucose pyrophosphorylase family protein	0	0	0	0	0	0	0	-1	0

Calvin cycle, starch and sucrose production

As explained before CO_2 fixation in Calvin cycle is catalyzed by Rubisco, which comprises two subunits, small (*rbcS*) and large (*rbcL*), which are encoded by nuclear and chloroplast genes respectively. The *rbcL* gene is positively significant at three timepoints from 3 to 9h (figure 5.21). Among the four *rbcS* gene family only 1A and 1B were found from analysis and the other two subunits were missing. Both the genes were negatively significant at 24 and30h timepoints similar to CO_2 and NaCl stress response. Rubisco small subunit 1A gene was also down-regulated at 1 and 9h. Rubisco activase gene was also missing in the analysis. Two gene encoding phosphoglycerate kinase, the enzyme catalyzing the conversion of 3PG to 1,3-bis-phosphoglycerate, is negatively significant at 1h, similar to Rubisco small subunit 1A. Other genes were down-regulated at 6 and 18h timepoints (figure 5.21).

Triose-phosphates transported from the chloroplasts to the cytoplasm are converted to fructose-1,6-bisphosphate by fuctose-bisphosphate aldolase. Though the chloroplast copy of this gene was negatively significant, but cytoplasmic gene was positively significant from paired SAM analysis. This underscores the importance of analyzing gene expression in the context of cellular components.

The genes encoding Phosphoglucomutase and UDP-glucose pyrophosphorylase, enzymes involved in production of UDP-glucose from glucose-6-phosphate, is significantly under-expressed at 6-18h of perturbation (Figure 5.21). One of the SPS genes was negatively significant at only 6h timepoint while the other one was differentially expressed at none of the timepoints. SPS, which is potentially the main regulatory enzyme and activated by glucose-6P and inhibited by inorganic phosphate [Buchanan et al., 2001]. One of the sucrose phsosphate phosphatase genes corresponding to the next reaction follows exactly same significance profile as that of SPS, as it becomes negatively significant at 6h. However, the other sucrose phsosphate phosphatase gene was positively significant from 9-24h period. The differential response of the sucrose phsosphate phosphatase genes from this and also from other experiments raises the question which one is actually regulating the processes. SS gene is considered to be mainly active in the breakdown of sucrose was positively significant [Buchanan et al., 2001] at all timepoints and also from paired-SAM. The other copy of SS gene was non-significant at all the timepoints and hence from paired SAM. This again shows the need to identify the right gene from analysis and incomplete nature of the current annotation. It was found before that sucrose synthase and invertase activities are affected by trehalose in soybeans [Muller et al., 1998]. Here I see the similarity in *A. thaliana* as well. Four genes encoding intervase At4g25250, At5g51520, At5g62340 and At5g64620 were also found negatively significant from paired SAM in this experiment.

Starch, produced in chloroplast, and also serves as a transient sink to accommodate excess photosynthate that cannot be converted to sucrose and exported (Figure 5.21) [Lea and Leegood, 1993]. Hence, when sucrose synthesis is restricted, starch synthesis is promoted. ADP-glucose pyrophosphorylase (AGPase) is a key enzyme catalyzing ADP-glucose formation and genes encoding this enzyme was found negatively significant at 1 and 6h timepoints. This result is inconsistent with previous finding of increase in ADP-glucose pyrophosphorylase gene under trehalose stress. However, the gene encoding starch synthase was found positively significant at 5 out of 8 timepoints. Starch synthase catalyses the starch production from ADP glucose and this gene’s over-expression is possibly increasing the starch synthesis, as expected in case of trehalose stress. Other genes involved in starch degradation were mostly non-significant.

In conclusion, the rate of carbon fixation decreases especially at the later stage of the experiment. Sucrose synthase gene is over-expressed leading to possible degradation of sucrose and starch synthesis is also increasing, both being consistent with past results.

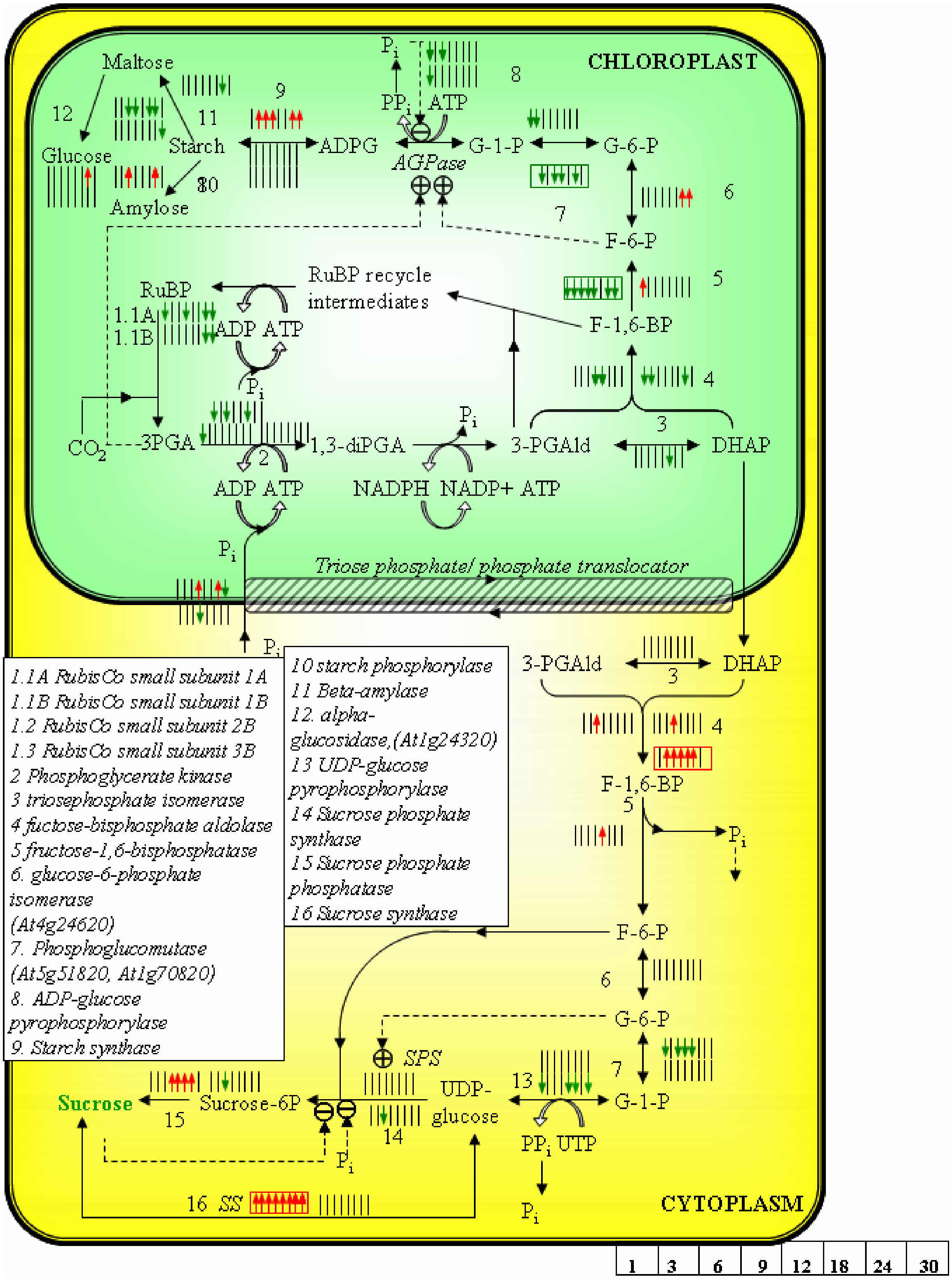

CHLOROPLAST
Maltose
12
Glucose
11
Starch
9
ADPG
10
Amylose
P_i
PP_i ATP
8
AGPase
G-1-P
G-6-P
7
6
F-6-P
RuBP
RuBP recycle intermediates
1.1A
1.1B
ADP ATP
P_i
F-1,6-BP
5
4
3
CO_2
3PGA
2
1,3-diPGA
3-PGAld
DHAP
ADP ATP
NADPH NADP+ ATP
P_i
Triose phosphate/ phosphate translocator
P_i
1.1A RubisCo small subunit 1A
1.1B RubisCo small subunit 1B
1.2 RubisCo small subunit 2B
1.3 RubisCo small subunit 3B
2 Phosphoglycerate kinase
3 triosephosphate isomerase
4 fuctose-bisphosphate aldolase
5 fructose-1,6-bisphosphatase
6. glucose-6-phosphate isomerase (At4g24620)
7. Phosphoglucomutase (At5g51820, At1g70820)
8. ADP-glucose pyrophosphorylase
9. Starch synthase
10 starch phosphorylase
11 Beta-amylase
12. alpha-glucosidase,(At1g24320)
13 UDP-glucose pyrophosphorylase
14 Sucrose phosphate synthase
15 Sucrose phosphate phosphatase
16 Sucrose synthase
3-PGAld
3
DHAP
4
F-1,6-BP
5
P_i
F-6-P
6
G-6-P
SPS
7
13
G-1-P
UDP-glucose
14
PP_i UTP
P_i
Sucrose
15
Sucrose-6P
P_i
16 SS
CYTOPLASM
1
3
6
9
12
18
24
30

Figure 5.21 Observed effect of the applied trehalose perturbation on the physiology of Calvin cycle, starch and sucrose biosynthesis pathways at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Sucrose degradation is increasing with a possible increase in starch synthesis.

Photorespiration

Most of the photorespiratory genes were found negatively significant. Some of them were negatively significant from paired SAM as well showing strong down regulation over the 30h of the experiment. Figure 5.22 shows the significance level of the genes involved in the photorespiration pathway at individual timepoints and from paired SAM analysis.

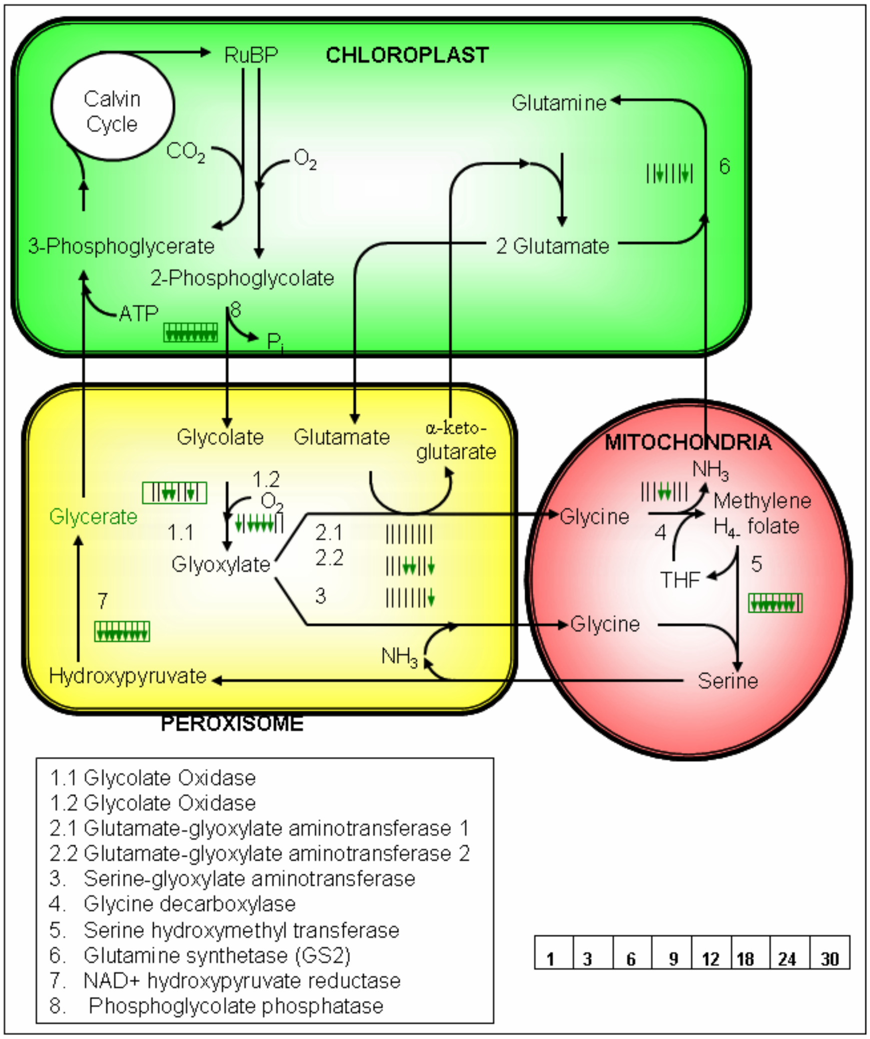


Figure 5.22 Observed effect of the applied trehalose perturbation on the physiology of photorespiration pathways at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Photorespiratory pathway was negatively significant as it was in other stress responses.

Nitrogen assimilation and amino-acid biosynthesis

Apart from CO_2 stress response, only under the trehalose stress NR1 gene is positively significant (Fgiure 5.23). Nitrate reductase 2 (NR2) is positively significant at all the timepoints and also from paired SAM. Being consistent with NR1 gene, nitrite reductase and ferredoxin-dependent glutamate synthase (Fd-GOGAT 1) genes are positively significant 30h timepoint. This provides a compelling indication that nitrogen assimilation is increasing under trehalose stress. GS2 is also a part of nitrogen assimilation through GOGAT mechanism is not showing similar significance profile and the gene is negatively significant at 3 and 24h timepoints (figure 5.23). Recycle of NH_4^+ from photorespiration is 5-10 times more than NH_4^+ assimilated from nitrate reduction. In spite of strong down-regulation of photorespiratory pathway GS2 gene was down regulated at only two timepoints. In the overall effect increase in NH_4^+ assimilation from nitrate reduction is more than compensated by decrease in NH_4^+ release from photorespiration, hence we see a moderate decrease in GS2 gene.

TCA cycle flux is possibly also going towards production of aspartate and other amino acids for which it is a precursor. This becomes evident as we see aspartate aminotransferase gene is positively significant at half of the timepoints. This is indeed true as glutamate decarboxylase gene was also positively significant at 2 of the timepoints. The flux towards threonine production is possibly decreasing as both the genes encoding threonine synthase are under-expressed, one of them is negatively significant from paired SAM as well. The flux towards methionine production is possibly increasing as methionine synthase gene is positively significant at two of the timepoints and negatively significant in none. Similarity of CO_2 stress response and trehalose stress response in the context of nitrogen assimilation is quite perceptible. In both the stresses nitrate reduction was increasing coupled with rise in flux towards beta-alanine and methionine production pathway.

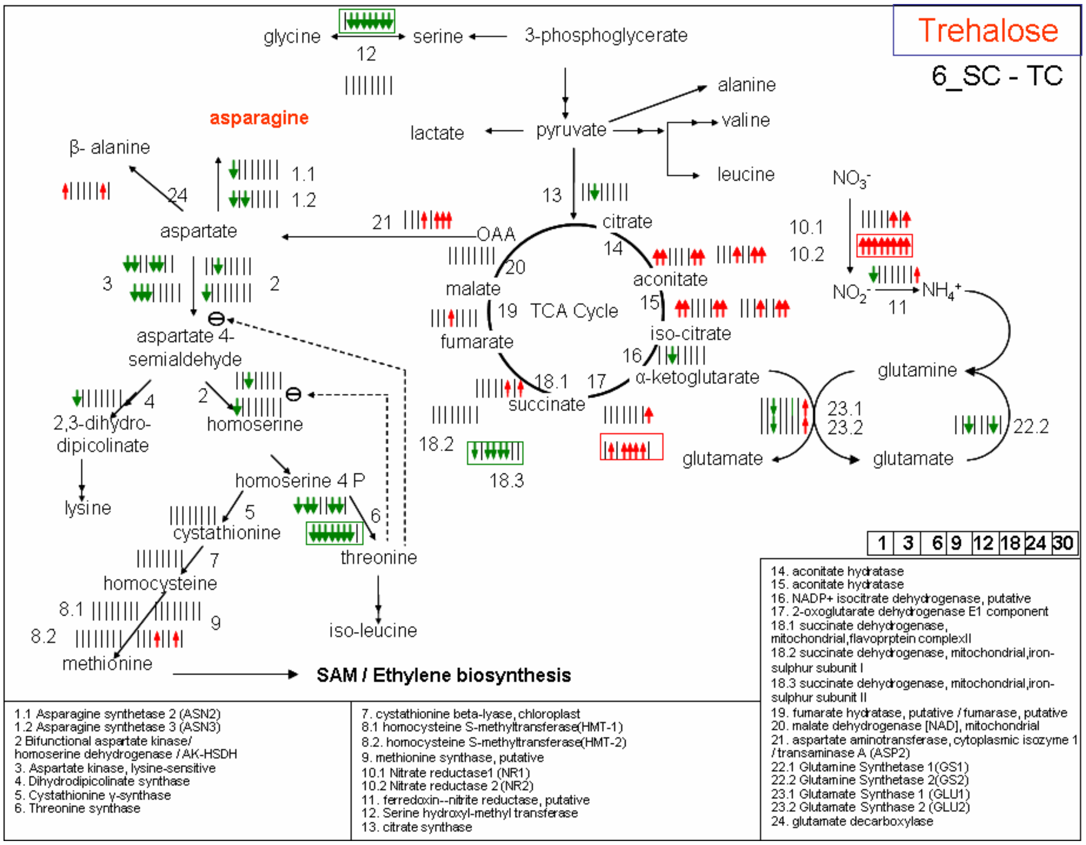

Trehalose
6_SC - TC
glycine
12
serine
3-phosphoglycerate
alanine
pyruvate
valine
leucine
lactate
asparagine
β- alanine
24
1.1
1.2
aspartate
21
OAA
13
citrate
14
aconitate
15
iso-citrate
16
α-ketoglutarate
17
succinate
18.1
18.2
18.3
fumarate
19
TCA Cycle
malate
20
NO3-
10.1
10.2
NO2-
11
NH4+
glutamine
23.1
23.2
22.2
glutamate
glutamate
3
2
aspartate 4-semialdehyde
4
2,3-dihydro-dipicolinate
lysine
2
homoserine
homoserine 4 P
5
6
cystathionine
threonine
iso-leucine
7
homocysteine
8.1
8.2
9
methionine
SAM / Ethylene biosynthesis
1 3 6 9 12 18 24 30
14. aconitate hydratase
15. aconitate hydratase
16. NADP+ isocitrate dehydrogenase, putative
17. 2-oxoglutarate dehydrogenase E1 component
18.1 succinate dehydrogenase, mitochondrial,flavoprptein complexII
18.2 succinate dehydrogenase, mitochondrial,iron-sulphur subunit I
18.3 succinate dehydrogenase, mitochondrial,iron-sulphur subunit II
19. fumarate hydratase, putative / fumarase, putative
20. malate dehydrogenase [NAD], mitochondrial
21. aspartate aminotransferase, cytoplasmic isozyme 1 / transaminase A (ASP2)
22.1 Glutamine Synthetase 1(GS1)
22.2 Glutamine Synthetase 2(GS2)
23.1 Glutamate Synthase 1 (GLU1)
23.2 Glutamate Synthase 2 (GLU2)
24. glutamate decarboxylase
1.1 Asparagine synthetase 2 (ASN2)
1.2 Asparagine synthetase 3 (ASN3)
2 Bifunctional aspartate kinase/ homoserine dehydrogenase / AK-HSDH
3. Aspartate kinase, lysine-sensitive
4. Dihydrodipicolinate synthase
5. Cystathionine γ-synthase
6. Threonine synthase
7. cystathionine beta-lyase, chloroplast
8.1 homocysteine S-methyltransferase(HMT-1)
8.2. homocysteine S-methyltransferase(HMT-2)
9. methionine synthase, putative
10.1 Nitrate reductase1 (NR1)
10.2 Nitrate reductase 2 (NR2)
11. ferredoxin--nitrite reductase, putative
12. Serine hydroxyl-methyl transferase
13. citrate synthase

Figure 5.23 Observed effect of the applied trehalose stress on Nitrogen assimilation and amino-acid biosynthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Nitrogen assimilation is possibly increasing with increase in some of the amino-acids biosynthesis flux, unlike NaCl stress response where the amino-acid biosynthesis reactions were going down.

Similar to the other stress response pathways, genes involved in this pathway are mostly up-regulated. All the genes except two in the pathway starting from Chorismate to tryptophan are positively significant at 6 or 8 timepoints and also from paired SAM (figure 5.24). Anthranilate phosphoribosyltransferase and indole-3-glycerol-phosphate synthase genes, which are not up-regulated in other stress responses were positively significant here (figure 5.24). Most up-regulation of gene belonging to this pathway was obtained under trehalose stress. Overall increase in nitrogen assimilation is possibly allowing a greater flux towards tryptophan biosynthesis.

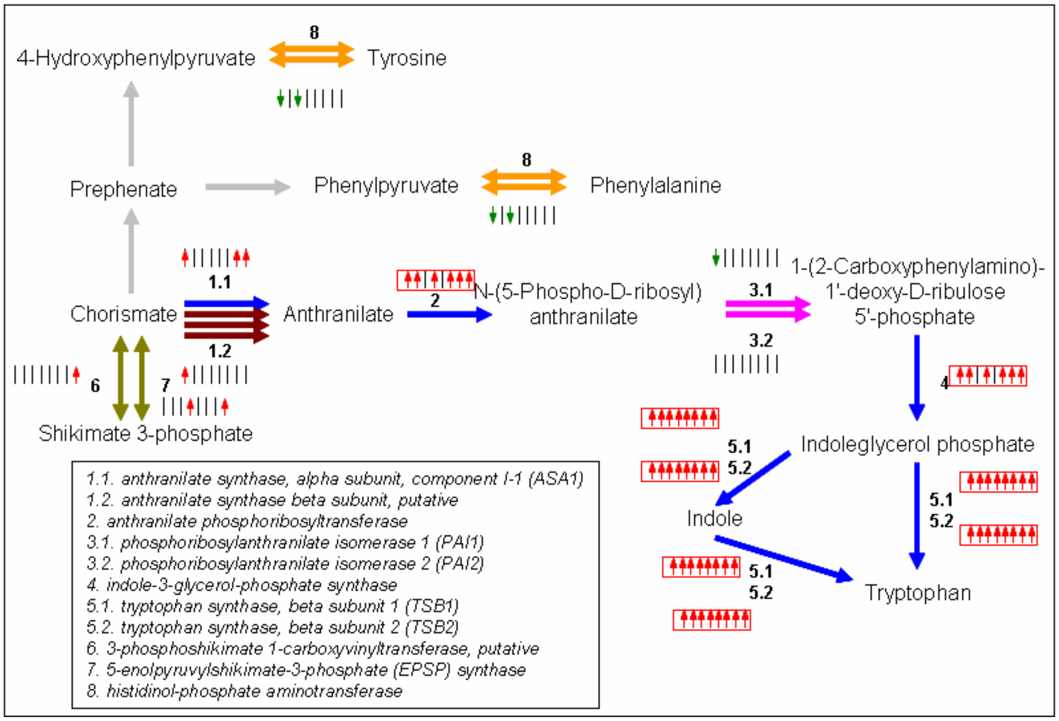

4-Hydroxyphenylpyruvate
8
Tyrosine
Prephenate
Phenylpyruvate
8
Phenylalanine
1.1
Chorismate
1.2
Anthranilate
2
N-(5-Phospho-D-ribosyl) anthranilate
3.1
3.2
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5'-phosphate
6
7
Shikimate 3-phosphate
4
5.1
5.2
Indoleglycerol phosphate
Indole
5.1
5.2
5.1
5.2
Tryptophan
1.1. anthranilate synthase, alpha subunit, component I-1 (ASA1)
1.2. anthranilate synthase beta subunit, putative
2. anthranilate phosphoribosyltransferase
3.1. phosphoribosylanthranilate isomerase 1 (PAI1)
3.2. phosphoribosylanthranilate isomerase 2 (PAI2)
4. indole-3-glycerol-phosphate synthase
5.1. tryptophan synthase, beta subunit 1 (TSB1)
5.2. tryptophan synthase, beta subunit 2 (TSB2)
6. 3-phosphoshikimate 1-carboxyvinyltransferase, putative
7. 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
8. histidinol-phosphate aminotransferase

Figure 5.24 Observed transcriptional response of the trehalose stress on tryptophan synthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. The rate of tryptophan biosynthesis is possibly increasing, similar to other stress responses.

Ethylene Biosynthesis and Signaling

Under the trehalose stress condition we see an interesting phenomenon in ethylene biosynthesis pathway. One of the SAM synthase and two ACC oxidase genes are positively significant at almost all the timepoints and also from paired SAM, implying a possible increase in ethylene biosynthesis. However, ACC synthase, putative gene (At4g08040) was negatively significant at 4 out of 8 timepoints, while the ACC synthase 2 (ACS2) gene (At1g01480) was positively significant only at 30h. ETR2 gene which is activated in presence of ethylene was positively significant, also indicating and increase in ethylene production. From all the three observations taken together no convincing conclusion could be derived about response of ethylene biosynthesis pathway under trehalose stress.

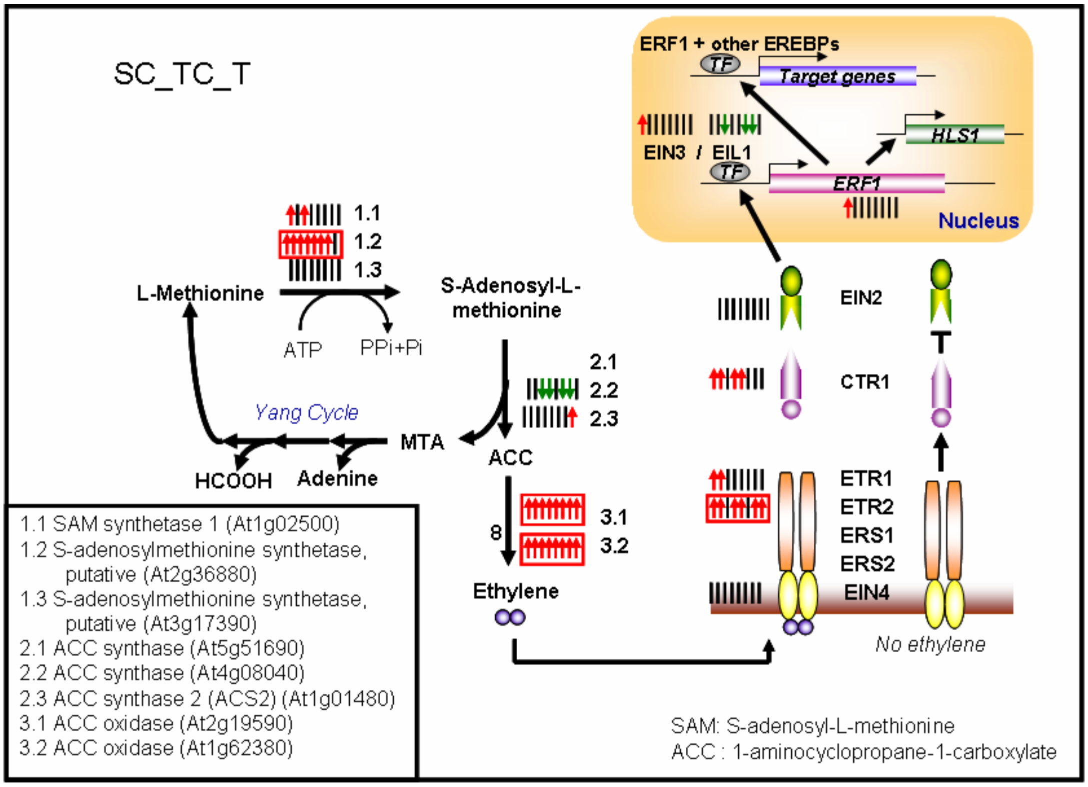

SC_TC_T
ERF1 + other EREBPs
TF
Target genes
EIN3 / EIL1
TF
HLS1
ERF1
Nucleus
1.1
1.2
1.3
L-Methionine
S-Adenosyl-L-methionine
ATP
PPi+Pi
2.1
2.2
2.3
Yang Cycle
MTA
ACC
HCOOH
Adenine
8
3.1
3.2
Ethylene
EIN2
CTR1
ETR1
ETR2
ERS1
ERS2
EIN4
No ethylene
1.1 SAM synthetase 1 (At1g02500)
1.2 S-adenosylmethionine synthetase, putative (At2g36880)
1.3 S-adenosylmethionine synthetase, putative (At3g17390)
2.1 ACC synthase (At5g51690)
2.2 ACC synthase (At4g08040)
2.3 ACC synthase 2 (ACS2) (At1g01480)
3.1 ACC oxidase (At2g19590)
3.2 ACC oxidase (At1g62380)
SAM: S-adenosyl-L-methionine
ACC : 1-aminocyclopropane-1-carboxylate

Figure 5.25 Observed effect of the trehalose stress on ethylene biosynthesis and signaling at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Both ERT1 and CTR1 genes were positively significant at first 2 timepoints (figure 5.25). Though CTR1 gene was positively significant at 4 timepoints, but EIN2 gene in downstream of CTR1 was non-significant all timepoints. No positive or negative correlation was observed between CTR1 and EIN2 genes from previous comparisons as well, which is feasible if the regulation is not taking place at the transcriptional level. EIN3 and ERF1 genes also show similar significance profile of being positively significant at 1h timepoint (figure 5.25). Theses genes have shown similar profiles under NaCl stress and NaCl and CO_2 combined stress as well.

Significance levels of different EREBP genes, which are affected by ethylene response are shown in table 5.10. Unlike NaCl and combined response most of the genes' significance level is not changing. The only similarity with previous stress comparisons is most of the genes are negatively significant at 9h timepoint. This gives some indication that ethylene synthesis is not increasing though number of genes of this pathway are significantly over-expressed.

Table 5.10 Significance level of ethylene-responsive element-binding family genes under trehalose stress. Overall stress response these genes were not significant as it was in case of NaCl or NaCl and CO_2 stress. Negative response was observed predominantly at 9h timepoint.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
ethylene-responsive element-binding factor 4 (ERF4)	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding family protein	0	0	0	-1	0	0	0	-1	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	0	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	0	0	0	0
ethylene-responsive element-binding family protein	1	0	0	0	1	0	1	0	0
ethylene-responsive element-binding protein, putative	0	0	0	0	0	0	0	0	0

5.2 Combined Stress Responses

5.2.1 Transcriptional response of Arabidopsis thaliana liquid cultures subjected to NaCl and CO_2 stress

I have discussed the transcriptional response of NaCl and CO_2 stress applied individually. Now I will discuss the response when the stresses are applied in combination. For an effective comparison in this combined stress experiment the strength of NaCl and CO_2 stresses was same as that when applied individually. Samples were also harvested at the same timepoints. The combine stress was found to create physiological change to the culture within first 30 hours of its application. Table 5.11 shows the weight of the plant samples and the corresponding media pH measurements for control experiment (SC) and combined stress experiment (NP).

Table 5.11 Weights and media pH readings of plant samples harvested in control and NaCl with CO_2 stress.

		Sucrose control		NaCl and CO_2 stress (NP)	
Time Pt.	Sample No	Weight	pH	Weight	pH
0	20	14	6.16	21.3	6.73
0	19	16.9	6.21	22.0	6.88
0	18	19	6.18	20.8	6.60
0	17	17.8	6.45	16.8	6.60
1	1	13.7	6.13	21.3	6.58
1	2	12.8	6.43	19.3	6.60
3	3	14.7	6.09	18.7	6.40
3	4	16.3	6.23	21.4	6.59
6	5	9.2	6.32	21.7	6.56
6	6	15.1	6.3	23.2	6.67
9	7	18	6.35	18.8	6.52
9	8	21	6.24	19.7	6.48
12	9	12	6.28	21.5	6.44
12	10	14.5	6.4	19.7	6.50
18	11	22.9	6.3	25.3	6.44
18	12	21.9	6.36	20.3	6.47
24	13	21.8	6.27	22.3	6.48
24	14	20.1	6.45	23.7	6.58
30	15	28.3	6.51	27.3	6.75
30	16	30.6	6.4	27.6	6.64

Plants were immediately frozen in liquid nitrogen and kept at -80oC until they were ground in liquid nitrogen. During hand grinding of the frozen plants in liquid nitrogen it was experienced that these plants are much easier to grind compared to frozen plants like the NaCl stress experiment. This is only a qualitative observation and can not be quantified; nevertheless it is an important observation as it implies there have been some physiological changes.

5.2.1.1 Multivariate statistical analysis

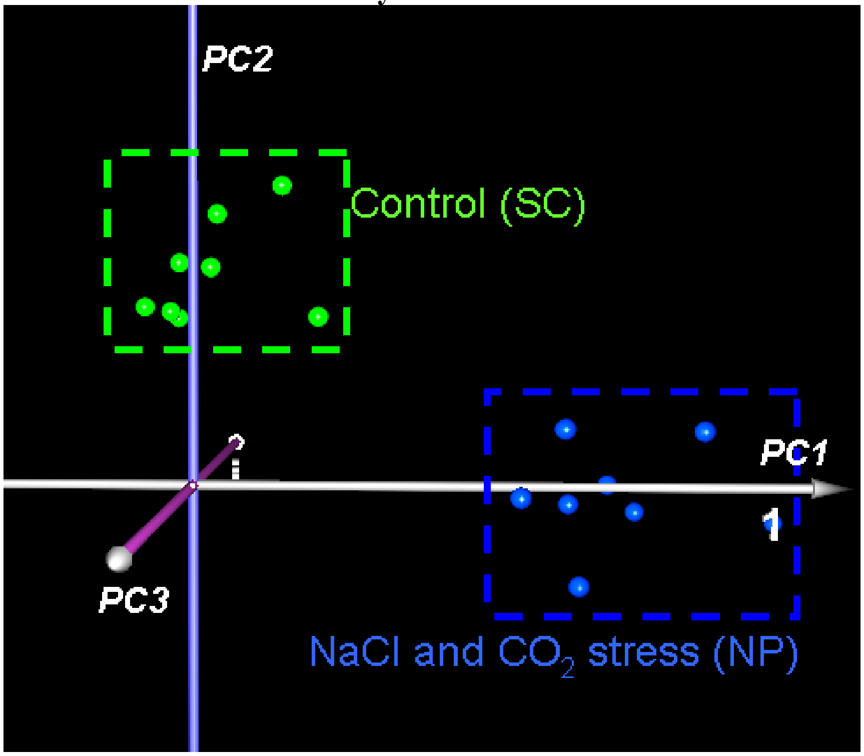


Figure 5.26 PCA analysis shows the control and NaCl with CO_2 stress experimental timepoints on reduced gene space. The timepoints are clearly separated implying the combined stress is producing a significant change in *A. thaliana* physiology. Combined stress is moving the timepoints towards PC1 which accounts for maximum variance in the data.

Similar to the previous stress analysis first a common repository of genes is selected. The selected 11080 genes have non-zero expression values for at least 12 out of 16 timepoints. From the Principal Component Analysis (PCA), the control transcriptomic profiles can be clearly differentiated from their perturbed counterparts (Figure 5.26). This implies that the physiology of the plant liquid cultures is affected by the applied perturbation at transcriptional level, even during the first 30h of treatment. First 3 principal components were found to capture 60, 16 and 6% of the information, interestingly these numbers are same as the of NaCl stress comparison. Hence, when the experiments are viewed at 3-D space it can account for most of the variance (82%). It can also be seen due to the application of NaCl stress timepoints have moved along principal component 1, which accounts for maximum variability.

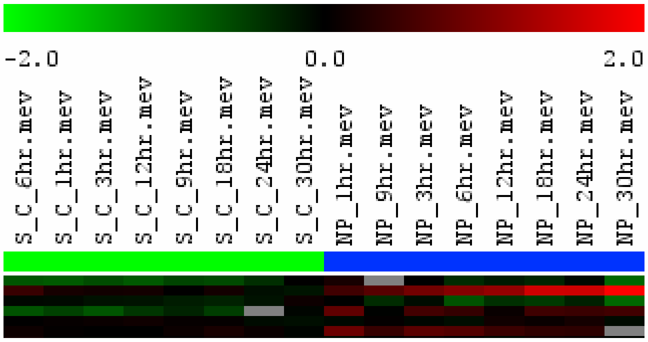


Figure 5.27 Hierarchical clustering of the samples using Pearsons' correlation distance shows a clear separation as it was in PCA analysis. This implies salt stress is causing a significant change in the transcriptional level of *A. thaliana*.

Experimental timepoints were also clustered using hierarchical clustering and it also shows a clear separation between them (Figure 5.27).

Both paired SAM and MiTimeS were used for significance analysis based on overall and individual timepoints as explained in the previous stress experiments. Delta value of 2.4 was selected for paired SAM, as this delta value has highest number of significant genes with minimum (0 in this case) FDR.

There were 1729 and 1616 genes found positively and negatively significant from paired SAM with this delta value, which constitutes around 16% and 15% of genes respectively used for analysis (11080). The delta value used for MiTimS analysis was same as that of paired SAM. Number of genes positively, negatively and non-significant at individual timepoints obtained from MiTimeS was plotted with that of paired SAM results in figure 5.28. All the time-points including paired SAM shows almost equal number of significant genes of both types. It is clear 6 and 18h timepoints have maximum and minimum number of significant genes respectively. The profile of number of significant genes with time is very similar to that of NaCl stress response.

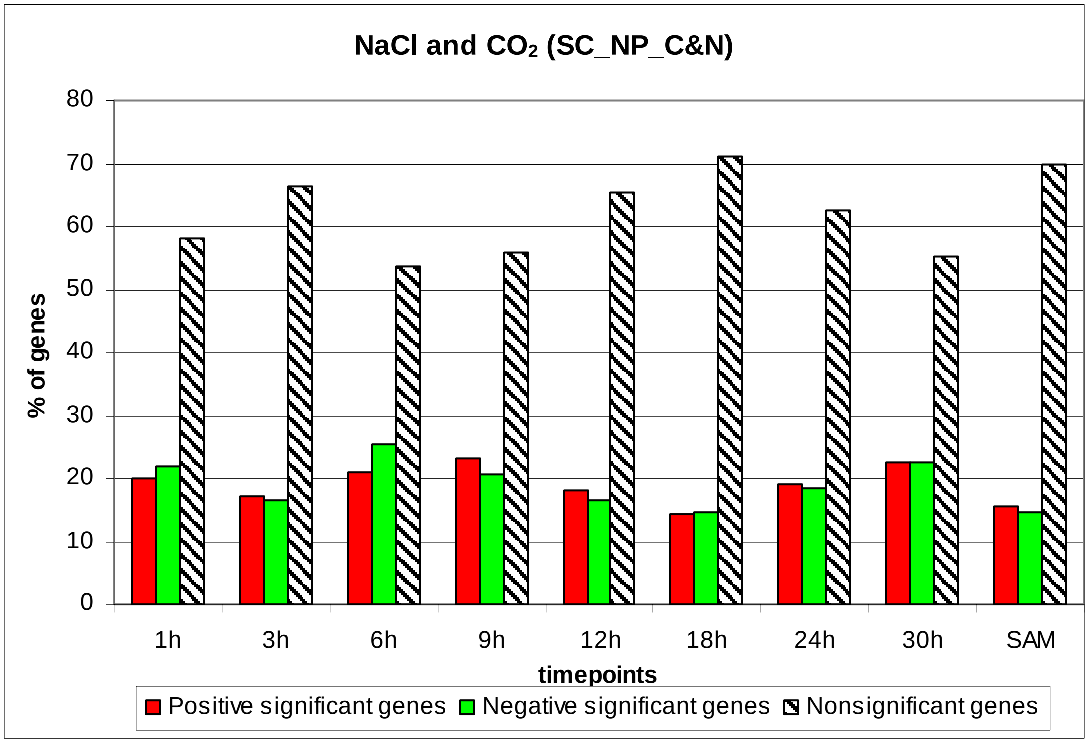


Figure 5.28 Percentage of positively, negatively and non-significant genes at individual timepoints and from paired SAM. Significant gene numbers were comparable between timepoints and also with paired SAM.

5.2.1.2 Data validation and interpretation in the context of plant physiology

As a result of genetic, molecular and biochemical analysis combined salt and CO_2 stress response pathway is well studied. Figure 5.12 shows a schematic diagram of the salt stress response pathway, also called SOS pathway. Calcium signal is induced by salt stress which is further sensed by calcium binding protein SOS3. SOS3 interacts and activated SOS2 which is a serine threonine protein kinase. SOS1 is a salt tolerance effector gene encoding a plasma membrane Na+/H+ antiporter is regulated be combined activity of SOS2 and SOS3 [Zhu JK, 2000] . SOS1 gene regulates transcriptionally and post-transcriptionally expression of other genes related to salt tolerance [Shi et al, 2000]. SOS pathway may also be involved in stimulation or suppression the activities of other transporters involved in ion homeostasis under salt stress, such as vacuolar H+-ATPases and pyrophosphatases (PPase), vacuolar Na+/H+ exchanger (NHX), and plasma membrane K+ and Na+ transporters [Zhu JK, 2002] .

From the hypothesis testing results of this experiment it was observed SOS3 gene (sodium proton exchanger, putative (NHX7)) was positively significant at 9h and 18-30h timepoints, similar to NaCl stress response. Information about SOS1 and SOS2 gene was missing in this analysis. Responses of several other sodium proton exchanger genes (NHX) were also studied. Among them NHX2, NHX3 and NHX6 genes were positively significant at most of the timepoints and also from paired SAM. However other NHX genes were non-significant from paired SAM. Several genes (At4g19960, At2g30070, At4g13420, At1g70300, At2g26650, At4g23640) coding for potassium transporter were found to be positively significant and paired SAM analysis, again congruous with the results from the literature. It was also observed Na+/Ca2+ antiporter gene (At2g47600) is positively significant at 1-6h and 12h timepoints and also from paired SAM analysis. In general it was observed that SOS pathway responses to NaCl stress and the combines stress are very similar. Hence this pathway of salt stress response is conserved whether or not CO_2 stress is applied in conjunction.

Calvin cycle, sucrose and starch production pathway

In the combined stress response of NaCl and CO_2 most of the Calvin cycle pathway genes are strongly under-expressed. NaCl and CO_2 stress responses applied individually causes a down regulation of RuBisCO and other genes in chloroplast mostly at last two timepoints [Figure 5.5 and 5.13]. In case of combined stress response all the four sub-units of RuBisCO genes are negatively significant from paired SAM analysis (figure 5.29). In the following reaction photosynthetic product 3-PGA is converted to 1,3-diPGA. This reaction is catalyzed by phosphoglycerate kinase which was also found under-produced. Genes encoding other chloroplast reactions catalyzed by fructose-1,6-bisphosphatase and Phosphoglucomutase are also negatively significant from paired SAM analysis. Starch synthase genes, involved in starch production from ADP-glucose are non-significant at most of the timepoints and also from paired SAM analysis. One of the beta-amylase genes were positively significant from paired SAM analysis implying starch degradation is increasing under this combined stress condition. This beta amylase gene was also positively significant from NaCl stress response, but only at few timepoints. In case of combined stress the cumulative response was higher and the expression level crossed the threshold of significance for paired SAM analysis. Sucrose is possibly over-produced as sucrose phosphate synthase is gene is over-expressed. However unlike NaCl stress, sucrose phosphate phosphatase and sucrose synthase genes are non-significant from paired SAM analysis implying sucrose production is increasing, but not as much as it was in case of NaCl stress alone

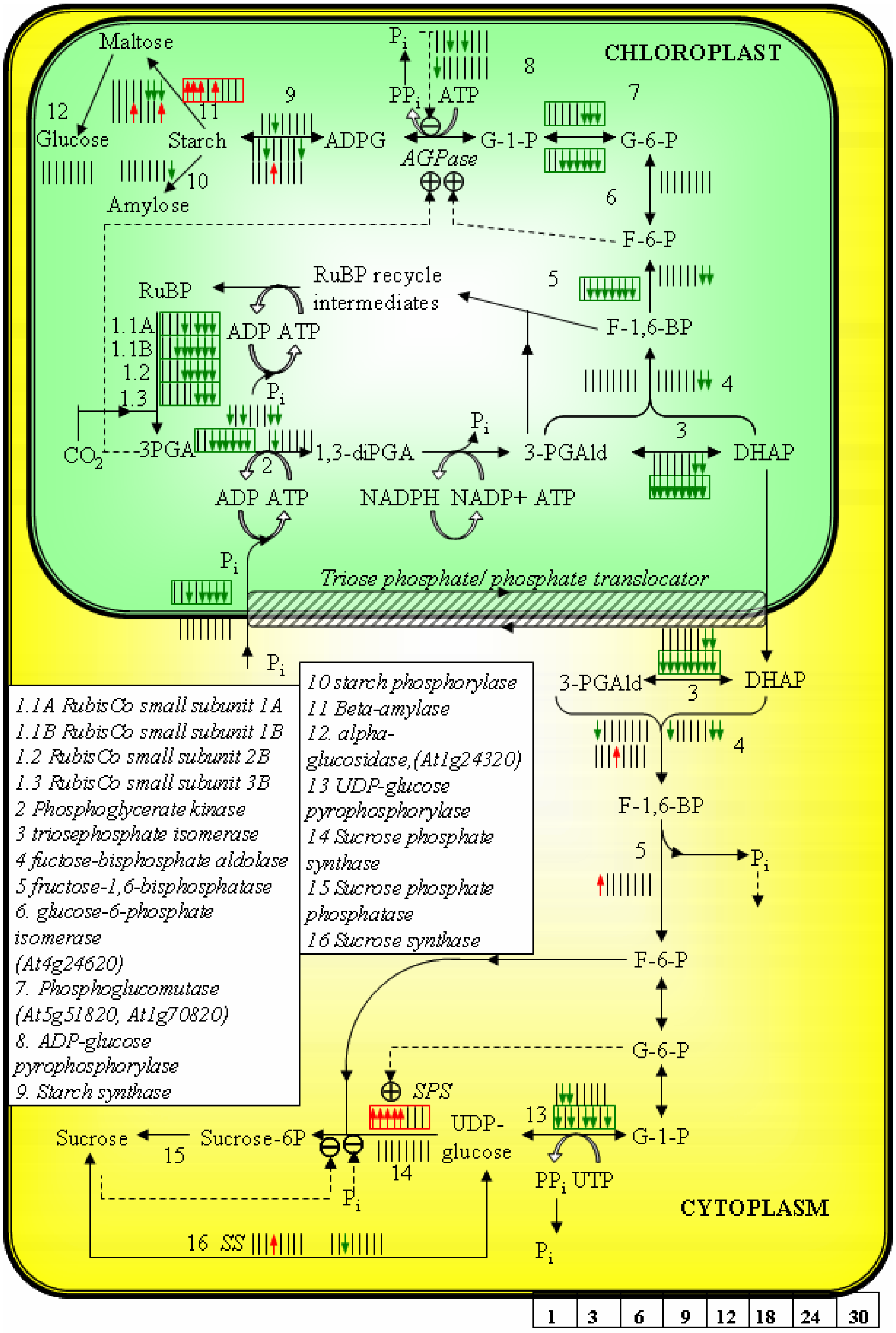

CHLOROPLAST
Maltose
12
Glucose
11
Starch
9
ADPG
P_i
PP_i
ATP
8
AGPase
G-1-P
7
G-6-P
10
Amylose
6
F-6-P
RuBP
RuBP recycle intermediates
5
1.1A
1.1B
1.2
1.3
ADP ATP
P_i
F-1,6-BP
4
CO_2
-3PGA
2
1,3-diPGA
P_i
3-PGAld
3
DHAP
ADP ATP
NADPH NADP+ ATP
P_i
Triose phosphate/ phosphate translocator
P_i
3-PGAld
3
DHAP
4
F-1,6-BP
5
P_i
F-6-P
G-6-P
SPS
13
UDP-glucose
G-1-P
Sucrose
15
Sucrose-6P
14
P_i
PP_i UTP
P_i
16 SS
CYTOPLASM
1.1A RubisCo small subunit 1A
1.1B RubisCo small subunit 1B
1.2 RubisCo small subunit 2B
1.3 RubisCo small subunit 3B
2 Phosphoglycerate kinase
3 triosephosphate isomerase
4 fuctose-bisphosphate aldolase
5 fructose-1,6-bisphosphatase
6. glucose-6-phosphate isomerase (At4g24620)
7. Phosphoglucomutase (At5g51820, At1g70820)
8. ADP-glucose pyrophosphorylase
9. Starch synthase
10 starch phosphorylase
11 Beta-amylase
12. alpha-glucosidase,(At1g24320)
13 UDP-glucose pyrophosphorylase
14 Sucrose phosphate synthase
15 Sucrose phosphate phosphatase
16 Sucrose synthase
1
3
6
9
12
18
24
30

Figure 5.29 Observed effect of the applied perturbation on the physiology of Calvin cycle, starch and sucrose biosynthesis pathways at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Rubisco genes were negatively significant from paired SAM analysis implying a reduction in carbon fixation. Gene encoding SPS, the regulating enzyme in sucrose synthesis pathway was positively significant along with sucrose phosphate phosphatase.

Photorespiration

Similar to NaCl stress response, most of the photorespiratory pathway genes are negatively significant from paired SAM analysis (figure 5.30). Genes encoding glutamate-glyoxylate aminotransferase 2, serine hydroxymethyl transferase, NAD+ hydroxyl pyruvate reductase, phosphoglycolate phosphatase are all negatively significant at most of the timepoints and also from paired SAM analysis, implying a net reduction in photorespiratory flux.

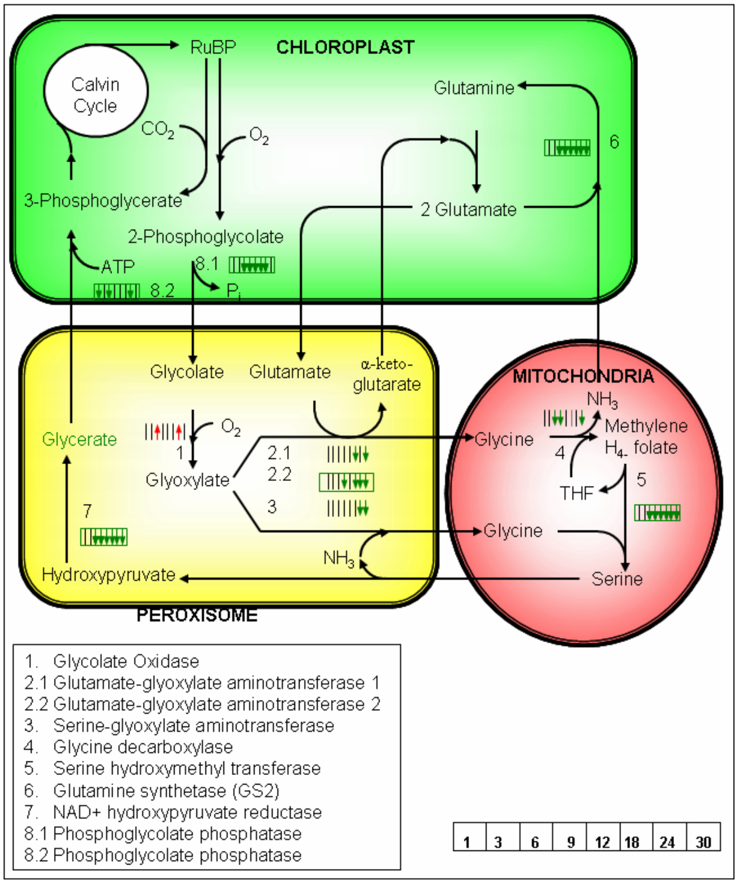


Figure 5.30 Observed effect of the applied perturbation on the physiology of photorespiration, at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Photorespiratory pathway flux is possibly decreasing significantly as most of the genes are negatively significant from paired SAM analysis. Glutamate-glyoxylate aminotransferase 2 gene was not negatively significant in any other stress, unlike this combined stress.

Nitrogen Assimilations and amino-acid biosynthesis

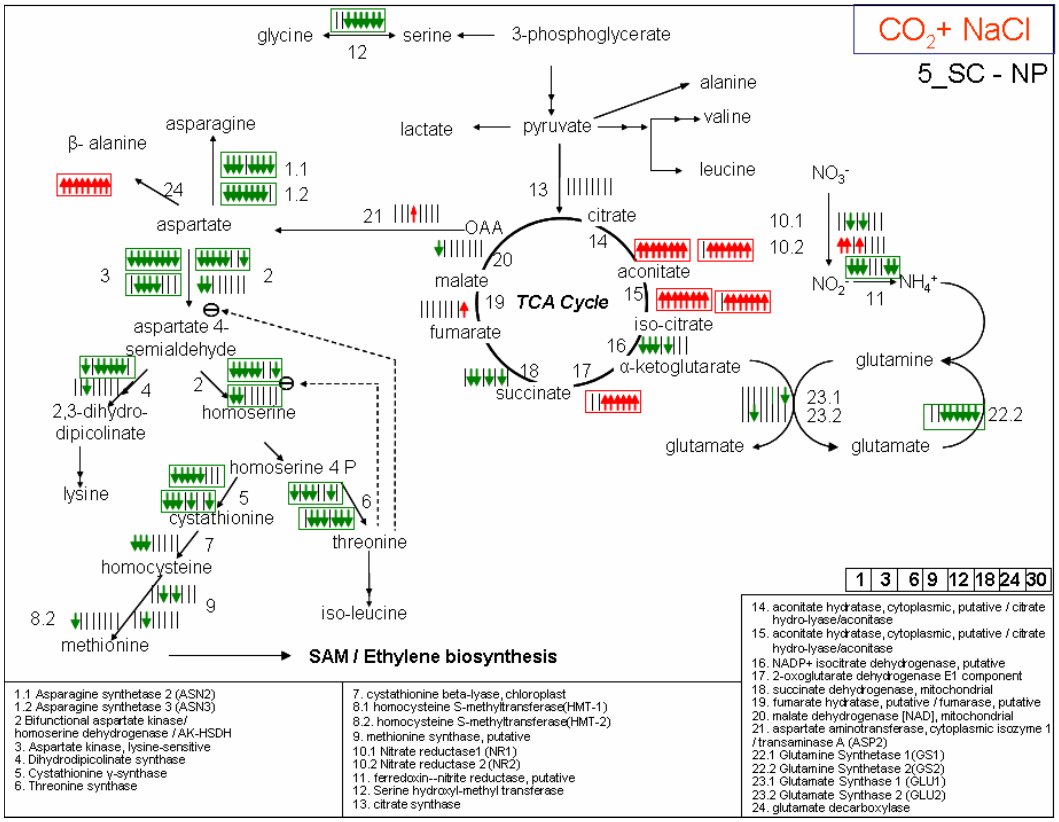

CO2+ NaCl
5_SC - NP
glycine
12
serine
3-phosphoglycerate
alanine
valine
leucine
lactate
pyruvate
asparagine
β- alanine
1.1
1.2
24
aspartate
21
OAA
13
citrate
14
aconitate
15
iso-citrate
16
α-ketoglutarate
17
succinate
18
fumarate
19
TCA Cycle
malate
20
NO3-
10.1
10.2
NO2-
11
NH4+
glutamine
23.1
23.2
22.2
glutamate
glutamate
3
2
aspartate 4-semialdehyde
4
2,3-dihydro-dipicolinate
lysine
2
homoserine
homoserine 4 P
5
cystathionine
6
threonine
7
homocysteine
9
8.2
methionine
iso-leucine
SAM / Ethylene biosynthesis
1 3 6 9 12 18 24 30
1.1 Asparagine synthetase 2 (ASN2)
1.2 Asparagine synthetase 3 (ASN3)
2 Bifunctional aspartate kinase/ homoserine dehydrogenase / AK-HSDH
3. Aspartate kinase, lysine-sensitive
4. Dihydrodipicolinate synthase
5. Cystathionine γ-synthase
6. Threonine synthase
7. cystathionine beta-lyase, chloroplast
8.1 homocysteine S-methyltransferase(HMT-1)
8.2. homocysteine S-methyltransferase(HMT-2)
9. methionine synthase, putative
10.1 Nitrate reductase1 (NR1)
10.2 Nitrate reductase 2 (NR2)
11. ferredoxin--nitrite reductase, putative
12. Serine hydroxyl-methyl transferase
13. citrate synthase
14. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
15. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
16. NADP+ isocitrate dehydrogenase, putative
17. 2-oxoglutarate dehydrogenase E1 component
18. succinate dehydrogenase, mitochondrial
19. fumarate hydratase, putative / fumarase, putative
20. malate dehydrogenase [NAD], mitochondrial
21. aspartate aminotransferase, cytoplasmic isozyme 1 / transaminase A (ASP2)
22.1 Glutamine Synthetase 1(GS1)
22.2 Glutamine Synthetase 2(GS2)
23.1 Glutamate Synthase 1 (GLU1)
23.2 Glutamate Synthase 2 (GLU2)
24. glutamate decarboxylase

Figure 5.31 Observed effects of the applied NaCl and CO_2 combined stress on Nitrogen assimilation and amino-acid biosynthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. This pathway was significantly affected due to the applied perturbation (more than any other perturbation shown here) as most of the gene of this pathway are significantly up or down regulated.

Figure 5.31 shows significance profiles of genes involved in nitrate reduction and its assimilation for amino-acid biosynthesis. Nitrate reductase 1 and 2 shows differenetial expression, with NR1 and NR2 becoming negatively and positively significant respectively at first few timepoints. Nitrite reductase was negatively significant at most of the timepoints and also from paired SAM. Its bemusing why nitrite reductase is strongly down-regulated while nitrate reductase isoenzymes are not, as this might lead to accumulation of toxic nitrite ions inside cells. This is the only gene annotated as nitrite reductase in the latest *Arabidopsis* annotation.

There could be another nitrite reductase gene currently unannotated. The regulation of nitrite reductase might not be at the transcriptional level; hence its mRNA abundance doesn't truly represent its active enzyme concentration. GS2, glutamine synthetase 2 is active in chloroplast was from 6-30h period providing a strong indication of the down regulation of nitrogen assimilation. This down regulation could also be because of decrease in photorespiration pathway flux. NH_4^+ released from photorespiration is 5-10 times more than that of NH_4^+ fixation from nitrate reductase.

As mentioned before, aspartate is precursor for number of amino acids and is produced from OAA, a TCA cycle intermediate. Glutamate decarboxylase, catalyzing reaction from aspartate to beta-alanine is positively significant at all time points and also from paired SAM analysis. Most of the other genes of this pathway [Figure 5.31] involved in producing other amino acids are negatively significant.

Similar to NaCl stress in case of combined stress it was observed that flux through tryptophan biosynthesis pathway is increasing. Genes encoding anthranilate phosphoribosyltransferase, both the subunits of anthranilate synthase and tryptophan synthase are positively significant at all timepoints and also from paired SAM analysis. However phosphoribosylanthranilate isomerase isoenzymes are under-produced at first 24hs, which is contradictory to the other genes from the same pathway. It should be noted that these phosphoribosylanthranilate isomerase are annotated as putative and the gene coding this enzyme can not be verified from established biochemical databases like KEGG as the latter shows this enzyme as unidentified for *Arabidopsis* pathway. Again it is clear that transcriptional response of the trytophan biosynthesis pathway under combined stress is very similar to that of NaCl stress response. The only difference is in case of combined stress anthranilate phosphoribosyltransferase and histidinol-phosphate aminotransferase genes are strongly up and down regulated respectively. Hence it can be speculated that in case of combined stress tryptophan production is increasing at the cost of tyrosine and phenylalanine, the other two amino-acids that share the same precursor chorismate. Tryptophan could be related to general stress response, hence in case of combined stress higher stress level is possibly causing increased tryptophan production.

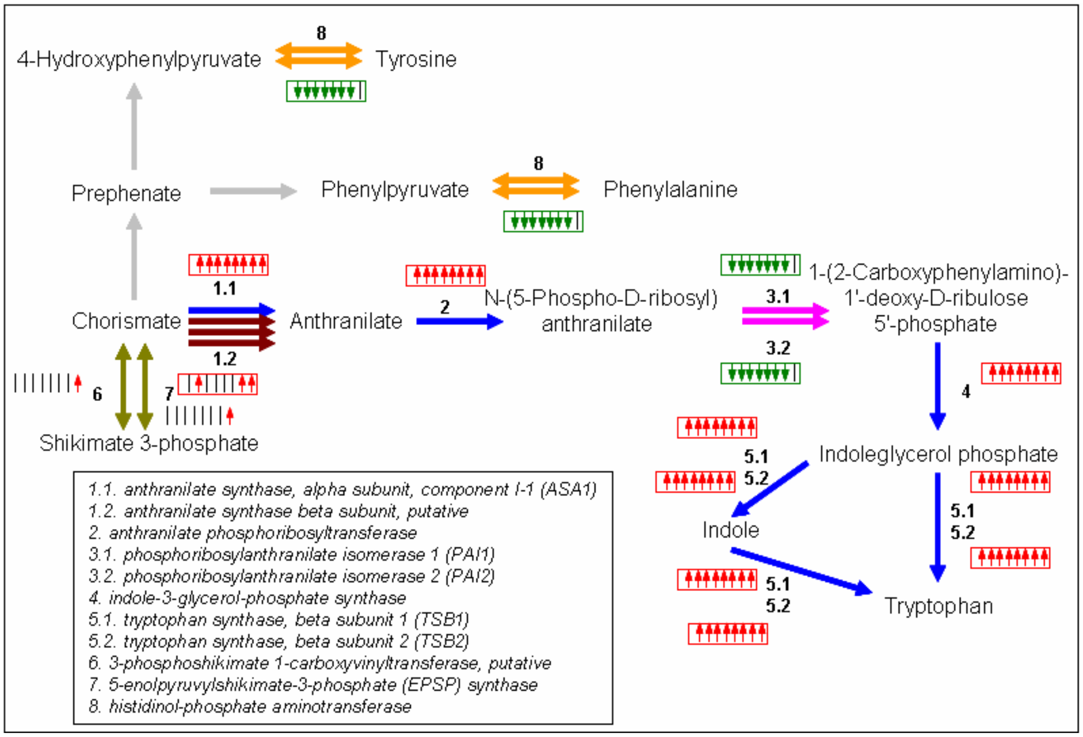

4-Hydroxyphenylpyruvate
8
Tyrosine
Prephenate
Phenylpyruvate
8
Phenylalanine
Chorismate
1.1
1.2
Anthranilate
2
N-(5-Phospho-D-ribosyl) anthranilate
3.1
3.2
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5'-phosphate
6
7
Shikimate 3-phosphate
4
Indoleglycerol phosphate
5.1
5.2
Indole
Tryptophan
1.1. anthranilate synthase, alpha subunit, component I-1 (ASA1)
1.2. anthranilate synthase beta subunit, putative
2. anthranilate phosphoribosyltransferase
3.1. phosphoribosylanthranilate isomerase 1 (PAI1)
3.2. phosphoribosylanthranilate isomerase 2 (PAI2)
4. indole-3-glycerol-phosphate synthase
5.1. tryptophan synthase, beta subunit 1 (TSB1)
5.2. tryptophan synthase, beta subunit 2 (TSB2)
6. 3-phosphoshikimate 1-carboxyvinyltransferase, putative
7. 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
8. histidinol-phosphate aminotransferase

Figure 5.32: Observed transcriptional response of the NaCl and CO_2 combined stress on tryptophan synthesis pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5. Similar to amino acid biosynthesis pathway, most of these genes were significantly up or down regulated due to applied perturbation.

Ethylene biosynthesis and signaling cascade

In case of both CO_2 and NaCl stress response it was observed that ACC synthase gene was positively significant at most of the timepoints and also from paired SAM analysis. When combined stress is applied, only ACS2 gene was identified from analysis as the other two ACS genes were missing (figure 5.33). Similar to NaCl stress response ACS2 gene was positively significant at all timepoints and from paired SAM. ACC oxidase, the enzyme for the next reaction in ethylene biosynthesis was positively significant at 4 timepoints and also from paired SAM. But, the other ACC oxidase gene was non-significant at most of the timepoints. As ACCS is the most regulating enzyme, up- regulation of this gene could cause a increase in ethylene biosynthesis.

Similar to NaCl stress response, ETR1 and CTR1 gene was positively significant at all timepoints and also from paired SAM analysis (figure 5.33). As explained before, ETR1 gene is up-regulated by ethylene and here we see its up-regulation where ACS2 gene, which plays a regulatory role in ethylene biosynthesis, is also positively significant. EIN3, EIL1 and ERF1 genes show similar response as that of NaCl stress alone. These genes are regulated by CTR1 gene, as the CRT1 gene expression is similar in NaCl and combined stress responses, so are the expression of its downstream genes.

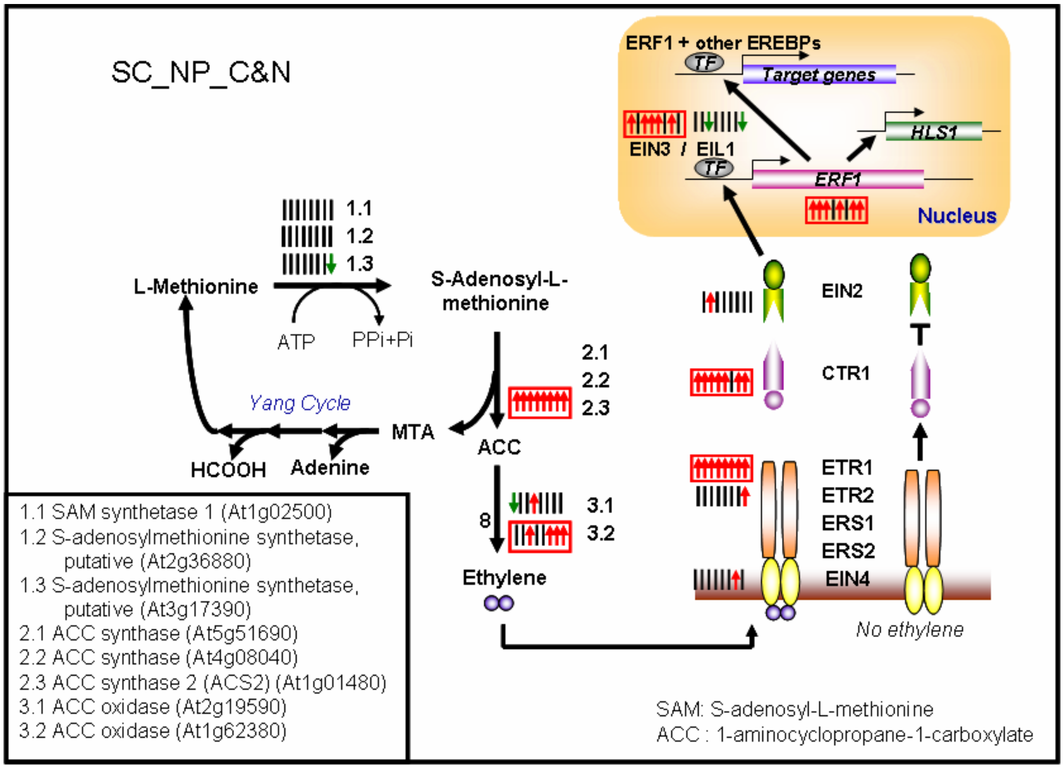

SC_NP_C&N
ERF1 + other EREBPs
TF
Target genes
EIN3 / EIL1
TF
HLS1
ERF1
Nucleus
1.1
1.2
1.3
L-Methionine
ATP
PPi+Pi
S-Adenosyl-L-methionine
Yang Cycle
MTA
HCOOH
Adenine
2.1
2.2
2.3
ACC
8
3.1
3.2
Ethylene
EIN2
CTR1
ETR1
ETR2
ERS1
ERS2
EIN4
No ethylene
1.1 SAM synthetase 1 (At1g02500)
1.2 S-adenosylmethionine synthetase, putative (At2g36880)
1.3 S-adenosylmethionine synthetase, putative (At3g17390)
2.1 ACC synthase (At5g51690)
2.2 ACC synthase (At4g08040)
2.3 ACC synthase 2 (ACS2) (At1g01480)
3.1 ACC oxidase (At2g19590)
3.2 ACC oxidase (At1g62380)
SAM: S-adenosyl-L-methionine
ACC : 1-aminocyclopropane-1-carboxylate

Figure 5.33 Observed effect of the NaCl and CO_2 combined stress on ethylene biosynthesis and signaling at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

It was observed before expression of EREBP genes are affected by ethylene. Significance results of *Arabidopsis* EREBP genes obtained in this comparison are shown in table 5.12. At5g07580 and At5g25190 genes are negatively significant at most of the timepoints and from paired SAM. Other EREBP genes are also negatively significant at 9 and 18h timepoints. At5g25190 gene was also negatively significant from NaCl stress response and most of the other EREBP genes were also negatively significant at 9h timepoint.

In general it was observed that NaCl stress and NaCl and CO_2 stress shown considerable similarity in their transcriptional response of ethylene signaling cascade genes and ethylene response genes, apart from two genes of this pathway ETR2 and EIN2. Possibly these two genes follow different regulatory mechanism then the rest of the genes.

Table 5.12: Significance level of ethylene-responsive element-binding family genes under NaCl and CO_2 combined stress. Overall stress response was through down regulation and almost all the genes were negatively significant at 9 and 18h timepoints.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
ethylene-responsive element-binding family protein	-1	-1	0	-1	-1	-1	-1	-1	-1
ethylene-responsive element-binding protein, putative	-1	-1	-1	-1	-1	0	-1	0	-1
ethylene-responsive element-binding factor 4 (ERF4)	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	-1	0	0	0

Fatty acid biosynthesis and metabolism

Vast majority of fatty acid biosynthesis in plants occurs in plastids. Fatty acid is also produced in mitochondria but rates of this activity are very small in comparison to those found in plastid [Lea and Leeegood, 1993]. Fatty acid biosynthesis is a multistep process involving number of enzymes. It was also observed that under the salt stress several genes involved in fatty acid biosynthesis from acetyl CoA are negatively significant. genes related to fatty acid elongation were also found negatively significant. Following table 5.13 shows the list of genes involved in fatty acid biosynthesis and their significance level at individual timepoints and also from paired SAM.

Table 5.13: Significance level of genes at individual timepoints and also from paired SAM related to fatty acid biosynthesis under salt stress. Notation used is same as that of table 5.2. Most of the genes response was through down regulation which is similar to NaCl stress response. Maximum number of negatively significant genes was observed at 6h timepoint.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
acyl carrier family protein / ACP family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
beta-hydroxyacyl-ACP dehydratase, putative	-1	-1	-1	-1	-1	0	-1	0	-1
acyl carrier protein, chloroplast, putative / ACP, putative	-1	-1	0	-1	-1	0	-1	-1	-1
acyl carrier protein, chloroplast, putative / ACP, putative	-1	-1	0	-1	-1	0	-1	-1	-1
acyl carrier protein 3, chloroplast (ACP-3)	-1	-1	0	0	0	0	0	0	0
acetyl-CoA C-acyltransferase 1 / 3-ketoacyl-CoA thiolase 1 (PKT1)	-1	0	-1	-1	-1	-1	-1	-1	-1
acetyl-CoA C-acyltransferase 1 / 3-ketoacyl-CoA thiolase 1 (PKT1)	-1	0	-1	-1	-1	-1	-1	-1	-1
acetyl-CoA carboxylase, biotin carboxylase subunit (CAC2)	-1	0	-1	-1	-1	0	-1	0	-1
beta-hydroxyacyl-ACP dehydratase, putative	-1	0	-1	-1	-1	0	-1	0	-1
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	-1	0	-1	-1	0	0	0	0	0
3-oxoacyl-[acyl-carrier-protein] synthase I	-1	0	-1	-1	0	0	0	0	0
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	-1	0	-1	-1	0	0	0	0	0
3-oxoacyl-[acyl-carrier-protein] synthase III, chloroplast / beta-ketoacyl-ACP synthase III / 3-ketoacyl-acyl carrier protein synthase III (KAS III)	-1	0	-1	0	0	0	0	-1	0
3-oxoacyl-[acyl-carrier protein] reductase, chloroplast / 3-ketoacyl-acyl carrier protein reductase	-1	0	-1	0	0	0	0	0	0
acyl carrier protein, mitochondrial / ACP / NADH-ubiquinone oxidoreductase 9.6 kDa subunit	0	0	-1	-1	-1	0	-1	0	-1
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	0	-1	0	0	0	0	0	0
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	0	-1	0	0	0	0	0	0
3-ketoacyl-ACP synthase, putative	0	0	0	0	0	0	0	-1	0
3-ketoacyl-CoA thiolase	0	0	0	0	0	0	0	1	0
acetyl-CoA synthetase, putative / acetate-CoA ligase, putative	0	0	0	0	0	0	0	1	0
3-oxoacyl-[acyl-carrier-protein] synthase II, putative	0	0	0	0	0	0	1	0	0
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase,	0	0	0	1	0	0	0	0	0

putative									
acetyl-CoA C-acyltransferase, putative / 3-ketoacyl-CoA thiolase, putative	0	0	0	1	0	0	0	0	0
acetyl-CoA carboxylase 2 (ACC2)	1	0	0	0	0	0	0	0	0
acetyl-CoA carboxylase 1 (ACC1)	1	1	0	1	1	0	0	0	0

While there is a possible indication of reduction in fatty acid biosynthesis, it was also observed that genes related to fatty acid metabolism are mostly positively significant. List of the genes involved in fatty acid metabolism and their significance level from paired SAM and MiTimeS are shown in table 5.14.

Table 5.14: Significance level of genes at individual timepoints and also from paired SAM related to fatty acid metabolism under salt and CO_2 combined stress. Notation used is same as that of table 5.2. Most of the genes response was through up regulation implying increased degradation of fatty acid, similar to NaCl stress response. Maximum number of negatively significant genes was observed at 6h timepoint.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
acyl-activating enzyme 18 (AAE18)	1	1	1	1	1	1	1	1	1
acyl-CoA oxidase (ACX2)	1	1	1	1	1	1	1	1	1
acyl-CoA oxidase (ACX1)	1	1	1	1	1	1	1	1	1
malonyl-CoA decarboxylase family protein	0	0	1	0	1	1	1	0	0
acyl-CoA dehydrogenase-related	0	1	1	0	1	1	1	1	1

A general conclusion can be drawn that under the stressed condition, when photosynthesis, the source of energy, is decreasing in plants fatty acid degradation is possibly increasing to release energy from it which can be used for plant's survival and to make possible changes in physiology for stress acclimation. There is no excess photosynthetic product which needs to be stored, hence fatty acid biosynthesis reactions are negatively significant.

Table 5.15: Significance levels of Universal Stress Protein (USP) genes at individual timepoints and also from paired SAM under salt and CO_2 stress. Notation used is same as that of table 5.2. Similar to NaCl stress most of the gene's response was through down-regulation.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	0	-1	-1	-1	-1	0	-1	0	-1
universal stress protein (USP) family protein	0	-1	-1	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	-1	-1	0	-1	0	-1	-1	-1	-1
universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	0	0	0	-1	-1	-1	-1	-1	-1
universal stress protein (USP) family protein	-1	-1	-1	-1	0	-1	-1	-1	-1
universal stress protein (USP) family protein / responsive to dessication protein (RD2)	0	0	0	0	1	0	0	0	0
universal stress protein (USP) family protein	-1	0	0	0	0	0	0	0	0
universal stress protein (USP) family protein	-1	0	-1	-1	0	-1	0	0	0
universal stress protein (USP) family protein	-1	0	0	0	0	0	-1	-1	0
universal stress protein (USP) family protein	0	0	1	1	1	0	0	0	0
universal stress protein (USP) family protein	-1	-1	0	-1	0	-1	-1	0	0
universal stress protein (USP) family protein	0	0	0	0	0	0	-1	0	0

Genes annotated as Universal Stress Protein (USP) are mostly significantly down regulated both from paired SAM and MiTimeS analysis. These genes are of special interest because they are supposed to be affected by stress responses. However to what extent they are affected and if they are regulated by up or down regulation is can be analyzed from stress response studies.

5.2.2 Transcriptional response of Arabidopsis thaliana liquid cultures subjected to trehalose and CO_2 stress

The transcriptional response of trehalose and CO_2 stress applied individually were discussed above. Now I will discuss the response when the stresses are applied in combination. For an effective comparison in this combined stress experiment the strength of trehalose and CO_2 stresses was same as that when applied individually. Samples were also harvested at the same timepoints. The combines stress was found to create physiological change to the culture within first 30 hours of its application. Table 5.16 shows the weight of the plant samples and the corresponding media pH measurements for control experiment (SC) and combined stress experiment (TP).

Table 5.16 Comparison of weight and media pH of control and trehalose and CO_2 combined stress response.

		Control (SC)		trehalose and CO_2 stress(TP)	
Time Pt.	Sample No	Weight	pH	Weight	pH
0	20	14	6.16	22.7	6.93
0	19	16.9	6.21	23.0	6.84
0	18	19	6.18	20.7	6.68
0	17	17.8	6.45	16.2	6.51
1	1	13.7	6.13	25.2	6.76
1	2	12.8	6.43	22.6	6.72
3	3	14.7	6.09	20.9	6.50
3	4	16.3	6.23	27.0	6.64
6	5	9.2	6.32	20.1	6.58
6	6	15.1	6.3	24.7	6.57
9	7	18	6.35	25.8	6.81
9	8	21	6.24	21.8	6.76
12	9	12	6.28	25.8	6.75
12	10	14.5	6.4	27.8	6.63
18	11	22.9	6.3	24.0	6.48
18	12	21.9	6.36	24.3	6.33
24	13	21.8	6.27	29.4	6.67

24	14	20.1	6.45	30.7	6.77
30	15	28.3	6.51	31.9	6.95
30	16	30.6	6.4	32.1	6.90

Plants were immediately frozen in liquid nitrogen and kept at -80°C until they were ground in liquid nitrogen. 2 grams of ground sample was used for transcriptional profiling analysis. Experimental protocol for RNA extraction, RNA amplification, hybridization are explained in detail in materials and methods section.

5.2.2.1 Multivariate statistical analysis

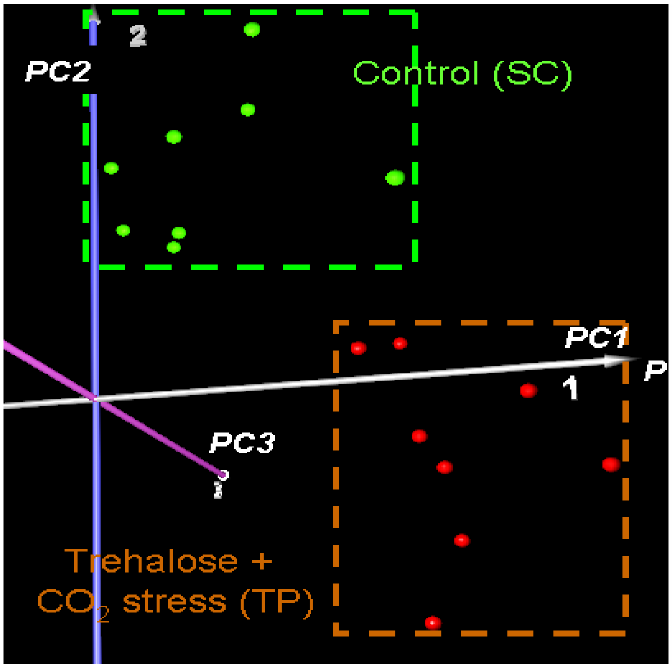


Figure 5.34 PCA analysis of the experimental timepoints of control and trehalose and CO_2 stress experiments. The experiments are separated in reduced gene space as the combined stress is moving the timepoints along PC1.

Similar to the previous stress analysis first a common repository of genes is selected. The selected 11025 genes have non-zero expression values for at least 12 out of 16 timepoints. From the Principal Component Analysis (PCA), the control transcriptomic profiles can be clearly differentiated from their perturbed counterparts (Figure 5.34). This implies that the physiology of the plant liquid cultures is affected by the applied perturbation at transcriptional level, even during the first 30h of treatment. First 3 principal components were found to capture 38, 25 and 8% of the information. Hence, when the experiments are viewed at 3-D space it can account for most of the variance (71%). It can also be seen due to the application of trehalose stress timepoints have moved along principal component 2, whereas in case of combined stress they moved along principal component 1, which accounts for maximum variability.

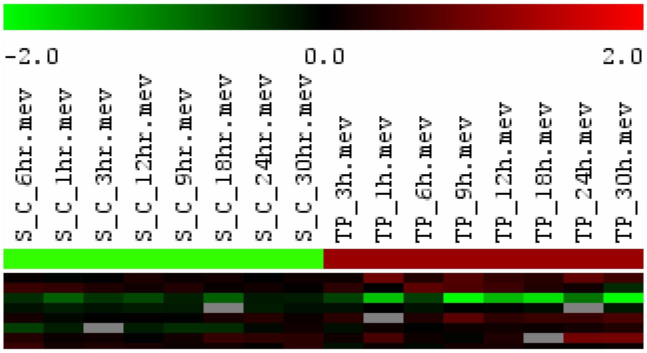


Figure 5.35 Hierarchical clustering of the samples shows two experimental groups form two distinct clusters.

Experimental timepoints were also clustered using hierarchical clustering and it also shows a clear separation between them (Figure 5.35).

Significance analysis of the combined stress was carried out in a similar fashion as that of previous comparisons. Paired SAM and MiTimeS were used for significance analysis. Delta value of 2.11 was selected for paired SAM, as this delta value has highest number of significant genes with minimum (0 in this case) FDR. There were 784 and 632 genes found positively and negatively significant from paired SAM with this delta value, which constitutes around 7 and 6% respectively of genes used for analysis (11025). The delta value used for

MiTimeS analysis was 2.2655. The delta value for individual timepoints was higher than the one used for paired SAM because of stringent multiple test correction. Number of genes positively, negatively and non-significant at individual timepoints obtained from MiTimeS was plotted with that of paired SAM results in figure 5.36. Timepoint 1h shows maximum number of significant genes of both the significant types due to strong initial response of the trehalose stress. Strong initial response was also observed in case of trehalose stress, and is possibly conserved for any type of trehalose stress response.

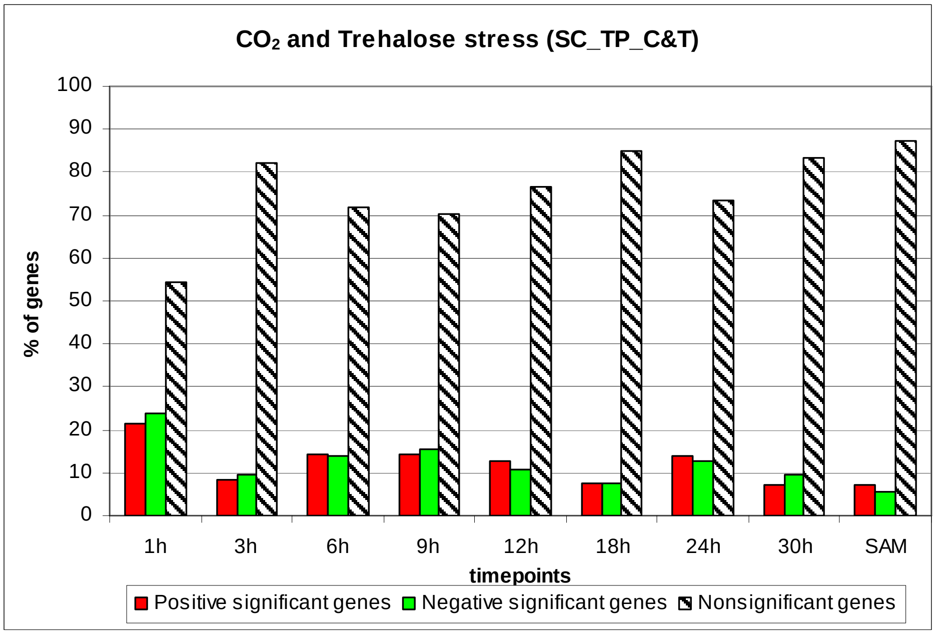


Figure 5.36 The bar diagram show the percentage of positively, negatively and non-significant genes at individual timepoints and also from paired SAM under trehalose and CO_2 combined stress. Timepoint 1h has maximum number of positive and negative significant genes, similar to trehalose stress response.

5.2.2.2 Data validation and interpretation in the context of plant physiology

Similar to the response of the trehalose stress alone, it was found the trehalase gene is positively significant at 6 of the 8 timepoints and also from paired SAM analysis. However, other genes involved in this pathway TPP and

TPS show response different from when trehalose stress is applied individually. TPP gene At4g17770 was negatively significant at five timepoint and also from paired SAM analysis [please see table 5.17]. Another TPP gene At4g22590, which was up-regulated at 6 timepoints from trehalose stress was non-significant in case of combined stress. Comparison of significance profile of the of the TPP and TPS genes shows that trehalose biosynthesis pathway is responding differently in individual and combined stress responses. Significance level of these genes at individual timepoints and from paired SAM are shown in table 5.17.

Table 5.17 Significance profile of genes related to trehalose synthesis and degradation. Color-code used was same as table 5.2.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
glycosyl hydrolase family protein 37 / trehalase, putative	1	0	1	1	0	1	1	1	1
glycosyl transferase family 20 protein / trehalose-phosphatase family protein	1	0	1	0	0	0	0	0	0
glycosyl transferase family 20 protein / trehalose-phosphatase family protein	0	0	0	0	1	0	0	0	0
glycosyl transferase family 20 protein / trehalose-phosphatase family protein	-1	-1	-1	0	-1	-1	0	0	-1
trehalose-6-phosphate phosphatase (TPPA)	0	0	0	-1	0	0	0	0	0
trehalose-6-phosphate phosphatase (TPPB)	0	1	0	1	0	0	0	0	0
trehalose-6-phosphate phosphatase, putative	0	0	0	0	0	0	0	0	0
trehalose-6-phosphate phosphatase, putative	0	0	1	0	0	0	0	0	0
trehalose-6-phosphate phosphatase, putative	-1	0	0	-1	0	0	0	0	0

Calvin cycle, sucrose and starch biosynthesis

It is explained before that RuBosCO catalyses the carbon fixation of Calvin cycle. The *rbcL* gene, which encodes the large subunit of rubisco, was positively significant at 9 and 12h timepoints. The *Arabidopsis rbcS* gene family consists of four members, namely 1A, 1B, 2B and 3B, as mentioned before. In the present study, all four subunits were identified as negatively significant at 30h of perturbation. In all other stress comparisons, *rbcS* genes were negatively significant at more than one timepoints. Hence, trehalose and CO_2 combined stress shows least negative response of rubisco genes. Two genes were found from analysis encoding phosphoglycerate kinase, one of them being negatively significant at most of the timepoints while the other one was down-regulated at only 1 and 9h timepoints. Triose-phosphates transported from the chloroplasts to the cytoplasm are converted to hexose-phosphates. Sucrose synthesis takes place with in a series of reaction starting from triose phosphates transported in cytosol.

In the pathway of sucrose synthesis starting from triose phosphates several genes encoding the following enzymes fuctose-bisphosphate aldolase, glucose-6-phosphate isomerase, sucrose phosphate synthase and sucrose synthase shows similar significance profiles. All of them are positively significant at 9h timepoint. None of the genes in this pathway are strongly up or down regulated as none of them are identified as differentially expressed from paired SAM analysis. This response is very different from what is obtained in case of trehalsoe stress alone, where sucrose synthase and fuctose-bisphosphate aldolase genes were positively significant from paired SAM. Four genes encoding invertase At2g01610, At4g25250, At5g38610, At5g64620 were negatively significant at 5-6 timepoints and also from paired SAM (figure 5.37). Two of these genes At4g25250 and At5g64620 were also negatively significant under trehalose stress alone.

Increase in starch synthesis and up-regulation of ADP-glucose pyrophosphorylase was observed in past when trehalose stress is applied. Under trehalose and CO_2 combined stress two ADP-glucose pyrophosphorylase genes were negatively significant at four timepoints during 1-12h period. When trehalose stress was applied individually it was significant at only at one or two timepoints. Unlike trehalose stress, starch synthase gene was negatively or non-significant at most of the timepoints. These observations together provide indication that starch synthesis is decreasing in case of combined stress.

In conclusion, the decrease in rate of carbon fixation is possibly minimal under this stress response. Contrary to trehalsoe stress alone, reduction in rate of starch synthesis was observed in case of combined stress response. Sucrose synthase gene was not significantly up-regulated as observed in individual stress.

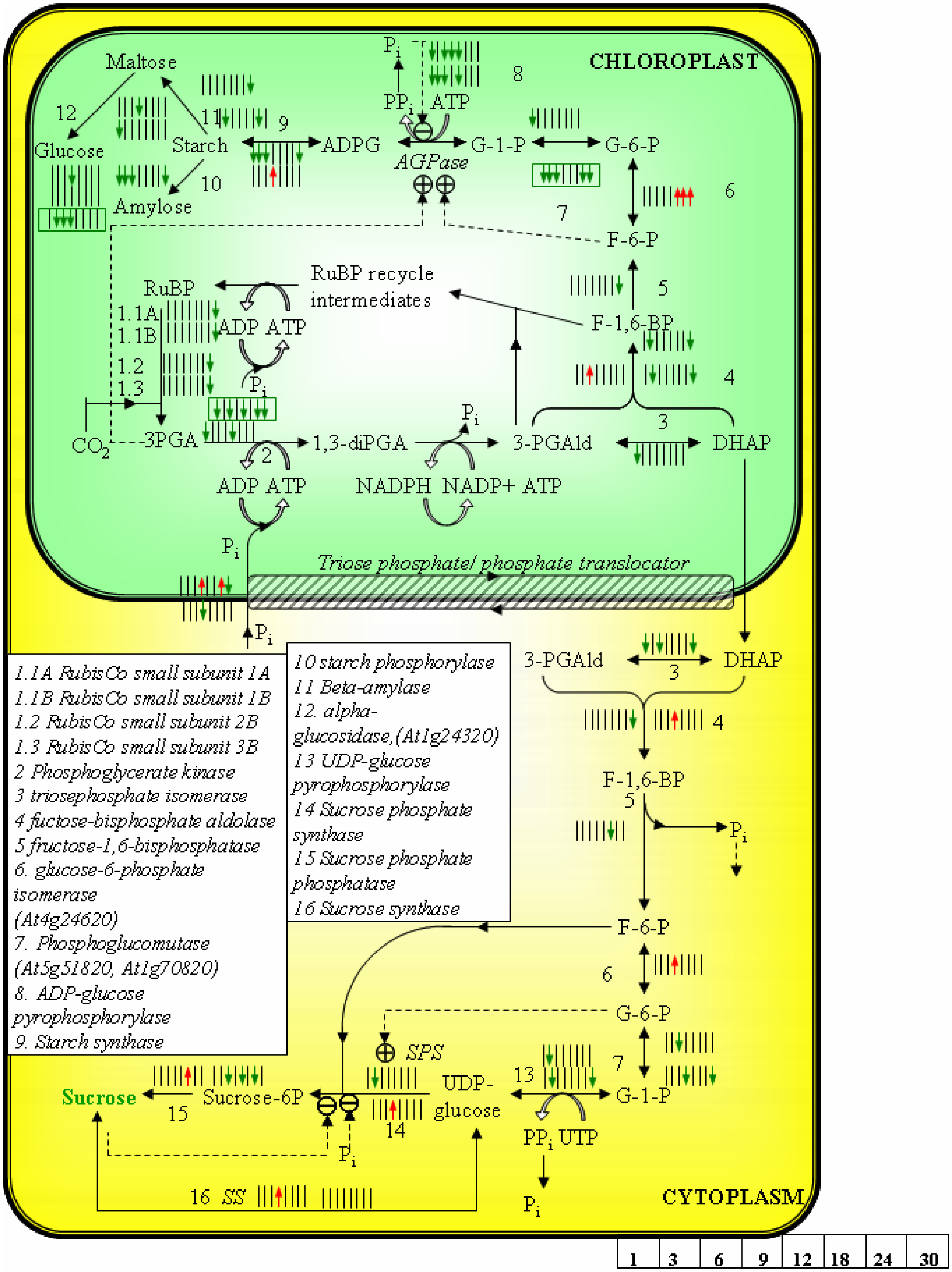

CHLOROPLAST
Maltose
12
Glucose
11
Starch
9
ADPG
10
Amylose
P_i
PP_i
ATP
8
AGPase
G-1-P
G-6-P
7
6
F-6-P
RuBP
RuBP recycle intermediates
5
1.1A
1.1B
1.2
1.3
ADP ATP
P_i
F-1,6-BP
4
CO2
3PGA
2
1,3-diPGA
P_i
3-PGAld
3
DHAP
ADP ATP
NADPH NADP+ ATP
P_i
Triose phosphate/ phosphate translocator
P_i
1.1A RubisCo small subunit 1A
1.1B RubisCo small subunit 1B
1.2 RubisCo small subunit 2B
1.3 RubisCo small subunit 3B
2 Phosphoglycerate kinase
3 triosephosphate isomerase
4 fuctose-bisphosphate aldolase
5 fructose-1,6-bisphosphatase
6. glucose-6-phosphate isomerase (At4g24620)
7. Phosphoglucomutase (At5g51820, At1g70820)
8. ADP-glucose pyrophosphorylase
9. Starch synthase
10 starch phosphorylase
11 Beta-amylase
12. alpha-glucosidase,(At1g24320)
13 UDP-glucose pyrophosphorylase
14 Sucrose phosphate synthase
15 Sucrose phosphate phosphatase
16 Sucrose synthase
3-PGAld
3
DHAP
4
F-1,6-BP
5
P_i
F-6-P
6
G-6-P
SPS
Sucrose
15
Sucrose-6P
14
UDP-glucose
13
7
G-1-P
PP_i UTP
P_i
P_i
16 SS
CYTOPLASM
1
3
6
9
12
18
24
30

Figure 5.37: Observed effect of the applied perturbation on the physiology of Calvin cycle, starch and sucrose biosynthesis pathways at the transcriptional level. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Photorespiration

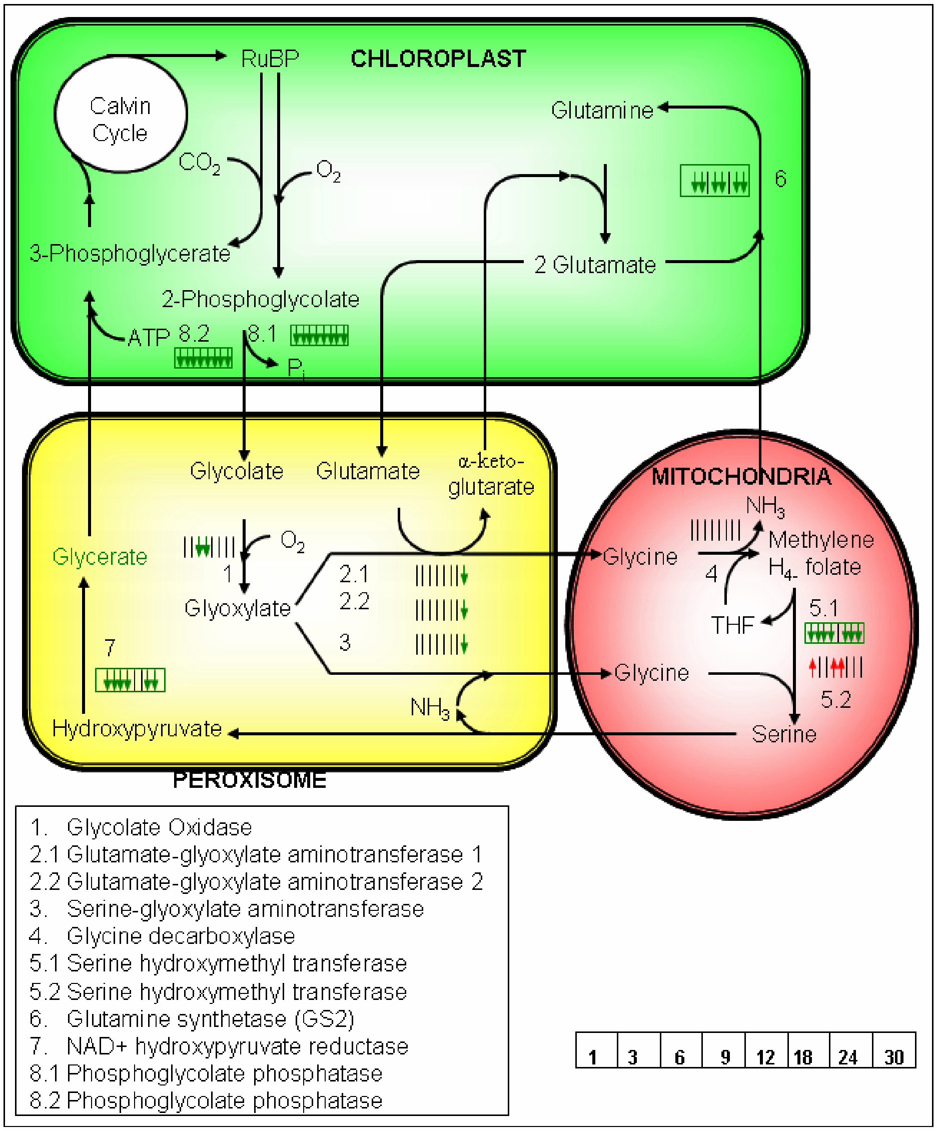


Figure 5.38_Observed effect of the applied perturbation on the physiology of photorespiration pathway. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

As explained before Carbon fixation and photorespiration "compete" for Rubisco activity, hence with increase in CO_2 concentration expression of photorespiratory genes were suppressed. When only trehalose stress was applied, photorespiratory genes were also down-regulated. In case of combined stress several genes involved in photorespiration pathway like Phosphpglyconate phosphatase, NAD+ hydroxypyruvate reductase, Glutamine synthetase (GS2), Serine hydroxymethyl transferase were negatively significant at most of the timepoints and also from paired SAM. Only one gene encoding Serine hydroxymethyl transferase was positively significant at 3 timepoints.

The suppression of photorespiration was observed in all stress responses individual or combination hence could be a general stress response behavior rather than specific to any particular stress. Please see the last chapter for more details.

Nitrogen assimilation and amino-acid biosynthesis

Consistent with other stress responses NR1 and NR2 genes show differential response under combined stress response. NR1 was negatively significant at 12h timepoint while NR2 gene was positively significant from 6-12h period. Nitrite reductase was negatively significant at 1h timepoint. Though NR1 and nitrite reductase genes were down-regulated at only one of the timepoints, but GS2 gene, involved in assimilation of reduced NH_4^+ was negatively significant at 6 out of 8 timepoints and also from paired SAM. This apparent inconsistency can be attributed to the following reason. NH_4^+ released from photorespiration is higher than NH_4^+ produced from nitrate reduction. It is already observed, under this stress condition photorespiration is suppressed, reducing the reflux of NH_4^+. Hence, what down-regulation of GS2 genes is the combined response of photorespiration and nitrate reduction, where photorespiration is the predominant effect. Ferredoxin-dependent glutamate synthase (Fd-GOGAT 1) gene is negatively significant at 3 and 6h timepoints, consistent with the hypothesis that nitrogen assimilation is moderately suppressed, which is different from trehalose stress alone where nitrogen assimilation was increasing.

TCA cycle genes aconitate hydratase, NADP+ isocitrate dehydrogenase and succinate dehydrogenase are positively significant at more than one timepoints. Gene involved in biosynthesis of amino acids produced from aspartate was mostly negatively significant at several timepoints, with a very few exceptions. Like one of the glutamate decarboxylase genes is non-significant at all timepoints while the other one was missing. Other genes like homocysteine S-methyltransferase(HMT-1), methionine synthase were positively significant at one of the timepoints. These enzymes catalyze resactions in biosynthesis of methionine.

The response of tryptophan biosynthesis pathway was quite similar to other stress response. Tryptophan synthase, beta subunit 1 was positively significant at 7 out of 8 timepoints and also from paired SAM. No information about tryptophan synthase, beta subunit 2 (TSB2) was obtained as this gene was missing in the analysis. Interestingly, genes encoding phosphoribosylanthranilate isomerase 1 (PAI1), phosphoribosylanthranilate isomerase 2 (PAI2) and indole-3-glycerol-phosphate synthase were negatively significant at 6 of the 8 timepoints and also from paired SAM. Anthranilate phosphoribosyltransferase gene was also down-regulated at 3 timepoints. competing reactions producing phenylalanine and tyrosine are suppressed as histidinol-phosphate aminotransferase gene is negatively significant at 7 timepoints and as expected, also from paired SAM. Histidinol-phosphate aminotransferase gene was also negatively significant in other combined stress response of NaCl and CO_2, though it was non-significant at individual stress responses. In conclusion, tryptophan synthesis is possibly increasing like other stress responses, but the flux through this pathway might not be as high as it is in case of other stresses.

Though tryptophan synthase, beta subunit 2 (TSB2) gene is over-expressed in trehalose and CO_2 combined stress, but increase in tryptophan biosynthesis flux is possible less than increase due to other stress responses (figure 5.40). In no other stress response 3 genes from this pathway was negatively significant from paired SAM analysis. In this context this stress response is quite unique.

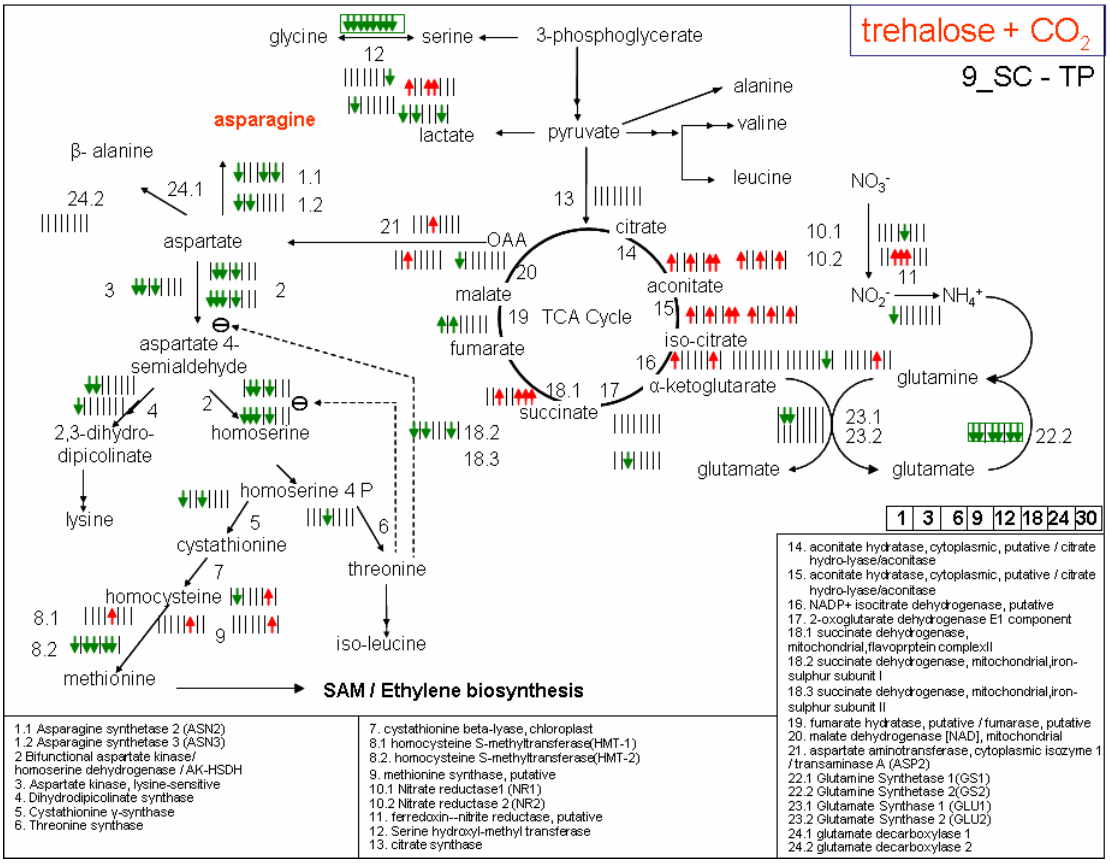

trehalose + CO$_2$
9_SC - TP
glycine
serine
3-phosphoglycerate
12
alanine
asparagine
lactate
pyruvate
valine
leucine
β- alanine
1.1
1.2
24.1
24.2
13
NO$_3^-$
21
OAA
citrate
10.1
10.2
aspartate
14
20
aconitate
11
3
2
malate
NO$_2^-$
NH$_4^+$
19
TCA Cycle
15
fumarate
iso-citrate
aspartate 4-semialdehyde
16
glutamine
4
2
18.1
17
α-ketoglutarate
succinate
23.1
23.2
22.2
2,3-dihydro-dipicolinate
homoserine
18.2
18.3
glutamate
glutamate
lysine
homoserine 4 P
5
6
1
3
6
9
12
18
24
30
cystathionine
threonine
7
homocysteine
8.1
8.2
9
iso-leucine
methionine
SAM / Ethylene biosynthesis
1.1 Asparagine synthetase 2 (ASN2)
1.2 Asparagine synthetase 3 (ASN3)
2 Bifunctional aspartate kinase/ homoserine dehydrogenase / AK-HSDH
3. Aspartate kinase, lysine-sensitive
4. Dihydrodipicolinate synthase
5. Cystathionine γ-synthase
6. Threonine synthase
7. cystathionine beta-lyase, chloroplast
8.1 homocysteine S-methyltransferase(HMT-1)
8.2. homocysteine S-methyltransferase(HMT-2)
9. methionine synthase, putative
10.1 Nitrate reductase1 (NR1)
10.2 Nitrate reductase 2 (NR2)
11. ferredoxin--nitrite reductase, putative
12. Serine hydroxyl-methyl transferase
13. citrate synthase
14. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
15. aconitate hydratase, cytoplasmic, putative / citrate hydro-lyase/aconitase
16. NADP+ isocitrate dehydrogenase, putative
17. 2-oxoglutarate dehydrogenase E1 component
18.1 succinate dehydrogenase, mitochondrial,flavoprptein complexII
18.2 succinate dehydrogenase, mitochondrial,iron-sulphur subunit I
18.3 succinate dehydrogenase, mitochondrial,iron-sulphur subunit II
19. fumarate hydratase, putative / fumarase, putative
20. malate dehydrogenase [NAD], mitochondrial
21. aspartate aminotransferase, cytoplasmic isozyme 1 / transaminase A (ASP2)
22.1 Glutamine Synthetase 1(GS1)
22.2 Glutamine Synthetase 2(GS2)
23.1 Glutamate Synthase 1 (GLU1)
23.2 Glutamate Synthase 2 (GLU2)
24.1 glutamate decarboxylase 1
24.2 glutamate decarboxylase 2

Figure 5.39 Observed effect of the applied perturbation on nitrogen assimilation and amino-acid biosynthesis pathway under trehalose and CO_2 combined stress. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

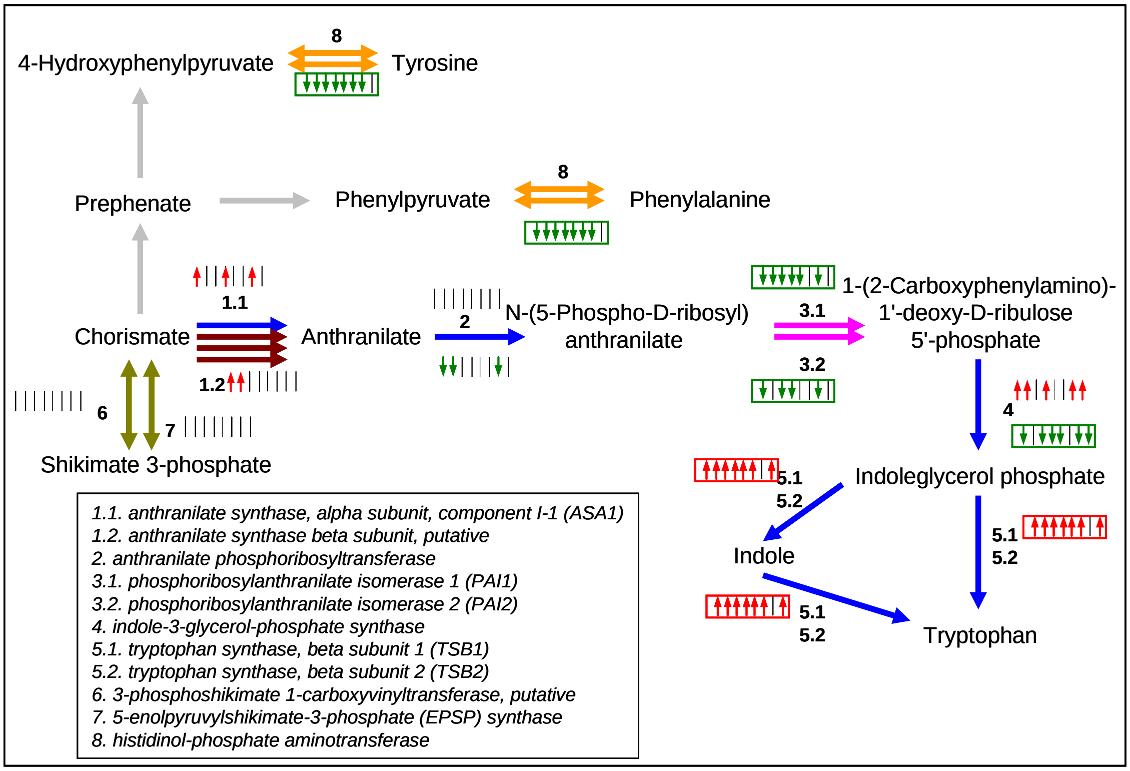

4-Hydroxyphenylpyruvate
8
Tyrosine
Prephenate
Phenylpyruvate
8
Phenylalanine
1.1
Chorismate
1.2
Anthranilate
2
N-(5-Phospho-D-ribosyl) anthranilate
3.1
3.2
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5'-phosphate
4
6
7
Shikimate 3-phosphate
Indoleglycerol phosphate
5.1
5.2
Indole
5.1
5.2
5.1
5.2
Tryptophan
1.1. anthranilate synthase, alpha subunit, component I-1 (ASA1)
1.2. anthranilate synthase beta subunit, putative
2. anthranilate phosphoribosyltransferase
3.1. phosphoribosylanthranilate isomerase 1 (PAI1)
3.2. phosphoribosylanthranilate isomerase 2 (PAI2)
4. indole-3-glycerol-phosphate synthase
5.1. tryptophan synthase, beta subunit 1 (TSB1)
5.2. tryptophan synthase, beta subunit 2 (TSB2)
6. 3-phosphoshikimate 1-carboxyvinyltransferase, putative
7. 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
8. histidinol-phosphate aminotransferase

Figure 5.40 Observed transcriptional response of tryptophan biosynthesis pathway under trehalose and CO_2 combined stress. Though TSB1 gene was positively significant at 7 out of 8 timepoints and also from paired SAM, indole-3-glycerol-phosphate synthase, PAI1 and PAI2 genes from the same pathway are negatively significant. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

Ethylene biosynthesis and signaling

The response of ethylene signaling cascade genes were very similar in NaCl stress and NaCl and CO_2 stress, hence they were conserved. But in case of trehalose stress and trehalose and CO_2 similarity was not obvious. SAM synthase gene was significant at only one or two initial timepoints. One of the ACC synthase genes were significant at 9 and 24h timepoints, while the other one was positively and negatively significant at 1 and 9h respectively (figure 5.41). Responses of ACC oxidase genes were very similar to trehalose stress as they were strongly up-regulated.

ETR2 gene, which is up-regulated at ethylene stress, is positively significant at only 1h timepoint. Interesting genes involved in part of this signaling cascade i.e.

ETR1, ETR2, CTR1, EIN2 shows exactly same significance profile. Under previous stress response comparisons EIN3 and ERF1 were showing similar expression profiles. But here EIN3 is positively significant from paired SAM but ERF1 is done-regulated at two timepoints. The regulations of these genes are not always at the transcriptional level, still they show similar profile which is quite surprising. It could be merely a coincidence that all of them becoming positively significant at 1h timepoint. Nevertheless co-expression of these genes needs to be validates and elucidated at the molecular level.

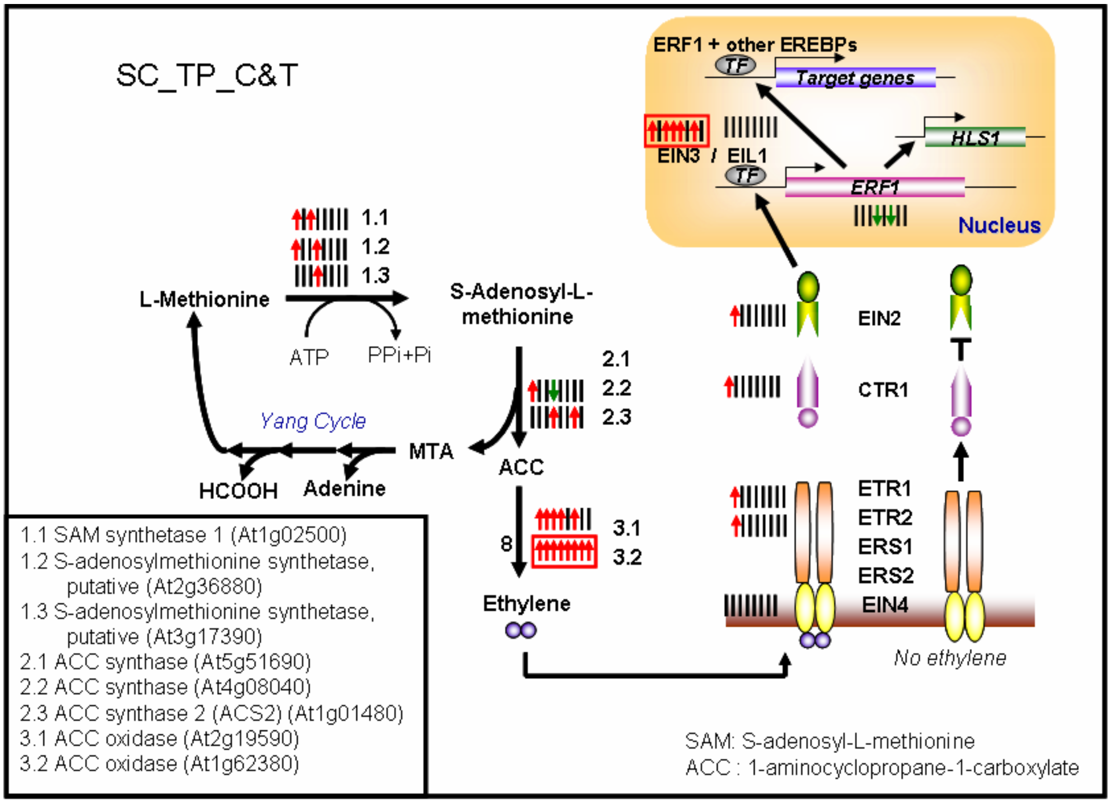

SC_TP_C&T
ERF1 + other EREBPs
TF
Target genes
HLS1
EIN3 / EIL1
TF
ERF1
Nucleus
1.1
1.2
1.3
L-Methionine
S-Adenosyl-L-methionine
ATP
PPi+Pi
Yang Cycle
MTA
HCOOH
Adenine
2.1
2.2
2.3
ACC
8
3.1
3.2
Ethylene
EIN2
CTR1
ETR1
ETR2
ERS1
ERS2
EIN4
No ethylene
1.1 SAM synthetase 1 (At1g02500)
1.2 S-adenosylmethionine synthetase, putative (At2g36880)
1.3 S-adenosylmethionine synthetase, putative (At3g17390)
2.1 ACC synthase (At5g51690)
2.2 ACC synthase (At4g08040)
2.3 ACC synthase 2 (ACS2) (At1g01480)
3.1 ACC oxidase (At2g19590)
3.2 ACC oxidase (At1g62380)
SAM: S-adenosyl-L-methionine
ACC : 1-aminocyclopropane-1-carboxylate

Figure 5.41 Observed response of the ethylene bio-synthesis and signaling pathway genes under trehalose and CO_2 combined stress. Genes encoding signaling cascade proteins ETR1, ETR2, CRT1 and EIN2 shows same significance profile. Positively and negatively significant genes are color-coded as described in the caption of Figure 5.5.

The response of the EREBP genes under this stress condition was also quite different from that of trehalose stress alone (table 5.18). Here most of the genes are negatively significant at 9 and 18h timepoints which is similar to NaCl and CO_2 response. One of these genes At5g25190 was also negatively significant at 5 out of 8 timepoints and also from paired SAM. This is the same gene that was negatively significant at CO_2 stress, NaCl stress and NaCl and CO_2 combined stress. The response of this gene in the context of ethylene stress could a subject of future study.

Table 5.18: Response of the EREBP genes to combined trehalose and CO_2 stress shows almost all the genes are negatively significant at 9 and 12h timepoints.

Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
ethylene-responsive element-binding protein, putative	-1	0	-1	-1	-1	0	0	0	-1
ethylene-responsive element-binding factor 4 (ERF4)	0	0	0	-1	0	-1	-1	-1	0
ethylene-responsive element-binding family protein	0	0	0	0	0	-1	0	0	0
ethylene-responsive element-binding family protein	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding protein, putative	0	0	0	-1	0	-1	0	0	0
ethylene-responsive element-binding family protein	0	0	0	0	0	0	-1	0	0

5.3 Comparison of combined stresses with their constitutive ones

The unique experimental design used in this research not only allows us to study the response of the system by applying different perturbations, but also can compare the effect of combined perturbation with individual ones. For each of the NaCl and trehalose experiment separately, we can compare the combined stress response (CO_2 and NaCl, CO_2 and trehalose) with individual ones (CO_2, NaCl or trehalose). One way to compare these responses is by comparing the genes that are differentially expressed in response to the stresses applied. To check if the stress responses are additive in nature it's important to investigate this question: "are the genes significant at both the individual stresses necessarily significant at combined stress?" Or a question opposite to that can also be asked "Do the genes significant at combined stress are significant in at least one of the corresponding individual stress?" The data reveals the answers to some of these questions.

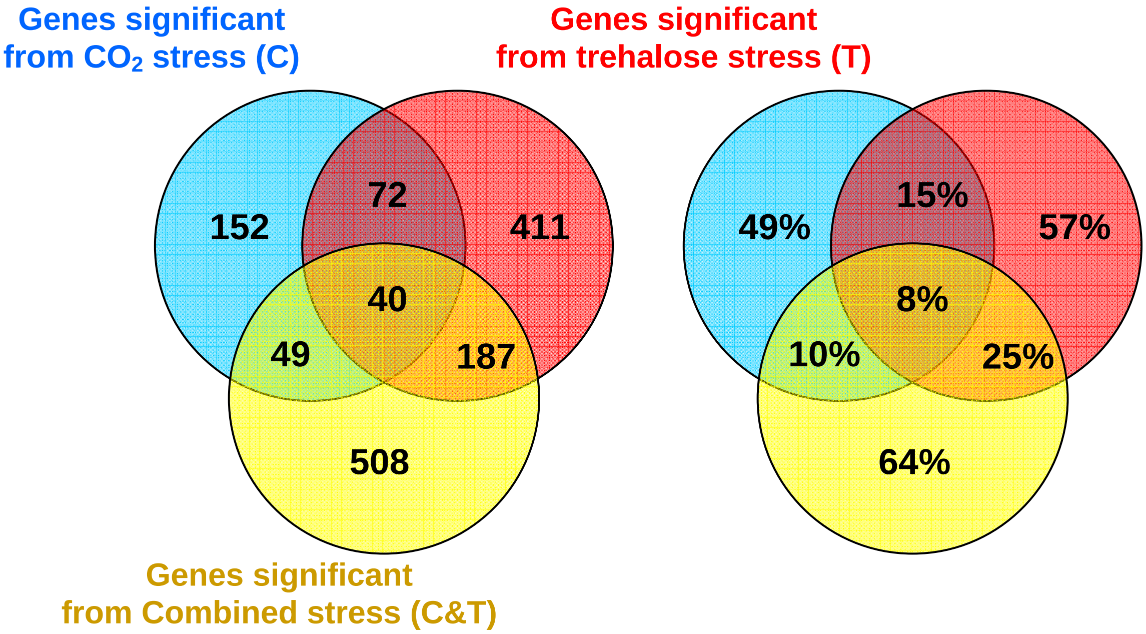

Genes significant from CO_2 stress (C)
Genes significant from trehalose stress (T)
152
72
411
40
49
187
508
49%
15%
57%
8%
10%
25%
64%
Genes significant from Combined stress (C&T)

Figure 5.42: Comparison of positively significant genes from individual and combined stress responses. Blue, red and yellow circles in the Venn diagram represent the genes that are differentially expressed in response to elevated CO_2, trehalose and combined stress. In the Venn diagram on the right, gene numbers are normalized with respect to total genes of that category. The overlap of significant genes between two (or three) stress responses are normalized with respect to their geometric means.

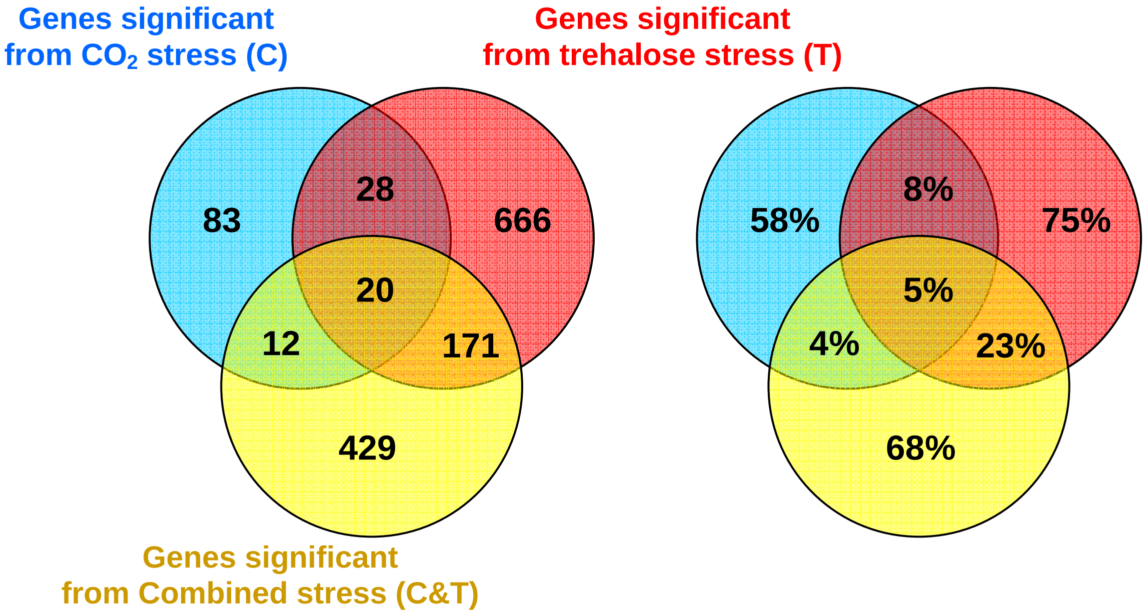


Figure 5.43 Comparison of negatively significant genes from individual and combined stress responses. Rest of the notations is same as figure 5.42.

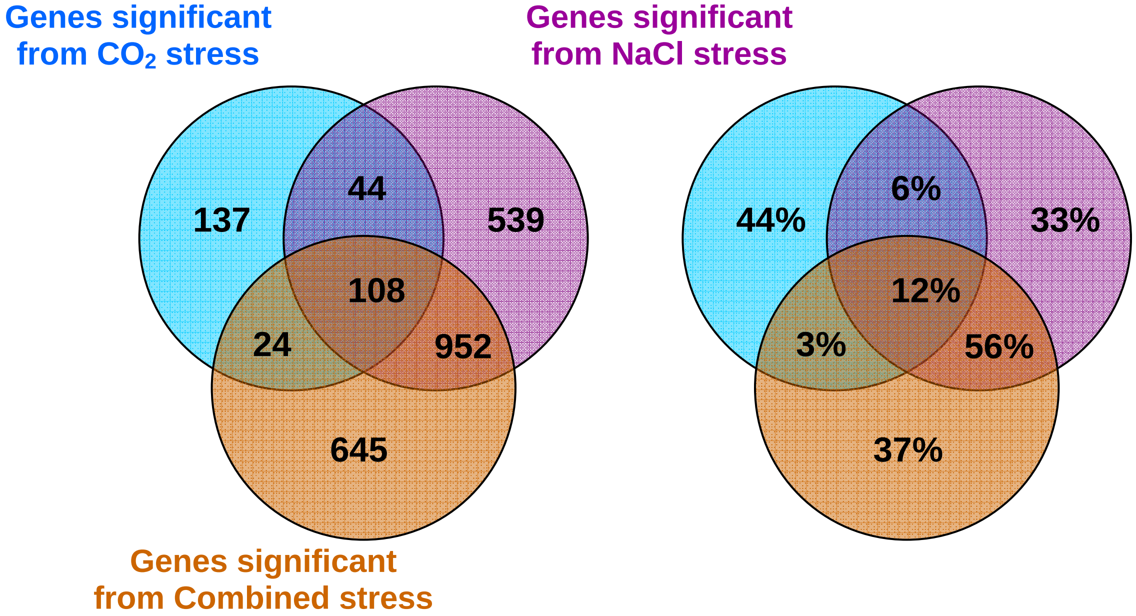


Figure 5.44 Comparison of positively significant genes from individual and combined stress responses. Blue, violet and brown circles in the Venn diagram represent the genes that are differentially expressed in response to elevated CO$_2$, NaCl and combined stress. Rest of the notations is same as Figure 5.42.

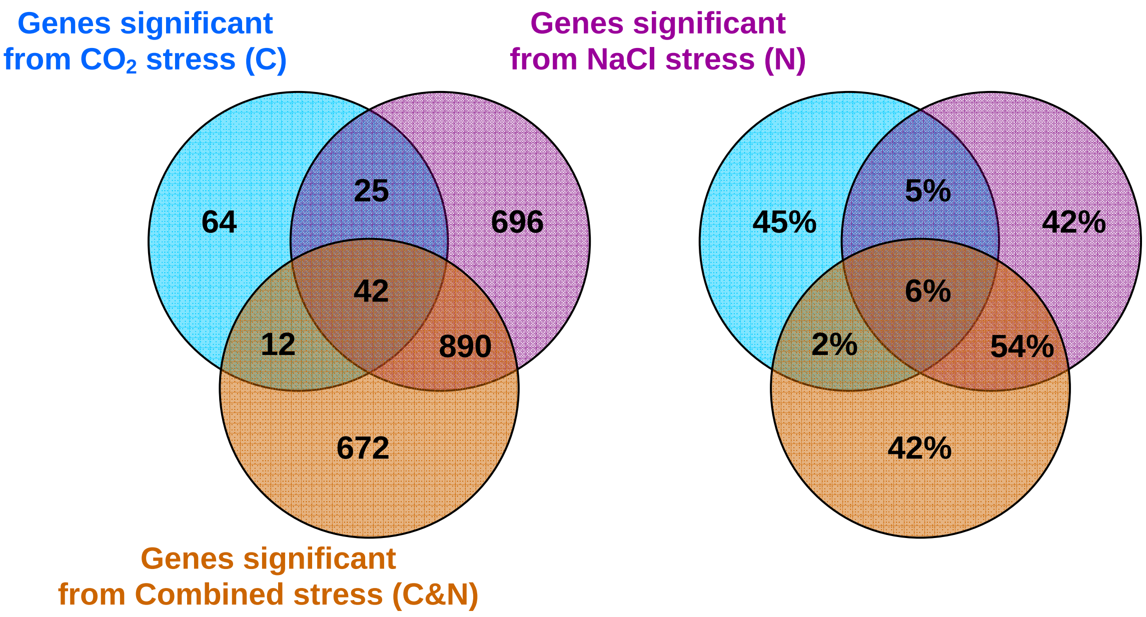


Figure 5.45 Comparison of negatively significant genes from individual and combined stress responses. Rest of the notations is same as figure 5.42.

Analysis of all the four Venn diagrams (figure 5.42 to 5.45) leads us to following conclusions:

- In all the four cases positively significant genes have higher overlap (between any two or all the stresses) compared to negatively significant genes.
- Trehalose stress (T) response is more conserved compared to CO_2 stress (C).
- NaCl stress (N) response is more conserved compared to CO_2 stress (C).
- Between NaCl (N) and trehalose stress (T) NaCl is more conserved.
- Between NaCl and trehalose stress responses, the later one has higher similarity with CO_2 stress response.
- Fraction of genes that are uniquely significant are highest at trehalose stress (T) and lowest at NaCl stress (N)

From analyzing the Venn diagrams, it can be concluded that stress responses are not additive. There are some genes (higher in case of trehalose compared to NaCl stress, please see Figure 5.42 to 5.45) that are significant at both the individual stress responses but non-significant from combined stress response. List of these genes are available in supplementary table S5.1 and S5.2.

5.4 Study of CO_2 stresses with or without other stresses

The response of the only elevated CO_2 stress (C) can be compared with elevated CO_2 stress response with NaCl (C(N)) and trehalose stress (C(T)). This comparison is important as it can tell us to what extent CO_2 stress is conserved at the transcriptional level along with other stresses. The approach used here was to compare the genes that are differentially expressed in three different stress responses. Finally pool of genes that are significant in all the three possible comparison (C, C(N) and C(T)), if any, were identified. Positively and negatively significant gene pools were compared separately. Table 5 shows the comparison of the positively significant genes. There was no gene that is positively significant at elevated CO_2 stress with or without other stresses. Interestingly C(N) and C(T) shown maximum number of common genes while C and C(N) shows the minimum number.

Table 5.19 The table shows the list of genes that are found common from pair-wise comparison of positively significant genes of stress responses.

C and C(N)	
1	copper-binding family protein
2	zinc finger protein-related
3	kelch repeat-containing protein / serine/threonine phosphoesterase family protein
C and C(T)	
1	CACTA-like transposase family (Tnp2/En/Spm)
2	cleavage stimulation factor, putative
3	copia-like retrotransposon family
4	DNA-binding protein-related
5	eukaryotic translation initiation factor 2 family protein / eIF-2 family protein
6	expressed protein

7	expressed protein
8	expressed protein
9	expressed protein
10	expressed protein
11	importin beta-2 subunit family protein
12	MATE efflux family protein
13	mRNA capping enzyme family protein
14	nuclear transport factor 2 (NTF2) family protein / RNA recognition motif (RRM)-containing protein
15	phosphatidyl serine synthase family protein
16	preprotein translocase secA subunit, putative
17	pseudo-response regulator 2 (APRR2) (TOC2)
18	ubiquitin system component Cue domain-containing protein
19	urease, putative / urea amidohydrolase, putative
20	zinc finger (CCCH-type) family protein
21	zinc finger (Ran-binding) family protein
C(N) and C(T)	
1	auxin efflux carrier family protein
2	basic helix-loop-helix (bHLH) family protein
3	calmodulin-binding protein-related
4	cold-acclimation protein, putative (FL3-5A3)
5	COP1-interactive protein 1 / CIP1
6	cytochrome P450 71A16, putative (CYP71A16)
7	cytochrome P450 family protein
8	cytochrome P450 family protein
9	dormancy/auxin associated family protein
10	expressed protein
11	expressed protein
12	expressed protein
13	expressed protein
14	expressed protein
15	expressed protein
16	glycosyl hydrolase family 3 protein
17	haloacid dehalogenase-like hydrolase family protein
18	laccase, putative / diphenol oxidase, putative
19	leucine-rich repeat transmembrane protein kinase, putative
20	major intrinsic family protein / MIP family protein
21	major latex protein-related / MLP-related
22	myrcene/ocimene synthase, putative
23	neurofilament protein-related
24	nodulin MtN3 family protein
25	nodulin MtN3 family protein
26	Null
27	phenylalanine ammonia-lyase, putative
28	plasma membrane intrinsic protein 1C (PIP1C) / aquaporin PIP1.3 (PIP1.3) / transmembrane protein B (TMPB)
29	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
30	protein kinase family protein
31	PWWP domain-containing protein
32	senescence-associated protein-related

33 U-box domain-containing protein
34 zinc finger (C3HC4-type RING finger) family protein

C and C(T) and C(N)

None

The results from the comparison of negatively significant genes are shown in table 6. In this case as well there was no gene that was significant at all the stresses C, C(N) and C(T). However, unlike positively significant genes there were only two genes that were common between C and C(T) and unfortunately none of them are annotated. There are not many genes in this list that are metabolically related. In both positive and negatively significant case I find highest number of genes that are common between C(T) and C(N). This implies, when CO_2 stress is applied with NaCl or trehalose stress it is more conserved compared to when it is applied alone. The rationale behind this observation is apparently imperceptible, and this could just be a numerical artifact.

Table 5.20 the table shows the list of genes that are found common from pair-wise comparison of negatively significant genes of stress responses.

C and C(N)

1 hypothetical protein
2 expressed protein
3 hypothetical protein
4 hypothetical protein
5 oxygen-evolving enhancer protein 3, chloroplast, putative (PSBQ2)
6 pseudogene, hypothetical protein

C and C(T)

1 expressed protein
2 hypothetical protein

C(N) and C(T)

1 CBL-interacting protein kinase 25 (CIPK25)
2 early nodule-specific protein, putative
3 ethylene receptor, putative (ETR2)
4 expressed protein
5 expressed protein
6 expressed protein
7 expressed protein
8 Glucose-6-phosphate/phosphate translocator, putative
9 glycosyl hydrolase family 17 protein
10 major intrinsic family protein / MIP family protein
11 monodehydroascorbate reductase, putative
12 peroxidase, putative

13 short-chain dehydrogenase/reductase (SDR) family protein
14 transferase family protein

C, C(N) and C(T)

None

5.5 Pathways and gene families of general interest

5.5.1 Auxin biosynthesis and regulation

Up-regulation of tryptophan biosynthesis pathway from all the stress comparisons opens up a question "why this pathway is up regulated under all the different stress conditions?" To search for this answer, biochemical pathways that consume tryptophan were analyzed. Tryptophan is used as precursor for biosynthesis of a plant hormone auxin which plays an important role in regulation. In this section biosynthesis and regulation of auxin will be discussed at length.

Auxin is an essential plant hormone that influences many aspects of plant growth and development, including cell division and elongation, differentiation, tropisms, apical dominance, senescence, abscission, and flowering [Davies, 1995]. Although auxin has been studied for over 100 years, the mechanisms of its biosynthesis remain elusive. Multiple pathways have been proposed [Cohen *et al.* 2003] for the biosynthesis of indole-3-acetic acid (IAA) (the main auxin), including two tryptophan-dependent pathways and a tryptophan-independent one (figure 5.46). YUCCA, a flavin monooxygenase (FMO)–like enzyme, catalyzes a key step in Arabidopsis tryptophan-dependent auxin biosynthesis [Zhao et al., 2001]. YUCCA catalyzes the N-oxygenation of tryptamine, and that this transformation is a rate-limiting step in auxin biosynthesis in many plants [Zhao et al., 2001]. The conversion of tryptophan to IAOx is known to be catalyzed by two cytochrome P450s, CYP79B2 and CYP79B3 in Arabidopsis [Cohen *et al.* 2003]. Another gene encoding the cytochrome P450 CYP83B1, results in increased indolic glucosinolate levels, a class of secondary compounds [Cohen *et al.* 2003]. It has been suggested that CYP83B1 serves regulates the branch point between IAA and indolic glucosinolate biosynthesis [Cohen *et al.* 2003]. Over-expression of CRY79B2 yielded plants with an IAA-overproduction phenotype, elevated IAA levels and increased expression of IAA-inducible genes similar to that seen in YUCCA over-expression [Cohen *et al.* 2003].

CYP79B2 and CYP83B1 are differentially localized within the cell. CYP79B2 is chloroplastic and CYP83B1 resides in the endoplasmic reticulum (ER). YUCCA appears to be cytoplasmic [Cohen *et al.* 2003]. The disparate localizations for these enzymes rule out their involvement in an IAA-synthase enzyme complex. The differential subcellular localizations suggest that a great deal of internal indolic trafficking is involved in the use and control of these pathways [Teale *et al.,* 2006].Although several proteins with clear binding specificities were identified, the functional characterization focused on one of them, AUXIN-BINDING PROTEIN-1 (ABP1), as it binds auxins with high specificity and affinity [Teale *et al.,* 2006].ABP1 is a soluble, ER-located, dimeric glycoprotein, which forms a-jellyroll barrel that carries auxin in a central hydrophobic pocket [Teale *et al.,* 2006].

So, auxin influences aspects of cell division, cell elongation and cell differentiation, although exactly how it is involved in each process (and to what extent they are intertwined) is not completely understood [Teale *et al.,* 2006]. Whereas the levels of some mRNAs decrease many fold in response to auxin, those of other mRNAs increase many fold (for example, Aux/IAA, *GRETCHENHAGEN-3*(*GH3*) and members of the small auxin up RNA (SAUR) gene family) [Teale *et al.,* 2006]. The complex auxin responses are mediated by two groups of well-studied genes: the Aux/IAA genes, which consist of 29 members, and the auxin response factor (ARF) genes with 23 members, in Arabidopsis thaliana. Aux/IAA proteins have been shown to function as negative regulators of gene expression [Teale *et al.*, 2006].

Figure 5.46 shows the auxin biosynthesis pathway dependent or independent of tryptophan. All the 5 stress responses are shown in the figure represented by 5 different background colors. Eight arrows or line signifies the significance level of the corresponding gene at 8 timepoints and a box around it signifies the significance level from paired SAM. Red and green signifies positively and negatively significant.

In the indole dependent auxin biosynthesis pathway CYP79B2 is positively significant at stress comparisons except CO_2 stress. CYP83B1 gene is also shows very similar significance response. However, nitrilase 1 and 3 (NIT1 and NIT3) genes are positively significant at only NaCl stress. I don't know how important role NIT1 and NIT3 plays in regulation of auxin production. If it does, then auxin synthesis is increasing under salt stress only. Over-expression of CYP83B1, which plays an important regulatory role at the branch point, also indicates possible increase in indolic glucosinolates production from typtophan under all the stress conditions.

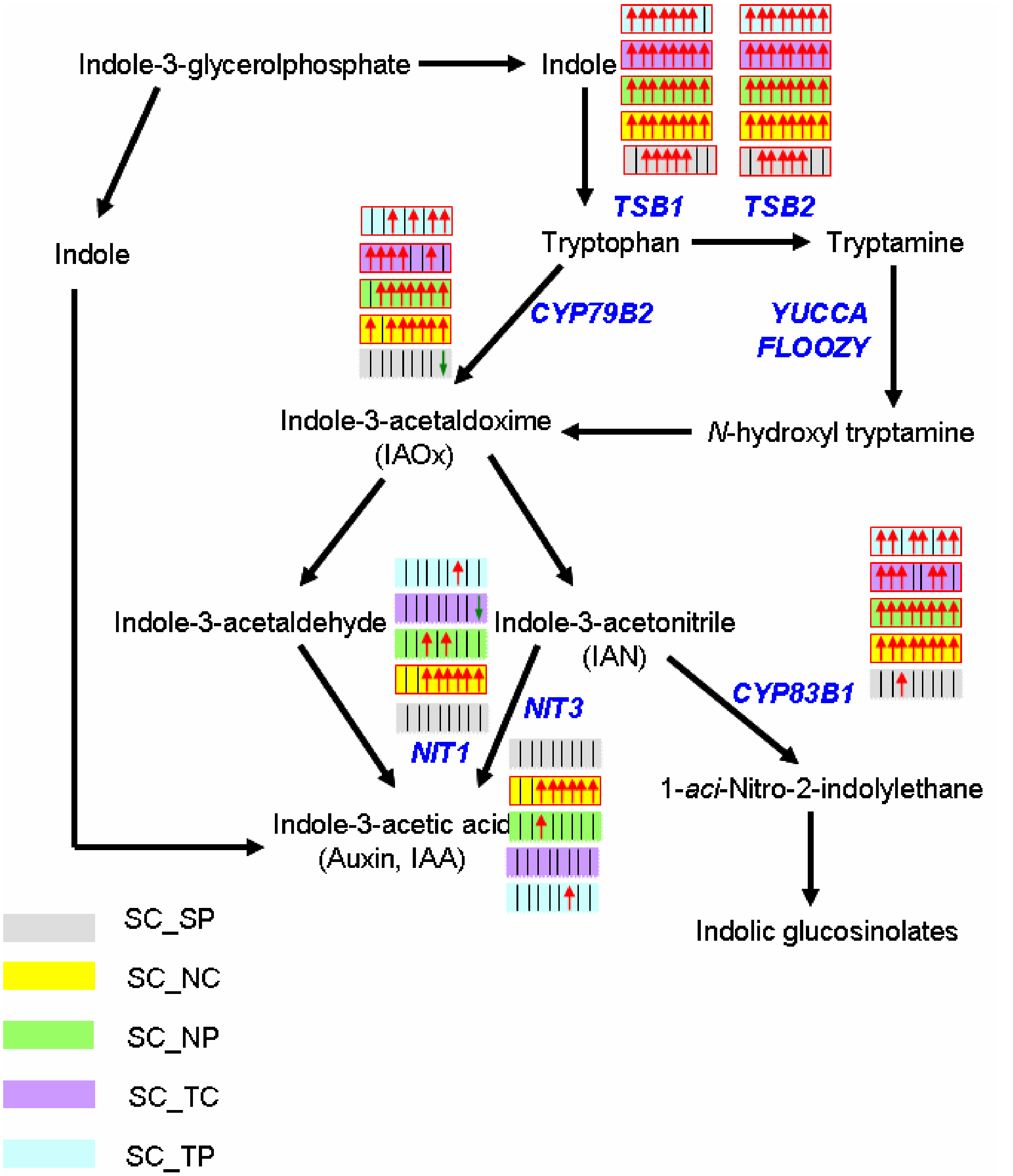
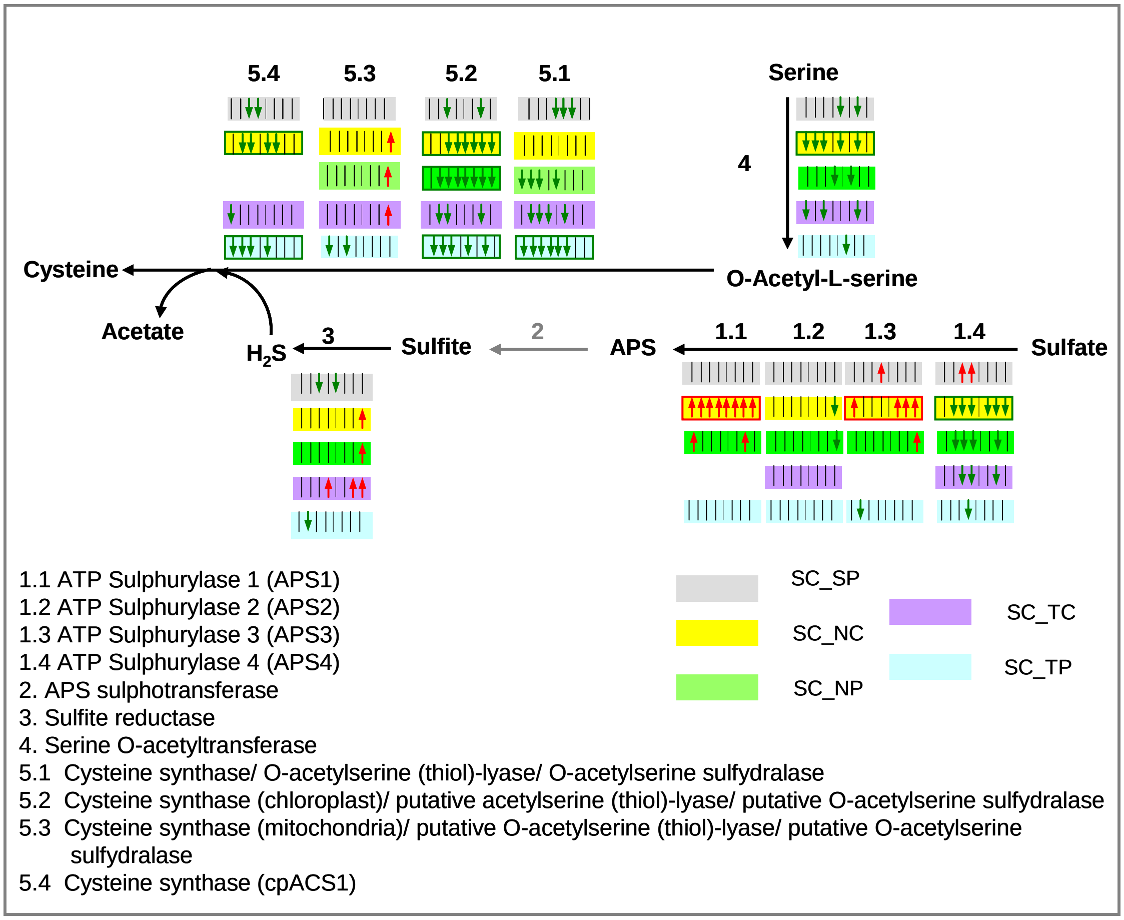


Figure 5.46 Pathway for tryptophan dependent and independent auxin synthesis. Significance profiles of the genes encoding enzymes of this pathway are shown using notation and color code same as figure 5.5.

5.5.2 Sulfate reduction and sulfur assimilation

As sulfur is an important constituent of cysteine, methionine and glutathione, it is essential to study its metabolism. Sulfate is initially activated in the presence of ATP to form adenosine-5-sulphatophosphate or adenosine phosphosulphate (APS) catalyzed by ATP sulphurylase (figure 5.47). Electrons required for sulfate reduction are derived from reduced ferredoxin, which may be formed in the chloroplast directly from photosystem I. However prior to reduction APS is bound to a carrier molecule, which is probably glutathione, and contains a free thiol group. The enzyme APS sulphotransferase catalyzes this reaction and the gene encoding this enzyme is not identified in *A. thaliana.* Sulfite is reduced to sulfide by sulfite reductase (Figure 5.47). Free sulfide reacts with O-acetylserine to form cysteine catalyzed by enzyme cysteine synthase. O-acetylserine is synthesized by the acetylation of serine, using acetyl CoA as substrate.

Comparison of significance profiles from different comparisons show NaCl stress is affecting the genes of this pathway most significantly. Four isoenzymes of APS are catalyzed by four different genes and they show differential expression. Multiple copies of cysteine synthase genes were also identified which are active in different cellular components and they show distinctly different significance profiles.



1.1 ATP Sulphurylase 1 (APS1)
1.2 ATP Sulphurylase 2 (APS2)
1.3 ATP Sulphurylase 3 (APS3)
1.4 ATP Sulphurylase 4 (APS4)
2. APS sulphotransferase
3. Sulfite reductase
4. Serine O-acetyltransferase
5.1 Cysteine synthase/ O-acetylserine (thiol)-lyase/ O-acetylserine sulfydralase
5.2 Cysteine synthase (chloroplast)/ putative acetylserine (thiol)-lyase/ putative O-acetylserine sulfydralase
5.3 Cysteine synthase (mitochondria)/ putative O-acetylserine (thiol)-lyase/ putative O-acetylserine sulfydralase
5.4 Cysteine synthase (cpACS1)

Figure 5.47 Significance profiles of sulfate reduction and sulfur metabolism pathway genes from all the stress comparisons

5.5.3 Histone Proteins

Analyses of significance profiles of genes encoding hisotne proteins are particularly important because of the important role they play in cellular physiology. None of these genes were differentially expressed in elevated CO2 stress. In case of NaCl stress and NaCl and CO_2 combined stress most (almost half) of these genes are negatively significant from paired SAM analysis [table 5.21]. Many of the genes that are non-significant from paired SAM are also significant at individual timepoints. When trehalose stress is applied 8 out of 28 genes were negatively significant from paired SAM analysis [table 5.21]. However when trehalose and CO_2 stress is applied in combination, the response was different from other stress responses and there were only two significant genes one of each significant types. The trehalose stress response was also found significantly different in pathways like Calvin cycle and photosynthesis, universal stress proteins.

Table 5.21 Significance profiles of genes encoding different histone proteins from all the five stress comparisons. Notation used is same as table 5.2.

Number	annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
SC_SP										
1	histone H2A, putative	-1	0	0	0	0	-1	0	0	0
2	histone H1, putative	0	-1	0	-1	-1	0	-1	0	0
3	histone H1.2	0	0	0	0	1	0	0	0	0
4	histone H1/H5 family protein	0	0	1	1	0	0	0	0	0
5	histone H2A, putative	0	0	0	0	0	0	0	0	0
6	histone H2A, putative	0	0	0	0	0	0	0	0	0
7	histone H2A, putative	0	0	1	1	0	1	0	0	0
8	histone H2A, putative	0	0	0	1	0	1	0	0	0
9	histone H2A, putative	0	0	0	0	0	0	0	0	0
10	histone H2A.F/Z	0	0	0	0	0	0	0	0	0
11	histone H2B	0	-1	0	0	-1	0	0	0	0
12	histone H2B, putative	0	-1	0	-1	0	0	0	0	0
13	histone H2B, putative	0	-1	0	-1	0	0	0	0	0
14	histone H2B, putative	0	-1	-1	0	0	0	0	0	0
15	histone H3	0	0	0	0	0	0	0	0	0
16	histone H3	0	-1	0	0	0	0	0	0	0
17	histone H3	0	0	0	0	0	0	0	0	0
18	histone H3	0	-1	0	-1	0	0	0	0	0
19	histone H3	0	-1	0	-1	0	1	0	0	0
20	histone H3, putative	0	0	1	0	0	0	0	-1	0
21	histone H3.2	0	0	1	0	0	0	0	0	0
22	histone H3.2	0	0	1	0	0	0	0	0	0
23	histone H3.2, putative	0	0	0	0	0	1	0	0	0
24	histone H4	0	0	0	0	0	0	0	0	0
25	histone H4	0	0	0	0	0	0	0	1	0
26	histone H4	0	0	1	1	0	0	0	0	0
27	histone H4	0	0	0	0	-1	0	0	0	0
28	histone H4	0	0	1	0	0	0	0	0	0
SC_NC										
1	histone H1, putative	-1	-1	-1	-1	-1	-1	-1	-1	-1

2	histone H1/H5 family protein	-1	0	0	0	0	0	-1	-1	-1
3	histone H1-3 (HIS1-3)	0	0	0	-1	0	-1	-1	-1	-1
4	histone H2A, putative	-1	-1	-1	-1	-1	0	-1	-1	-1
5	histone H2A, putative	-1	-1	-1	-1	0	0	-1	0	-1
6	histone H2A.F/Z	-1	-1	-1	-1	-1	-1	0	0	-1
7	histone H2B	-1	-1	-1	-1	-1	-1	0	0	-1
8	histone H2B, putative	-1	-1	-1	-1	-1	-1	0	0	-1
9	histone H3	-1	-1	0	-1	0	0	-1	0	-1
10	histone H4	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	histone H4	-1	-1	-1	-1	-1	-1	-1	0	-1
12	histone H4	-1	-1	-1	-1	-1	-1	-1	0	-1
13	histone H4	-1	-1	0	-1	-1	-1	-1	-1	-1
14	histone H1.2	0	1	0	0	0	0	0	0	0
15	histone H2A, putative	0	0	0	0	0	0	-1	-1	0
16	histone H2A, putative	-1	0	0	0	0	0	-1	-1	0
17	histone H2A, putative	0	0	1	0	0	0	0	0	0
18	histone H2A, putative	0	0	0	0	0	-1	-1	0	0
19	histone H2B, putative	-1	-1	-1	-1	0	0	0	0	0
20	histone H2B, putative	0	-1	-1	-1	0	0	0	0	0
21	histone H2B, putative	0	-1	0	-1	0	0	0	0	0
22	histone H2B, putative	-1	-1	0	-1	-1	0	-1	0	0
23	histone H3	0	-1	-1	-1	-1	0	0	0	0
24	histone H3	0	0	0	-1	0	0	0	0	0
25	histone H3	-1	-1	-1	-1	-1	0	-1	0	0
26	histone H3	-1	-1	0	-1	0	0	0	0	0
27	histone H3, putative	0	0	1	0	0	0	0	0	0
28	histone H3.2	0	0	1	0	0	0	0	0	0
29	histone H4	-1	0	-1	-1	0	0	-1	0	0
30	histone H1/H5 family protein	0	0	1	0	0	1	1	0	1
SC_NP										
1	histone H1, putative	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	histone H1.2	0	0	0	1	0	0	0	0	0
3	histone H1/H5 family protein	0	0	0	0	0	0	-1	0	0
4	histone H2A, putative	-1	-1	-1	0	-1	0	0	-1	-1
5	histone H2A, putative	-1	0	0	0	0	-1	-1	0	-1
6	histone H2A, putative	-1	-1	-1	0	-1	0	0	-1	-1
7	histone H2A, putative	-1	-1	0	0	0	0	0	0	0
8	histone H2A, putative	0	0	0	1	0	1	0	0	0
9	histone H2A,	-1	-1	-1	-1	-1	0	0	0	0

	putative									
10	histone H2A.F/Z	-1	-1	-1	-1	-1	0	0	0	-1
11	histone H2B	-1	-1	-1	-1	-1	0	0	0	-1
12	histone H2B, putative	-1	-1	-1	-1	-1	0	0	0	-1
13	histone H2B, putative	0	0	-1	-1	0	0	0	0	0
14	histone H2B, putative	-1	-1	-1	-1	-1	0	0	0	0
15	histone H2B, putative	0	0	0	0	-1	0	0	0	0
16	histone H3	-1	-1	0	-1	-1	0	-1	0	-1
17	histone H3	-1	-1	-1	-1	-1	0	0	0	0
18	histone H3	0	-1	0	-1	-1	0	0	0	0
19	histone H3	0	-1	-1	-1	-1	0	0	0	0
20	histone H3	0	-1	-1	-1	-1	0	0	0	0
21	histone H3, putative	-1	-1	0	0	-1	0	-1	-1	0
22	histone H3.2	0	0	1	0	0	0	0	0	0
23	histone H3.2	0	0	0	0	0	0	0	-1	0
24	histone H4	-1	-1	-1	-1	-1	0	0	0	-1
25	histone H4	-1	-1	-1	-1	-1	-1	-1	0	-1
26	histone H4	-1	-1	-1	-1	-1	-1	0	0	-1
27	histone H4	-1	-1	-1	-1	-1	0	0	0	-1
28	histone H4	-1	0	-1	-1	0	0	0	0	0
SC_TC										
1	histone H1, putative	-1	0	0	-1	-1	0	-1	0	-1
2	histone H1.2	-1	0	0	0	0	0	0	0	0
3	histone H1/H5 family protein	0	0	0	0	0	0	0	0	0
4	histone H2A, putative	-1	-1	-1	-1	-1	0	-1	-1	-1
5	histone H2A, putative	0	0	0	0	0	0	0	-1	0
6	histone H2A, putative	0	0	0	0	0	1	0	0	0
7	histone H2A, putative	-1	0	0	0	0	0	0	0	0
8	histone H2A, putative	-1	0	0	0	-1	0	-1	-1	0
9	histone H2A, putative	-1	0	0	0	0	1	0	0	0
10	histone H2A.F/Z	-1	0	0	0	-1	0	0	0	0
11	histone H2B	0	0	0	0	-1	0	0	0	0
12	histone H2B, putative	-1	0	0	0	0	0	0	0	0
13	histone H2B, putative	-1	0	0	0	-1	0	0	0	0
14	histone H2B, putative	-1	0	0	0	0	0	0	0	0
15	histone H2B, putative	-1	0	0	0	0	0	0	0	0
16	histone H2B, putative	-1	0	0	-1	-1	0	0	0	0
17	histone H3	-1	-1	0	-1	-1	0	-1	0	-1
18	histone H3	-1	0	0	-1	-1	0	0	-1	-1
19	histone H3	0	0	0	0	-1	0	0	0	0
20	histone H3	-1	0	0	0	-1	0	0	0	0

21	histone H3	0	0	0	-1	-1	0	0	0	0
22	histone H3.2	0	0	0	0	0	0	0	-1	0
23	histone H3.2	0	0	0	0	0	0	0	-1	0
24	histone H4	-1	-1	0	-1	-1	-1	-1	-1	-1
25	histone H4	-1	-1	0	-1	-1	-1	-1	-1	-1
26	histone H4	-1	0	0	-1	-1	0	-1	-1	-1
27	histone H4	-1	-1	0	-1	-1	-1	-1	0	-1
28	histone H4	0	0	0	-1	0	0	0	0	0
SC_TP										
1	histone H1, putative	-1	0	0	-1	-1	0	0	-1	-1
2	histone H1.2	0	0	0	0	0	0	0	-1	0
3	histone H1/H5 family protein	0	0	0	0	0	0	0	0	0
4	histone H2A, putative	-1	0	0	0	0	0	0	0	0
5	histone H2A, putative	-1	0	0	-1	0	0	0	0	0
6	histone H2A, putative	-1	0	0	0	0	0	-1	0	0
7	histone H2A, putative	-1	0	0	-1	0	0	-1	-1	0
8	histone H2A, putative	-1	0	0	-1	0	-1	0	0	0
9	histone H2A.F/Z	-1	0	0	-1	0	0	0	0	0
10	histone H2B	-1	0	0	-1	-1	0	0	0	0
11	histone H2B, putative	0	0	0	0	0	0	0	0	0
12	histone H2B, putative	-1	0	0	0	0	0	0	0	0
13	histone H2B, putative	-1	0	0	0	0	0	0	0	0
14	histone H2B, putative	-1	0	0	0	0	0	0	0	0
15	histone H3	0	0	0	0	0	0	0	0	0
16	histone H3	0	0	0	0	0	0	0	1	0
17	histone H3	-1	0	0	0	0	0	0	0	0
18	histone H3	-1	0	0	-1	-1	0	0	0	0
19	histone H3	0	0	1	0	0	0	0	1	0
20	histone H3, putative	0	0	0	0	0	0	0	0	0
21	histone H3.2	0	0	1	1	0	0	0	0	0
22	histone H3.2	1	1	1	1	1	1	1	0	1
23	histone H4	-1	0	0	0	0	0	0	0	0
24	histone H4	-1	0	0	-1	-1	0	0	0	0
25	histone H4	-1	0	0	-1	-1	0	0	0	0
26	histone H4	-1	0	0	-1	-1	0	0	0	0
27	histone H4	-1	0	0	-1	-1	0	0	0	0

5.5.4 Universal Stress Protein (USP) Family genes

It would be interesting to study the response of the genes annotated as universal stress protein (USP) under the different stress conditions. As the name imply do these genes really get differentially expressed under all the stress conditions? Do they respond by up-regulation or down regulation? Do these genes respond get positively significant under one stress, while negatively significant in the other? Or is there any preferential timepoint when they significant? To answer these questions significance level of USP genes were from all the stress responses were compared together (table 5.22). In general it was observed that in case of CO_2 stress, there was no gene significant from paired SAM and minimum number of significant genes at individual timepoints. While in case of NaCl and CO_2 combined stress, 7 out of 14 genes were negatively significant from paired SAM. Even the genes that are non-significant from paired SAM, were significant at multiple timepoints. Genes were predominantly down-regulated and there were only two genes At2g47710 and At2g21620 that are positively significant at three and one timepoints respectively. At3g53990 gene was found strongly negatively significant from in all the stress responses. In case of NaCl, NaCl and CO_2 and trehalose stress this gene was negatively significant from paired SAM. While in CO_2 stress and trehalose and CO_2 stress it was down-regulated at 3 and 4 timepoints respectively. Hence, it can be hypothesized that general response of At2g47710 gene is by its down-regulation. On the contrary At2g47710 gene was positively significant from paired SAM in NaCl stress and trehalose and CO_2 stress. It was also positively significant at multiple timepoints in other stress responses and negatively significant at none. Similar conclusion of At2g47710 genes up-regulation to general stress response can be formulated. Though among the differentially expressed genes most of them were negatively-significant, but in case of paired SAM analysis of trehalose and CO_2

combined stress two genes were significantly over-expressed but none were under-expressed, which seems to be different from other stress responses.

Table 5.22 Significance profiles of the universal stress protein genes from different stress comparisons. Notation used is same as table 5.2.

Locus	**annotation**	**1h**	**3h**	**6h**	**9h**	**12h**	**18h**	**24h**	**30h**	**Paired SAM**
SC_SP										
1	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
2	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
3	universal stress protein (USP) family protein	0	0	0	0	0	0	-1	0	0
4	universal stress protein (USP) family protein	0	-1	0	-1	0	-1	0	-1	0
5	universal stress protein (USP) family protein	0	0	0	1	0	0	0	0	0
6	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
7	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
8	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
9	universal stress protein (USP) family protein	0	0	0	-1	0	0	0	0	0
10	universal stress protein (USP) family protein	0	0	0	-1	-1	-1	0	0	0
11	universal stress protein (USP) family protein	0	0	0	0	0	-1	0	0	0
12	universal stress protein (USP) family protein / responsive to dessication protein (RD2)	0	0	0	1	0	0	0	0	0
SC_NC										
1	universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	universal stress protein (USP) family protein	0	0	-1	-1	-1	-1	-1	-1	-1
3	universal stress protein (USP) family protein	0	0	1	1	0	0	1	0	0
4	universal stress protein (USP) family protein	-1	0	-1	0	0	-1	0	0	0
5	universal stress protein (USP) family protein	0	0	-1	0	0	0	-1	0	0
6	universal stress protein (USP) family protein	-1	-1	-1	0	-1	-1	0	0	0
7	universal stress protein (USP) family protein	0	0	0	0	0	0	0	1	0
8	universal stress protein (USP) family protein	1	0	1	0	1	0	0	0	0
9	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
10	universal stress protein (USP) family protein	1	1	1	1	1	1	1	1	1
SC_NP										
Locus	Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
1	universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	universal stress protein	0	-1	-1	-1	-1	0	-1	0	-1

	(USP) family protein		🟩	🟩	🟩	🟩		🟩		🟩
3	universal stress protein (USP) family protein	0	-1	-1	-1	-1	-1	-1	-1	-1
4	universal stress protein (USP) family protein	-1	-1	0	-1	0	-1	-1	-1	-1
5	universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	universal stress protein (USP) family protein	0	0	0	-1	-1	-1	-1	-1	-1
7	universal stress protein (USP) family protein	-1	-1	-1	-1	0	-1	-1	-1	-1
8	universal stress protein (USP) family protein / responsive to dessication protein (RD2)	0	0	0	0	1	0	0	0	0
9	universal stress protein (USP) family protein	-1	0	0	0	0	0	0	0	0
10	universal stress protein (USP) family protein	-1	0	-1	-1	0	-1	0	0	0
11	universal stress protein (USP) family protein	-1	0	0	0	0	0	-1	-1	0
12	universal stress protein (USP) family protein	0	0	1	1	1	0	0	0	0
13	universal stress protein (USP) family protein	-1	-1	0	-1	0	-1	-1	0	0
14	universal stress protein (USP) family protein	0	0	0	0	0	0	-1	0	0
SC_TC										
Locus	Annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
1	universal stress protein (USP) family protein	-1	-1	-1	-1	0	-1	-1	-1	-1
2	universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	universal stress protein (USP) family protein	-1	-1	-1	-1	-1	-1	0	-1	-1
4	universal stress protein (USP) family protein	0	0	-1	0	0	0	0	0	0
5	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
6	universal stress protein (USP) family protein	-1	0	0	0	0	0	-1	0	0
7	universal stress protein (USP) family protein	-1	-1	0	0	0	0	0	0	0
8	universal stress protein (USP) family protein	0	0	0	1	0	0	0	0	0
9	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
10	universal stress protein (USP) family protein	0	0	0	-1	0	-1	-1	0	0
11	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
12	universal stress protein (USP) family protein	0	0	0	-1	0	0	0	0	0
13	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
14	universal stress protein (USP) family protein / responsive to dessication protein (RD2)	0	0	-1	-1	0	-1	-1	0	0
SC_TP										
1	universal stress protein (USP) family protein	0	0	-1	0	0	0	0	0	0
2	universal stress protein (USP) family protein	0	-1	-1	0	0	-1	0	0	0
3	universal stress protein	-1	0	0	-1	0	0	0	0	0

	(USP) family protein	🟩			🟩					
4	universal stress protein (USP) family protein	-1	0	0	-1	0	0	0	0	0
5	universal stress protein (USP) family protein	-1	0	0	-1	0	-1	0	0	0
6	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
7	universal stress protein (USP) family protein	0	1	1	1	1	1	1	0	1
8	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
9	universal stress protein (USP) family protein	1	1	1	1	1	0	1	0	1
10	universal stress protein (USP) family protein	0	0	0	0	0	0	0	0	0
11	universal stress protein (USP) family protein	0	0	0	0	0	0	1	0	0
12	universal stress protein (USP) family protein	-1	0	0	-1	-1	-1	0	-1	0
13	universal stress protein (USP) family protein	-1	0	0	-1	0	0	0	0	0
14	universal stress protein (USP) family protein / responsive to dessication protein (RD2)	0	0	0	0	1	0	0	0	0

5.5.5 Cellulose synthase family protein

Cellulose is an unbranched polymer consisting of D-glucose molecules which are connected to each other by glycosidic linkage. The biochemical basis for cellulose synthesis is not well understood. Sequencing of Arabidopsis thaliana genome has revealed the existence of multiple copies of cellulose synthase which shows 64% sequence similarity, however the functions of individual isoenzymes are not known. Cellulose synthase is posttranslationally regulated and is known to be phosphorylated, but the mechanisms that regulate activity are not yet known. The genes for cellulose synthase are developmentally regulated, but there is relatively little evidence for environmental regulation of expression. Here I see cellulose synthase family protein genes are differentially expressed due to applied stress (table 5.23). Differential responses of these genes are most prominent in case of NaCl stress and NaCl and CO_2 combined stress. In both the cases almost all the genes except two were significantly up or down regulated from paired SAM analysis. UDP-glucose is the building block of cellulose and UDP-glucose pyrophosphorylase catalyzes the UDP-glucose synthesis reaction. UDP-glucose pyrophosphorylase gene was down-regulated almost at all timepoints and also from paired SAM in response to NaCl stress and NaCl and CO_2 combined stress. In case of CO_2 stress none of these genes were differentially expressed from paired SAM. At4g24010 and At5g16910 genes were positively significant at most of the timepoints and also from paired SAM in NaCl stress and NaCl and CO_2 combined stress. In case of other stress responses they were also positively significant at individual timepoints, but didn't qualify as significant from overall analysis. On the other hand At2g32530 and At2g32540 genes were negatively significant in most of the timepoints and also from paired SAM analysis in all the stress responses except CO_2 stress. Interestingly enough these genes are closely located in the chromosome. It needs to be verifies if proximity on the chromosomal map possibly make these two genes to be co-regulated.

Table 5.23 Significance profiles of cellulose synthase family protein genes from different stress response experiments. Notation used is same as table 5.2.

Locus	annotation	1h	3h	6h	9h	12h	18h	24h	30h	Paired SAM
SC_SP										
1	cellulose synthase family protein	0	1	0	0	0	0	0	0	0
2	cellulose synthase family protein	0	0	0	0	0	0	0	0	0
3	cellulose synthase family protein	0	0	0	0	0	-1	0	0	0
4	cellulose synthase family protein	-1	0	0	0	-1	-1	0	0	0
5	cellulose synthase family protein (CslD3)	0	1	0	1	0	0	0	0	0
6	cellulose synthase family protein	0	1	1	1	0	0	0	0	0
7	cellulose synthase family protein	0	0	0	0	0	-1	0	0	0
8	cellulose synthase family protein	0	0	-1	-1	-1	-1	0	0	0
9	cellulose synthase family protein	0	0	0	0	0	0	0	0	0
10	cellulose synthase family protein	0	0	0	0	0	0	0	0	0
11	cellulose synthase family protein	0	0	-1	0	0	-1	0	0	0
SC_NC										

1	cellulose synthase family protein	1	1	1	1	1	0	0	1	1
2	cellulose synthase family protein	1	1	1	1	0	0	1	0	1
3	cellulose synthase family protein	0	1	1	1	1	1	0	0	1
4	cellulose synthase family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	cellulose synthase family protein (CslD3)	0	0	1	0	0	0	0	0	0
6	cellulose synthase family protein	0	0	-1	-1	0	-1	-1	0	-1
7	cellulose synthase family protein	0	0	-1	-1	-1	-1	-1	-1	-1
8	cellulose synthase family protein	0	0	-1	-1	-1	-1	-1	-1	-1
9	cellulose synthase family protein	0	0	0	1	0	1	1	1	0
SC_NP										
1	cellulose synthase family protein	1	1	1	1	1	0	1	1	1
2	cellulose synthase family protein	1	1	1	1	0	0	0	1	1
3	cellulose synthase family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
4	cellulose synthase family protein	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	cellulose synthase family protein	0	0	-1	-1	0	-1	-1	-1	-1
6	cellulose synthase family protein	0	0	-1	-1	-1	0	0	0	0
7	cellulose synthase family protein	0	0	0	0	0	0	0	1	0
SC_TC										
1	cellulose synthase family protein	0	1	0	1	0	0	0	0	0
2	cellulose synthase family protein	0	0	0	0	0	0	0	0	0
3	cellulose synthase family protein	0	0	-1	0	0	0	-1	0	0
4	cellulose synthase family protein	-1	0	-1	0	0	-1	-1	0	-1
5	cellulose synthase family protein (CslD3)	0	0	0	0	0	0	0	0	0
6	cellulose synthase family protein	0	0	0	0	0	0	0	0	0
7	cellulose synthase family protein	0	0	0	0	0	-1	0	0	0
8	cellulose synthase family protein	-1	0	-1	-1	-1	-1	-1	0	-1
9	cellulose synthase family protein	-1	0	-1	-1	-1	-1	-1	0	-1
10	cellulose synthase family protein	-1	0	-1	0	0	0	0	0	0
11	cellulose synthase family protein	0	0	0	1	1	0	1	1	0
SC_TP										
1	cellulose synthase family protein	1	0	0	1	0	0	0	0	0
2	cellulose synthase family protein	0	0	0	0	0	-1	0	0	0
3	cellulose synthase family protein	0	0	-1	0	0	-1	0	0	0
4	cellulose synthase family protein	-1	-1	-1	-1	-1	0	0	0	-1
5	cellulose synthase	1	0	1	1	0	0	0	0	0

	family protein (CslD3)	🟥		🟥	🟥					
6	cellulose synthase family protein	0	-1	-1	0	0	-1	0	0	0
7	cellulose synthase family protein	-1	0	-1	-1	-1	-1	0	0	-1
8	cellulose synthase family protein	-1	0	-1	-1	-1	-1	0	-1	-1
9	cellulose synthase family protein	-1	0	-1	-1	-1	-1	0	0	-1
10	cellulose synthase family protein	-1	0	0	0	0	0	0	0	0

5.6 Comparison of MiTimeS results from different stress responses

5.6.1 Comparison of significant gene numbers

Figure 5.48 and 5.49 show the percentage of genes that are positively and negatively significant at each timepoint over the 30 hours of experimental duration. Each curve on the figure corresponds to one of the pair-wise stress comparisons. Comparing the two figures it is apparent that positive and negatively significant genes corresponding to each comparison shows similar profiles. For all the stresses except CO_2 stress (SC_SP comparison), there is a distinct acute stress response marked by large number of significant genes at time1h. For all the four comparisons and for both the significance type, number of significant genes decrease from 1 to 3h. This decrease is most prominent in case of trehalose stress and trehalose with CO_2 stress. The curves corresponding to trehalose stress and trehalose and CO_2 combined stress are very similar, especially in case of negatively significant genes, implying a possible dominance of trehalose stress in the combined stress response. Same is true with NaCl stress and NaCl and CO_2 combined stress. In case of CO_2 stress, number of differentially expressed genes increase steadily for first 9 hours and falls again. It can be hypothesized that it takes longer time for plants to experience the CO_2 stress compared to the other stress responses shown here.

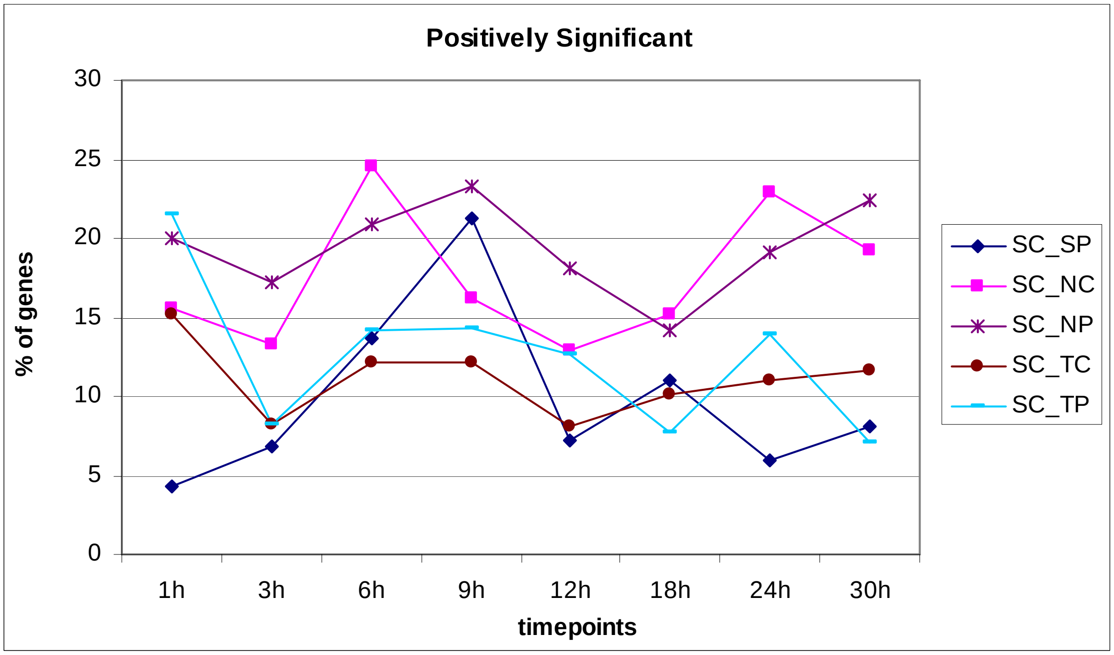


Figure 5.48 Percentage of genes that are identified as positively significant over the duration of the experiment.

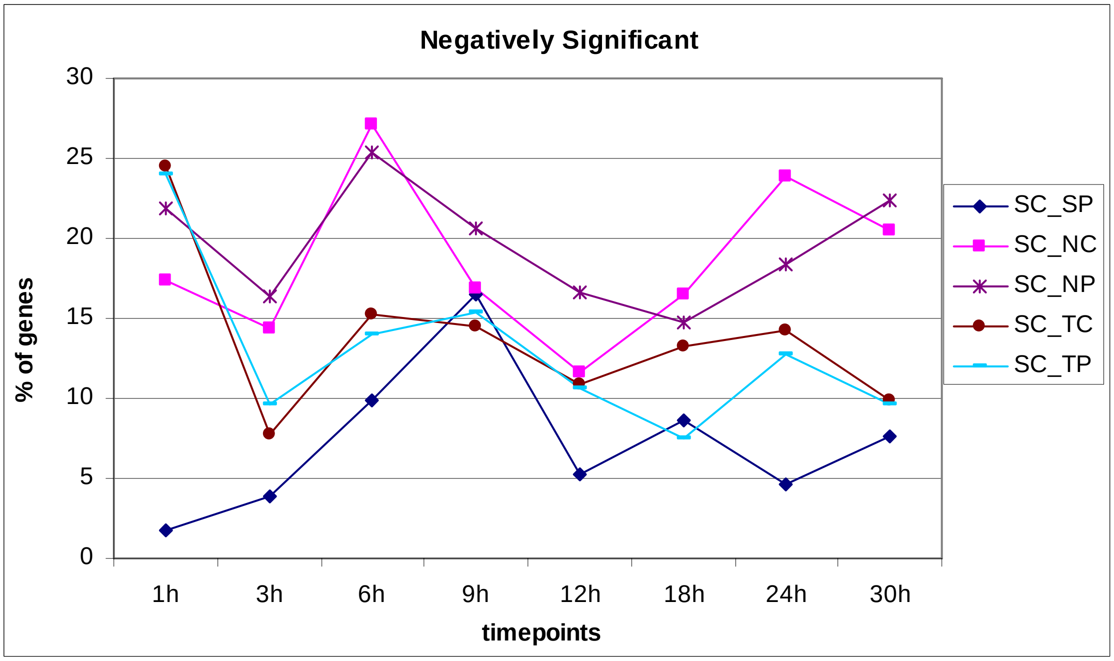


Figure 5.49 Percentage of genes that are identified as negatively significant over the duration of the experiment.

5.6.2 SV Score distribution

Figure 5.50 shows the SV score distribution of all the 5 pair-wise comparisons that is discussed in the previous section. It shows that sucrose stress has distinctly different transcriptional response, which was also observed when compared in the context of *A. thaliana* physiology. SC_SP and SC_TP have minimum and maximum average SV score of 0.26 and 0.3 respectively. In SC_SP nearly half of the genes had SV score of 0.29.

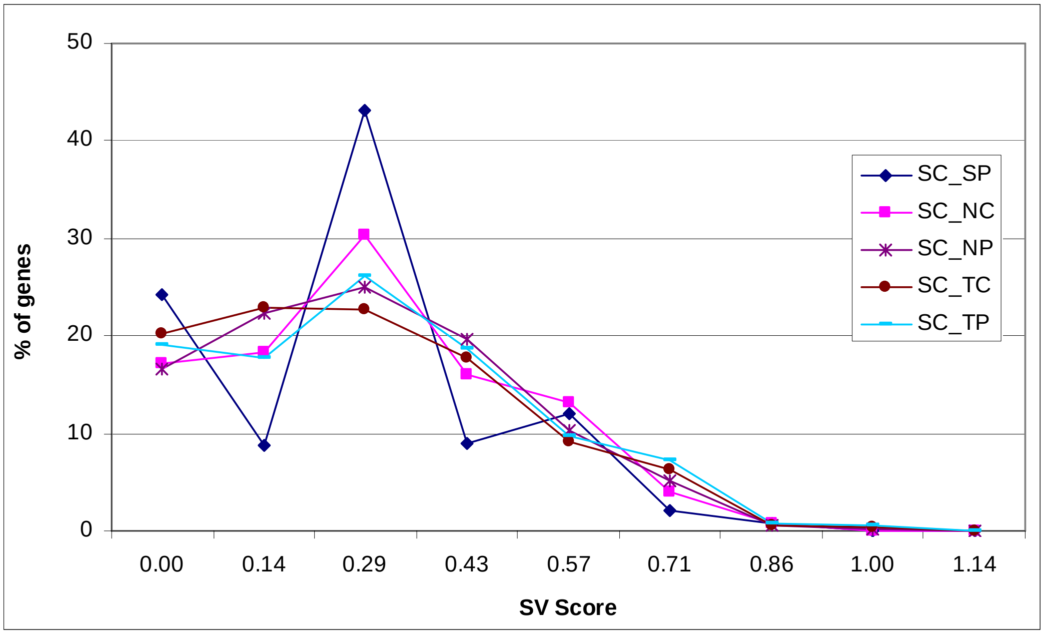


Figure 5.50 SV score distribution of five stress comparisons

When analyzed separately, trehalose stress and trehalose and CO_2 combined stress was following very similar SV score distributions, which again implies the dominance of trehalose stress in the combined stress response (figure 5.51).

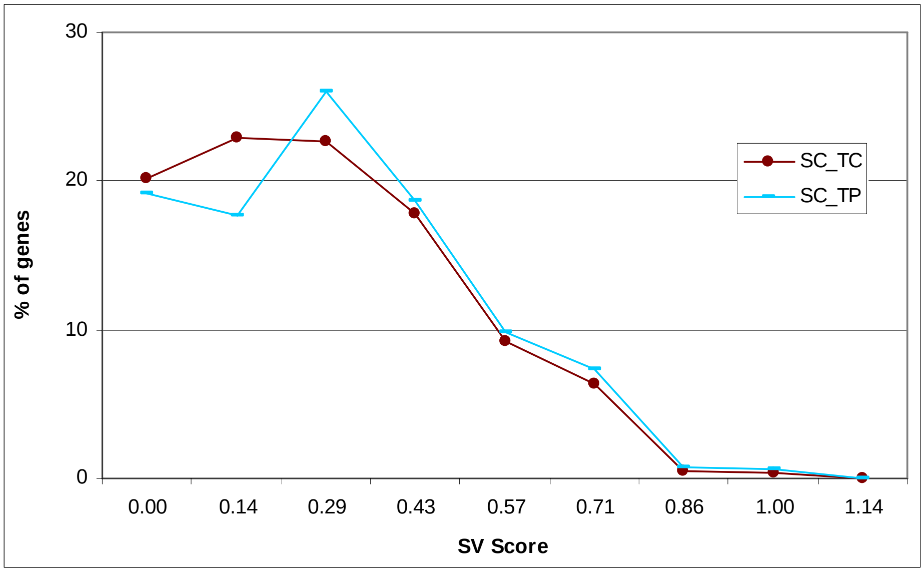


Figure 5.51 SV score distribution of trehalose stress and trehalose with CO_2 stress shows similarity.

Comparison of three CO_2 stress responses, with or without trehalose and NaCl stress shows similarity in their SV score distributions (Figure 5.52). This similarity in SV score distributions can be attributed to inherent response behavior of the CO_2 stress.

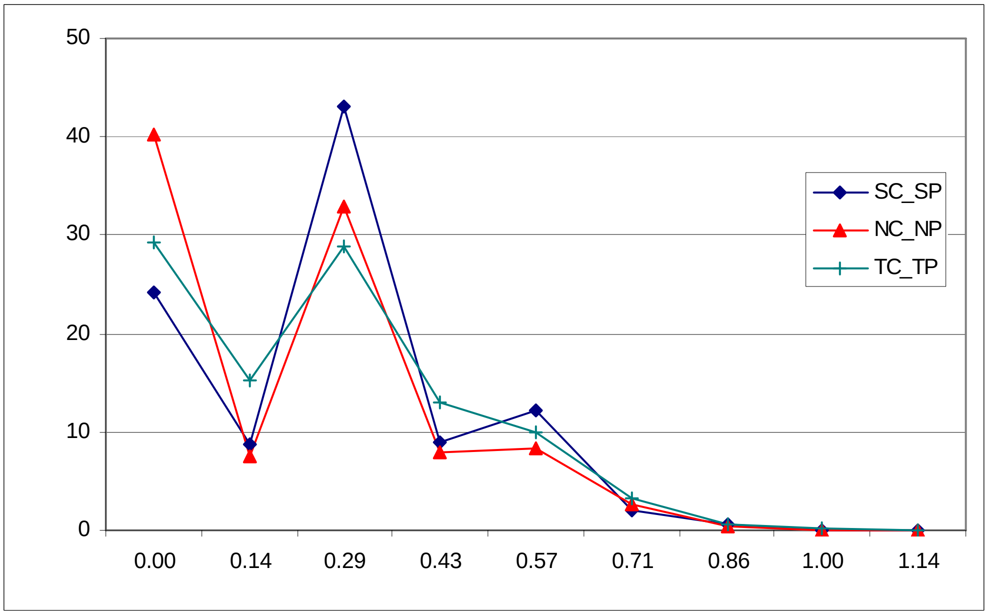


Figure 5.52 SV score distribution of three CO_2 stress responses show close similarity.

6 Comparison of different stress responses

In the previous chapter the individual stress responses were discussed independently in the context of *Arabidopsis thaliana* physiology. One of the main reasons for carrying out multiple stress experiment was indeed not just to study them independently but also to compare these responses. The comparison can be at several levels, like matching the significant genes for individual stress responses and find the ones that are common for all the stresses. At a higher level it can be studied how the combined stress response is varying from the individual stress responses or how it is conserved leading to answer the cross talk between different responses. Studying all the stress responses together will provide considerable number of samples (54 in this case) over which genes can be clustered and the ones clustered together in such wide physiological space most likely would lead to biological implications unobserved before. The unique experimental design and plethora of valuable data provided to the scientific community from this experiment will be basis for developing and validating novel framework for statistical analysis of high-throughput molecular fingerprint data.

6.1 Comparison of all stress responses

6.1.1 Clustering of all the experiments

An experimental design where multiple perturbations were applied to the same system in a similar fashion will allow us to compare them effectively. Studying how the transcriptional response changes under different stresses is an important motivation for this research. One of the ways to analyze it effectively is to cluster the experimental timepoints. After using 75% cutoff (i.e. genes that are present in 12 or more out of 16 experiments are selected) on all the experiments a total of 11204 genes were used for this analysis. Hierarchical clustering with Euclidean distance was used to cluster the sample timepoints and the results are shown in figure 6.1.

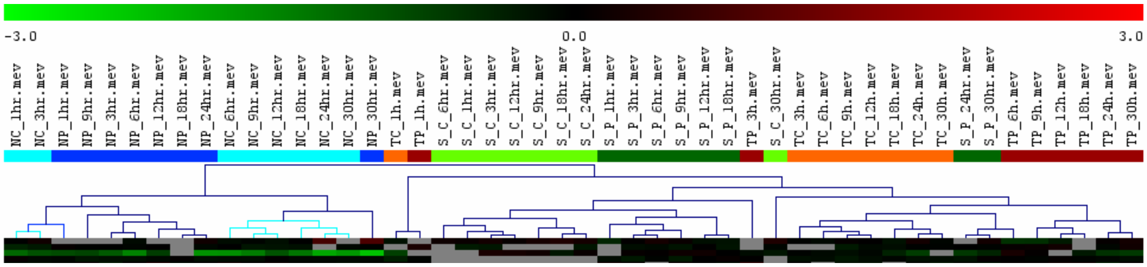


Figure 6.1: The figure shows the clustering profile of the experimental timepoints from hierarchical clustering using Euclidean distance. Stress responses are mostly separated from each other.

The clustering clearly creates three main clusters corresponding to sucrose experiment (SC and SP), NaCl experiment (NC and NP) and trehalose experiment (TC and TP). Within each main group there are two subgroups corresponding to presence or absence of elevated CO_2 stress. In spite of this overall trend there are few anomalies. 1 and 3 hr timepoints of NC was found to cluster with NP group, whereas 30h timepoint of NP is found to be closer to NC cluster. 1 h timepoints of TC and TP was found to cluster separately from rest of the timepoints. I speculate that initial response of trehalose stress is very strong and unique, causing it to show similar transcriptional response to each other but different from rest of the timepoints. It was also observed that 1h timepoint of TC and TP have maximum number of significant genes. Again I observed 30h timepoint of SC, 24 and 30h timepoitns of SP are clustering with the trehalose cluster. The rationale behind this observation is not very clear, but a general conclusion can be drawn that usually 30h timepoints are showing most distinct anomaly.

PCA analysis was carried out on the same data and using the same color code figure 6.1. In the figure 6.2 we see there is a separation between SC and SP, again between TC and TP. However NC and NP timepoints seems to mingle together and no clear separation was observed. Interestingly when only NC and NP are clustered together they were distinctly separated.

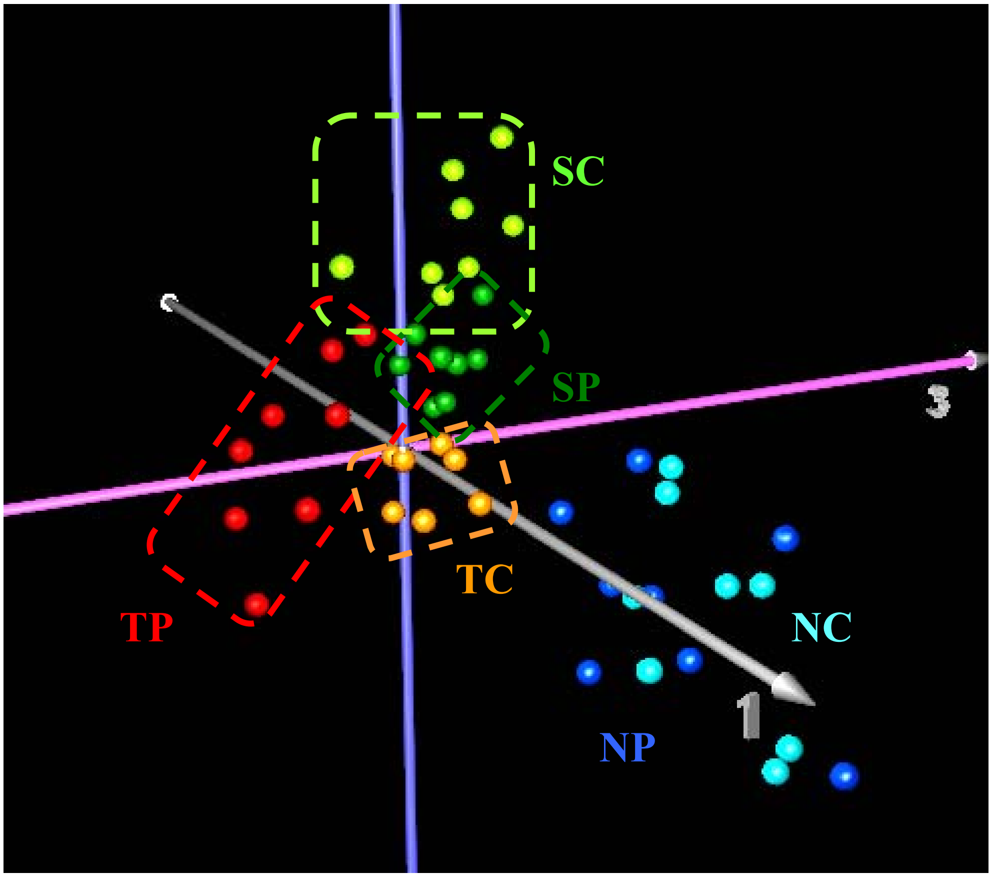


Figure 6.2: From the PCA analysis it is clear that most the stress responses are distinctly different except NC and NP, as they cluster together separated from rest of the samples.

The reason could be when we cluster all the experiments together, the principal components representing the maximum variance is different from that of the principal components created from only NC and NP experiments. Hence, as the coordinate system becomes different, their representation with respect to this new system also becomes different. NC and NP timepoints seems to have moved together along the principal component 1 which accounts for maximum variability. As explained before, the salt stress is perceived to create stronger transcriptional response than the CO_2 stress; hence the separation between NC and NP is not that prominent compared to the effect of salt stress. When only NC and NP are clustered together, the effect of salt stress is not present and the weaker difference between the NC and NP experiments becomes perceptible. Hence, to make the minute difference prominent, the stronger effect has to be removed. Another interesting thing that can be observed is the timepoints of the experiments without CO_2 stress, like SC and TC, are much widespread compare to the corresponding perturbed timepoints. A plausible reason could be, when elevated CO_2 stress is applied, the natural variation of the transcriptional state with time is reduced with respect to it's control state.

6.1.2 Identification of common significant genes from all stresses

All the different stresses were analyzed independently to find out genes that are significantly over or under expressed in each case. It would be important to study if there exist set of genes that are differentially expressed under all the stress conditions. This gene pool, being significant under variety of stress conditions, could take part in general abiotic stress response, if any. It would be important to study from literature the response of these genes under different other abiotic stress not considered in this particular project. Any finding of some of these genes involvement in other stress response will justify the existence of a common stress response gene regulatory network.

Positively and negatively significant genes from paired SAM for CO_2, NaCl, NaCl with CO_2, trehalose and trehalose with CO_2 were compared to find the common genes. There were 31 and 16 genes positively and negatively significant respectively belonging to this category (shown in table 6.2A and 6.2B). 9 out of these 16 negatively significant genes are related to metabolism. However, only 3 out of 31 positively significant genes were related to metabolism.

Table 6.1: List of genes positively significant from all the pair-wise comparisons

ABC transporter family protein
AP2 domain-containing transcription factor, putative
armadillo/beta-catenin repeat family protein / U-box domain-containing protein
bZIP family transcription factor
cytochrome P450, putative
DC1 domain-containing protein
disease resistance protein (TIR-NBS-LRR class), putative
DNA topoisomerase family protein
endoribonuclease L-PSP family protein
ethylene-responsive factor, putative
exportin1 (XPO1)
expressed protein
expressed protein
expressed protein
expressed protein
EXS family protein / ERD1/XPR1/SYG1 family protein
germin-like protein (GLP9)
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
lysine and histidine specific transporter, putative
O-acetyltransferase-related
peroxidase, putative
phytochrome B (PHYB)
protein kinase family protein
protein kinase family protein
protein kinase-related
protein phosphatase 2C, putative / PP2C, putative
putative endochitinase
tryptophan synthase, beta subunit 1 (TSB1)
tryptophan synthase, beta subunit 2 (TSB2)
Ubiquitin carboxyl-terminal hydrolase family protein

Table 6.2 List of genes negatively significant from all the pair-wise comparisons

ATP synthase protein I –related
carbonic anhydrase 2 / carbonate dehydratase 2 (CA2) (CA18)

chlorophyll A-B binding protein, putative / LHCI type II, putative
expressed protein
Glycerate dehydrogenase / NADH-dependent hydroxypyruvate reductase
glycine cleavage system H protein 1, mitochondrial (GDCSH) (GCDH)
glycine hydroxymethyltransferase / serine hydroxymethyltransferase / serine/threonine aldolase (SHM1)
Invertase/pectin methylesterase inhibitor family protein
nodulin, putative
pentatricopeptide (PPR) repeat-containing protein
phosphoglycolate phosphatase, putative
phosphoglycolate phosphatase, putative
polygalacturonase, putative / pectinase, putative
ribosomal protein L29 family protein
Transport protein-related
uridylyltransferase-related

Among the positively significant genes related to metabolism are tryptophan synthase, beta subunit 1 (TSB1) and beta subunit 2 (TSB2). This enzyme catalyzes the last reaction in the tryptophan biosynthesis pathway. Tryptophan biosynthetic pathway in plants is of particular importance because it is the source of precursors for many important indolic secondary products, in addition to its role in protein synthesis [Pruitt and Last, 1993]. These compounds include the plant growth regulator auxin [Wright et al., 1991], anti-microbial phytoalexins [Tsuji et al., 1992], and alkaloids and glucosinolates [Haughn et al., 1991]. Molecular biological and genetic approaches to the tryptophan pathway should provide insights into the regulation of metabolite flow through the pathway and the coordination of primary and secondary product biosynthesis in plants [Pruitt and Last, 1993]. It was also observed that in response to a punctual mechanical wound affected transcript level of many genes, including genes from tryptophan biosynthesis pathway was increased many folds within 90 to 120 min [Reymond et al., 2000], which shows important role tryptophan pathway plays in plant stress response in general.

Among the other genes that are positively significant under all the stress conditions are different protein kinase and protein phosphatase (PP2C) genes which are involved in different signaling pathway. There are many protein kinase genes and always we do not know their exact role in signaling pathway. However, there could be some core signaling pathway which involves many of the genes present in this list (table 2A). Another gene encoding ethylene-responsive factor is also possibly involved in cellular signaling.

It is stated before that among the genes that are negatively significant under all stresses; a significant fraction is related to metabolism. Interestingly most of these genes like phosphoglycolate phosphatase, glycine hydroxymethyltransferase / serine hydroxymethyltransferase / serine/threonine aldolase (SHM1), glycerate dehydrogenase / NADH-dependent hydroxypyruvate reductase are found to be involved in photorespiration. Photorespiration takes place in three different cellular components like chloroplast, mitochondria, peroxisome and these gene products catalyses reactions in those cellular components respectively. It is a significant observation that photorespiratory pathway genes are under-expressed not only under elevated CO_2 stress, but also in presence of other stresses, which apparently should not affect the carbon assimilation reaction directly. No verification of this postulate that down regulation of photorespiration under general stress response was obtained from literature hence needs to be studied in detail.

One gene encoding cytoplasmic phosphoglycolate phosphatase was also found to be negatively significant. I don't know if the under-expression of this gene is due to the overall under-expression of genes related to photorespiration. Phosphoglycolate phosphatase catalyses the conversion of phosphoglycolate to glycolate and according to the literature the reaction takes place in chloroplast [Heldt, 3rd Ed.], followed by transport of glycolate from chloroplast to peroxisomes through cytosol. I hypothesize that, some of the unconverted phosphoglycolate are also converted to glycolate in the cytoplasm when in transit. The other possibility could be glycolate that enters peroxisome is not only from chloroplast, but also from cytoplasm. Hence cytoplasmic conversion of phosphoglycolate to glycolate plays an important role in photorespiratory pathway.

Carbonic anhydrase (CA) encoding gene was also found to be negatively significant under all stress conditions. CA catalyzes the reversible hydration of CO_2 to bicarbonate and is one of the most abundant soluble proteins in the leaves of C3 higher plants, representing up to 1 to 2% of the soluble leaf protein [Fett and Coleman, 1994]. Within the C3 chloroplast it has been postulated that CA activity could maintain the supply of CO_2 for Rubisco by speeding the dehydration of HCO_3^- or by facilitating the diffusion of CO_2 across the chloroplast envelope via maintenance of the equilibrium between the inorganic carbon species (Reed and Graham, 1971). Two more gene encoding chlorophyll A-B binding protein, putative / LHCI type II, putative and uridylyltransferase-related proteins were also negatively significant. All the three genes mentioned are involved in regulation of photosynthesis and carbon fixation. Though enzyme involved in carbon fixation reaction, rubisco, is not present in the list of negatively significant genes, however, photosynthesis rate is possibly getting affected when the stresses are applied.

6.2 Clustering of genes from all the experiments

After observing how the different stresses are changing the transcriptional state of *Arabidopsis thaliana* liquid cultures it would obviously be interesting to analyze the expression profile of different genes under all the experimental conditions. In the context of gene regulation, it would be a good opportunity to find out genes that are showing similar expression profiles under all the different stress conditions. Co-expression of these genes could be due to their co-regulation. With this objective genes were clustered using hierarchical clustering implemented in TIGR MeV software with both Euclidean and Pearson correlation distances. Genes clustered together with Euclidean distance will have close expression values under all the experimental conditions. Whereas when Pearson's correlation is used genes showing similar expression profiles (not necessarily with similar expression values) will cluster together. Both the clustering results were studied extensively to test whether mathematical similarity has true biological relevance. In this context it was observed Pearson's correlation distance created clusters that seem to have more biological significance. Some of these clustering results are discussed here in the context of *Arabidopsis thaliana* physiology.

6.2.1 Genes related to Photosynthesis and carbon fixation:

From hierarchical clustering using Pearson's correlation distance I could identify a cluster of genes that are that are directly or indirectly related to photosynthesis and carbon fixation. This cluster contains 80 genes [Supplementary table 6.1] and is divided into different sub-clusters. Rubisco enzyme catalyses the carbon fixation reaction in Calvin cycle and genes coding for small subunits regulates the carbon fixation reaction. Four genes encoding all the four subunits of Rubisco small chain, 1A, 1B, 2B and 3B were found cluster together and form a sub-cluster. While Euclidian distance was used for clustering, three out of four genes clustered together, only subunit 1A clustered separately. Based on this observation I hypothesize that expression of the B subunits of the RubisCO small chain are probably regulated together. All the B subunits (1B, 2B and 3B) of RubisCO are physically close in the chromosome, which might facilitate their regulation. The subunit 1A, whose regulation is also synchronous to the other subunits, is possibly regulated in slightly different way. Five other genes coding for sedoheptulose-1,7-bisphosphatase, chloroplast (At3g55800), fructose-1,6-bisphosphatase (At3g54050), glyceraldehyde-3-phosphate dehydrogenase B (At1g42970), fructose-bisphosphate aldolase, putative (At2g21330) and phosphoribulokinase (PRK) (At1g32060) the enzymes that catalyze several other reactions of carbon fixation pathway was also belong to the same cluster. Phosphoribulokinase catalyses the reaction from Ribulose-5P to Ribulose-1,5-BP, which is used as substrate for carbon fixation reaction by RubisCO. Genes encoding the three Calvin cycle enzymes sedoheptulose-1,7-bisphosphatase, chloroplast (At3g55800), fructose-1,6-bisphosphatase (At3g54050), glyceraldehyde-3-phosphate dehydrogenase B (At1g42970) were found to form a sub-cluster. It is indeed an interesting observation to find the enzymes of the three consecutive reactions are produced in exactly synchronous way. Two other genes carbonic anhydrase 1, chloroplast (At3g01500) and inorganic carbon transport protein-related (At1g70760), which are also involved in carbon fixation indirectly, belongs to this cluster.

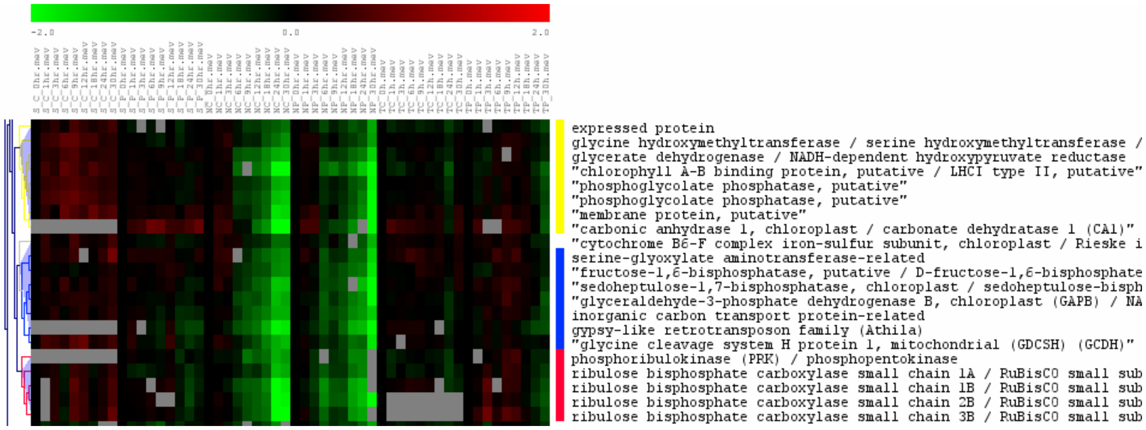


Figure 6.3: Clustering of all the experiments shows RubisCO subunits, Calvin cycle and photorespiration pathway genes form three distinct sub-clusters marked with different colors.

Another interesting observation is, the cluster mentioned above contains 6 genes related to photorespiration. 5 out of these 6 genes form a sub-cluster. Only the gene annotated as "serine-glyoxylate aminotransferase-related (At2g13360)" clustered with genes encoding Calvin cycle pathway enzymes. It was not obvious why this gene is expressed closer to Calvin cycle pathway compared to other photorespiratory genes. Three sub-clusters mentioned above coding for RuBisCO subunits and phosphoribulokinase, Calvin cycle pathway enzymes and photorespiratory pathway enzymes from a cluster of genes predominantly coding for enzymes. There were another cluster within the cluster of 80 genes [Figure 6.4] which are coding for photosystem I and II reaction center and chlorophyll A-B binding protein.

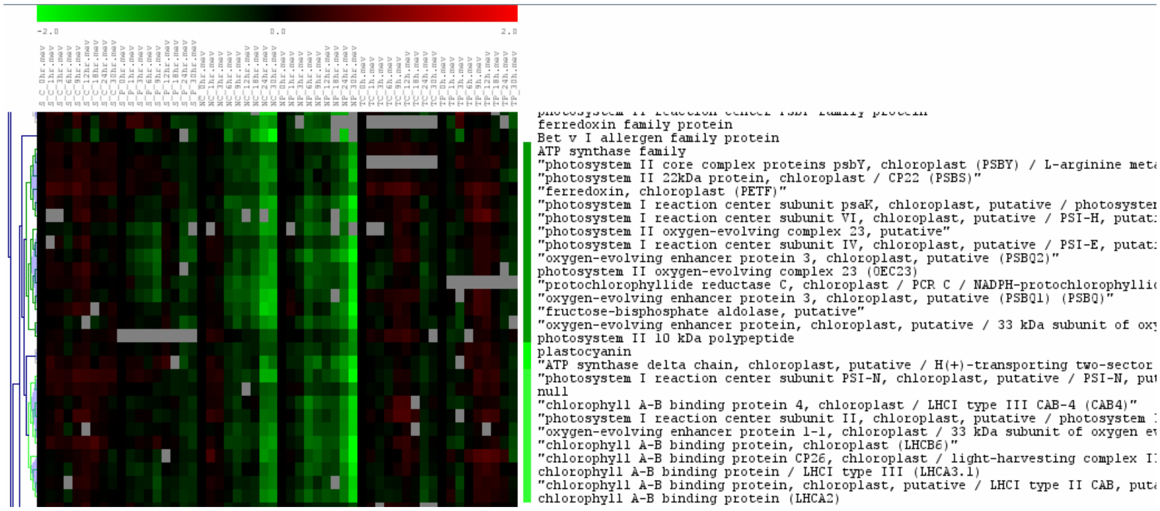


Figure 6.4 Clustering analysis shows genes related photosystem I and II and chlorophyll A-B binding proteins are clustering together. Both of these genes are involved in photosynthesis.

Few of other genes coding for enzymes which are not directly related to carbon fixation or photosynthesis were found to be present in this cluster of genes. One of these genes annotated as "glutathione S-transferase, putative". Glutathione S-transferase (GST) is differentially regulated under different form of biotic and abiotic stresses. Specific GSTs are reported to be induced upon infection, in response to treatment with ozone, hydrogen peroxide, glutathione and biotic elicitors, plant hormones, heavy metals, heat shock, dehydration, wounding and senescence, however little headway has been made in matching specific GST isozymes with either their preferred substrates or their function in vivo [Wagner et al., 2002]. Another gene with locus ID At1g65230, annotated as expressed protein was found to cluster closely with the genes coding for photorespiratory pathway genes. The coding sequence of this gene was not found to match closely with any other genes of *Arabidopsis* genome. Though I can't say for sure, but I believe protein coded by this gene may play some role in photorespiratory pathway.

6.2.2 Similarity between expression profiles and sequence alignment

Histones are the major structural proteins of chromosomes. The DNA molecule is wrapped twice around a Histone Octamer to make a Nucleosome [Albert et al. 4th Ed.]. Genes encoding different subunits of histones were found to form two different clusters. Out of 5 histone H4 genes used for analysis 4 clustered together along with two other genes. The fifth histone gene (locus At2g28740) had clustered with genes coding for different other subunits of histone. To investigate the different behavior of this anomalous histone gene, the nucleotide sequences of the hsitone genes from these two clusters were compared using sequence alignment software of European Bioinformatics Institute called ClustalW [http://www.ebi.ac.uk/clustalw/]. Phylogenic tree constructed from sequence similarity shows the outlier histone H4 gene (At2g28740) has diverged from rest of the rest of the H4 genes and is closer to the histone H2B and H3 genes it has clustered with [figure 6.5]. Though all the histone H4 proteins have same amino acid sequence, but they differ in their nucleotide sequence which gives rise to different regulation of these genes.

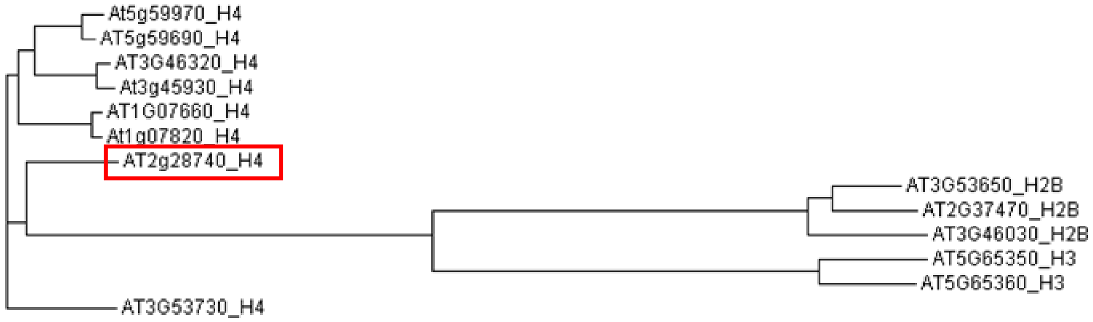


Figure 6.5: Sequence alignment shows At2g28740 (histone H4) gene is closer to the hsitone subunits H2B and H3 compared to other H4 subunits

A similar phenomenon was also observed in nitrilase genes. Nitrilases are enzymes that catalyze the hydroxylation of nitriles to carboxylic acid and ammonia. Two genes encoding nitrilase 1 (NIT1) (At3g44310) and nitrilase 3 (NIT3) (At3g44320) were found to cluster together. However another gene with a putative nitrilase annotation (At4g08790) remained separate from the above mentioned genes. A similar approach, as explained above was used which showsn NIT1 and NIT3 are closer in phylogenic tree [using ClustalW] according to their sequence similarity while the gene with the putative function is not (Figure 6.6).

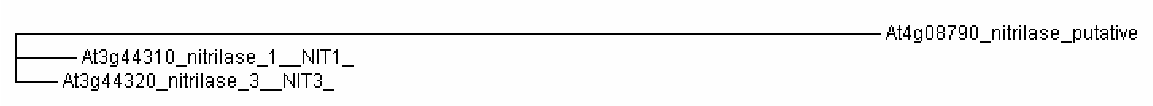

At4g08790_nitrilase_putative
At3g44310_nitrilase_1__NIT1_
At3g44320_nitrilase_3__NIT3_

Figure 6.6: Sequence alignment shows putative nitrilase gene is different from NIT1 and NIT3 genes which are clustering together from gene expression data.

Another example of sequence similarity of genes leading to a similar expression profile is shown here. The Arabidopsis genome contains many gene families that are not found in the animal kingdom. One of these is the multidrug and toxic compound extrusion (MATE) family, which has homology with bacterial efflux transporters [Andrew et al., 2001]. Multidrug transporters form a large class of membrane proteins present in the cells of most organisms. These proteins bind to a variety of potentially cytotoxic compounds and remove them from the cell in an ATP- or proton-dependent process. MATE family belongs to the family of multidrug transporter family and is characterized by the presence of 12 putative transmembrane segments [Andrew et al., 2001]. Figure 6.7 shown 6 genes belonging to MATE family where two are from chromosome 1 and four are from chromosome 2 (the first number after "*At*" in a gene's locus ID represent it's chromosome number). Their difference in the coding sequence of genes belonging to chromosome 1 and 2 are can be observed from the phylogenic tree. When clustered, these genes have shown similar clustering pattern consistent with the phylogenic tree.

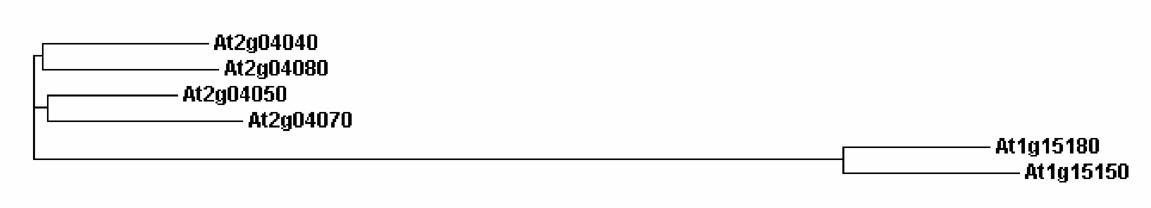


Figure 6.7: MATE family proteins from similar tree form sequence alignment and clustering based on gene expression.

6.2.3 Nitrate reduction pathway

In the nitrate assimilation pathway there are two nitrogen reduction reactions shown in figure 6.8.

$$\textbf{Nitrate} \xrightarrow[\text{Nitrate Reductase 2}]{\text{Nitrate Reductase 1}} \textbf{Nitrite} \xrightarrow[\text{Reductase}]{\text{Nitrite}} \mathbf{NH_4^+}$$

Figure 6.8: Nitrate reduction to nitrite takes place in two different steps catalyzed by NR1, NR2 and nitrite reductase.

Nitrate reduction reaction takes place in cytoplasm and is catalyzed by the enzyme Nitrate reductase, which has two subunits 1 and 2 encoded by two different genes At1g77760 and At1g37130 respectively. Nitrate reductase is an exceptionally short lived protein. Its half life is few hours, hence activity if nitrate reductase is regulated by its synthesis. The reduction of nitrite to ammonia is carried out in chloroplast [Heldt, 3rd Edition] using the reducing power of Ferredoxin and catalyzed by Nitrite reductase. Nitrate reduction is strictly controlled so that nitrate reduction doesn't proceed faster than nitrite reduction, since otherwise the toxic level of nitrite will accumulate in the cell. Here the expression of profiles of two nitrate reductase genes is plotted along with that of one nitrite reductase gene [figure 6.9]. Visually all three of them show almost similar expression profiles. NR1 gene shows comparatively less change in their expression level, while NR2 gene shows higher magnitude change. It implies response of NR2 gene is more amplified than that of NR1 gene. When the expression profiles of these genes were compared statistically, it was observed NR1 and ferredoxin-nitrite reductase genes are clustered closely both from Pearson's correlation and Euclidian distance. Correlation coefficient of these two genes expression values, a measure of how similar they are, was 0.82, which is a fairly high value. Correlation coefficient between NR2 and ferredoxin--nitrite reductase genes was found fairly low (-0.1). Again, correlation coefficient between NR1 and NR2 genes was 0.13, which implies almost no correlation. From comparison of individual stresses it can be observed that for most of the comparisons differential response of NR1 and ferredoxin--nitrite reductase genes are similar. When NaCl and trehalose stresses are applied individually, NR2 was found positively significant from paired SAM analysis while the other two genes were not. As it is known from biochemistry cell doesn't want to create a situation where nitrate reduction rate is higher than that of nitrite reduction. To keep the nitrate reduction at the same pace as that of nitrite reduction, possibly it is NR1 gene that actively regulates the overall Nitrate reduction reaction, not NR2 gene.

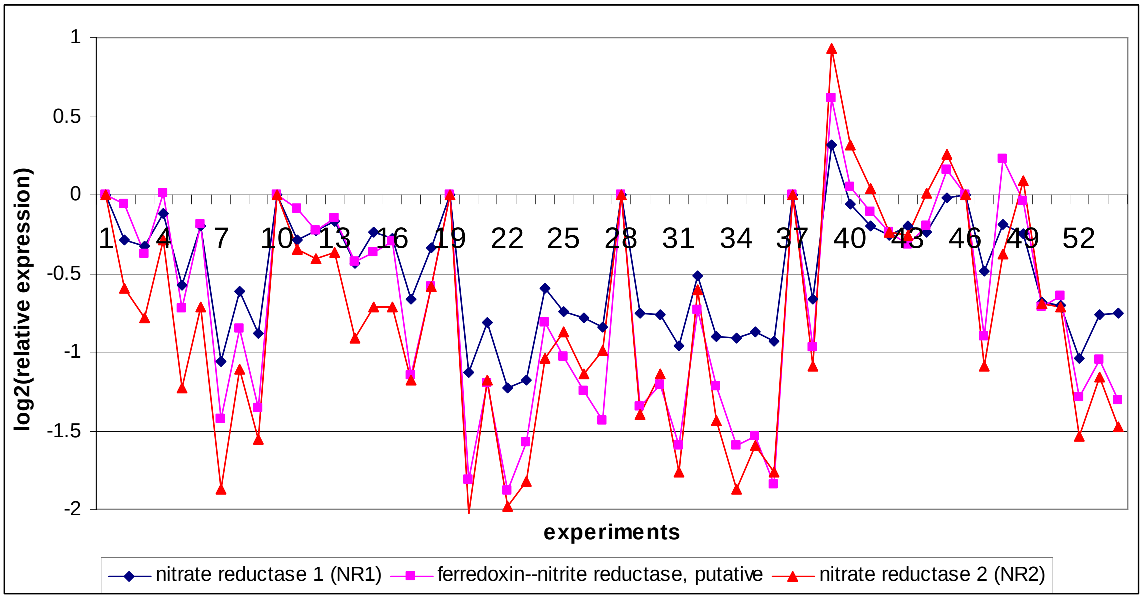


Figure 6.9: NR1 and nitrite reductase genes show high degree of co-expression.

6.2.4 Glutamine synthetase 2 (GS2) and glutamate synthase 1 (GLU1) are coexpressed

Both glutamine synthetase and glutamate synthase plays in important role in nitrogen assimilation for the biosynthesis of amino-acids. Both of these enzymes have two isoenzymes which are encoded by two different genes. These isoenzymes are active in different cellular compartments. In most of our analysis these isoenzymes have shown differential response under almost all the stress comparisons. Glutamine synthetase 1 and 2 (GS1 and GS2) are active in chloroplast and cytoplasm respectively. Glutamate synthase 1 (GLU1) mainly regulated the GOGAT mechanism of nitrogen assimilation in chloroplast and usually have higher level of expression. GLU2 is a housekeeping gene and shows lower expression level. GLU1 was also found to show similar expression profile as that of GS2 [Heldt, 3rd Edition]. In most of the pair-wise significance analysis GLU1 and GS2 have shown similar significance profiles. When gene expression data was clustered based on all the experimental timepoints, they were found to cluster closely. Pearson's correlation coefficient of GS2 and GLU1 was 0.83. However the correlation coefficient between GLU1 and GLU2 or GS1 and GS2 were quite poor. This observation validates the previous finding that GS2 and GLU1 work in concert and show similar expression profiles.

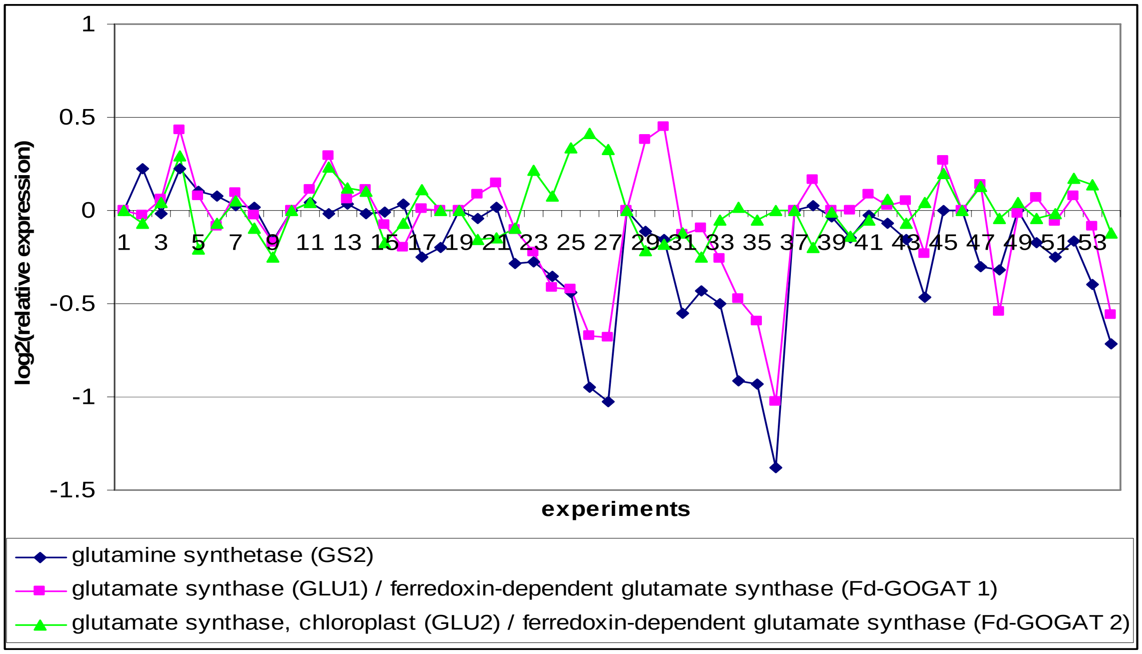


Figure 6.10: GS2 and GLU1genes show high degree of co-expression.

6.2.5 Tryptophan biosynthesis pathway

I noticed number of genes encoding enzymes for Phenylalanine, tyrosine and tryptophan biosynthesis pathway is also being co-expressed. These genes were found to form clusters. These genes form one big and multiple small clusters. These clustering results are pictorially represented in figure 6.11, where genes belonging to same cluster are colored same. If there are multiple genes corresponding to one enzyme then they are shown by multiple arrows.

I observed metabolic pathway branch of tryptophan biosynthesis from chorismate (4 out of 5 genes of this pathway) are co-regulated. These genes are colored blue. Two genes corresponding to phosphoribosylanthranilate isomerase 1 and 2 show similar expression profile with respect to each other but different from cluster of other genes marked in blue. Three anthranilate synthase beta subunit genes were clustering together (marked in brown), however show different expression profiles compared to corresponding alpha subunit gene. I do not have any information for the genes corresponding to the arrows colored gray, either because they are un-annotated or because they are missing in our analysis.

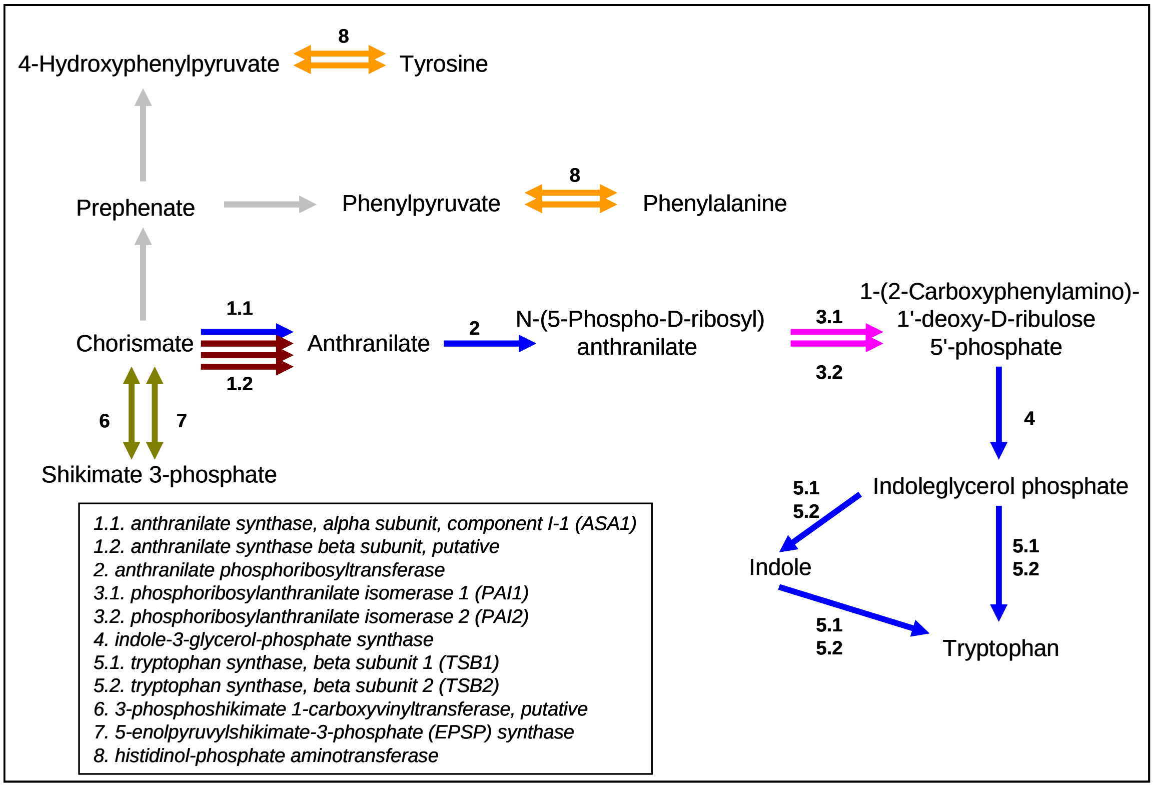

4-Hydroxyphenylpyruvate
8
Tyrosine
Prephenate
Phenylpyruvate
8
Phenylalanine
Chorismate
1.1
1.2
Anthranilate
2
N-(5-Phospho-D-ribosyl) anthranilate
3.1
3.2
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5'-phosphate
4
6
7
Shikimate 3-phosphate
Indoleglycerol phosphate
5.1
5.2
Indole
5.1
5.2
5.1
5.2
Tryptophan
1.1. anthranilate synthase, alpha subunit, component I-1 (ASA1)
1.2. anthranilate synthase beta subunit, putative
2. anthranilate phosphoribosyltransferase
3.1. phosphoribosylanthranilate isomerase 1 (PAI1)
3.2. phosphoribosylanthranilate isomerase 2 (PAI2)
4. indole-3-glycerol-phosphate synthase
5.1. tryptophan synthase, beta subunit 1 (TSB1)
5.2. tryptophan synthase, beta subunit 2 (TSB2)
6. 3-phosphoshikimate 1-carboxyvinyltransferase, putative
7. 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
8. histidinol-phosphate aminotransferase

Figure 6.11 Tryptophan biosynthesis genes show similar expression profiles. Arrows colored same are clustered together from hierarchical clustering.

From Alkaloid biosynthesis II pathway I observed 4, 2 and 2 genes coding tropinone reductase, putative / tropine dehydrogenase, copper amine oxidase, putative, and phenylalanine ammonia-lyase are coregulated. Each set of genes that are co-regulated are marked with different colors. Interestingly tropinone reductase, putative /

tropine dehydrogenase gene found from KEGG pathway were not showing up in the corresponding cluster of Alkaloid biosynthesis II pathway from EASE analysis.

Table 6.3 Genes from Alkaloid biosynthesis II pathway are clustered together

tropinone reductase, putative / tropine dehydrogenase, putative
short-chain dehydrogenase/reductase (SDR) family protein / tropinone reductase, putative
tropinone reductase, putative / tropine dehydrogenase, putative
tropinone reductase, putative / tropine dehydrogenase, putative
copper amine oxidase, putative
copper amine oxidase, putative
phenylalanine ammonia-lyase 2 (PAL2)
phenylalanine ammonia-lyase 1 (PAL1)

6.3 Relation between gene regulation and chromosomal location

I tried to investigate if there is any correlation between genes that are showing very similar expression values or profiles over the wide range of experiments can be correlated to their proximity in chromosome. To answer this question I identified gene pairs that are closest in their expressions values or profiles and also closest to each other physically on the chromosomal map. From a repository of 10963 genes that used for analysis, with Pearson's correlation distance I identified 327 such pairs, whereas with Euclidean distance 292 pairs were identified and they are listed in supplementary table S6.2A and S6.2B. In this case Pearson's correlation distance was found to perform slightly better than Euclidean distance. When these two list of genes were compared, to my surprise I noticed 248 genes are common between them. Hence conclusions are not very much dependent on the distance measure that we choose. Most of these correlated gene pairs in supplementary table S6.2A and S6.2B have same annotation. Possibly, these genes' expressions need to be regulated together and these regulations become more efficient when they are physically close on the chromosome.

7 FUTURE WORKS

The work presented in the thesis is one of beginning, though not completely perfect, steps in the long journey systems biology is about to go through. To the best of my knowledge, it was the first experimental design where multiple perturbations were applied to a eukaryotic system, individually or in combination and the integrated transcriptional and metabolic response was studied in a time-series manner.

Analysis of single cellular fingerprint can only study one of the facades of biology and is not always sufficient to derive conclusions. Simultaneous study of multiple cellular fingerprints and the data integration will lead to more comprehensive and convincing results. The metabolic profiling analysis was carried out by a fellow graduate student and only the results relevant in the context of transcriptional profiling were discussed here.

The need for studying a system from multiple perturbations is well talked about in systems biology research. However, studying the response of multiple perturbations and comparing them with individual ones is not very common. The importance of doing time-series high-throughput experiment is becoming more strengthened these days as regulatory mechanisms can only be derived from dynamic responses, not from snapshots.

Results obtained from this multiple stress high-throughput experiment lead to several biologically relevant conclusions. However in the process it opens up lot more questions than what it answers, which need to be addressed. Experience gained from carrying out the experiments and analyzing the data will help to develop better experimental design and analytical methods in future.

To formulate a mathematical framework for the integration of gene expression and metabolic data:

Transcriptomic and metabolomic data generated from the multiple perturbation time-series experiment can be used basis for development of mathematical models to integrate different data types. As part of future work a mathematical model is proposed with the following assumptions:

- All the reactions follow 0 order kinetics.
- Active enzyme concentration at any time point is same as the total enzyme concentration.
- Total enzyme concentrations at any time point are proportional to the corresponding gene expressions at that time point.

While this is definitely an oversimplified picture of *in vivo* reality, what it describes can be used to extract information about the activity of metabolic pathway from the transcriptional data. If a gene is over-expressed then this indicates (based on transcriptional data in absence of any flux analysis data) that the corresponding reaction rate is possibly higher as well.

The model is been demonstrated below in the context of a simple linear pathway around metabolite B, but it can be further extended to more complex reaction networks.

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \qquad (7.1)$$

Assuming k_1 and k_2 the rate constants of the reactions A->B and B->C respectively. The mass balance around metabolite B can be written as follows:

$$\frac{dB}{dt} = k_1 - k_2 \tag{7.2}$$

The rate constants k_1 and k_2 are proportional to the total concentrations of the enzymes catalyzing the corresponding reactions. At a particular time point if C_1 and C_2 are the proportionality constants, then equation 7.2 can be rewritten as

$$\frac{dB}{dt} = C_1E_1 - C_2E_2 \tag{7.3}$$

where E_1 and E_2 are the total enzyme concentrations of the corresponding reactions. Based on the initially stated assumptions, the total enzyme concentrations are proportional to the corresponding gene expressions. If D_1 and D_2 are the respective proportionality constants, then equation 7.3 can be rewritten as

$$\frac{dB}{dt} = C_1D_1G_1 - C_2D_2G_2 \tag{7.4}$$

Equation 7.4 can be integrated with proper boundary conditions to establish a preliminary relation between gene expression and metabolite concentrations

$$B_t - B_0 = \int_0^t (C_1D_1G_1 - C_2D_2G_2)dt \tag{7.5}$$

Where B_0 and B_t are the concentration of metabolite B at time zero and time t, respectively. On dividing both sides of the equation 7.5 by B_0

$$(B_t / B_0) - 1 = \int_0^t ((C_1D_1 / B_0)G_1 - (C_2D_2 / B_0)G_2)dt \tag{7.6}$$

Similarly the gene expression values in equation 7.6 can be normalized with respect to the gene's expression at time zero.

$$(B_t / B_0) - 1 = \int_0^t ((C_1 D_1 G_1{}^0 / B_0)(G_1 / G_1{}^0) - (C_2 D_2 G_2{}^0 / B_0)(G_2 / G_2{}^0))dt \qquad (7.7)$$

Where $G_1{}^o$ and $G_2{}^o$ are the expression values of gene 1 and 2 respectively at time zero and they are constants with respect to time. The constants involved in the right hand side of equation 7.7 are replaced by a single positive constant α_i corresponding to each gene and the variables G_1, G_2 and B_t are replaced by g_1, g_2 and b_t which represent the expression of gene 1 and 2 and the concentration of metabolite B at time t normalized with respect to time zero.

$$(b_t) - 1 = \int_0^t (\alpha_1 g_1 - \alpha_2 g_2)dt \qquad (7.8)$$

Equation 7.8 is independent of the physiological state of any biological system, hence valid for both control and perturbed systems. If c and p superscripts refer to the control and perturbed system respectively, the following equation is also true:

$$b_t{}^p - b_t{}^c = \alpha_1 \int_0^t (g_1{}^p - g_1{}^c)dt - \alpha_2 \int_0^t (g_2{}^p - g_2{}^c)dt \qquad (7.9)$$

α_1 and α_2 values are nonnegative as all the proportionality constants that are contained in α are positive. If gene 1 is over-expressed and gene 2 is under-expressed the integral over gene 1 is positive while that over gene 2 is negative. Hence the right hand side of the equation 7.9 will be positive, independent of α_1 and α_2 values. This leads to the conclusion that metabolite B is going to be over-produced in the perturbed vs the control system. On the contrary, if gene 2 and gene 1 are over-expressed and under-expressed respectively, then the right hand side of equation 7.9 is going to be always negative implying under-production of metabolite B. If both the genes are over or under-expressed, then it's the relative values of α_1 and α_2 that will determine which of the two terms of the right hand side will dominate. Figure 7.1 shows a pictorial representation of four possible physiological states and the conclusion that could be derived from equation 7.9.

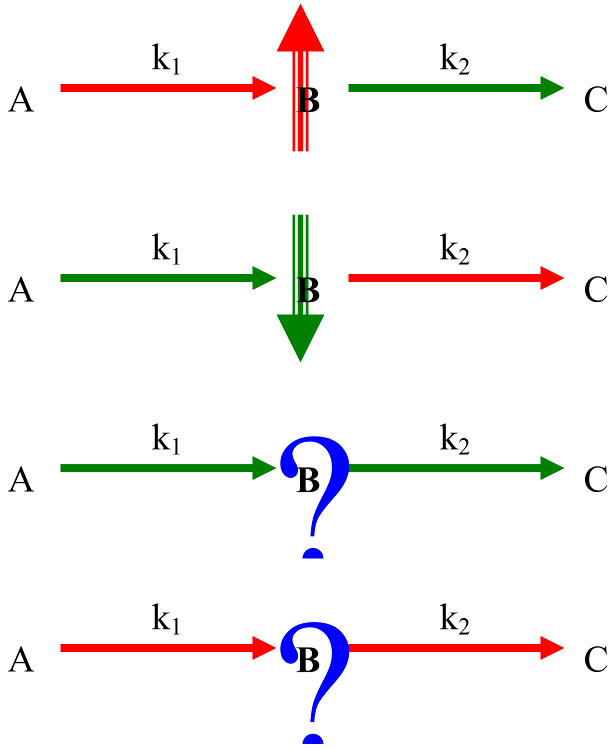


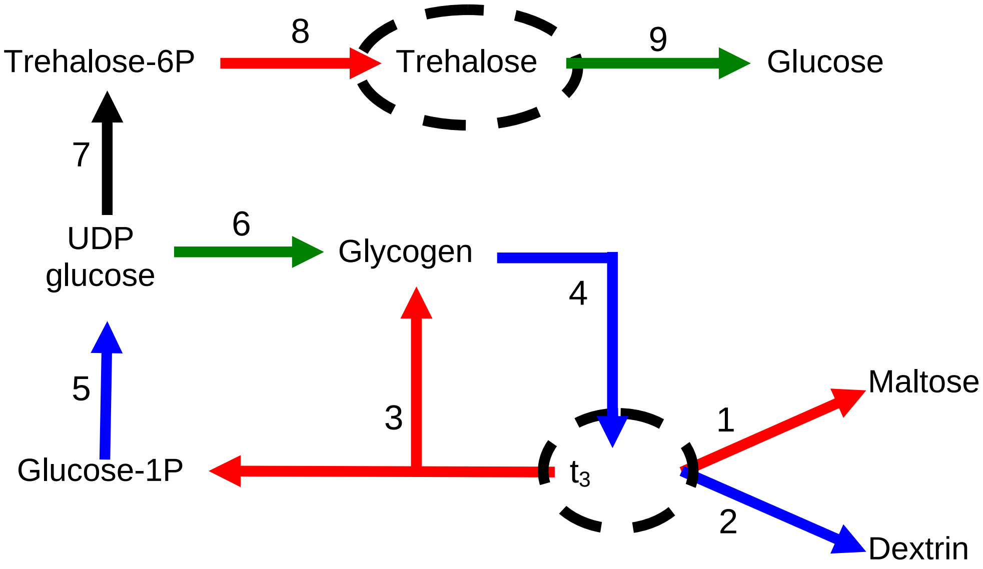
Figure 7.1 Four potential transcriptional states and the conclusions about the concentration of metabolite B. In the first two cases it can be concluded from gene expression analysis that the metabolite B is over or under-produced. Other two cases doesn't give us any conclusive information about metabolite B

Equation 7.9 can be written for each metabolite of the network. For a network of m metabolites and n genes, the system of these equations can be written in a compact matrix form as follows

$$\underline{M} = \underline{Z} * \underline{G} \tag{7.10}$$

Where the vector M has a dimension of m and it contains difference in metabolite concentrations between perturbed and control system. G is a vector of dimension n and it contains difference of the integrated gene expression values between perturbed and control. Z is a matrix that contains the constants (α_i values with associated signs) in equation 7.10.

In the following case study the explained modeling concept was used in real life data [Dutta et al., 2004]. Figure 7.2 shows biochemical pathway for trehalose and starch synthesis.



1. *beta-amylase*
2. *alpha-amylase*
3. *starch phosphorylase*
4. *1,4-alpha-glucan branching enzyme / starch branching enzyme*
5. *UDP-glucose pyrophosphorylase*
6. *glycogen synthase*
7. *trehalose-phosphatase family protein*
8. *trehalose-6-phosphate phosphatase*
9. *trehalase*

Figure 7.2 The reactions involved in starch and trehalose bio-synthesis pathway. The flux of the reactions colored red and green are increasing and decreasing respectively. Blue and black colored arrows represent reactions that are not undergoing significant change in flux and the reactions for which no information is available.

In the following case study the explained modeling concept was used in real life data [Dutta 2004]. Figure 7.2 shows biochemical pathway for trehalose and starch synthesis. A mathematical model for starch production could be derived in the way explained above

$$Starch_t{}^p - Starch_t{}^c = -\alpha_1\int_0^t (g_1{}^p - g_1{}^c)dt - \alpha_2\int_0^t (g_2{}^p - g_2{}^c)dt - \alpha_3\int_0^t (g_3{}^p - g_3{}^c)dt + \alpha_4\int_0^t (g_4{}^p - g_4{}^c)dt \tag{7.11}$$

The reactions for starch dissociation to maltose and glucose-1P are found to be increasing in perturbed compared to control. So the 1st and 3rd term in the right hand side of the equation 7.11 are positive, but they have a negative sign associated with it. However, the reactions from glycogen to starch and from starch to dextrin are not changing significantly. So the 2nd and 4th term being insignificant the net result is dictated by 1st and 3rd term, which makes the right hand side negative. This leads to the conclusion that starch is under-produced in perturbed compared to control.

Trehalose production pathway is another example where relative concentration of trehalose can be predicted from the gene expression data. Trehalose-6-phosphate phosphatase catalyses the reaction from trehalose-6P to trehalose and is over-produced. Trehalase which catalyses the reaction from trehalose to glucose, is under-produced.

$$Trehalose_t{}^p - Trehalose_t{}^c = \alpha_8\int_0^t (g_8{}^p - g_8{}^c)dt - \alpha_9\int_0^t (g_9{}^p - g_9{}^c)dt \tag{7.12}$$

In the right hand side of equation 7.12, the first term will be positive and the second term will be negative making the right hand side positive. So it can be predicted that trehalose is going to be over-produced in perturbed compared to control.

General systems biology experimental design and data analysis protocol:

In this section I will explain a general systems biology experimental design with the help of a schematic diagram (Figure 7.3). This is a generic design based on my knowledge and experience and can be tailored according to individual experimental requirements.

It starts with a systems biology experiment where multiple cellular fingerprints are measured in a time-series manner. Measurement techniques of this cellular fingerprints use different technologies hence are susceptible to different experimental biases which require them to be processed or normalized separately. Heterogeneous data obtained from multiple cellular fingerprints will be used to develop and validate mathematical and statistical models which can elucidate or even predict biological processes. Apart from the data obtained from the experimental measurements, a good model should be able to use existing biological information, like in the form of additional constraints or initial conditions. Model should also be able to include relevant data from other research groups. This doesn't sound feasible right now, because experimental design, measurement and analytical techniques vary significantly between research groups. But it's extremely important because this mammoth experimental and computational work needs expertise from different disciplines hence can not be carried out by a single lab.

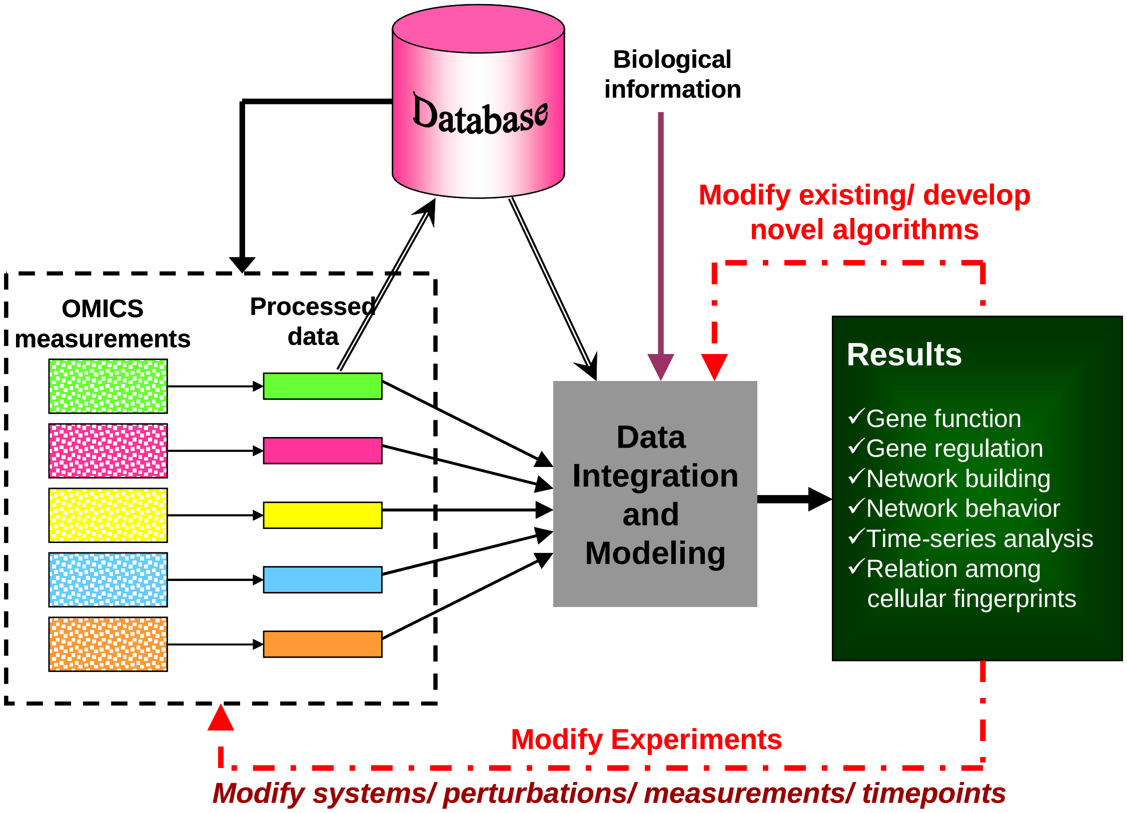


Figure 7.3 Schematic diagram of a future systems biology experiment.

Results and conclusions derived could be in several forms. It could be discrete information like function of an unknown gene or knowledge about its regulation. It could also be more generic in the form of network and sub-networks, network properties or even terms of mathematical equations consisting of different molecular fingerprint variables.

Based on the results, experimental design and data analysis methods can be modified in a feed back process. Modification in experimental design could be in the form of selecting proper system, applying correct perturbations, identifying right variables for measurement and the correct timescale. Biological system is inherently different from other systems where mathematical and statistical methods are extensively used at present. The huge amount of data that are or will be generated will pose a new challenge to identify biologically relevant conclusions. This challenge cannot be met without development of suitable mathematical and statistical analysis techniques, which closes the second feedback loop of the diagram.

Appendix

List of significant genes, SV scores and TDSM files for pair-wise comparisons were not included in the Appendix because of their large size. Soft copy of these files will be provided upon request**.**

Supplementary Table 5.19 Genes that are positively significant in both CO_2 and trehalose stress, but non-significant in combined stress. Genes marked in yellow are metabolically important. Number of genes from Glycolysis and Gluconeogenesis, Phosphatidylinositol signaling, Glycerolipid metabolism and Glycerophospholipid metabolism pathways are present in this list.

1-aminocyclopropane-1-carboxylate oxidase, putative / ACC oxidase, putative

2,3-biphosphoglycerate-independent phosphoglycerate mutase-related / phosphoglyceromutase-related

alcohol dehydrogenase (ADH)

bifunctional dihydrofolate reductase-thymidylate synthase, putative / DHFR-TS, putative

CACTA-like transposase family (Tnp1/En/Spm)

casein kinase, putative

curculin-like (mannose-binding) lectin family protein

DEAD/DEAH box helicase, putative (RH15)

diacylglycerol kinase, putative

dihydrolipoamide dehydrogenase 2, mitochondrial / lipoamide dehydrogenase 2 (MTLPD2)

disease resistance family protein

disease resistance family protein / LRR family protein

disease resistance protein (TIR-NBS-LRR class), putative

DNA-directed RNA polymerase, mitochondrial (RPOMT)

enhanced disease susceptibility 5 (EDS5) / salicylic acid induction deficient 1 (SID1)

ethylene receptor, putative (ETR2)

exocyst complex component-related

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

expressed protein

FAD-binding domain-containing protein

F-box family protein

ferrochelatase I

GTP-binding family protein

heat shock transcription factor family protein

HECT-domain-containing protein / ubiquitin-transferase family protein
kelch repeat-containing protein / serine/threonine phosphoesterase family protein
kelch repeat-containing serine/threonine phosphoesterase family protein
KH domain-containing RNA-binding protein (HEN4)
leucine-rich repeat family protein / protein kinase family protein
lipocalin, putative
molybdenum cofactor synthesis protein 3 / molybdopterin synthase sulphurylase (CNX5)
NOT2/NOT3/NOT5 family protein
nucleotidyltransferase family protein
Null
O-acetyltransferase family protein
oligopeptide transporter OPT family protein
oxidoreductase, 2OG-Fe(II) oxygenase family protein
Peroxidase, putative
phosphatidylinositol 3- and 4-kinase family protein
phox (PX) domain-containing protein
protein kinase family protein
protein kinase family protein
protein kinase family protein
protein kinase, putative
pseudogene, Ulp1 protease family
putative sterol dehydrogenase
pyruvate decarboxylase, putative
pyruvate decarboxylase, putative
Respiratory burst oxidase protein D (RbohD) / NADPH oxidase
ribose-phosphate pyrophosphokinase, putative / phosphoribosyl diphosphate synthetase, putative
SEC14 cytosolic factor, putative / phosphoglyceride transfer protein, putative
SEC14 cytosolic factor, putative / polyphosphoinositide-binding protein, putative
sucrose synthase, putative / sucrose-UDP glucosyltransferase, putative
tic20 family protein
transducin family protein / WD-40 repeat family protein
UbiA prenyltransferase family protein
ubiquitin-conjugating enzyme-related
vacuolar sorting protein 9 domain-containing protein / VPS9 domain-containing protein
zinc finger (C2H2 type) family protein
zinc finger (MYND type) family protein

Supplementary Table S5.2: Genes that are negatively significant in both CO_2 and trehalose stress, but non-significant in combined stress.

1	arabinogalactan-protein (AGP1)
2	ATP-dependent protease La (LON) domain-containing protein
3	auxin efflux carrier family protein
4	basic helix-loop-helix (bHLH) family protein
5	calmodulin-binding family protein
6	cytochrome P450 family protein
7	disease resistance-responsive protein-related / dirigent protein-related
8	expressed protein

9	expressed protein
10	expressed protein
11	hydroxyproline-rich glycoprotein family protein
12	NADP-dependent oxidoreductase, putative
13	NADP-dependent oxidoreductase, putative
14	no apical meristem (NAM) family protein
15	Null
16	Null
17	Null
18	photosystem II oxygen-evolving complex 23 (OEC23)
19	plastocyanin-like domain-containing protein
20	polygalacturonase, putative / pectinase, putative
21	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
22	proton-dependent oligopeptide transport (POT) family protein
23	Ran-binding protein 1, putative / RanBP1, putative
24	sugar transporter family protein
25	transducin family protein / WD-40 repeat family protein
26	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
27	Zinc finger (C3HC4-type RING finger) family protein
28	Zinc finger (DNL type) family protein

Supplementary Table S6.1A: List of genes that are clustering together using Pearson's correlation distance and are neighbor in their chromosomal location (total of 326 genes)

Gene Name
LIM domain-containing protein
LIM domain-containing protein
LIM domain-containing protein
histone H3
histone H3
MADS-box protein (MAF3)
MADS-box protein (MAF2)
lectin protein kinase, putative
lectin protein kinase family protein
xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
IAA-amino acid hydrolase 2 (ILL2)
IAA-amino acid hydrolase 3 (IAR3) (ILL1)
heat shock protein, putative
heat shock protein 81-4 (HSP81-4)
SUMO activating enzyme 1b (SAE1b)
SUMO activating enzyme, putative
expressed protein
60S ribosomal protein L32 (RPL32B)
integral membrane family protein
tapetum-specific protein-related
FAD-binding domain-containing protein

FAD-binding domain-containing protein
inorganic phosphate transporter (PHT2)
inorganic phosphate transporter (PHT1) (PT1)
kelch repeat-containing F-box family protein
kelch repeat-containing F-box family protein
germin-like protein (GER2)
germin-like protein (GLP2a) (GLP5a)
protein kinase family protein
protein kinase family protein
germin-like protein, putative
germin-like protein, putative
jacalin lectin family protein
jacalin lectin family protein
ribulose bisphosphate carboxylase small chain 1B / RuBisCO small subunit 1B (RBCS-1B) (ATS1B)
ribulose bisphosphate carboxylase small chain 2B / RuBisCO small subunit 2B (RBCS-2B) (ATS2B)
ribulose bisphosphate carboxylase small chain 3B / RuBisCO small subunit 3B (RBCS-3B) (ATS3B)
SAR DNA-binding protein, putative
SAR DNA-binding protein, putative
hypothetical protein
expressed protein
expressed protein
expressed protein
cytochrome P450 family protein
cytochrome P450 family protein
cytochrome P450 family protein
integral membrane transporter family protein
integral membrane transporter family protein
isochorismatase hydrolase family protein
isochorismatase hydrolase family protein
tubulin alpha-3/alpha-5 chain (TUA5)
tubulin alpha-3/alpha-5 chain (TUA3)
lipase class 3 family protein
lipase class 3 family protein
hypothetical protein
rubredoxin family protein
expressed protein
hypothetical protein
NADP-dependent oxidoreductase, putative
NADP-dependent oxidoreductase, putative
NADP-dependent oxidoreductase, putative (P1)
stress-responsive protein (KIN2) / stress-induced protein (KIN2) / cold-responsive protein (COR6.6) / cold-regulated protein (COR6.6)
stress-responsive protein (KIN1) / stress-induced protein (KIN1)
ubiquinol-cytochrome C reductase iron-sulfur subunit, mitochondrial, putative / Rieske iron-sulfur protein, putative
ubiquinol-cytochrome C reductase iron-sulfur subunit, mitochondrial, putative / Rieske iron-sulfur protein, putative

auxin-responsive GH3 family protein
auxin-responsive GH3 family protein
auxin-responsive GH3 family protein
DEAD/DEAH box helicase, putative
DEAD/DEAH box helicase, putative (RH15)
DEAD box RNA helicase (RH25)
DEAD box RNA helicase (RH26)
transferase family protein
transferase family protein
sulfotransferase family protein
sulfotransferase family protein
laccase family protein / diphenol oxidase family protein
laccase family protein / diphenol oxidase family protein
patatin, putative
patatin, putative
transcription factor IIB (TFIIB) family protein
expressed protein
aspartyl protease family protein
aspartyl protease family protein
coatomer beta subunit, putative / beta-coat protein, putative / beta-COP, putative
coatomer beta subunit, putative / beta-coat protein, putative / beta-COP, putative
expressed protein
expressed protein
cellulose synthase family protein
cellulose synthase family protein
protein kinase family protein
protein kinase family protein
cytochrome P450 family protein
cytochrome P450 family protein
hypothetical protein
protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
methionine sulfoxide reductase domain-containing protein / SelR domain-containing protein
methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein
subtilase family protein
subtilase family protein
FAD-binding domain-containing protein
FAD-binding domain-containing protein
glutaredoxin family protein
glutaredoxin family protein
pyruvate phosphate dikinase family protein
pyruvate phosphate dikinase family protein
peroxidase, putative
peroxidase, putative
mitogen-activated protein kinase, putative
mitogen-activated protein kinase, putative
equilibrative nucleoside transporter, putative (ENT3)

equilibrative nucleoside transporter, putative (ENT6)
expressed protein
serine protease inhibitor, Kazal-type family protein
expressed protein
F-box family protein
splicing factor, putative
splicing factor, putative
proline-rich extensin-like family protein
proline-rich extensin-like family protein
strictosidine synthase family protein
strictosidine synthase, putative (YLS2)
copia-like retrotransposon family
broad-spectrum mildew resistance RPW8 family protein
ADP-ribosylation factor, putative
ADP-ribosylation factor, putative
disease resistance protein (EDS1)
lipase class 3 family protein / disease resistance protein-related
ubiquitin carboxyl-terminal hydrolase family protein
ubiquitin carboxyl-terminal hydrolase-related
disease resistance protein (TIR-NBS-LRR class), putative
disease resistance protein (TIR-NBS-LRR class), putative
nitrilase 3 (NIT3)
nitrilase 1 (NIT1)
short-chain dehydrogenase/reductase (SDR) family protein
short-chain dehydrogenase/reductase (SDR) family protein
H+-transporting two-sector ATPase, putative
H+-transporting two-sector ATPase, putative
cytochrome P450 family protein
cytochrome P450 family protein
cytochrome P450 family protein
cytochrome P450 family protein
cytochrome P450 71B20, putative (CYP71B2)
cytochrome P450 71B19, putative (CYP71B19)
pseudogene, cytochrome P450
cytochrome P450 family protein
allene oxide cyclase, putative / early-responsive to dehydration protein, putative / ERD protein, putative
early-responsive to dehydration stress protein (ERD12)
disease resistance family protein
disease resistance family protein
expressed protein
expressed protein
cyclopropane fatty acid synthase-related
cyclopropane-fatty-acyl-phospholipid synthase family protein
pseudogene, hypothetical protein
pseudogene, cytochrome P450
hypothetical protein
expressed protein
jacalin lectin family protein
jacalin lectin family protein

cytochrome P450, putative
cytochrome P450, putative
forkhead-associated domain-containing protein / FHA domain-containing protein
transcriptional activator, putative
acyl-[acyl-carrier-protein] desaturase, putative / stearoyl-ACP desaturase, putative
acyl-[acyl-carrier-protein] desaturase, putative / stearoyl-ACP desaturase, putative
short-chain dehydrogenase/reductase (SDR) family protein
short-chain dehydrogenase/reductase (SDR) family protein
FAD-binding domain-containing protein
FAD-binding domain-containing protein
cyclic nucleotide-regulated ion channel, putative (CNGC11)
cyclic nucleotide-regulated ion channel / cyclic nucleotide-gated channel (CNGC3)
DC1 domain-containing protein
DC1 domain-containing protein
putative endochitinase
putative endochitinase
hypothetical protein
auxin-regulated protein
expressed protein
expressed protein
putative cinnamoyl-CoA reductase
putative cinnamoyl-CoA reductase
putative synaptobrevin
putative synaptobrevin
cellulose synthase family protein
cellulose synthase family protein
short-chain dehydrogenase/reductase (SDR) family protein / tropinone reductase, putative
tropinone reductase, putative / tropine dehydrogenase, putative
cytochrome P450 family protein
cytochrome P450 family protein
60S acidic ribosomal protein P2 (RPP2A)
60S acidic ribosomal protein P2 (RPP2B)
epoxide hydrolase, putative
epoxide hydrolase, soluble (sEH)
putative membrane transporter
unknown protein
expressed protein
expressed protein
hydrolase, alpha/beta fold family protein
hydrolase, alpha/beta fold family protein
putative MYB family transcription factor
unknown protein
sinapoylglucose:malate sinapoyltransferase (SNG1)
serine carboxypeptidase S10 family protein
mannose 6-phosphate reductase (NADPH-dependent), putative

mannose 6-phosphate reductase (NADPH-dependent), putative
expressed protein
expressed protein
UDP-glucoronosyl/UDP-glucosyl transferase family protein
UDP-glucoronosyl/UDP-glucosyl transferase family protein
leucine-rich repeat family protein / protein kinase family protein
leucine-rich repeat family protein / protein kinase family protein
ligase, putative
pseudogene, C-1-tetrahydrofolate synthase
glycine-rich protein
glycine-rich protein (GRP)
DNAJ heat shock N-terminal domain-containing protein
DNAJ heat shock N-terminal domain-containing protein
MATE efflux family protein
MATE efflux family protein
coenzyme Q biosynthesis Coq4 family protein / ubiquinone biosynthesis Coq4 family protein
expressed protein
glutathione S-transferase zeta 1 (GSTZ1) (GST18)
glutathione S-transferase, putative
major latex protein-related / MLP-related
major latex protein-related / MLP-related
expressed protein
protein kinase family protein
curculin-like (mannose-binding) lectin family protein
curculin-like (mannose-binding) lectin family protein / PAN domain-containing protein
isoflavone reductase, putative
isoflavone reductase, putative
major latex protein-related / MLP-related
Bet v I allergen family protein
expressed protein
remorin family protein
pseudogene, putative receptor serine/threonine kinase
protein kinase family protein / glycerophosphoryl diester phosphodiesterase family protein
serine/threonine protein kinase, putative
protein kinase, putative
S-adenosyl-L-methionine:carboxyl methyltransferase family protein
S-adenosyl-L-methionine:carboxyl methyltransferase family protein
cytochrome P450, putative
cytochrome P450, putative
S-locus protein kinase, putative
S-locus protein kinase, putative
S-locus protein kinase, putative
S-locus protein kinase, putative
aldo/keto reductase family protein
aldo/keto reductase family protein
glycosyl transferase family 20 protein / trehalose-phosphatase family protein
gypsy-like retrotransposon family

expressed protein (CW7)
expressed protein (CW7)
disease resistance protein (CC-NBS-LRR class), putative
xylan endohydrolase, putative
pseudogene, disease resistance protein [fragment]
PAPA-1-like family protein / zinc finger (HIT type) family protein
myrosinase-associated protein, putative
myrosinase-associated protein, putative
eukaryotic translation initiation factor 2B family protein / eIF-2B family protein
eukaryotic translation initiation factor 2B family protein / eIF-2B family protein
jacalin lectin family protein
jacalin lectin family protein
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
IAA-amino acid hydrolase 5 / auxin conjugate hydrolase (ILL5)
IAA-amino acid hydrolase 3 / IAA-Ala hydrolase 3 (IAR3)
expressed protein
pseudogene, cytochrome P450 family
hypothetical protein
chlorophyll A-B binding protein, putative (LHCA5)
expressed protein
expressed protein
leucine-rich repeat family protein
disease resistance protein-related / LRR protein-related
subtilase family protein
subtilase family protein
FAD-binding domain-containing protein
FAD-binding domain-containing protein
eukaryotic translation initiation factor 4E, putative / eIF-4E, putative / eIF4E, putative / mRNA cap-binding protein, putative
eukaryotic translation initiation factor 4E, putative / eIF-4E, putative / eIF4E, putative / mRNA cap-binding protein, putative
lipase
lipase, putative
lipase, putative
glutathione S-transferase, putative
glutathione S-transferase, putative
FAD-binding domain-containing protein
FAD-binding domain-containing protein
expressed protein
expressed protein
UDP-glucoronosyl/UDP-glucosyl transferase family protein
UDP-glucoronosyl/UDP-glucosyl transferase family protein
pseudogene, putative UDP-glucose glucosyltransferase

UDP-glucoronosyl/UDP-glucosyl transferase family protein
O-methyltransferase, putative
O-methyltransferase, putative
dehydroascorbate reductase, putative
dehydroascorbate reductase, putative
mitogen-activated protein kinase, putative / MAPK, putative (MPK8)
laccase family protein / diphenol oxidase family protein
MATE efflux family protein
MATE efflux family protein
2-oxoglutarate-dependent dioxygenase, putative
2-oxoglutarate-dependent dioxygenase, putative
sulfotransferase family protein
sulfotransferase family protein
fatty acid desaturase family protein
fatty acid desaturase family protein
polygalacturonase, putative / pectinase, putative
polygalacturonase, putative / pectinase, putative
UDP-glucose transferase (UGT75B2)
UDP-glucoronosyl/UDP-glucosyl transferase family protein
extracellular dermal glycoprotein, putative / EDGP, putative
extracellular dermal glycoprotein, putative / EDGP, putative
glutathione S-transferase, putative
glutathione S-transferase, putative
multidrug resistance P-glycoprotein, putative
multidrug resistance P-glycoprotein, putative

Supplementary Table S6.1B: List of genes that are clustering together using Euclidean distance and are neighbor in their chromosomal location (total of 292 genes)

Gene Name
LIM domain-containing protein
LIM domain-containing protein
histone H3
histone H3
MADS-box protein (MAF3)
MADS-box protein (MAF2)
xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative
protein kinase family protein
serine O-acetyltransferase (SAT-52)
IAA-amino acid hydrolase 2 (ILL2)
IAA-amino acid hydrolase 3 (IAR3) (ILL1)
SUMO activating enzyme 1b (SAE1b)
SUMO activating enzyme, putative
pentacyclic triterpene synthase, putative
cytochrome P450 family protein
Lon protease homolog 1, mitochondrial (LON)
glycine-rich protein
expressed protein
60S ribosomal protein L32 (RPL32B)
integral membrane family protein
tapetum-specific protein-related
FAD-binding domain-containing protein
FAD-binding domain-containing protein
inorganic phosphate transporter (PHT2)
inorganic phosphate transporter (PHT1) (PT1)
nodulin MtN21 family protein
nodulin-related
germin-like protein (GER2)
germin-like protein (GLP2a) (GLP5a)
protein kinase family protein
protein kinase family protein
germin-like protein, putative
germin-like protein, putative
jacalin lectin family protein
jacalin lectin family protein
ribulose bisphosphate carboxylase small chain 1B / RuBisCO small subunit 1B (RBCS-1B) (ATS1B)
ribulose bisphosphate carboxylase small chain 2B / RuBisCO small subunit 2B (RBCS-2B) (ATS2B)
ribulose bisphosphate carboxylase small chain 3B / RuBisCO small subunit 3B (RBCS-3B) (ATS3B)
SAR DNA-binding protein, putative
SAR DNA-binding protein, putative
expressed protein

expressed protein
cytochrome P450 family protein
cytochrome P450 family protein
cytochrome P450 family protein
integral membrane transporter family protein
integral membrane transporter family protein
integral membrane family protein
SOH1 family protein
tubulin alpha-3/alpha-5 chain (TUA5)
tubulin alpha-3/alpha-5 chain (TUA3)
hypothetical protein
rubredoxin family protein
expressed protein
hypothetical protein
NADP-dependent oxidoreductase, putative
NADP-dependent oxidoreductase, putative
stress-responsive protein (KIN2) / stress-induced protein (KIN2) / cold-responsive protein (COR6.6) / cold-regulated protein (COR6.6)
stress-responsive protein (KIN1) / stress-induced protein (KIN1)
dentin sialophosphoprotein-related
dentin sialophosphoprotein-related
transferase family protein
transferase family protein
laccase family protein / diphenol oxidase family protein
laccase family protein / diphenol oxidase family protein
patatin, putative
patatin, putative
coatomer beta subunit, putative / beta-coat protein, putative / beta-COP, putative
coatomer beta subunit, putative / beta-coat protein, putative / beta-COP, putative
expressed protein
expressed protein
protein kinase family protein
protein kinase family protein
cytochrome P450 family protein
cytochrome P450 family protein
methionine sulfoxide reductase domain-containing protein / SelR domain-containing protein
methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein
subtilase family protein
subtilase family protein
disease resistance protein (TIR-NBS-LRR class), putative
disease resistance protein (TIR-NBS-LRR class), putative
glutaredoxin family protein
glutaredoxin family protein
multi-copper oxidase, putative (SKU5)
auxin-responsive family protein
peroxidase, putative
peroxidase, putative
mitogen-activated protein kinase, putative
mitogen-activated protein kinase, putative

equilibrative nucleoside transporter, putative (ENT3)

equilibrative nucleoside transporter, putative (ENT6)

expressed protein

serine protease inhibitor, Kazal-type family protein

expressed protein

F-box family protein

short-chain dehydrogenase/reductase (SDR) family protein

short-chain dehydrogenase/reductase (SDR) family protein

splicing factor, putative

splicing factor, putative

proline-rich extensin-like family protein

proline-rich extensin-like family protein

copia-like retrotransposon family

broad-spectrum mildew resistance RPW8 family protein

ADP-ribosylation factor, putative

ADP-ribosylation factor, putative

ubiquitin carboxyl-terminal hydrolase family protein

ubiquitin carboxyl-terminal hydrolase-related

disease resistance protein (TIR-NBS-LRR class), putative

disease resistance protein (TIR-NBS-LRR class), putative

nitrilase 3 (NIT3)

nitrilase 1 (NIT1)

short-chain dehydrogenase/reductase (SDR) family protein

short-chain dehydrogenase/reductase (SDR) family protein

H+-transporting two-sector ATPase, putative

H+-transporting two-sector ATPase, putative

cytochrome P450 family protein

cytochrome P450 family protein

cytochrome P450 71B20, putative (CYP71B2)

cytochrome P450 71B19, putative (CYP71B19)

pseudogene, cytochrome P450

cytochrome P450 family protein

cytochrome P450 71B16, putative (CYP71B16)

allene oxide cyclase, putative / early-responsive to dehydration protein, putative / ERD protein, putative

early-responsive to dehydration stress protein (ERD12)

disease resistance family protein

disease resistance family protein

expressed protein

expressed protein

pseudogene, hypothetical protein

pseudogene, cytochrome P450

hypothetical protein

expressed protein

cytochrome P450, putative

cytochrome P450, putative

no apical meristem (NAM) family protein

no apical meristem (NAM) family protein

forkhead-associated domain-containing protein / FHA domain-containing protein

transcriptional activator, putative

acyl-[acyl-carrier-protein] desaturase, putative / stearoyl-ACP desaturase, putative

acyl-[acyl-carrier-protein] desaturase, putative / stearoyl-ACP desaturase, putative
hydroxyproline-rich glycoprotein family protein
expressed protein
short-chain dehydrogenase/reductase (SDR) family protein
short-chain dehydrogenase/reductase (SDR) family protein
cytochrome b5, putative
hypothetical protein
DC1 domain-containing protein
DC1 domain-containing protein
putative endochitinase
putative endochitinase
auxin-regulated protein
auxin-regulated protein
aldo/keto reductase family protein
aldo/keto reductase family protein
expressed protein
expressed protein
chlorophyll A-B binding protein / LHCII type I (LHB1B1)
chlorophyll A-B binding protein / LHCII type I (LHB1B2)
putative cinnamoyl-CoA reductase
putative cinnamoyl-CoA reductase
putative synaptobrevin
putative synaptobrevin
cellulose synthase family protein
cellulose synthase family protein
oxysterol-binding family protein
oxysterol-binding family protein
cytochrome P450 71A13, putative (CYP71A13)
cytochrome P450 71A12, putative (CYP71A12)
short-chain dehydrogenase/reductase (SDR) family protein / tropinone reductase, putative
tropinone reductase, putative / tropine dehydrogenase, putative
cytochrome P450 family protein
cytochrome P450 family protein
60S acidic ribosomal protein P2 (RPP2A)
60S acidic ribosomal protein P2 (RPP2B)
epoxide hydrolase, putative
epoxide hydrolase, soluble (sEH)
unknown protein
unknown protein
expressed protein
expressed protein
hydrolase, alpha/beta fold family protein
hydrolase, alpha/beta fold family protein
putative MYB family transcription factor
unknown protein
mannose 6-phosphate reductase (NADPH-dependent), putative
mannose 6-phosphate reductase (NADPH-dependent), putative
UDP-glucoronosyl/UDP-glucosyl transferase family protein
UDP-glucoronosyl/UDP-glucosyl transferase family protein
leucine-rich repeat family protein / protein kinase family protein

leucine-rich repeat family protein / protein kinase family protein
ligase, putative
pseudogene, C-1-tetrahydrofolate synthase
glycine-rich protein
glycine-rich protein (GRP)
DNAJ heat shock N-terminal domain-containing protein
DNAJ heat shock N-terminal domain-containing protein
chlorophyll A-B binding protein / LHCII type II (LHCB2.1) (LHCB2.3)
chlorophyll A-B binding protein / LHCII type II (LHCB2.2)
MATE efflux family protein
MATE efflux family protein
MATE efflux family protein
expressed protein
expressed protein
glutathione S-transferase zeta 1 (GSTZ1) (GST18)
glutathione S-transferase, putative
curculin-like (mannose-binding) lectin family protein
curculin-like (mannose-binding) lectin family protein / PAN domain-containing protein
isoflavone reductase, putative
isoflavone reductase, putative
esterase/lipase/thioesterase family protein
hydrolase, alpha/beta fold family protein
remorin family protein
remorin family protein
pseudogene, putative receptor serine/threonine kinase
protein kinase family protein / glycerophosphoryl diester phosphodiesterase family protein
S-adenosyl-L-methionine:carboxyl methyltransferase family protein
S-adenosyl-L-methionine:carboxyl methyltransferase family protein
cytochrome P450, putative
cytochrome P450, putative
cytochrome P450, putative
rhomboid family protein
cell division cycle protein-related
S-locus protein kinase, putative
S-locus protein kinase, putative
aldo/keto reductase family protein
aldo/keto reductase family protein
glycosyl transferase family 20 protein / trehalose-phosphatase family protein
gypsy-like retrotransposon family
expressed protein (CW7)
expressed protein (CW7)
xylan endohydrolase, putative
glycosyl hydrolase family 10 protein / carbohydrate-binding domain-containing protein
pseudogene, disease resistance protein [fragment]
PAPA-1-like family protein / zinc finger (HIT type) family protein
myrosinase-associated protein, putative
myrosinase-associated protein, putative
eukaryotic translation initiation factor 2B family protein / eIF-2B family protein
eukaryotic translation initiation factor 2B family protein / eIF-2B family protein
jacalin lectin family protein

jacalin lectin family protein
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
leucine-rich repeat protein kinase, putative
IAA-amino acid hydrolase 5 / auxin conjugate hydrolase (ILL5)
IAA-amino acid hydrolase 3 / IAA-Ala hydrolase 3 (IAR3)
expressed protein
pseudogene, cytochrome P450 family
expressed protein
expressed protein
leucine-rich repeat family protein
disease resistance protein-related / LRR protein-related
subtilase family protein
subtilase family protein
FAD-binding domain-containing protein
FAD-binding domain-containing protein
eukaryotic translation initiation factor 4E, putative / eIF-4E, putative / eIF4E, putative / mRNA cap-binding protein, putative
eukaryotic translation initiation factor 4E, putative / eIF-4E, putative / eIF4E, putative / mRNA cap-binding protein, putative
lipase
lipase, putative
lipase, putative
FAD-binding domain-containing protein
FAD-binding domain-containing protein
expressed protein
expressed protein
pseudogene, putative UDP-glucose glucosyltransferase
UDP-glucoronosyl/UDP-glucosyl transferase family protein
O-methyltransferase, putative
O-methyltransferase, putative
dehydroascorbate reductase, putative
dehydroascorbate reductase, putative
laccase family protein / diphenol oxidase family protein
laccase family protein / diphenol oxidase family protein
MATE efflux family protein
MATE efflux family protein
pseudogene, similar to CmE8
2-oxoglutarate-dependent dioxygenase, putative
fatty acid desaturase family protein
fatty acid desaturase family protein
polygalacturonase, putative / pectinase, putative
polygalacturonase, putative / pectinase, putative
extracellular dermal glycoprotein, putative / EDGP, putative
extracellular dermal glycoprotein, putative / EDGP, putative
multidrug resistance P-glycoprotein, putative
multidrug resistance P-glycoprotein, putative

Reference

.

Abeles, F.B., Morgan, P.W. and Saltveit, Jr., M.E. 1992. Ethylene in Plant Biology, Academic Press, San Diego, CA.

Adams-Phillips L., Barry C., Kannan P., Leclercq J., Bouzayen M., Giovannoni J. 2004. Evidence that CTR1-mediated ethylene signal transduction in tomato is encoded by a multigene family whose members display distinct regulatory features. Plant Mol. Biol. 54: 387-404.

Alberts B, Jhonson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell, 4th edition.

Andrew C. Diener, Roberto A. Gaxiola, and Gerald R. Fink. 2001. Arabidopsis ALF5, a Multidrug Efflux Transporter Gene Family Member, Confers Resistance to Toxins Plant Cell 13: 1625-1638

Baldi P and Long AD. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17:509-519.

Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci U S A. 100(18):10146-51

Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I. 2003. Continuous Representations of Time Series Gene Expression. J Comput Biol 10:341-356(b).

Bar-Joseph Z. 2004. Analyzing time series gene expression data. Bioinformatics. 20(16):2493-503.

Bassi P.K, Spencer M.S. 1982. Effect of Carbon Dioxide and Light on Ethylene Production in Intact Sunflower Plants, Plant Physiol. 69: 1222-1225.

Bland M. 2000. An Introduction to Medical Statistics. Oxford University Press.

Blazquez, M.A., Santos, E., Flores, C.-L., Martnez-Zapater, J.-M., Salinas, J., Gancedo, C., 1998. Isolation and molecular characterization of the Arabidopsis TPS1 gene, encoding trehalose-6-phosphate synthase. Plant Journal 13: 685-689.

Brown PO and Botstsein D. 1999. Exploring the new world of the genome with DNA microarrays. Nature Genetics. 21:33-37.

Buchanan B., Gruissem W., Jones R.(Eds.), 2001. Biochemistry & Molecular Biology of Plants, American Society of Plant Physiologists, Rockville, pp. 630-675.

Chatfield C. 2003. The Analysis of Time Series: An Introduction. Chapman and Hall.

Chen D, Liu Z, Ma X, Hua D. 2005. Selecting genes by test statistics. J Biomed Biotechnol 2:132-138

Chen YF, Etheridge N, Schaller GE. 2005. Ethylene signal transduction. Ann Bot (Lond). 95(6):901-15. Epub 2005 Mar 7

Chen YF, Randlett MD, Findell JL, Schaller GE. 2002. Localization of the ethylene receptor ETR1 to the endoplasmic reticulum of Arabidopsis. J Biol Chem. 31;277(22):19861-6.

Cheng C.L., Acedo G.N., Dewdney J, Goodman H.M., Conkling M.A., 1991. Differential Expression of the Two Arabidopsis Nitrate Reductase Genes, Plant Physiol. 96: 275-279.

Cheong H. Y., Chang H. S., Gupta R., Wang X., Zhu T., Luan S., 2002. Transcription profiling reveals novel interactions between wounding, pathogen, abiotic stress and hormonal response in *Arabidopsis*. Plant Physiology, Vol 129, pp 661-677

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. 1998. A Genome wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2: 65-73

Chu G, Narasimhan B, Tibshirani R, Tusher V. SAM User guide and Technical Document [http://www-stat.stanford.edu/~tibs/SAM/].

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. 1998. The transcriptional program of sporulation in budding yeast. Science, Vol 282, 699-705

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Stat. Assoc. 74, 829-836.

Cohen JD, Slovin JP, Hendrickson AM. 2003. Two genetically discrete pathways convert tryptophan to auxin: more redundancy in auxin biosynthesis. Trends Plant Sci. May;8(5):197-9.

Coschigano K.T., Melo-Oliveira R., Lim J, Coruzzi G.M. 1998. Arabidopsis gls Mutants and Distinct Fd-GOGAT Genes: Implications for Photorespiration and Primar y Nitrogen Assimilation. The Plant Cell 10: 741-752

Crowe, J.H., Carpenter, J.F., Crowe, L.M. 1998. The role of vitrification in anhydrobiosis. Annual Review of Physiology. 60: 73-103.

Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6:59-75

Davies PJ, 1995. Ed., in Plant Hormones: Physiology, Biochemistryand Molecular Biology, (Kluwer Academic, Dordrecht, Netherlands, pp. 1-12.

Denis D.T., Blakely S.D.,Carbohydrate Metabolism, in: B. Buchanan, W. Gruissem, R. Jones (Eds.), 2001. Biochemistry & Molecular Biology of Plants, American Society of Plant Physiologists, Rockville, pp. 630-675.

Dhawan K.R., Bassi P.K., Spencer M.S., 1981. Effects of Carbon Dioxide on Ethylene Production and Action in Intact Sunflower Plants, Plant Physiol. 68: 831-834.

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA. 2003. Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 19:1348 - 1359.

Dutta B, Kanani HH, Quackenbush J, Klapa MI. 2007a. Dynamic transcriptomic and metabolomic short-term response to elevated CO_2 stress in *Arabidopsis thaliana*: a plant systems biology case *(Under review)*

Dutta B, Snyder R, Klapa MI. 2007b. Significance Analysis of Time-Series Transcriptomic Data: A methodology that enables the identification and further exploration of the differentially expressed genes at each time-point *(In Press).*

Dutta B. 2004. Time series transcriptional profiling analysis of A thaliana using full genome DNA microarray and metabolic information. Master's Thesis. University of Maryland, College Park.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA 95, 14863-14868.

Ernst J, Nau GJ, Bar-Joseph Z. 2005. Clustering short time series gene expression data. Bioinformatics Suppl 1:i159-i168

Essah PA, Davenport R, Tester M. 2003. Sodium influx and accumulation in Arabidopsis. Plant Physiol. 133(1):307-18

Fett JP and Coleman JR. 1994. Characterization and Expression of Two cDNAs Encoding Carbonic Anhydrase in Arabidopsis thaliana. Plant Physiol. 105: 707-713

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Wilmitzer L. 2000. Metabolite profiling for plant functional genomics. Nat Biotechnol 18:1157–1161.

Fodor SPA. 1997. Massively parallel genomics. Science 277:393– 395.

Fritzius, T., Aeschbacher, R., Wiemken, A., Wingler, A., 2001. Induction of ApL3 expression by trehalose complements the starchdeficient Arabidopsis mutant adg2-1 lacking ApL1, the large subunit of ADP-glucose pyrophosphorylase. Plant Physiology 126, 883-889.

Gamborg O.L., Murashige T., Thorpe T.A., Vasil I.K., 1976. Plant tissue culture media, In Vitro 12: 473-478.

Gao Z., Chen Y.F., Randlett M.D., Zhao X.C., Findell J.L., Kieber J.J, Schaller G.E., 2003. Localization of the Raf-like kinase CTR1 to the endoplasmic reticulum of Arabidopsis through participation in ethylene receptor signalling complexes, J Biol. Chem. 278: 34725-34732.

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol, epub(ahead of print).

Graham D, Reed ML. 1971 Carbonic anhydrase and the regulation of photosynthesis. Nat New Biol. May 19;231(20):81-3.

Grodzinski B., Woodrow L., Leonardos E.D., Dixon M., Tsujita M.J., 1996. Plant responses to short- and long-term exposures to high carbon dioxide levels in closed environments. Adv. Space. Res. 18: 203-21.

Guo H., Ecker J.R., 2004. The ethylene signaling pathway: new insights, Curr. Opin. Plant Biol. 7: 40-49.

Hall AE, Findell JL, Schaller GE, Sisler EC, Bleecker AB. 2000. Ethylene perception by the ERS1 protein in Arabidopsis. Plant Physiol. Aug;123(4):1449-58

Haughn GW, Dmin L, Giblin M, Underhill EW .1991. Biochemical genetics of plant secondary metabolites in Arabidopsis thaliana. The glucosinolates. Plant Physiol97: 217-226

Heldt HW. Plant Biochemistry. 2005. Elsevier Academic Press.

Heyer, L. J., Kruglyak, L. & Yooseph, S. 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9, 1106-1115.

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M,Fujiwara T, Saito K. 2004. Integration of transcriptomics and metabolomics for understanding of globalresponses to nutritional stresses in Arabidopsis thaliana. PNAS. 101(27):10205-10.

Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. 2001. Dynamic modeling of gene expression data. Proceeding for National Academy of Sciences USA, Vol 98, pp 1693-1698

Hua J.,Sakai H., Nourizadeh S., Chen Q.J., Bleecker A.B., Ecker J.R., Meyerowitz E.M., 1998. Ein4 and ERS2 are members of the putative ethylene receptor family in Arabidopsis. Plant Cell. 10: 1321-1332.

Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H. 2005a. A data integration methodology for systems biology, Proc Natl Acad Sci USA 102, 17296-17301.

Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, de Atauri P, Siegel AF, Bolouri H, Aitchison JD, Hood L, 2005b. A data integration methodology for systems biology: experimental verification, Proc Natl Acad Sci USA 102, 17302-17307.

Hwang D., Rust A.G., Ramsey S., Smith J.J, Leslie D.M., Weston A.D., Atauri P., Aitchison J.D.,Hood L.,Siegel A.F., Bolouri H., 2005. A data integration methodology for systems biology, Proc Natl Acad Sci USA 102: 17296-17301.

Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersold R., Hood L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934.

Kamimura RT, Bicciato S, Shimizu H, Alford J, Stephanopoulos G. 2000. Mining of biological data I: identifying discriminating features via mean hypothesis testing. Metabolic Engineering 2:218-27.

Kanani H, Dutta B, Quackenbush J, Klapa MI. 2005. Time-series integrated metabolomic and transcriptional profiling analyses: Short term response of Arabidopsis thaliana primary metabolism to elevated CO_2 - Case study. In, Nikolaou B, editor. Proceedings of the 3rd International Congress on Plant Metabolomics. Netherlands: Kluwer Academic Publishers.

Kanani H. 2007. High-throughput time-series metabolic analysis of a systematically perturbed system. PhD Thesis. University of Maryland.

Kieber J.J., Rothenberg M., Roman G.,Feldman K.A., Ecker J.R., 1993. CTR1, a negative regulator of the ethylene response pathway in Arabidopsis, encodes a member of the Raf family of protein kinases, Cell 72 427-441.

Kim H, Snesrud EC, Haas B, Cheung F, Town CD, Quackenbush J. 2003. Gene expression analyses of Arabidopsis chromosome 2 using a genomic DNA amplicon microarray. Genome Res. 13:327-340

Klapa MI and Quackenbush J. 2003. The quest for the mechanisms of life. Biotechnology and Bioengineering 84:739-742.

Korn EL, Troendle JF, McShane LM, and Simon R. 2001. Controlling the number of false discoveries: application to high-dimensional genomic data Technical report 003, Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/~brb/TechReport.htm

Krebbers E., Seurinck J., Herdies L., Cashmore A.R., Timko M.P., 1988. Four genes in two diverged subfamilies encode the ribulose-1,5-bisphosphate carboxylase small subunit polypeptides of Arabidopsis thaliana, Plant Mol Biol. 11 745-759.

Kreps J. A., Wu Y., Chang H. S., Zhu T., Wang X., Harper J. F., 2002. Transcriptome changes for Arabidopsis in response to salt, osmotic and cold stress. Plant Physiology, Vol 130, pp 2129-2141.

Larsson O, Wahlestedt C, Timmons JA. 2005. Considerations when using the significance analysis of microarrays (SAM) algorithm. BMC Bioinformatics 6: 129.

Lea P, Leegood RC. 1993. Plant Biochemistry and Molecular Biology. John Wiley & Sons.

Liu F, Vantoai T, Moy LP, Bock G, Linford LD, Quackenbush J. 2005. Global Transcription Profiling Reveals Comprehensive Insights into Hypoxic Response in Arabidopsis. Plant Physiol. 137:1115-1129.

Lopez F, Pichereaux C, Burlet-Schiltz O, Pradayrol L, Monsarrat B, Esteve JP. 2003. Improved sensitivity of biomolecular interaction analysis mass spectrometry for the identification of interacting molecules. Proteomics 3:402–12.

Manabe T. 2003. Analysis of complex protein-polypeptide systems for proteomic studies. J Chromatogr B Analyt Technol Biomed Life Sci 787:29–41.

Meyer S L, 1975. Data Analysis for scientists and engineers, John Wiley and Sons.

Moore B, Zhou L, Rolland F, Hall Q, Cheng WH, Liu YX, Hwang I, Jones T, Sheen J. 2003. Role of the Arabidopsis glucose sensor HXK1 in nutrient, light, and hormonalsignaling. Science. 300(5617):332-6.

Muller J., Aeschbacher R., Wingler A., Boller T., Wiemken A. 2001. Trehalose and Trehalase in Arabidopsis. Plant Physiology 125:1086–1093

Müller, J., Boller, T., Wiemken, A., 1998. Trehalose affects sucrose synthase and invertase activities in soybean (Glycine max [L.] Merr.) roots. J. Plant Physiol. 153, 255-257

Nelson D.L.,Cox M.M., 2002. Lehninger Principles of Biochemistry, Worth Publishers, New York,.

Oksman-Caldentey KM, Inze D. 2004. Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites, *Trends. Plant. Sci.* **9.** 433-440.

Oliveira I.C., Coruzzi G.M., 1999. Carbon and amino acids reciprocally modulate the expression of Glutamine Synthetase in Arabidopsis. Plant Physiology, Vol. 121, pp. 301-309,

Orlando C, Raggi CC, Bianchi S, Distante V, Simi L, Vezzosi V, Gelmini S, Pinzani P, Smith MC, Buonamano A, Lazzeri E, Pazzagli M, Cataliotti L, Maggi M, Serio M. 2004. Measurement of somatostatin receptor subtype 2 mRNA in breast cancer and corresponding normal tissue. Endocr Relat Cancer. Jun;11(2):323-32.

Pan W. 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18:546-554.

Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, Lee YS. 2003. Statistical tests for identifying differentially expressed genes in time-course microarray experiments Bioinformatics 19: 694 - 703

Paul M.J.,Foyer C.H., 2001. Sink regulation of photosynthesis. J. Exp. Bot. 52: 1383-1400

Philippe Reymond, Hans Weber, Martine Damond, and Edward E. Farmer. 2000. Differential Gene Expression in Response to Mechanical Wounding and Insect Feeding in Arabidopsis. Plant Cell 12: 707-720.

Pruitt KD and R. L. Last. 1993. Expression Patterns of Duplicate Tryptophan Synthase [beta] Genes in Arabidopsis thaliana. Plant Physiol. 102: 1019-1026

Quackenbush J, 2001, Computational analysis of microarray data, Nature Genetics, Vol 2, 418-427

Quackenbush, J. 2002 Microarray data normalization and transformation. Nature Genetics, Supplement 2, Vol. 32 Issue 4, p496,

Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ Jr, Hallett JP, Leak DJ, Liotta CL, Mielenz JR, Murphy R, Templer R, Tschaplinski T. 2006. The path forward for biofuels and biomaterials. Science. Jan 27;311(5760):484-9

Raychaudhuri, S., Stuart, J. M. & Altman, R. B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 2000, 455–466.

Renné P.; Dreßen U.; Hebbeker U.; Hille D.; Flügge U-I.; Westhoff P.; Weber A.P.M. 2003. The Arabidopsis mutant dct is deficient in the plastidic glutamate/malate translocator DiT2 The Plant Journal. 35(3) 316-331

Roessner U., Wagner C., Kopka J., Trethewey R., Willmitzer L. 2000. Simultaneous analysis of metabolites in potato tuber by gas hromatographymass spectrometry. Plant J. 23:131-142

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34:374-8

Saidi SA, Holland CM, Kreil DP, MacKay DJ, Charnock-Jones DS, Print CG, Smith SK. 2004. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene. Aug 26;23(39):6677-83.

Schaller GE, Ladd AN, Lanahan MB, Spanbauer JM, Bleecker AB. 1995. The ethylene response mediator ETR1 from Arabidopsis forms a disulfide-linked dimer.J Biol Chem. 270(21):12526-30

Schaller, G.E. and Kieber, J.J. 2002 in: The Arabidopsis Book, Vol. DOI/10.1199/tab.0071 (Somerville, C. and Meyerowitz, E., Eds.), American Society of Plant Biologists, Rockville, MD.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270(5235):467-70.

Schmitt W. A., Stephanopoulos G., 2003. Prediction of transcription profiles of Synechocystis PCC6803 by dynamic autoregressive modeling of DNA microarray data. Biotechnology and Bioengineering, Vol 84, No 7, pp 855-863

Shi H, Ishitani M, Kim C, Zhu JK. 2000. The Arabidopsis thaliana salt tolerance gene SOS1 encodes a putative NaC/HC antiporter.Proc. Natl. Acad. Sci. USA 97: 6896-901

Slater S, Mitsky TA, Houmiel KL, Hao M, Reiser SE, Taylor NB, Tran M, Valentin HE, Rodriguez DJ, Stone DA, Padgette SR, Kishore G, Gruys KJ. 1999 Metabolic engineering of Arabidopsis and Brassica for poly(3-hydroxybutyrate-co-3-hydroxyvalerate) copolymer production, Nat. Biotechnol. 17. 1011-1016.

Smith C.J.,Carbohydrate Chemistry, in: P.J. Lea, R.C. Leegood (Eds.), 1993. Plant Biochemistry and Molecular Biology, John Wiley and Sons, West Sussex, England, pp. 73:113.

Sommerville C., Dangl J., 2000. Genomics. Plant biology in 2010, Science 290 2077-2078.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Ansers K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell-cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of Cell, Vol 9, pp 3273-3297

Stephenopoulous G., Aristidou A.A., Nielsen J., 1998. Metabolic Engineering: Principals and Methodologies, Academic Press, San Diego.

Stitt M. 1991 Rising CO_2 levels and their potential significance for carbon flow in photosynthetic cells, Plant Cell and Environment. 14: 741-762

Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. 2005. Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA 102:12837-12842

Taji T, Seki M, Satou M, Sakurai T, Kobayashi M, Ishiyama K, Narusaka Y, Narusaka M, Zhu JK, Shinozaki K. 2004. Comparative genomics in salt tolerance between Arabidopsis and arabidopsis-related halophyte salt cress using Arabidopsis microarray. Plant Physiol. 135(3):1697-709.

Taiz, L., and Zeiger, E., 2002. Plant Physiology. Sinauer Associates, Inc., MA

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

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. 1999. Systematic determination of genetic network architecture. Nature Genet. 22, 281-285.

Taylor J, King RD, Altmann T, Fiehn O. 2002. Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics 2:S241–S248

Teale WD, Paponov IA, Palme K. 2006. Auxin in action: signalling, transport and the control of plant growth and development. Nat Rev Mol Cell Biol. Nov;7(11):847-59. Epub 2006 Sep 20. Review

Thrower J.S., Blalock R. 3rd, Klinman J.P., 2001 Steady-state kinetics of substrate binding and iron release in tomato ACC oxidase, Biochemistry 40 9717-9724.

Tian J, Ishibashi K, Handa JT. 2004. The expression of native and cultured RPE grown on different matrices. Physiol Genomics. Apr 13;17(2):170-82

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520-525.

Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18: 1454-1461.

Tsuji J, Jackson EP, Gage DA, Hammerschmidt R, Somerviille SC .1992. Phytoalexin accumulation in Arabidopsis thaliunu during the hypersensitive reaction to Pseudomonas syringae pv syringue. Plant Physiol98: 13.04-1309

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl Acad. Sci 98: 5116-5121.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. 1995. Serial analysis of gene expression. Science 270, 484-487 (1995).

Verala C, Agosin E, Baez M, Klapa M, Stephenopoulos G. 2003. Metabolic flux redistribution in Corynebacterium glutamicum in response to osmotic stress. Appl Microbiol Biotchnol. 60:547-55

Vivian G. Cheung, Michael Morley, Francisco Aguilar, Aldo Massimi, Raju Kucherlapati2 & Geoffrey Childs, 1999. Making and reading microarrys. Nature Genetics, Vol 21, 15-19

Vogel, G., Aeschbacher, R.A., Mu¨ ller, J., Boller, T., Wiemken, A., 1998. Trehalose-6-phosphate phosphatases from Arabidopsis thaliana: identification by functional complementation of the yeast tps2 mutant. Plant Journal 13, 673-683.

Wagner, U., Edwards, R., Dixon, D.P. & Mauch, F. 2002. Probing the diversity of the Arabidopsis glutathione S-transferase gene family.

Wang H, Hanash S. 2003. Multi-dimensional liquid phase based separations in proteomics. J Chromatogr B Analyt Technol Biomed Life Sci. 787: 11–18.

Wang S and Chen JJ. 2004. Sample size for identifying differentially expressed genes in microarray experiments. J Comput Biol 11:714-726.

Webber N. Andrew, Nie G and Long S. P. 1994. Acclimation of photosynthetic proteins to rising atmospheric CO_2 , Photosynthesis Research 39: 413-425

Wiemken, A., Boller, T., Wingler, A. 2002. Induction of trehalase in Arabidopsis plants infected with the trehalose-producing pathogen Plasmodiophora brassicae. Molecular Plant-Microbe Interactions.

Wiemken, A., 1990. Trehalose in yeast, stress protectant rather than reserve carbohydrate. Antonie van Leeuwenhoek. 58, 209-217.

Wingler A, Fritzius T, Wiemken A, Boller T, Aeschbacher RA. 2000. Trehalose induces the ADP-glucose pyrophosphorylase gene, ApL3, and starch synthesis in Arabidopsis. Plant Physiol. 124(1):105-14

Wright AD, Sampson MB, Neuffer MG, Michalczuk L, Slovin JP, Cohen JD. 1991. Indole-3-acetic acid biosynthesis in the :mutant maize orange pericarp, a tryptophan auxotroph. Science 254: 998-1000

Wu B. 2005. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. Bioinformatics 21: 1565-71.

Yang Y H, Xiao Y, Segal R M. 2004. Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics. Apr 1;21(7):1084-93

Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J. 2002. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol. 3, research0062.1-0062.12.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, e15.

Zar JH. 1999. Biostatistical Analysis. 4th edition. New Jersey: Prentice Hall.

Zhao Y, Chen BP, Miao H, Yuan S, Li YS, Hu Y, Rocke DM, Chien S. 2002. Improved significance test for DNA microarray data: temporal effects of shear stress on endothelial genes. Physiol. Genomics 12: 1-11.

Zhao Y, Christensen SK, Fankhauser C, Cashman JR, Cohen JD, Weigel D, Chory J. 2001. A role for flavin monooxygenase-like enzymes in auxin biosynthesis. Science 291, 306–309

Zhong G.V., Burns J.K., 2003. Profiling ethylene-regulated gene expression in Arabidopsis thaliana by microarray analysis, Plant Mol Biol. 53: 117-131.

Zhu JK., 2000. Genetic analysis of plant salt tolerance using Arabidopsis thaliana. Plant Physiol. 124:941-48

Zhu,G., Spellman,P.T., Volpe,T., Brown,P.D., Botstein,D., Davis,T.N. and Futcher,B. 2000. Two yeast forkhead genes regulate cell cycle and pseudohyphal growth. *Nature*, 406, 90–94.