INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE BIOLOGÍA

INGENIERÍA EN BIOTECNOLOGÍA

Senior Graduation Project Report

EXPRESSION ANALYSIS OF MICROARRAY DATA FOR IDENTIFICATION OF EXPRESSION PATTERNS RELATED TO THERMODORMANCY BETWEEN TWO VARIETIES OF LETTUCE (*Lactuca sativa* and *Lactuca serriola*)

Sebastian Reyes Chin-Wo

University of California, Davis
Seed Biotechnology Center

Cartago, Costa Rica, 2010

_____	_____
Allen Van Deynze PhD	Giovanni Garro Monge MSc.
UCDavis Advisor	ITCR Advisor

INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE BIOLOGÍA

INGENIERÍA EN BIOTECNOLOGÍA

Senior Graduation Project Report

EXPRESSION ANALYSIS OF MICROARRAY DATA FOR IDENTIFICATION OF EXPRESSION PATTERNS RELATED TO THERMODORMANCY BETWEEN TWO VARIETIES OF LETTUCE (*Lactuca sativa* and *Lactuca serriola*)

Sebastian Reyes Chin-Wo

University of California, Davis
Seed Biotechnology Center

Cartago, Costa Rica, 2010

Allen Van Deynze PhD	Giovanni Garro Monge MSc.
UCDavis Advisor	ITCR Advisor

REVIEW

Thermodormancy is a phenomenon that affects lettuce seed and increase the productions cost due to specific treatments needed to overcome germination. Because of this, is important to understand the main factors involved in the establishment and maintaining of thermodormancy in this crop. To achieve this goal a wide research of the transcriptome using the microarray analysis is needed to find candidate genes related with thermodormancy. Data from this experiment was analyzed using several approaches to filter the genes looking for biological significance related with thermotolerance. Through the analysis of two varieties of lettuce (*Lactuca sativa* var Salinas and *Lactuca serriola)* were able to find 63 candidate genes that can play an important role. These genes are related with metabolism, regulation or response of the plant hormones ABA, GA and ethylene and should be taken in further investigations to determine their specific function in germination.

RESUMEN

Termodormancia es un proceso que afecta el cultivo de lechuga e incrementa los costos de producción debido a tratamientos necesarios para lograr la germinación de la semilla. A causa de esto, es necesario entender los factores involucrados en la termodormancia de este cultivo. Estudios del transcriptoma a través del análisis de microarreglos es un método para lograr entender y encontrar genes involucrados con este fenómeno. Datos generados de microarreglos fueron analizados utilizando diferentes técnicas para filtrar genes en busca de significancia biológica relacionado con termotolerancia. Analizando dos variedades de lechuga (*Lactuca sativa* var Salinas y *Lactuca serriola*) 63 genes candidatos fueron encontrados que pueden ser fundamentales en la regulación de la termodormancia. Estos genes están relacionados con el metabolismo, señalización y respuesta a ABA, GA y etileno, estos genes deberán ser estudiados posteriormente para determinar su relación y función en la termodormancia.

EXPRESSION ANALYSIS OF MICROARRAY DATA FOR IDENTIFICATION OF EXPRESSION PATTERNS RELATED TO THERMODORMANCY BETWEEN TWO VARIETIES OF LETTUCE (*Lactuca sativa* and *Lactuca serriola*)

Informe presentado a la Escuela de Biología del Instituto Tecnológico de Costa Rica como requisito parcial para optar al título de Bachiller en Ingeniería en Biotecnología.

Miembros del Tribunal/ Members of the Committe

MSc. Giovanny Garro Monge
Asesor ITCR/ ITCR Advisor

PhD. Allen E. Van Deynze
Asesor UCDavis/ UCDavis Advisor

PhD. Hamid Ashrafi
Lector/ Reviewer

DEDICATORIA

A mis padres y hermanos
sin cuya ayuda y apoyo
no hubiera sido posible
lograr mis metas y sueños

Gracias

Sebas

DEDICATION

To my parents and family
without which help and support
i wouldn´t be able
to achieve my goals and dreams

Thanks

Sebas

ACKNOWLEGMENTS

I want to thank the Seed Biotechnology Center of the University of California at Davis, especially to the Dr. Allen E. Van Deynze and Dr. Kent J. Bradford for let me work with them, also to the Dr. Hamid Ashrafi for mentor me and help me develop the investigation, to the Dr. Peetambar Dahal for allow work with his project, and to both research teams for all their help and support. I want to thank to MSc. Giovanny Garro Monge for be my advisor and all the help that he bring to me. I want to be grateful to the Ministry of Science and Technology (MICIT) and the National Committee of Scientific and Technologic Investigations (CONICIT) of Costa Rica for their financial support without which I wouldn´t be able to develop my senior project. And most important I want to thank my family, my mother and father, for their help and support, financially and emotionally without which it had been impossible to accomplish this investigation.

SUMMARY

REVIEW iii

RESUMEN iv

ACKNOWLEGMENTS..... vii

SUMMARY viii

FIGURES AND CHARTS x

INTRODUCTION.....1

BACKGROUND.....3

Lettuce to study thermoinhibition 4

Transcriptomic analysis of seed dormancy 5

An overview of the microarray technology 7

Microarray Data Preprocessing and Quality Assessment 10

Microarray Expression Analysis.....12

Main Characteristics of the Lettuce Chip 13

JUSTIFICATION AND SIGNIFICANCE.....15

OBJECTIVES 16

General Objective 16

Specific Objectives.....16

MATERIALS AND METHODS.....17

Plant materials and treatments 17

Preprocessing of the Microarray Data.....18

Quality Assessment 18

Diagnostic Plots and Images 20
Principal Component Analysis and Sample Clustering 20
Expression Analysis 20
CLUSTERING METHODS 21
GENEByGENE ANALYSIS 23
RESULTS 26
Chips Quality Assessment 26
Differential expression analysis show basic relations between treatments 32
Main sources of differences between treatments determined by Cluster of Samples and PCA 33
Combined methods to identify putative genes related with thermodormancy 38
DISCUSSION 45
Delay of Germination due to High Temperature 45
Gene clustering related with traits 46
Genes involved in maintain and release of dormancy 48
RECOMMENDATIONS 50
REFERENCES CITED 51

FIGURES AND CHARTS

Figure 1. Minimal Abscisic Acid Regulation Pathway 6
Figure 2. Overview of array and target preparation for cDNA Microarray and High-Density Oligonucleotide Microarray 9
Figure 3. Affymetrix Custum Lettuce GeneChip® 14
Figure 4. Germination curves for treatments..... 17
Figure 5. Workflow for generate the WGCNA..... 23
Figure 6. Average Expression per Treatment for Affymetrix hybridization controls 26
Figure 7. Boxplot of Hybridization Intensities for Probes across chips. 29
Figure 8. Histogram of Hybridization Intensities across chips. The perpendicular red line is the background level. 30
Figure 9. Results from the Probe Level Model fitting 31
Figure 10. TAGGIT categories across treatments for above background genes..... 34
Figure 11. Dendrogram of Samples based in Pearson Correlation 35
Figure 12. 2D plot of principal components 36
Figure 13. Hierarchical clustering of 15309 genes with module assignment..... 38
Figure 14. Example of gene network from Module Gold with an interaction cut off of 0.413..... 41
Figure 15. Heatmap of correlation between module eigengenes with experimental conditions..... 42

Table 1. Description of samples used in microarray analysis 17
Table 2. Comparisons used for generate VENN diagrams with differentially expressed genes..... 21
Table 3. Genes involved in the ABA regulation or metabolism pathway represented in the microarray..... 24
Table 4. Data of Technical Replicates genes in the chip 27
Table 5. Differentially expressed genes above background..... 32
Table 6. Genes result of filtering microarray dataset for relation with thermodormancy 43

INTRODUCTION

Understanding of the regulation of complex metabolic processes, such as seed germination, requires the application of new technologies including microarray analysis. Seed germination is a control point in plant development that involves both internal and external conditions to start seedling growth. It is necessary to identify the different factors that regulate germination because of its importance in the annual cropping cycle.

Seed dormancy has been used as a selection trait to avoid pre-harvest sprout and also to obtain a homogeneous germination population. Nevertheless some species present problems, for example, lettuce (*Lactuca sativa*) fails to germinate at temperatures above 25 to 30ºC, this kind of dormancy is known as thermodormancy or thermoinhibition. A QTL, High Temperature Germination 6.1 (HTG6.1) has been identified related to the thermodormancy, and a candidate gene was identified to collocate with this locus, the LsNECD4; and can be the responsible for his effect over the thermodormancy. The failure in germination of lettuce brings problems to the producers that need to applied expensive measures to avoid thermodormancy.

In order to study the thermoinhibition processes in lettuce, microarray analysis allows the characterization of the expression profile of thousands of genes in only one assay. The Lettuce genechip developed by A. Van Deynze, R.W. Michelmore and colleagues at UC Davis (Caldwell, 2008, van Leeuwen, 2008) permits the identification of differential expression between two lettuce genotypes, and this allows the recognition of genes related to thermoinhibition.

The differences among expression patterns, of seeds of two lettuce varieties, *L. sativa* cv. Salinas and *L. serriola*, in different temperatures and times of imbibition were analyzed to determine which genes and modules of genes could be related with the thermodormancy. Also with the utilization of the network analysis, we identified the relationships between each other. To develop this analysis several bioinformatics tools were applied that allow processing of the data generated for the microarray experiments, in order to obtain the information necessary to determine the expression patterns and the network interconnections. This investigation was developed between February 2009 to January 2010.

This investigation is part of the project developed for the Dr. Peetambar Dahal, Dr. Kent J. Bradford and collaborators of the UC Davis, “Combined Genetic and Transcriptomic Analysis of Lettuce Seed Dormancy” founded by the United States Department of Agriculture Initiative for Future Agriculture and Food Systems, and the National Science Foundation.

BACKGROUND

Seed dormancy is the inability of viable seed to germinate under environmental conditions normally favorable for germination. These conditions consist of a complex combination of factors such as water, light, temperature, gasses, mechanical restrictions, seed coats and hormones (Baskin, 2004). In nature, dormancy is a mechanism that allows seeds to survive in a an unfavorable environment until the optimal conditions for germination and establishment have been fulfilled (Allen, 2007, Benech-Arnold*, et al.*, 2000). There are different ways to stimulate dormancy of seed. High temperature is a known factor that can induce seed to enter to a state of dormancy; this phenomenon is called thermodormancy or thermoinhibition. For example, lettuce seeds do not germinate when they are imbibed at temperatures above 25 to 30 ºC (Lefebvre, 2006).

Arabidopsis thaliana has been used as a model species, in order to study seed dormancy. *Arabidopsis* mutant plants affecting the dormancy processes have allowed for the identification of some of the regulators of this phenomenon, such as the plant hormones abscisic acid (ABA), gibberellic acid (GA), ethylene and auxin (Feurtado, 2007, Kucera, 2005), along with external factors such as light and temperature (Heggie, 2005). When Arabidopsis seeds were exposed to high temperatures, the expression of genes involved in the ABA synthesis was elevated whereas the expression of genes in the GA synthesis pathway was inhibited, both changes in gene expression promote the establishment and maintenance of the thermodormancy (Tarquis, 1992). Recently, with the advent of new technologies, dozens of genes involved in seed dormancy have been identified in *Arabidopsis*. Examples of these genes are AtNECD6 and AtNCED9, of the 9-*cis*-

epoxycarotenoid dioxygenase (NCED) family that are genes related with the synthesis of ABA (Lefebvre, 2006, Tamura, 2006). These studies have lead the path to discovering key factors involved in thermodormancy, and give very informative data for directing future studies.

Lettuce to study thermoinhibition

Before the emergence of Arabidopsis as a model plant system, lettuce seed was widely used for germination and dormancy studies, allowing the description of phenotypes across varieties and species that show differences in germination under various conditions (Kozarewa, 2006). The advantage of using lettuce to study a dormancy process like thermoinhibition instead of Arabidopsis, is that lettuce is a commercially important crop, thus the results obtained in this type of research can be readily applied to the seed industry.

Argyris *et al* (2005) studied a phenotype obtained by crossing *Lactuca sativa* cv. Salinas, a commercial variety of lettuce that is sensitive to temperature and *L. serriola*, the wild progenitor of lettuce that is thermo-tolerant, to identify significant Quantitative Trait Loci (QTL) for thermodormancy. This study led to the discovery of the High Temperature Germination 6.1 (HTG6.1) locus, which has a major influence over this trait. In order to correlate genes with the HTG6.1 locus, 24 genes were mapped and found to have effects in seed dormancy. A relevant gene described was the LsNCED4 gene which is co-localized with the HTG6.1 locus and is implicated in the ABA synthesis; it also has a homology with the Arabidopsis gene AtNCED6. AtNCED6 in Arabidopsis controls the first committed step in ABA biosynthesis, and is required to induce dormancy (Lefebvre, 2006, Tamura, 2006). This suggests that LsNCED4 may regulate ABA

biosynthesis and is the gene responsible for the effect of the HTG6.1 locus in the induction and maintenance of the thermodormancy.

In addition to the balance between ABA and GA content along with the capability to perceive these hormones, dormancy and germination in lettuce seeds are regulated by (Finch-Savage, 2007), ethylene production (Riefler*, et al.*, 2006). Expression analysis of several genes involved in the biosynthesis, perception or metabolism of these hormones was done in the Salinas and L. serriola lines. These lead to generate expression profiles for a group of genes that can be related to this trait. Some are being up-regulated by high temperature in Salinas, such as NCED4, ABI5 and SNF4. On the other hand a set of genes were down-regulated in Salinas at high temperature including Ls3h1 and LsACS1, which are ones related to the GA and ethylene metabolism. (Argyris, 2005)**.**

Transcriptomic analysis of seed dormancy

For the analysis of metabolic processes, such as dormancy and germination, the classical transcript profiling using candidate genes can be used as a preliminary analysis to elucidate the general regulatory pathways involved. However, in order to better understand how these changes influence seed metabolism it is necessary to perform a comprehensive study of expression profiles for a large number of genes. Transcriptomic methods, based on the analysis of a wider range of genes, such as with microarray , provide more opportunities to decipher regulatory networks (Holdsworth, Finch-Savage*, et al.*, 2008). This approach has already been applied to identify candidate genes related to dormancy. For instance Cadman *et al.* (2006) identified 442 genes that had higher expression in dormant stages compared to the after-ripen stage in Arabidopsis seeds. Bove *et al.* (2005) found 1020

differentially expressed genes between dormant and non-dormant seeds using cDNA from *Nicotiana plumbaginifolia*. In Arabidopsis after sequencing 400 genes related to dormancy, 83 were assigned to known functions, of which 30 were related to the ABA synthesis (Schwember, 2005).

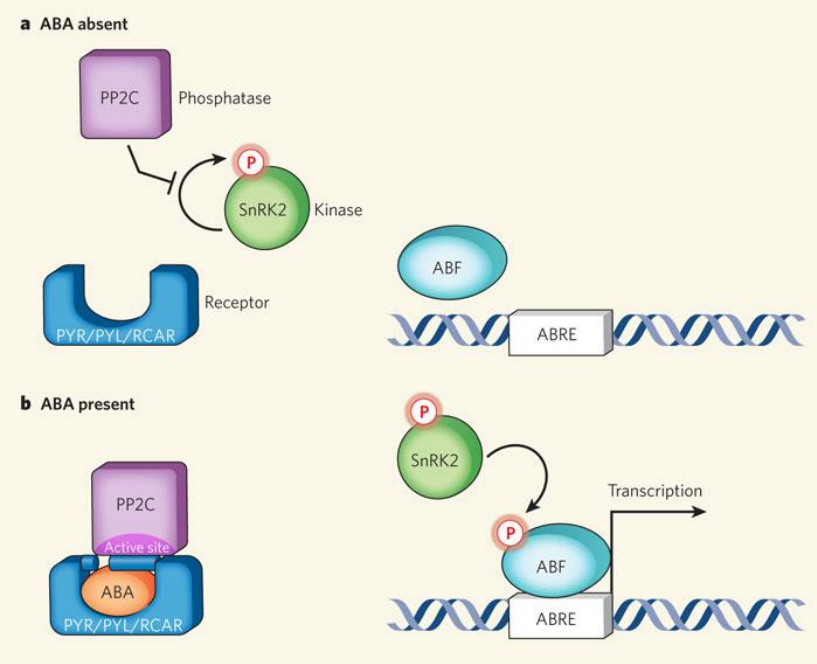


Figure 1. Minimal Abscisic Acid Regulation Pathway (Sheard, et al., 2009)

Further studies in Arabidopsis have identified additional genes related to the establishment and maintenance of dormancy. Genes associated with the perception and signaling of ABA were found to be important, such as the SNF (Sucrose non-fermenting) related protein kinase 2 (SnRK2) family. Using Arabidopsis mutants, it was determined that SnRK2's proteins are involved in the first steps of ABA perception and can play an important role triggering the ABA chain reaction (Nakashima*, et al.*, 2009). These proteins also seem to be regulated by Protein Phosphatases 2C (PP2C). In this study, it was determined that the PP2C are highly involved in the inhibition of the SnRK2's by a protein-protein interaction that blocks the active site of the kinase (Vlad*, et al.*, 2009). In the presence of ABA the PYR/PYL/RCAR receptor proteins bind to the

PP2C and repress it's activity, that enables the SnRK2's to phosphorylate transcription factors and trigger the transcription of ABA-responsive genes (Figure 1b). But without ABA, PP2C remains active and phosphorylates SnRK2, inactivating it and preventing downstream signaling (Figure 1a) (Park*, et al.*, 2009, Sheard, et al., 2009)

Thermodormancy regulation has been widely studied in other species, however this process is still not well understood in lettuce seeds. Since 2004, the Seed Biotechnology Center (SBC) of UC Davis, has developed a project to investigate the principles of thermodormancy in lettuce by standard transcript profiling and microarray analysis in order to gain insight into these processes.

An overview of the microarray technology

Microarray technology, developed in the mid 90's, is based on a multiplex assay. The fundamental principle of this technology is the capacity of the nucleic acids to recognize complementary sequences within complex mixtures of nucleic acids. It utilizes a microchip which consists in a series of thousands of oligonucleotides (called probes or features) arrayed over a surface to detect and quantify specific sequences of DNA or RNA. Usually the surface where the probes are attached is a small plaque of glass or silicon that can be mounted in a support to facilitate their manipulation. In most of the arrays several probes are designed to recognize multiple sequences within the same gene, and this group is called a probeset (Lipshutz*, et al.*, 2009, Schulze*, et al.*, 2001).

Before hybridization of the microarray, an extraction of RNA (preferably mRNA) or DNA sample needs to be done, and then the nucleic acids are labeled with a fluorescent dye and applied to the

microarray. After several washes, only the target-probe complex remains bound. If hybridization occurs between the probe in the microchip and the nucleic acid in the sample (target) a strong fluorescent signal will emit and that is the basis of detection technology to quantify the hybridization intensity. The fluorescence is detected with phospho-imaging or fluorescence scanning and intensity of the signal is related to the hybridization of target sequence which translates to the amount of target sequence in the sample (Bunney, 2003).

There are two basic microarray methods, the first one is the cDNA Spotted Array Method in which the arrayed material is cDNA sequences obtained through a Reverse Transcription PCR (Schulze, et al., 2001). These cDNA fragments are spotted by robots in dots of 100-300µm to control the positioning and produce an accurate high-density array (Figure 2. a) (Bunney, 2003). The major problem of these arrays is the lack of reproducibility, because the spotting process is not highly accurate. For these, control and test sample need to be assayed together, applying different dyes that allow the reading of the samples by scanning with different wavelengths the same slide (Schulze, et al., 2001).

The second microarray system is the Oligonucleotide Probe Array Method, this is mainly used by Affymetrix for their GeneChip® arrays. The most important difference in the Affymetrix GeneChip®, is that the synthesis of probes is done directly over the glass slide, rather than created and later placed on the slide (Lockhart, 2000).The *in situ* synthesis of the probes allows the precise positioning of millions of probes per array. This method is more accurate than the spotted array and permits comparisons between assays performed on multiple chips (Bunney, 2003, Schulze, et al., 2001). Therefore, the sample is labeled

with only one dye, however the reading of the chip is similar to the spotted arrays (Figure 2. b).

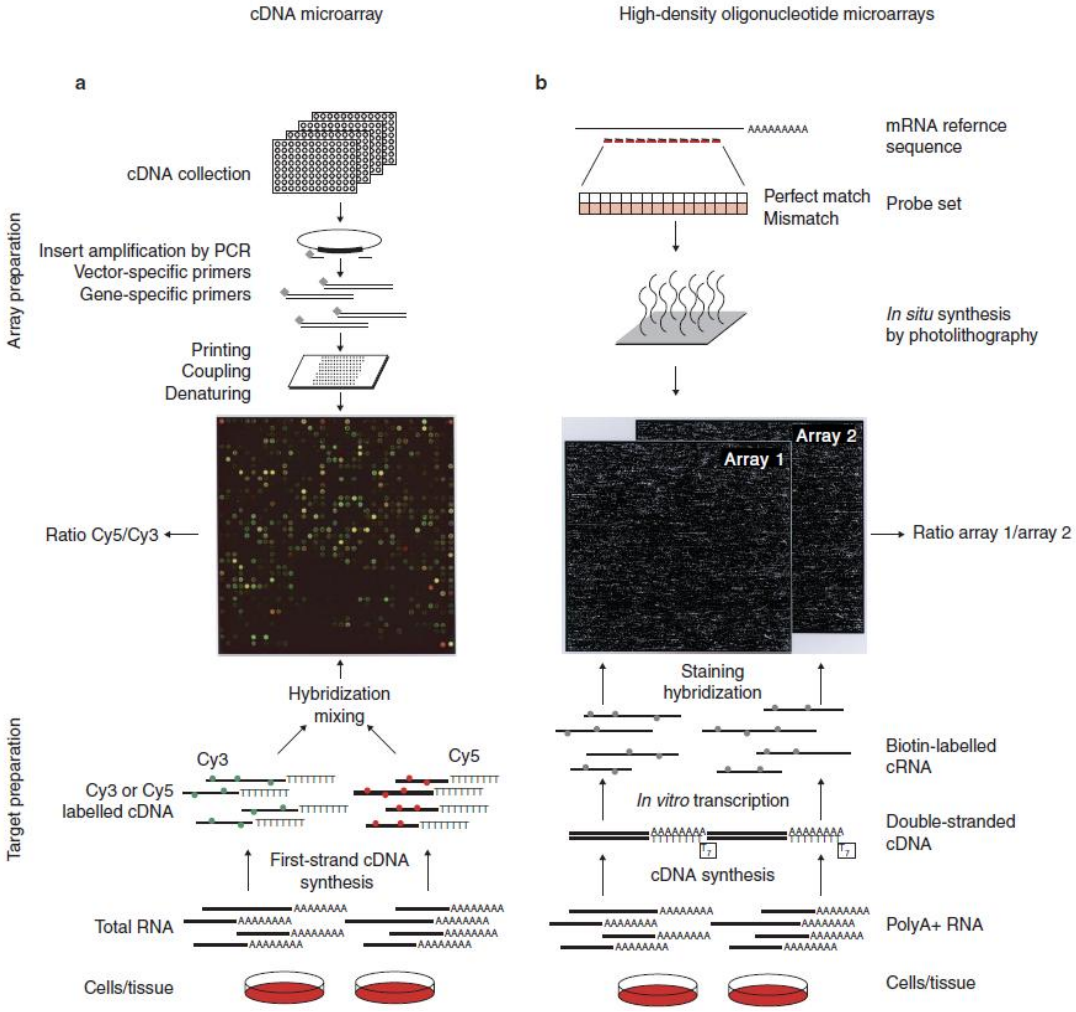


Figure 2. Overview of array and target preparation for cDNA Microarray and High-Density Oligonucleotide Microarray (Schulze, et al., 2001)

Experiments performed with these high density microarrays, generate very large amounts of data. In expression assays, the hybridization of one chip, can produce quantitative results, for as many as 40000 genes (Lipshutz, et al., 2009).

Microarray Data Preprocessing and Quality Assessment

For the large amount of information generated by the microarrays, the analysis of the data becomes a critical component for useful interpretation, because in this section it is here that errors on the chip or outliers in the data are identified. The analyses of the microarray data are performed in a sequence of steps that lead from the raw information used in the design of the array to the expression data. Major steps during the analysis of the data are to eliminate the sources of variation and identify possible defects in the chips or hybridizations, it is also necessary to perform several analyses to extract information useful to the goals of the investigation (Tang*, et al.*, 2009). These processes can be separated in three basics steps: preprocessing of the data, quality assessment and analysis of expression values. During preprocessing minor tasks are performed, including background adjustment, which is indispensable to reduce the total error owed to unspecific binding inherent to probe and for noise of the optical system within a chip. In order to compare all the chips together, normalization across all chips in the set is done to eliminate variation produced by external factors such as differences in transcription efficiencies, labeling or physical problems with the arrays. With the probe data adjusted and normalized it is possible to obtain the expression value of the probeset through summarization. This is designed to convert probe intensities to probeset hybridization values, which is accomplished by taking the intensities of all the probes in the probeset and averaging them. However, depending on the summarization method used, this procedure varies. In general, the idea is to apply weight to each probe in the probeset and eliminate or give less weight to probes that are out of the acceptable range within

the group in the final calculation of the probeset expression value. (Gentleman, 2005).

A corrupted chip can increase the number of differentially expressed genes resulting in more false positives. Therefore, quality assessment is done to detect irregularities in the arrays that can mean major problems or if the array is out of the acceptable random fluctuation range. A few low quality chips in a set can lead to wrong conclusions and render a complete experiment. Taking into account the high monetary, time and human costs of these kind of studies, quality control is a crucial step in data analysis (Heber*, et al.*, 2006).

Quality control is directed to find possible outlier arrays in a set where the acceptable range will be dependent on the experimental design, goals and context of the studies (Bolstad*, et al.*, 2003). Quality control starts with an evaluation of the chip images and Affymetrix spikes controls. Diagnostic plots are useful to the analysis of general hybridization in the chip and identify chips that may have problems. Additional measures help to identify defective chip hybridization at each individual probe. In the case that to detect a corrupted or erroneous chip, all the quality control analyses should be repeated without it and the results compared. If the analysis shows the existence of an outlier, then the corrupted chip has to be repeated and ideally the source of error identified (Heber, et al., 2006). Once a good quality and clean set of chips is obtained, it is possible to continue with the next step of the analysis.

Microarray Expression Analysis

For the analysis of the expression data generated by the previous steps there are several methodologies that can be applied to extract useful information. The main goal in each analysis is filter and select candidate genes that present an interesting expression profile or selected from any of the applied filters.

Across the literature two tests can be found as a basis for any expression analysis, the background filtering, and the differential expression analysis. During a microarray expression analysis we would expect that not all the genes can be detected in the sample due to tissue and developmental stage specific expression. For a given tissue or developmental stage there will be a set of genes where its expression is not going to be detectable. To filter these genes we establish an expression threshold called the background level. This threshold can be static or dynamic. A static threshold is an arbitrary value determine by the investigator, above which the expression can be called significant. Dynamic thresholds are considered more accurate because they are based on the information take from the microarrays. These kinds of thresholds are determined with spike in controls in the chip that enable us to identify the detection limit of the assay for a specific microarray(Gentleman, 2005).

After identification of the genes above background the next involves identification of the genes that have significant differences in expression between the treatments or samples. This is determined by statistical analysis such as the F or T tests. Statistical packages are available for microarray data analysis where they perform thousands of

hypothesis tests iteratively. Due to the high number of tests carried out the possibility of finding false positives is greatly increased. Several approaches have been studied to adjust the calculation of p-values in order to reduce the percentage of false positives to strengthen the biological significance of the analysis. Methods developed since the 80's to control multiple testing problems (HOMMEL, 1988, SIMES, 1986) were focused only on the probability of finding errors among the accepted tests but this approach can be too severe. Alternative methods to the Family Wise Error Rate (FWER) procedures were needed (Benjamini, 1995). The False Discovery Rate (FDR) method was presented by Benjamini and Hochberg (1995). This approach takes into consideration the number of false positives and also the erroneous rejections to determine the probability that a specific test be a false positive among all the comparisons. FDR allows for the control of error from multiple tests with a decrease in false negatives in contrast to the FWER methodology (Benjamini, 2000).

The final part of the microarray analysis is the analysis of the expression itself. Several approaches can be used to extract information from the probeset data, such as cluster analysis, the differentially expressed analysis, the correlation analysis and the network connection analysis.

Main Characteristics of the Lettuce Chip

A custom GeneChip array of lettuce (*Lactuca sativa*) was designed at UC Davis and developed by Affymetrix. This chip was designed to detect Single Feature Polymorphisms (SFP) in over 35,000 lettuce genes, using an Affymetrix high density GeneChip® technology

(Figure 3) (van Leeuwen, 2009). The purpose of this project was to improve marker discovery and generate an ultra-high density genetic and map made available for characterization of important agronomic traits (Caldwell, 2008).

The Lettuce Chip posesses 6,553,600 cells of 5µm and 6,482,479 of them contain probes. These probes are divided in to six categories; the first category in which majority of the probes falls into consists of the lettuce tiling probes with 6,410,923 probes derived from express sequence tags (ESTs). These probes were staggered by 4 base pairs across the length of a contig and were offset by 2 base pairs on both sense and antisense DNA strand in order to interrogate every 2 base pair position of a contig. The other five types of probes are controls and, the expression probes. The control probes are include the technical replicate probes, the Affymetrix control probes, the Affymetrix anti-genomic probes and the Affymetrix B2 Oligo grid probes, more information can be found in the Lettuce Chip WebSite (http://chiplett.ucdavis.edu/) (van Leeuwen, 2009).

Figure 3. Affymetrix Custum Lettuce GeneChip® (Seed Biotechnology Center, 2010)

JUSTIFICATION AND SIGNIFICANCE

Lettuce is one of the most abundant fresh vegetables commercialized in the United States, as well as in Central and South America, with a production of 5.4 million tons in 2007 (FAO, 2008). Lettuce seeds present thermodormancy and this is a problem in the agricultural industry because expensive measurements or seed treatments need to be done in order to make seeds to germinate. Some of these measures are pre-hydration and drying of the seeds (Toh, 2004); and these practices bring shorter life span of seeds and more sensitivity to the storage conditions (Tan, 2003). This situation gains importance during fall, because in order to supply the winter markets, lettuce seeds need to be planted in desert regions where the temperatures are high enough to cause thermoinhibition. It is necessary to apply different treatments to achieve seed germination, that increase the production costs (Toh, 2004).

The understanding of thermodormancy in lettuce as a model crop and the regulation pathways is relevant to develop new methods or varieties that improve the life time of seeds and decrease the production costs. This should improve the efficiency, increase the reliability and reduce the expenses (Tan, 2003). The use of the microarrays to study expression patterns brings the advantage to screen complete pathways and be able to identify other genes that are not only the ones that are directly related to germination. Also during the progress of this investigation developed new bioinformatics tools, protocols to determine expression patterns and extraction of microarray data.

OBJECTIVES

General Objective

Assess the global transcriptome analysis and interconnections among networks controlling thermodormancy using microarray data in lettuce (*Lactuca sativa* var Salinas and *Lactuca serriola*)

Specific Objectives

-Identify genes that present differences in expression across the treatments using an ANOVA analysis

-Establish clusters or modules of genes that have similar expression with Weighted Gene Correlation Network Analysis (WGCNA)

-Determine thermodormancy related genes using annotation

MATERIALS AND METHODS

Plant materials and treatments

Microarray expression assays were carried out by an Affymetrix high density Gene Chip® (Affymetrix, Santa Clara, CA, USA) in *Lactuca sativa* cv. Salinas (Salinas) and *Lactuca serriola* (Serriola) seeds, in two different imbibition temperature treatments (20ºC and 35ºC) and two time points (0 and 24 hours). In addition a Serriola sample treated with 3µm of ABA in order to test effect of a inhibit concentration for dormancy of ABA in this variety (Table 1).

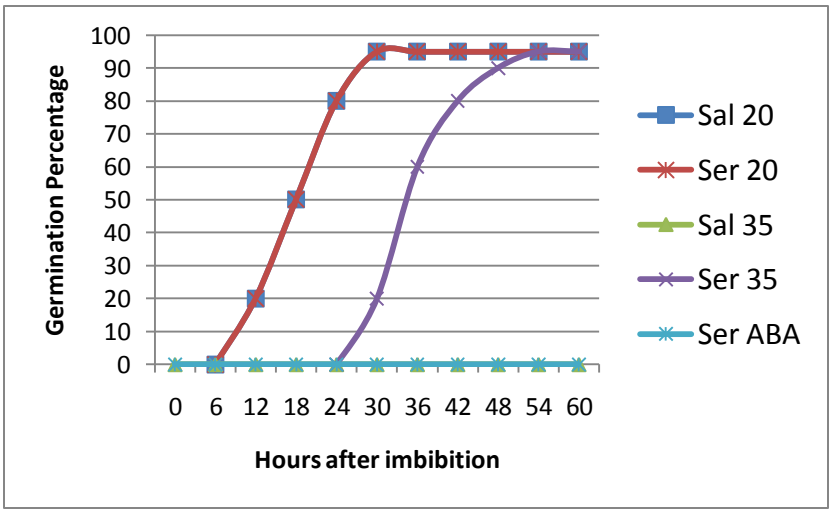


Figure 4. Germination curves for treatments

Table 1. Description of samples used in microarray analysis

Genotype	Treatment ID	Reps	Time After Imbibition	Temp	ABA	Germination
Salinas	DrySeed Sal	2	0h	N/A	N/A	N/A
Serriola	DrySeed Ser	2	0h	N/A	N/A	N/A
Salinas	Sal 20	3	24h	20°C	-	Yes
Serriola	Ser 20	3	24h	20°C	-	Yes
Sali	Sal 35	3	24h	3	-	No

nas				5°C			
Serriola	Ser 35	3	24h	3 5°C	-	Yes	
Serriola	Ser ABA	3	24h	3 5°C	+	No	

Preprocessing of the Microarray Data

Preprocessing of the microarray data was done by Bioconductor software (Gentleman*, et al.*, 2004) with the Robust Multi-Array Average expression measure (RMA) algorithm (Irizarry, Bolstad*, et al.*, 2003, Irizarry, Hobbs*, et al.*, 2003) found in the R-Package affy 2.6.2 (Irizarry R.A, 2010) under the function "rma". This method combines a convolution background correction, quantile normalization and a summarization by median polish algorithm that returns the probeset hybridization value from the raw probe intensity value. The preprocessing was performed in four separate batches due to differences among the samples, Dry Seed (no imbibition), 24h imbibitions at 20°C and 35°C without ABA and the 24h imbibitions at 35°C with ABA. The four different sets were then joined for further analyses.

Quality Assessment

In order to asses the quality of chip hybridization by having access to probe level data instead of using "rma" function that outputs summarized data for the entire probe sets, two separate functions were used to obtain probe level data. A step by step normalization, using convolution background correction as well as quantile normalization with the R-Package affy 2.6.2 under the functions "bg.correct" and "normalize" respectively. As an the R-Package affyPLM 2.6.2 (Bolstad, 2010) was used to obtain another set of probe level data. The function

“rmaPLM” applies a regular RMA normalization but returns the probe intensity value instead of the probeset hybridization value and fit the data to a Probe Level Model (PLM). It also calculates the Standard Error, Estimate Errors, Covariances, Residuals and Weights for each probe.

Affymetrix spike-in controls were checked to assess general quality of the chip. The average value for each replicate of the bioB, bioC, bioD and Cre was calculated to set up the expression scale and compare with affymetrix information (Affymetrix, 2004). When the lettuce GeneChip® was designed, in addition to spike-in probes another set of probes were synthesized on the chip. These probes correspond to ten conserved genes in the lettuce which are replicated indentified as Technical Replicate genes (TR) (van Leeuwen, 2009), for TR average, standard deviation and coefficient of variation to assess the homogeneity within each chip and across the replicates were calculated.

With data from the PLM fitting plots of Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) were generated using R package graphics with the functions “mbox” and “boxplot”. A final way to assess quality in a set of chips is by the correlation between them. Among replicates is expected to have a correlation coefficient over 95%, below this can mean problems with the samples. To assess this, a Pearson correlation coefficient was calculated in pairwise comparisons between all the chips (Gentleman, 2005).

Diagnostic Plots and Images

To generate box plots and histograms “boxplot” and “hist” functions of R package Graphics 2.9.0, was applied to raw data, normalized probe data and normalized probeset data sets.

Very useful tools during a microarray quality assessment are the chip images. They were generated directly from the reading of the chip data and to obtain an overall picture of the hybridization. Using data generated for the PLM fitting, it is possible to plot an image of signs of the residuals, which allowed the visualization of artifacts in the chip that could not be detected in the raw image (Gentleman, 2005).

Principal Component Analysis and Sample Clustering

For analysis the relationships between the samples a Principal Component Analysis was carried out with the algorithm “prcomp” of the based package “Stats” for R. 2D plots of the first, second and third component were done to visualize the possible sources of the variation in the dataset.

Sample clustering was done using Pearson correlation; distances were analyzed with hierarchical clustering using the average method provided in based package stats “hclust” for R.

Expression Analysis

After RMA background correction, the next filter that was applied to the data was eliminating the genes with expression values that were below background. Background level was determined based on the AntiGenomic Probesets (van Leeuwen, 2009), calculating the 90th percentile of these values for each chip and using the maximum like

expression threshold. Filtering was done with R-Package Genefilter 2.9.0 (Gentleman R, 2010) applying the “pOverA” algorithm.

The list of above background genes was analyzed with the TAGGIT macro to determine changes in the percentage of genes per category, for this the average, standard deviation and coefficient of variance (CV) per each category across treatments were calculated. The categories with an average percentage of genes higher that 1% and a CV higher that 15% are considered to have real variations.

For genes with above background signal statistical test were carried out to determine significant differences in expression in comparisons between treatments. An F-test of hypothesis was performed using the R-Maanova package 2.9.0 (WU, 2010), pair wise comparison between all the treatments to obtain any possible combination. F-test was performed with 5000 permutations, with “adaptative” (Benjamini, 2000) method for adjustment of the P-values with a thresholds of 0.0005 of adjusted P-value for determine significance.

CLUSTERING METHODS

VENN Diagrams

After differentially expressed genes were determined VENN diagrams were constructed to calculate the number of overlapping genes among comparisons, three per each diagram. Eight diagrams were done using the most significant sets (Table 2).

Table 2. Comparisons used for generate VENN diagrams with differentially expressed genes

VENN ID	Comparison

Venn1	DrySeed_Sal-Sal20C	DrySeed_Sal-Sal35C	Sal20C-Sal35C
Venn2	DrySeed_Ser-Ser20C	DrySeed_Ser-Ser35C	Ser20C-Ser35C
Venn3	DrySeed_Sal-DrySeed_Ser	Sal20C-Ser20C	Sal35C-Ser35C
Venn4	Sal20C-Sal35C	Sal35C-Ser35C	Ser20C-Ser35C
Venn5	Sal20C-Sal35C	Sal35C-Ser35C	DrySeed_Sal-Sal35C
Venn6	Ser20C-Ser35C	Sal35C-Ser35C	DrySeed_Ser-Ser35C
Venn7	Ser20C-SerABA	Sal35C-SerABA	Ser35C-SerABA
Venn8	Ser35C-SerABA	Sal20C-Sal35C	Ser35C-Sal35C

Results from the diagrams were analyzed to select interesting groups of genes to take them into further analysis. Selection was based on finding groups of genes related with the Sal35 or SerABA treatment where there could be some association with temperature, germination and genotype.

Weighted Gene Co-Expression Network Analysis

Another method to cluster the gene in groups is the Network Analysis. In this analysis genes (called nodes) are assigned to modules based on in their expression profile using a correlation coefficient to measure the distance between them. R-Package WGCNA 2.9.0 (Weighted Correlation Network Analysis) (Langfelder, 2008) was used to construct a network using all genes that their expression were above background (15309 genes) using the methodology described in the website (Langfelder, 2009).

Briefly, the steps to generate the WGCNA are as follows: (1) Calculate adjacency between the nodes, this matrix is generated with correlation analysis to determine concordance between nodes, (2) for the dissimilarity matrix the adjacency is taken to a power of 15 to calculate similarity and then it is transformed into dissimilarity (3) identify modules of nodes based in height from an hierarchical clustering, (4) determine eigengene values for each module determine likelihood the first principal component of the data (5) calculate Module Membership (MM) among the nodes and distance between the modules, (6) calculate the relation of the modules against the external factors involved in the analysis (Figure 5) (Langfelder, 2008, Zhang, 2005).

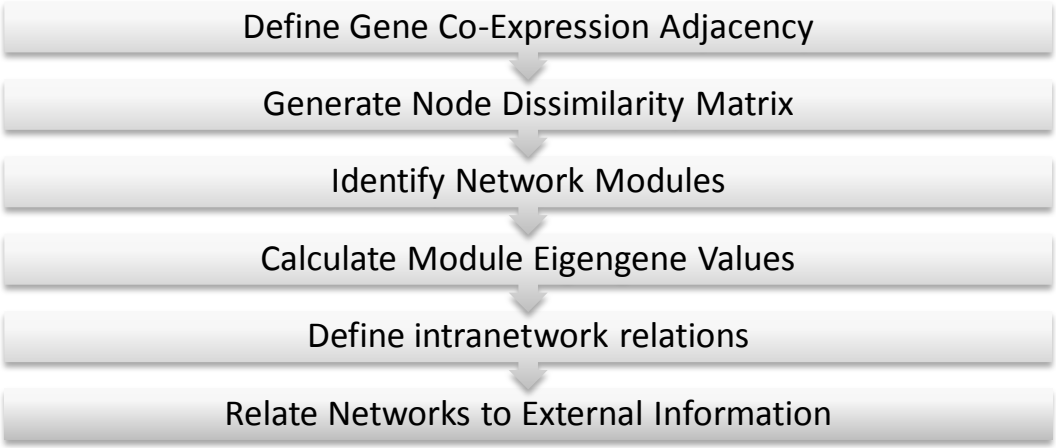


Figure 5. Workflow for generate the WGCNA
Based in (Zhang, 2005)

GENEByGENE ANALYSIS

Determination of possible related genes with thermodormancy was done base in annotational information and the expression profile of the genes tag by the previous methods. For annotation purposes the closest Arabidopsis homolog was used. Annotations of genes in Arabidopsis were used because the lettuce GeneChip is not annotated yet.

A set of genes was extracted directly to determine their their relationship with the ABA regulation and metabolism (Table 3). Using the information from the ultra saturated lettuce map (http://chiplett.ucdavis.edu/index.php) (Michelmore *et al.*, unpublished) these genes were located on the lettuce linkage groups and mapped contigs in their vicinity were identified, correlation between expression profile of the genes and the contigs were calculated, and contigs with a correlation value higher than %90 selected as candidate genes.

Table 3. Genes involved in the ABA regulation or metabolism pathway represented in the microarray

Gene	Contig	Gene	Contig
LsNCED4	GB_84579411	**HAB1**	CLS_S3_Contig527
SnrK2.2	CLS_S3_Contig3188	**ABF2**	QGI11J05.yg.ab1
SnrK2.6	CLS_S3_Contig8687	**ABF2**	CLS_S3_Contig7421
PYL2	CLSY2969.b1_A24.ab1	**ABF2**	CLS_S3_Contig7163
PYL2	CLS_S3_Contig6183	**ABF2**	QGI6O04.yg.ab1
AHG1/PP2C/HAB1/ABI1	CLS_S3_Contig9663	**PIL5**	CLSL1469.b1_J08.ab1
AHG1/PP2C/HAB1/ABI1	CLS_S3_Contig9696	**PIL5**	CLS_S3_Contig6666
AHG1/HAB1	QGB8B20.yg.ab1	**RGL2_DELLA**	CLSX3670.b1_K06.ab1
PP2C	CLS_S3_Contig8135	**RGL2_DELLA**	CLSM11108.b1_H18.ab1
HAB1	CLS_S3_Contig4603	**RGL2_DELLA**	QGD6I10.yg.ab1

Using Microsoft Excel TAGGIT macro the genes under different set of labels related with dormancy and germination were categorized. Genes in the categories of Dormancy, Germination Related, Heat Shock, Stress, ABA, Auxin, Gibberellins and Ethylene were extracted and their expression profile was analyzed to determine their plausible function in thermodormancy.

ABA related genes were used to determine relevant modules in the network that they were assigned. Also modules which eigengene profile were considered important and were tagged as relevant modules. With the java application VisANT (Hu*, et al.*, 2004), relevant modules from network analysis were plotted to find highly connected nodes (hub genes) that are central parts of each module (Hu*, et al.*, 2008). These genes show high correlation with experimental information and a next filter was apply to select the ones that their expression pattern indicates a possible relation with thermodormancy.

To determine the final set of genes with more relation with thermodormancy, the genes selected for the Venn diagrams, from network analysis, TAGGIT macro and VisAnt were analyzed together in the MapMan application (Thimm*, et al.*, 2004).

RESULTS

Chips Quality Assessment

Using the spike in controls, minimum background hybridization value was determined to be 7.478 and left 15309 genes with detectable expression. First insight in the quality of the chips was with the Affymetrix spike-in controls. The concentration scale of the hybridization controls is fine for all the treatments, with a lower value for BioB and increasing until Cre (Figure 6).

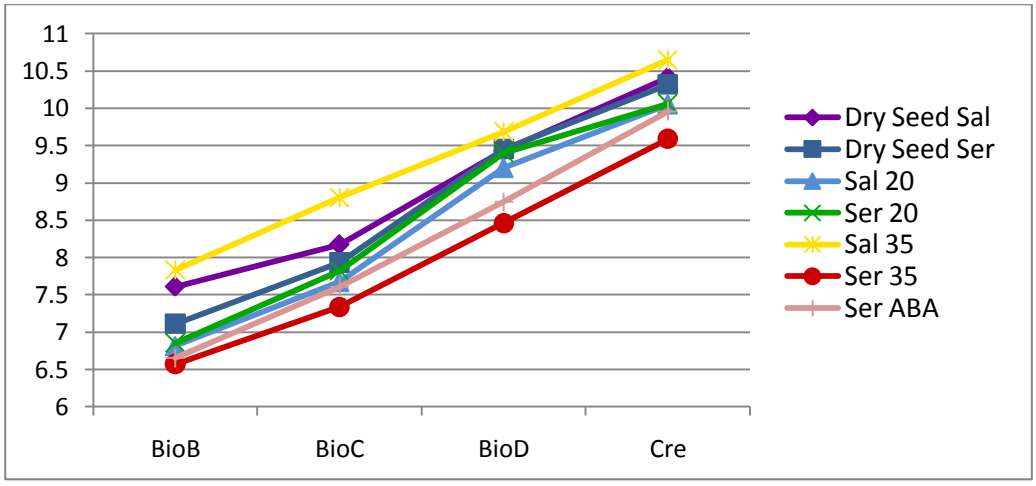


Figure 6. Average Expression per Treatment for Affymetrix hybridization controls

Hybridization intensities of TR's, 5 TR (TR 2, TR 6, TR 7, TR 8, TR 9) were above background. These 5 genes had low CV (< 4%), indicating consistency of the data as well as integrity of the chips. On the other hand, the remaining 5 genes which expression was below background had higher CV due to bigger variation at low intensity values, product of the noise in the chip (Table 1).

Table 4. Data of Technical Replicates genes in the chip

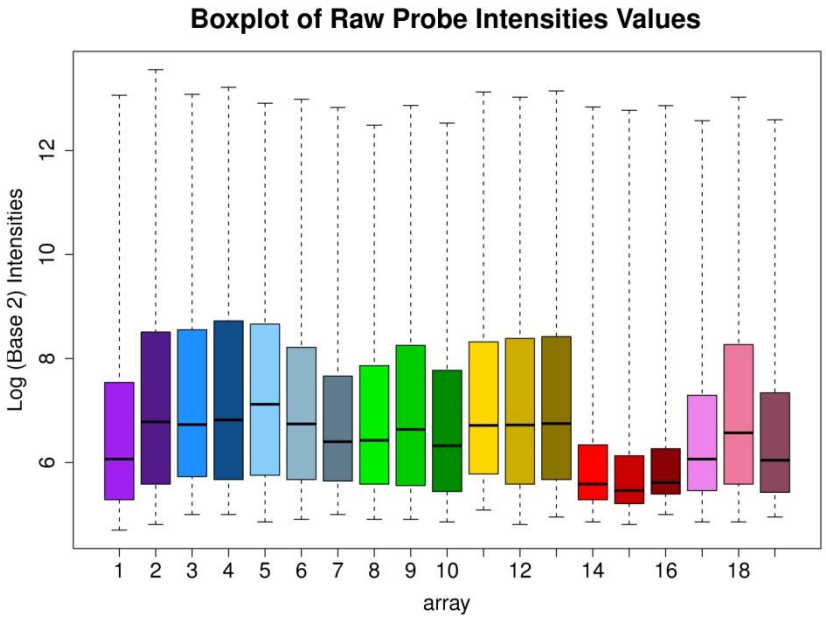
Gene	Average Expression	Standard Deviation	Coefficient of Variance
TR_1	3.80339	0.502599	15.15%
TR_2	9.618371	0.133886	1.42%
TR_3	2.571024	0.486395	22.30%
TR_4	4.742193	0.411715	9.57%
TR_5	3.887269	0.465064	13.33%
TR_6	11.08908	0.092765	0.84%
TR_7	11.3522	0.080172	0.72%
TR_8	9.682641	0.111291	1.22%
TR_9	7.948218	0.221691	3.22%
TR_10	2.350573	0.456102	20.99%

Due to the high complexity and amount of data generated for the microarray, exploratory visualization is needed to identify quality problems. Looking the data of all the arrays at the same time can help to detect irregular chips. In this the boxplot and the histogram are an appropriate tool to accomplish it (Gentleman, 2005). Raw data obtained from the microarray is very variable due to internal and external factors that affect the intensity value. For this reason a preprocessing of the sample data is needed before further analyzing the data. Boxplots of raw and normalized probe intensities values show the effect of the normalization process over the data (Figure 7).

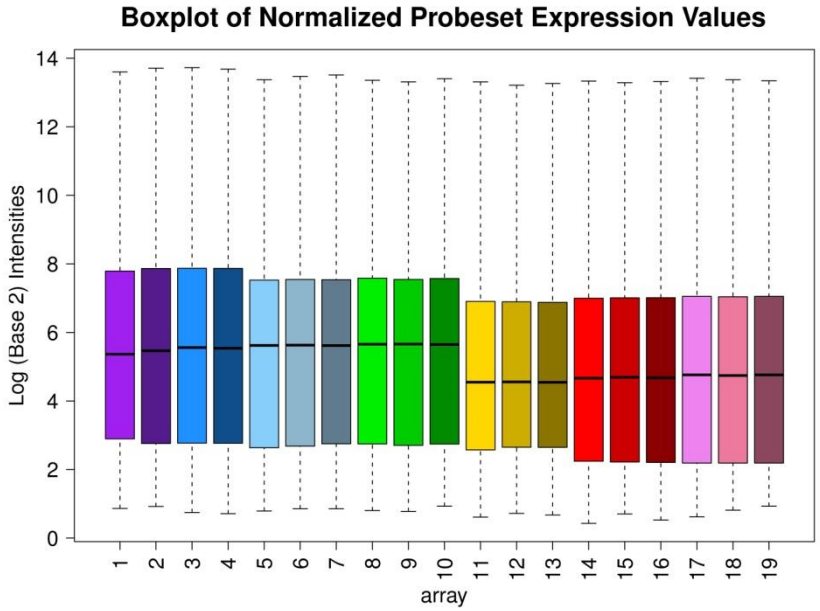
The chips of the same treatment were normalized together to avoid leveling off the hybridization intensities of differentially expressed genes between the treatments. We can see at probe level the fitting of the curves by subset of normalizations (Figure 8. A), at the top the 20°C

samples, follow by dry seed samples and at the bottom 35°C samples, regardless that this last group was done in two different normalizations. An oscillation is visible in 35°C samples in the lower part of the curve. This seen to be a result of lower hybridization in a sample but is not a problem if the curve stabilized before the background level. In a probeset level, curves fit by treatment remaining close by normalization. There is not oscillation at this level because it is a result of summarize information of all the probe by probeset, also 35°C treatment has the lowest expression value (Figure 8. B).

Regarding the plots from the Probe Level Model fitting, the RLE show that the data is in good conditions, the deviation of the relative expression is in the range of -0.04 to 0.04 and the replication is good (Figure 9. A). In case of the NUSE most of the samples are between 0.95 and 1.05, as expected according to the literature but the Sal 35 is off of this range (Figure 9. B). This problem probably is due to the lower expression that generates a higher variation in the data. Nevertheless, with this issue in the Sal35 the other analysis show that after processing the data is in proper condition and the variations are reduced to the expected levels. For this quality assessment concludes that the set of chips do not present any problem and the data generated is in good conditions for the expression analysis.

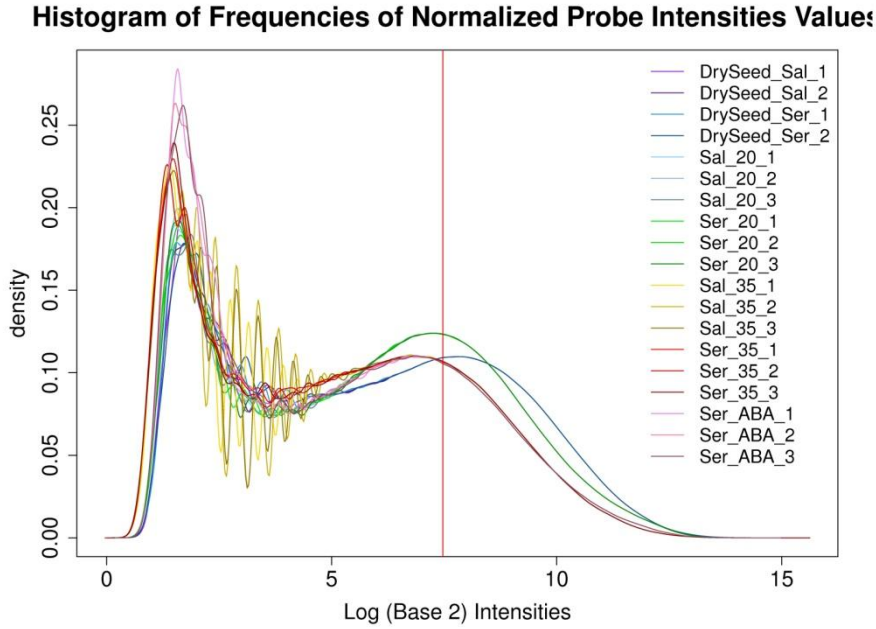


A.

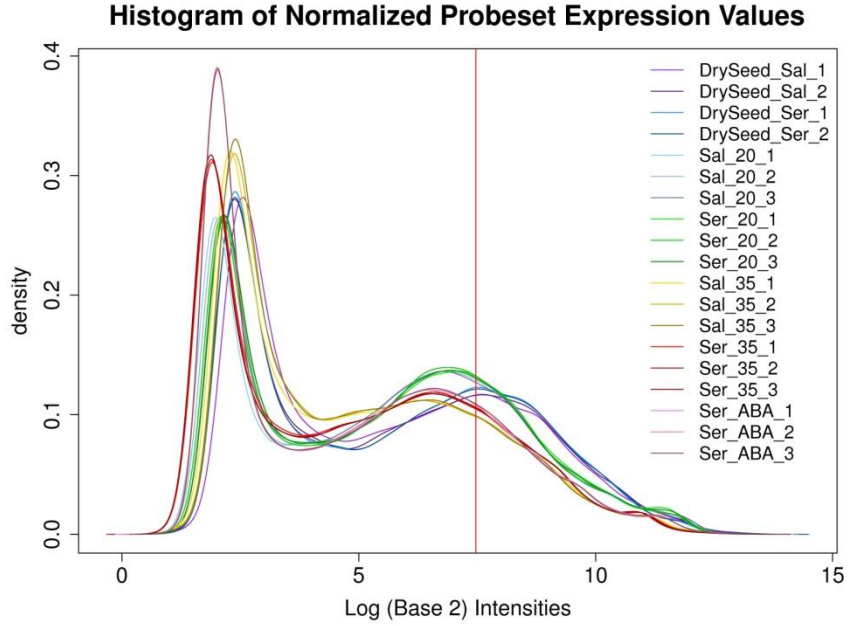


B.

Figure 7. Boxplot of Hybridization Intensities for Probes across chips.
By array 1-2 Dry Seed Sal, 3-4 Dry Seed Ser, Sal 20 5-7, Ser 20 8-10, Sal 35 11-13, Ser 35 14-16, Ser ABA 17-19
A. **Raw probe hybridization intensities, B. Normalized probe hybridization intensities**



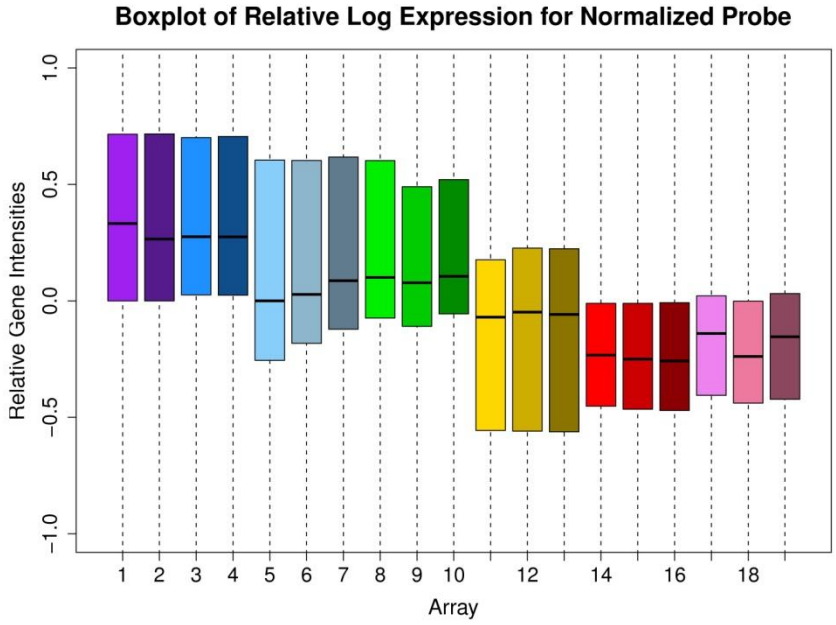
A.



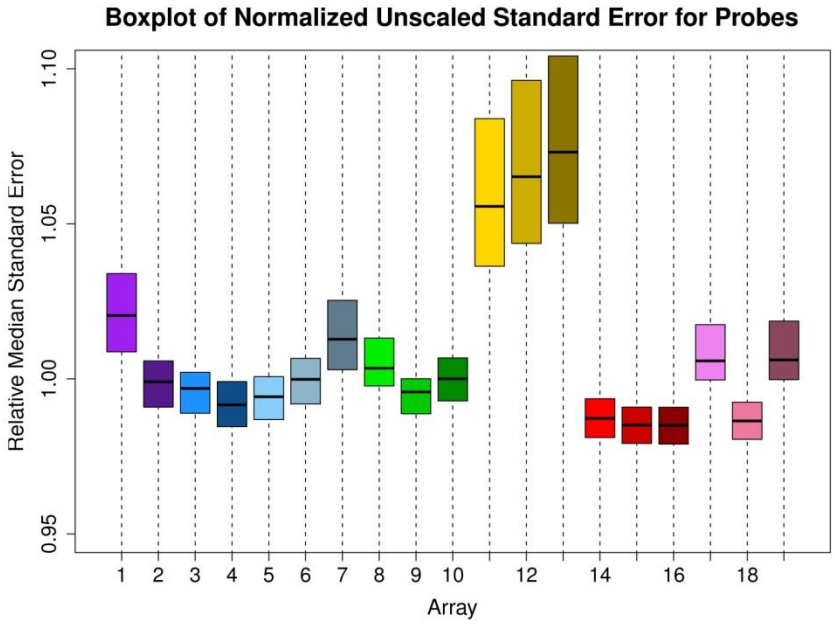
B.

Figure 8. Histogram of Hybridization Intensities across chips. The perpendicular red line is the background level.

A. **Raw probe hybridization intensities, B. Normalized probeset hybridization intensities**



A.



B.

Figure 9. Results from the Probe Level Model fitting

By array 1-2 DrySeed Sal, 3-4 Dry Seed Ser, Sal 20 5-7, Ser 20 8-10, Sal 35 11-13, Ser 35 14-16, Ser ABA 17-19

A. Relative Log Expression plot (RLE)

B. Normalized Unscaled Standard Error plot (NUSE)

Differential expression analysis show basic relations between treatments

Categorizing the genes with signal detected above background for all the treatments with the TAGGIT macro made it possible to find differences between the percentage of genes in the categories of ABA, auxin, DNA repair, cell cycle related, heat shock, cell wall modification and cytoskeleton (Figure 10).

Analyzing the number of genes detected as differentially expressed between the treatments indicated that the Ser 35 and Ser ABA treatments account for the minimum amount of differences. Also Dry Seed Sal with Dry Seed Ser and Sal 20 with Ser 20 had a low number of differences, that are the base differences between genotype. It is important to note that Dry seed Sal and Sal 35 have only 379 different genes showing an important repression of genes by the dormancy, compare to Dry Seed Sal with Sal 20 that had 2116 different genes. In the case of Ser 35 it seems to be more similar to the Dry Seed Ser than that of Ser 20, although Ser 35 was able to germinate. The major differences are between the germinating and non germinating treatments that in average each had ~1800 differentially expressed genes (Table 5).

Table 5. Differentially expressed genes above background

Comparison	Number of differentially expressed genes	
	With Fold Change >1.25	With Fold Change >1.5
Dry seed Sal-Dry seed Ser	711	357
Dry seed Sal-Sal 20	4404	2116
Dry seed Sal-Sal 35	2616	379
Dry seed Ser-Ser 20	3619	1536
Dry seed Ser-Ser 35	1917	334

Dry seed Ser-Ser ABA	2221	417
Sal 20-Ser 20	652	340
Sal 20-Sal 35	4855	2331
Ser 20-Ser 35	3174	1394
Ser 20-Ser ABA	2733	1194
Sal 35-Ser 35	1178	473
Sal 35-Ser ABA	1727	650
Ser 35-Ser ABA	374	54

Main sources of differences between treatments determined by Cluster of Samples and PCA

The cluster analysis indicated that samples that were treated at 20º C were out grouped compare to the rest of treatments. (Figure 11). The principal component partially support this result determining like the first component temperature separating the 20°C treatments with 60% of the total variance. Unless, in the second component was found genotype accounting for the 16% of the variation instead of treatment how it looks by the dendrogram (Figure 12. A). And in the third component imbibition with a good separation between dry seed and imbibed within genotype with a 15% of the variation (Figure 12. B).

Both methods indicate that seeds at 20°C are different than the other treatment due to activation of genes associated with germination. Nonetheless Ser 35, that is germinating, seems to have more weight the temperature factor than germination due to it cluster with Ser ABA. The rest of the variation can be accounted to an interaction between genotype and imbibition that requires a deeper analysis to be able to split it.

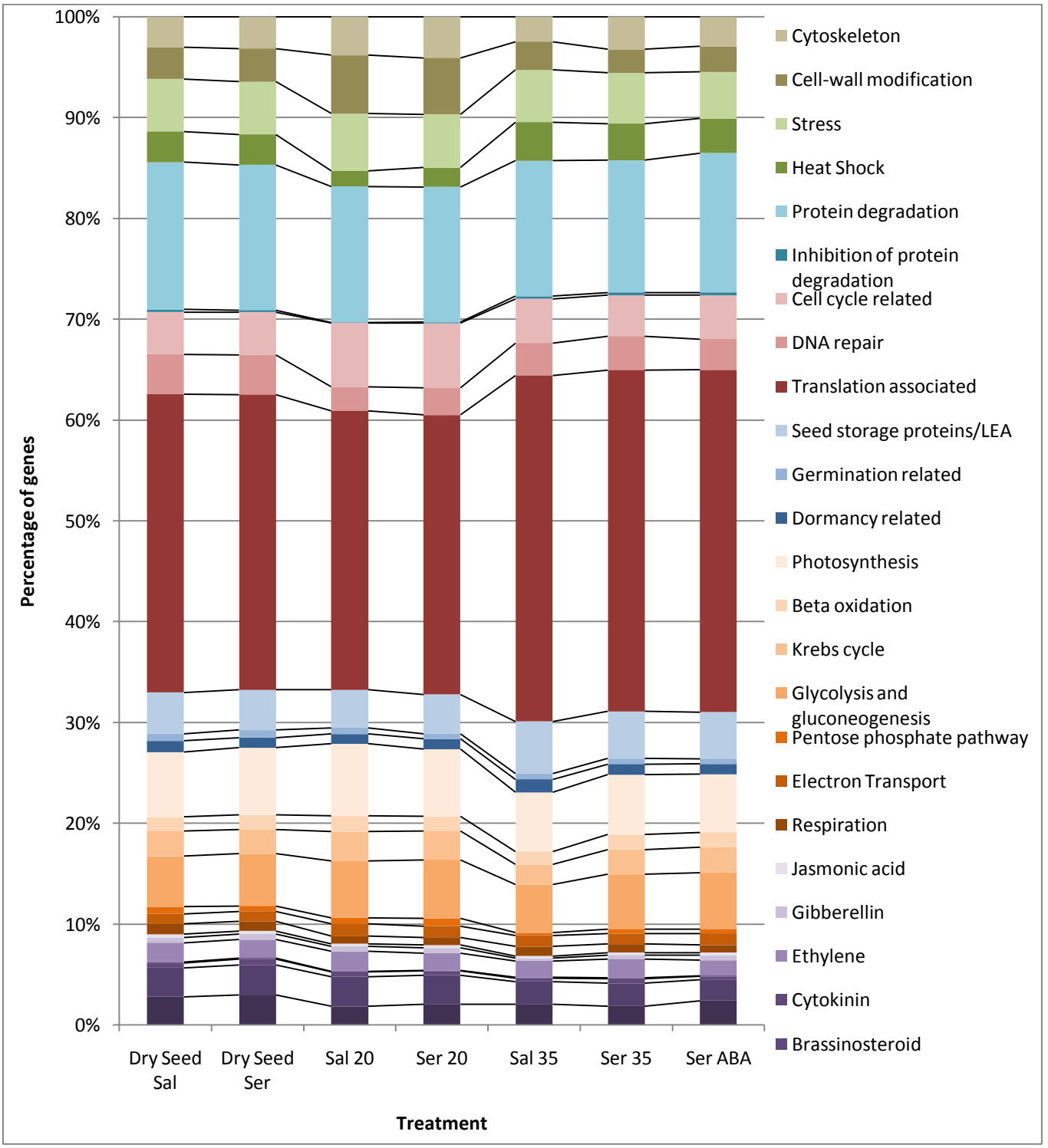


Figure 10. TAGGIT categories across treatments for above background genes

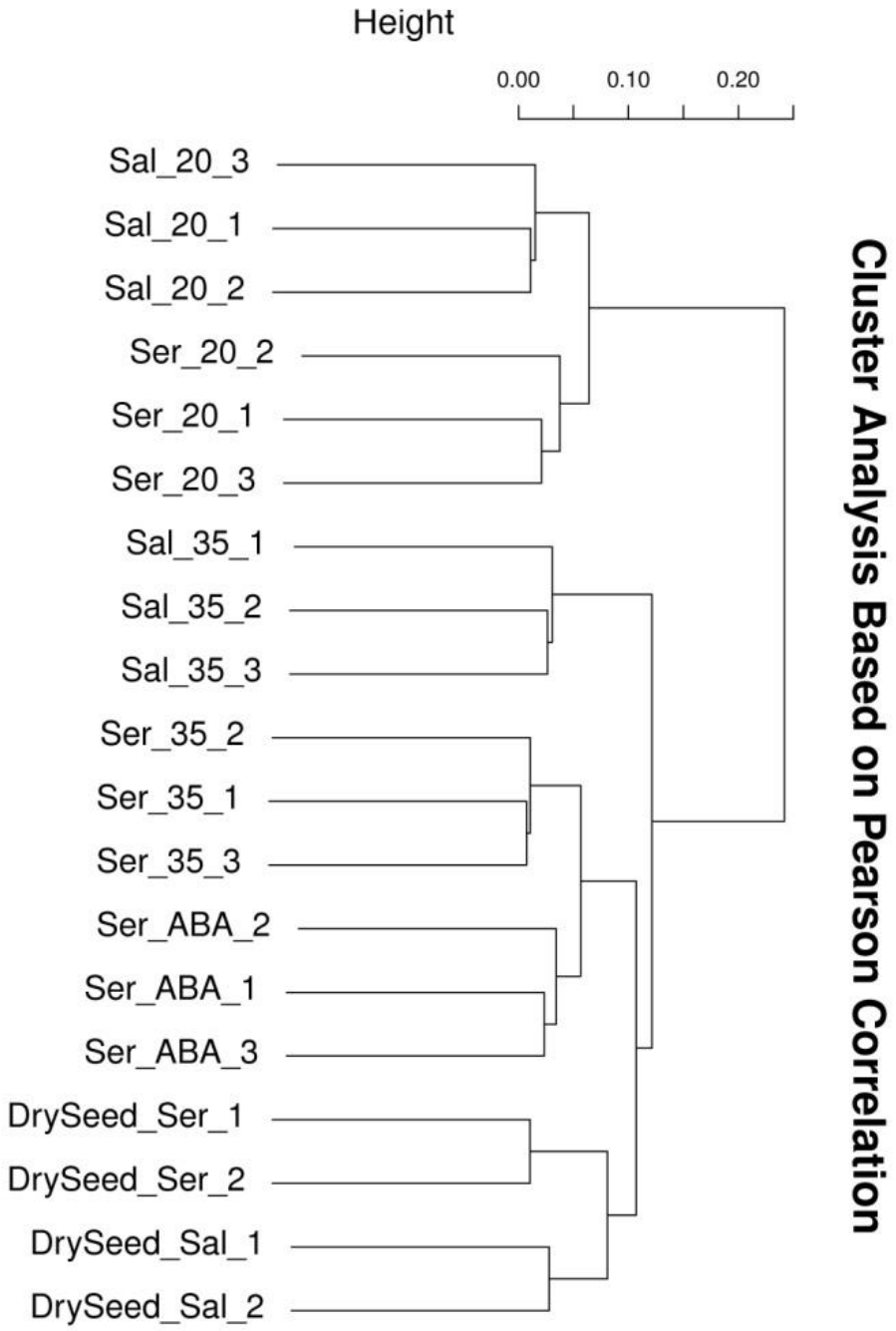
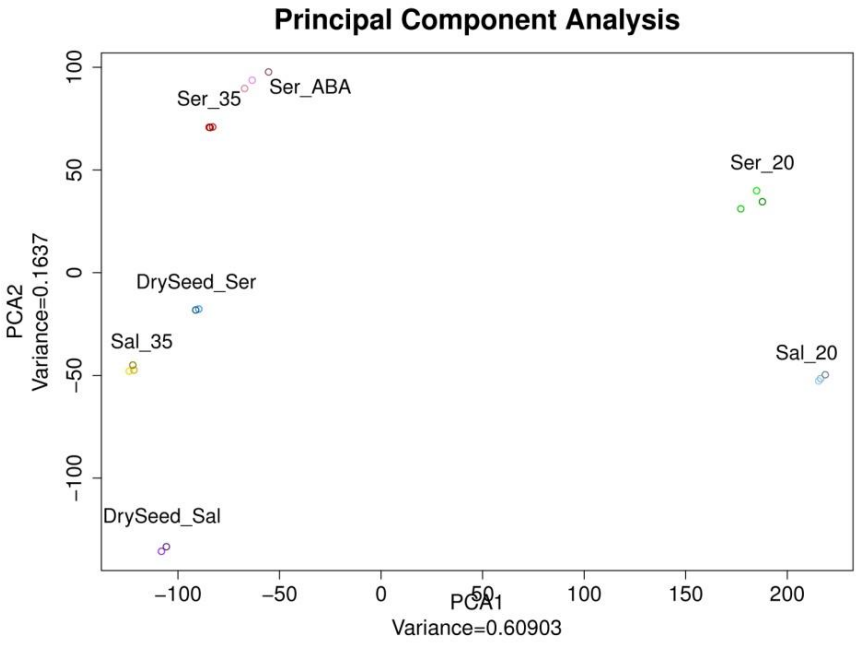


Figure 11. Dendrogram of Samples based in Pearson Correlation

A.



B.

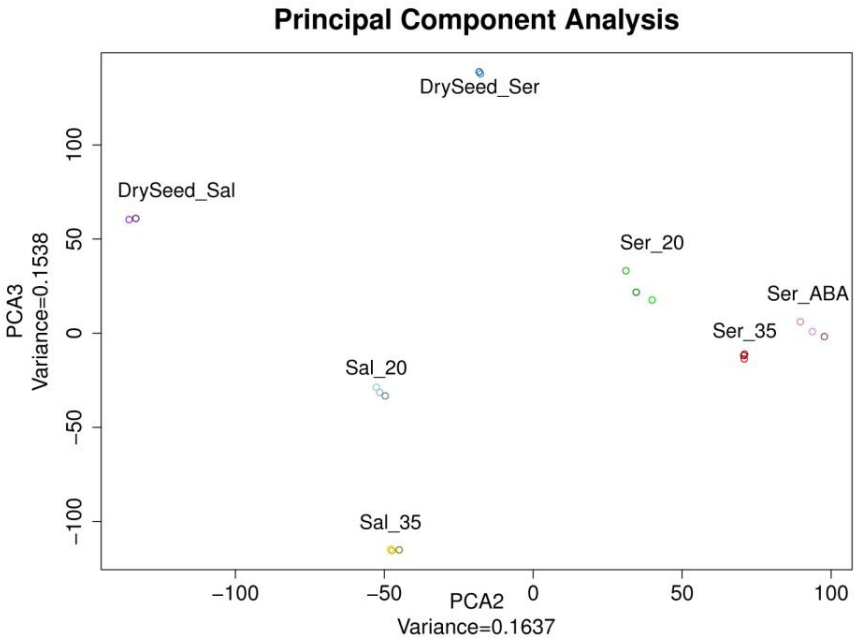


Figure 12. 2D plot of principal components
A. Plot of Principal Component 1 against Principal Component 2
A. Plot of Principal Component 2 against Principal Component 3

Combined methods to identify putative genes related with thermodormancy

The first approach to find possible genes related with thermodormancy was through the Venn diagrams, 16 Venn diagrams were developed 2 per each comparison. This resulted in 11 sets accounting for 441 interesting genes. After filtering based on it TAGGIT macro, positions in the map and expression profile the list were trim down to 90 genes, which were taken for further analysis.

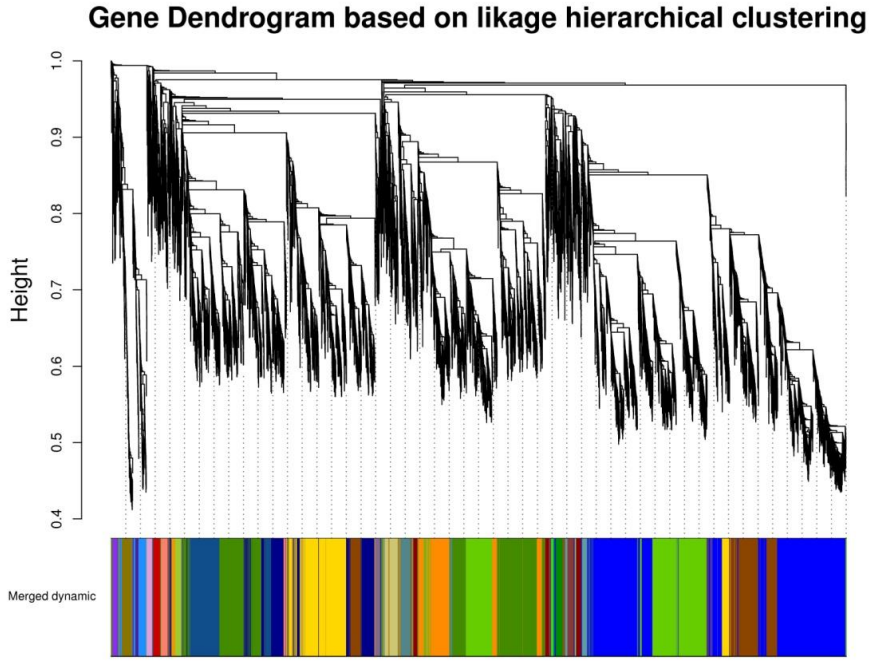


Figure 13. Hierarchical clustering of 15309 genes with module assignment

Starting with ~15000 genes WGCNA was able to make 30 modules, arbitrary color names were assigned to each of them for identification (Figure 13). The number of genes per module increased from 34 in Salmon4 to 3252 in Blue, module membership as an average go from 0.18 (Darkred) to 0.90 (Salmon4). To select relevant modules were used the ABA related genes and they were assigned to 7 different modules (Blue, Blue4, Chartreuse4, Darkorange, Darkorange4, Darkred

and Gold). Three more modules were selected (Orange4, Olivedrab4 and Khaki4) based in its eigengene profile that show relation with Sal 35 and/or Ser ABA treatments.

To determine biological significance the correlation between module eigengenes and four experimental conditions (Germination, Variety, Imbibition and Temperature) was calculated. Correlation analysis showed that the Blue4, Darkorange4, Gold and Orange4 are highly correlated with high temperature. Meanwhile the modules Blue, Darkorange, Chartreuse4 are highly correlated with low temperature and germination. Also the modules Darkred and Khaki4 were correlated with low temperature, germination and also have relation with the Serriola variety. The Olivedrab4 have relation with non-germination and a lower correlation with high temperature (Figure 15). For each of this modules a gene network with the VisAnt application was plotted which allowed the detection of highly connected genes. A total of 72 hub genes between the 10 modules, which 35 of them show an important expression pattern and where tag for further analysis.

The 10 modules contain 10079 genes, which for practical reasons it is necessary to filter them. Using the TAGGIT and the map information, the list was annotated and the interesting genes were extracted. After the first filter 1001 genes remain that can have some relation with thermodormancy, this set was later filtered based on the expression profile of each gene to generate the last set with 237 interesting genes.

A list with all of the genes tagged by the previous methods was prepared and this data was introduced into the MapMan application. Genes assigned to categories related with hormone metabolism or

relevant categories. This list contains 63 genes, 23 of them are up-regulated and the other 40 down-regulated in Sal 35.

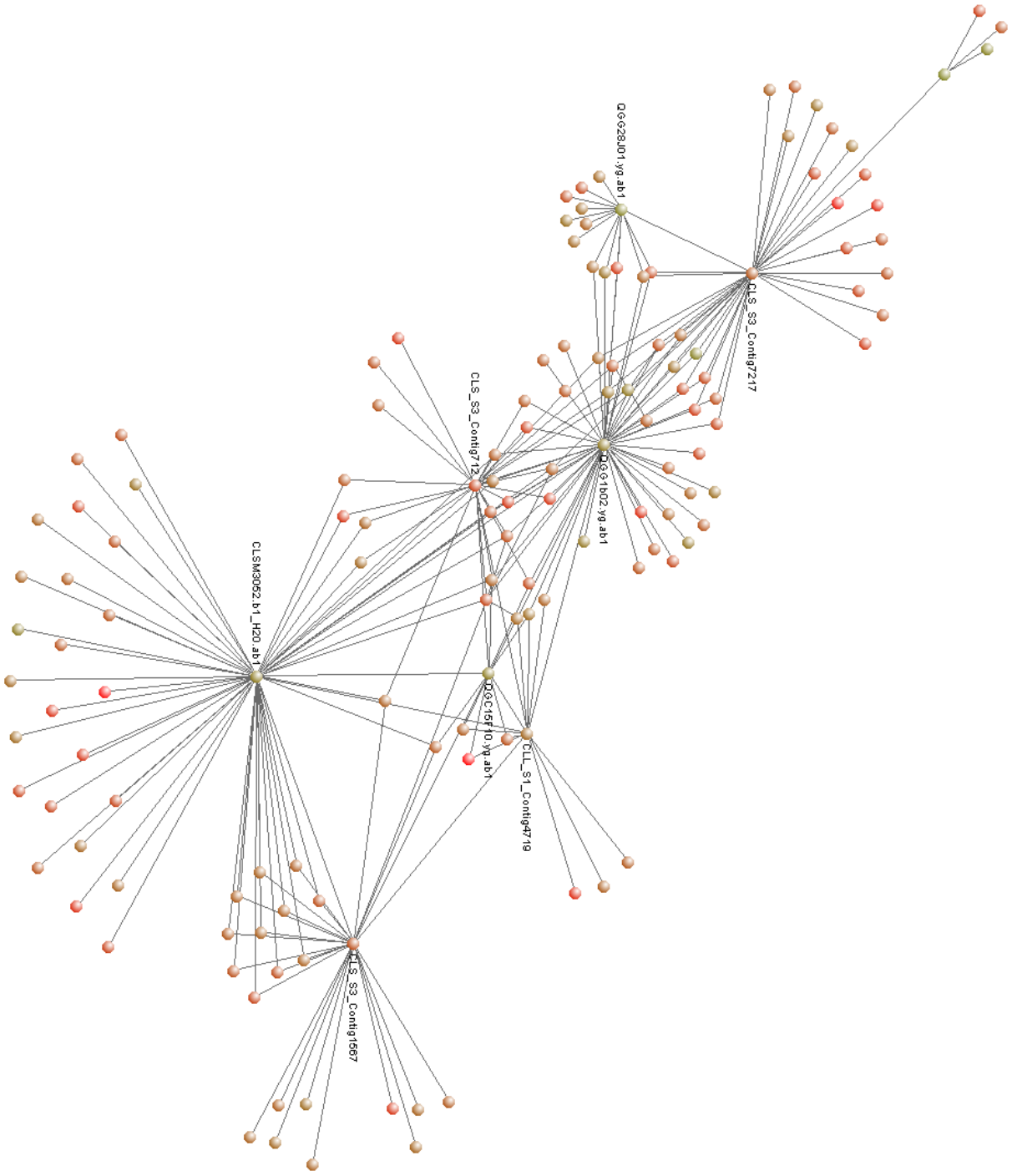


Figure 14. Example of gene network from Module Gold with an interaction cut off of 0.413

This is an example of the genes networks generated through VisAnt to visualize the interactions within the modules. Genes with name tag are the hubs in the module.

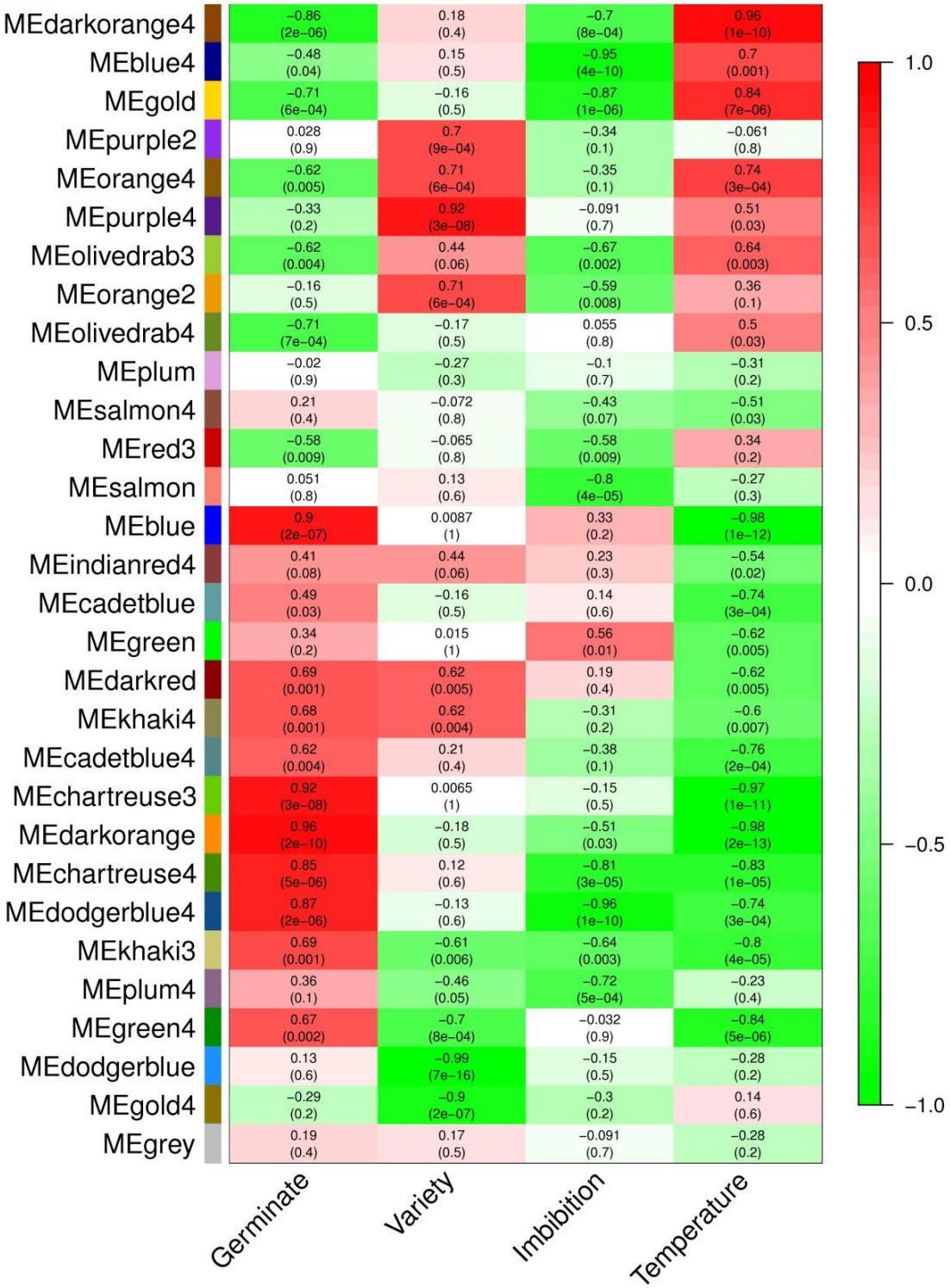


Figure 15. Heatmap of correlation between module eigengenes with experimental conditions

Table 6. Genes result of filtering microarray dataset for relation with thermodormancy

Upregulated Genes for Sal 35		Downregulated Genes for Sal 35
CLS_S3_Contig7163	QGI11J05.yg.ab1	CLS_S3_Contig527
CLS_S3_Contig7421	CLLY7887.b1_M04.ab1	CLS_S3_Contig6666
GB_84579411	CLR_S1_Contig3370	CLS_S3_Contig698
CLS_S3_Contig11334	CLS_S3_Contig10259	CLS_S3_Contig7420
QGI9I18.yg.ab1	CLS_S3_Contig10555	CLS_S3_Contig8135
QGC28G06.yg.ab1	CLS_S3_Contig11386	CLS_S3_Contig8687
CLS_S3_Contig201	CLS_S3_Contig1227	CLSL1469.b1_J08.ab1
QGB10H19.yg.ab1	CLS_S3_Contig2010	CLSM11108.b1_H18.ab1
CLSM19675.b1_E23.ab1	CLS_S3_Contig2093	CLSM3592.b1_P09.ab1
CLSM17919.b1_M16.ab1	CLS_S3_Contig2559	CLSS11062.b1_L06.ab1
CLSM8279.b1_M06.ab1	CLS_S3_Contig3188	CLSX1577.b1_A11.ab1
CLS_S3_Contig2075	CLS_S3_Contig3650	CLSX3670.b1_K06.ab1
CLVX12234.b1_D12.ab1	CLS_S3_Contig3661	CLSX8127.b1_M15.ab1
CLSS3671.b1_M06.ab1	CLS_S3_Contig3839	CLSY1248.b1_O23.ab1
CLS_S3_Contig667	CLS_S3_Contig4192	CLSY2969.b1_A24.ab1
QGB23D13.yg.ab1	CLS_S3_Contig4592	QGB22B23.yg.ab1
CLS_S3_Contig730	CLS_S3_Contig4	QGC25M21.yg.ab1

5		603		
.ab1	CLSY9332.b1_H05	683	CLS_S3_Contig4	QGC25N02.yg.ab1
	QGF17O19.yg.ab1	062	CLS_S3_Contig5	QGD6I10.yg.ab1
3	CLS_S3_Contig618	134	CLS_S3_Contig5	QGI10G13.yg.ab1
3	CLS_S3_Contig966			
6	CLS_S3_Contig969			
	QGB8B20.yg.ab1			

DISCUSSION

Delay of Germination due to High Temperature

Serriola seed has been known to be thermotolerant, but still its germination is affected by temperature producing a delay in its speed. At 20°C serriola germinate at 30 HAI (hours after imbibition), while at 35°C it takes up to 50 HAI to be fully germinated (Argyris, 2008a). The transcriptomic data support this information where are visible smaller differences between Ser 35 with Dry Seed Ser than Ser 35 and Ser 20 (Table 5). Statistical tests are not able to identify significant differences between Dry Seed Ser and Ser 35 but other approaches such as PCA or hierarchical clustering demonstrate that these two treatments have a different behavior. In a more comprehensive analysis between these samples, it was shown that Ser 35 is partially germinated where the transcriptional changes already starts but are not in the same state as Ser 20 or Sal 20 (Argyris, 2008b).

In lettuce a major regulator of dormancy is involved in ABA metabolism, was expected to reproduce the Salinas phenotype adding ABA to the Serriola seed. In *Arabidopsis* applying germination inhibiting concentrations of ABA an caused an important set of genes that are relate with the ABA response being detected as differentially expressed between the germinating and the inhibited seed. A part of the *Arabidopsis* genes detected in that experiment was expected to be found between Ser 35 and Ser ABA due to the similar conditions. However none of the genes found in *Arabidopsis* were detected in the current study. Nonetheless, Ser 35 was closely related with Ser ABA. As it can be seen in Figure 11 and Figure 12. These result are associated with the delay in germination presented in Ser 35 because these treatments shared all the same genotypic background and the

germination genes are in the process to be activated in Ser 35 but are not significant differences have not been detected yet (Argyris, 2008a).

This delay in germination of Serriola seeds at high temperature causes problems through the classical differential expression analysis because the differences presented between this sample and the inhibited seeds are large enough to be significantly detected. However other analyses such as WGCNA are able to separate these treatments due to the expression profile system applied in this method. This analysis is able to take into account the difference between the Serriola 35 seed with the Serriola ABA for clustering. Modules that are related to germination have been found which have a high expression value in all the germinating samples, even the Ser 35.

Conclusive results show that Serriola seeds at 35ºC are not in the same developmental state as Serriola seeds at 20ºC or Salinas seeds at 20ºC. This prevents to use it as germination control at high temperature. However useful information has been generated from these samples, but it seems to be necessary another sample of Serriola at 35ºC at 48 hours after imbibition that may present a better germination control.

Gene clustering related with traits

Gene clustering with WGCNA was able to unify genes with similar biological functions based on their expression profile. Using eigengenes it was possible to correlate each of the modules with seed characteristics and/or experimental conditions (Langfelder, 2008), as well as finding modules related with high temperature, germination, Salinas-specific and Serriola-specific. In other cases the modules present relation with several important conditions specially the ones related with high temperature and the Salinas variety were also

detected. This specific relation between the modules and the experimental conditions facilitated the selection of genes that potentially involved in thermodormancy. This behavior of the clustering has been detailed by the developers of the software because genes that have a very similar expression profile can be related to the same biological pathway (Zhang, 2005). According to previous microarray studies using WGCNA for analysis of data, showed that grouped genes trend to be coregulated or be related to the same metabolic or regulatory pathways (Ghazalpour*, et al.*, 2006, MacLennan*, et al.*, 2009). Using the annotational data of the genes, it was evident that some modules represent enrichment to a specific biological function. For instance, Darkred module represent a high percentage of Translation Associated genes related to the ribosome and tRNA, the darkorange module related to the genes involved in beta oxidation, photosynthesis, auxin and DNA repair. Both of these modules are associated with germination and have genes that are expected to be expressed during germination (Kucera, 2005). These genes cannot be directly affect the thermodormancy because of the differences in expression can only be a response to stress and not a part of the regulator systems that trigger this phenomenon, but they should further be analyzed to understand the metabolic pathways involved (Holdsworth, Bentsink*, et al.*, 2008).

In high temperature related modules we have the gold module, which contains a higher number of ethylene, DNA repair, cell wall modification and cytoskeleton. Modules with an elevated percentage of ABA genes are Blue4 and Darkorange4, the first one it also have a bigger number of auxin, cell cycle and protein degradation genes. For the darkorange4 module is associated with dormancy, seed storage protein and cell wall modification genes. The expression of these genes is triggered by high temperature and for this is more possible to be

closely related with the establishment and maintain of thermodormancy (Argyris, 2008b, Toh*, et al.*, 2008, Vlad, et al., 2009).

Genes involved in maintain and release of dormancy

Through the analysis of the microarray data, a set of genes were found to have a high possibility to be related with thermodormancy. These genes were selected using a filtering process that allowed to eliminate the noise in the data. The last group of 63 genes were is related to the three major hormones involved in germination and release of dormancy, ABA, GA and ethylene; and with specific transcription factors associated with this process. According to the annotational data the ABA genes involved in response (CLS_S3_Contig7163, QGI11J05.yg.ab1), signaling (CLS_S3_Contig7421) and synthesis (GB_84579411, CLS_S3_Contig4683, CLS_S3_Contig11334, QGI9I18.yg.ab1). Most of these genes are up-regulated in the Sal 35 treatments, showing and increase in ABA activity during dormancy (Seo, 2006, Sheard, et al., 2009, Toh, et al., 2008). In the case of the gibberellins associated genes they are down-regulated in the same treatment. This finding was in agreement with previous studies that demonstrate a cut in GA activity in dormant seeds (Piskurewicz*, et al.*, 2008, Seo, 2006, Toh, et al., 2008); the same genes are related to signaling (CLSM11108.b1_H18.ab1, CLSX3670.b1_K06.ab1, QGD6I10.yg.ab1, CLLY7887.b1_M04.ab1) and response (CLR_S1_Contig3370, CLS_S3_Contig3839) to GA. Similar to GA, ethylene activity is being suppressed for dormancy in Sal 35 (Saini*, et al.*, 1986), with a general down-regulation of the genes found related with this hormone; we identified eight genes related to ethylene signaling and two with the biosynthetic pathway of this hormone.

In Sal 35 transcription factors could be divided in two groups based on their up- or down- regulation. Up-regulated transcription factors are Heat Shock transcription Factors (HSF), Ethylene Insensitive Like transcription factors (EIL) and HomeoBox transcription factors, all of them known to be associated with dormancy. Meanwhile down-regulated transcription factors are basically the auxin related groups (IAA and ARF). This genes should be analyzed further because the role of auxins in dormancy has not be understood yet. Although they do not seem to be a key factor in this process.

These results present the general overview of how the thermodormancy affect the seed metabolism. The hormone balance is influenced by the high temperature triggering a signaling chain reaction that lead to an increase in ABA and a decrease in GA and ethylene activities. Nonetheless, they were already known to be the major controllers in dormancy (Feurtado, 2007, Kucera, 2005). This study showed us the possible genes related to the thermodormancy in lettuce. Further analysis is needed in order to confirm the association of these genes with the high temperature germination trait found in the *Lactuca serriola* and a more important task remains as how to transfer these phenotypes to the cultivated lettuce.

RECOMMENDATIONS

Through the analysis of the microarray data were able to determine that WGCNA is able to cluster the genes with similar biological function. Also with the combination of several annotational methods 63 genes were selected like candidate genes relate with thermodormancy and they should be analyzed in other investigation to discover their function in thermodormancy.

In order to have a better experimental design another set of samples should be made it in a second time point (48h). This treatments will gave a full germinate Serriola seed at 35ºC that will be in the same developmental stage that Salinas seed at 20ºC. With this data the genes that change later in germination will be able to be detected giving a bigger picture of the thermodormancy in lettuce.

REFERENCES CITED

Affymetrix. 2004. Genechip Expression Analysis: Data Analysis Fundamentals.

Allen, P. S., Benech-Arnold, R.L., Batlla, D. and Bradford, K.J. 2007. Modeling of Seed Dormancy. Blackwell Publishing.

Argyris, J., Dahal, P., Hayashi, E., Still, D. W. and Bradford, K.J. 2008a. Genetic Variation for Lettuce Seed Thermoinhibition Is Associated with Temperature-Sensitive Expression of Abscisic Acid, Gibberellin, and Ethylene Biosynthesis, Metabolism, and Response Genes. Plant Physiol. 148 (2), 926-947.

Argyris, J., Dahal, P., Truco, M.J., Ochoa, O., Still, D.W., Michelmore, R.W. and Bradford, K.J. 2008b. Genetic Analysis of Lettuce Seed Thermoinhibition.

Argyris, J., Truco, M.J., Ochoa, O., Knapp, S., Still, D., Lenssen, G., Schut, J., Michelmore, R. and Bradford K.J. 2005. Quantitative Trait Loci Associated with Seed and Seedling Traits in Lactuca. TAG Theoretical and Applied Genetics 111 (7), 1365-1376.

Baskin, J. M., and Baskin, C.C. 2004. A Classification System for Seed Dormancy. Seed Science Research 14 (01), 1-16.

Benech-Arnold, R. L., Sánchez, R. A., Forcella, F., Kruk, B. C. and Ghersa, C. M. 2000. Environmental Control of Dormancy in Weed Seed Banks in Soil. Field Crops Research 67 (2), 105-122.

Benjamini, Y. H., Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal Of The Royal Statistical Society Series B 57 (1), 289-300.

Benjamini, Y. H., Y. 2000. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. Journal of Educational and Behavioral Statistics 25 (1), 60-83.

Bolstad, B. M. 2010. Affyplm: Methods for Fitting Probe-Level Models. Bioconductor.

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. 2003. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. Bioinformatics 19 (2), 185-193.

Bove, J., Lucas, P., Godin, B., Ogé, L., Jullien, M. and Grappin, P. 2005. Gene Expression Analysis by Cdna-Aflp Highlights a Set of New Signaling Networks and Translational Control During Seed

Dormancy Breaking in Nicotiana Plumbaginifolia. Plant Molecular Biology 57 (4), 593-612.

Bunney, W. E., Bunney, B.G., Vawter, M.P., Tomita, H., Li, J., Evans, S.J., Choudary, P.V., Myers, R.M., Jones, E.G., Watson, S.J. and Akil, H. 2003. Microarray Technology: A Review of New Strategies to Discover Candidate Vulnerability Genes in Psychiatric Disorders. Am J Psychiatry 160 (4), 657-666.

Cadman, C. S. C., Toorop, P.E., Hilhorst, H.W.M. and Finch-Savage, W.E. 2006. Gene Expression Profiles of Arabidopsis Cvi Seeds During Dormancy Cycling Indicate a Common Underlying Dormancy Control Mechanism. The Plant Journal 46 (5), 805-822.

Caldwell, D., van Leeuwen, H., Stoffel, K., Chen, F., Kozik, A., Truco, M.J., Michelmore, R. and Van Deynze, A.E. . 2008. Development and Exploitation of a Lettuce (Lactuca Sativa) 6.6 Million Feature Affymetrix Genechip for Massively Parallel Genotyping and Gene Expression Analysis.

FAO, F. a. A. O. o. t. U. N. 2008. Faostat (On Line). Visited: may 21, 2009. Available in <http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor>.

Feurtado, J. A., and Kermode, A.R. 2007. A Merging of Paths: Abscisic Acid and Hormonal Cross-Talk in the Control of Seed Dormancy Maintenance and Alleviation. Blackwell Publishing.

Finch-Savage, W. E., Cadman, C.S.C., Toorop, P.E., Lynn, J.R. and Hilhorst, H.W.M. 2007. Seed Dormancy Release in Arabidopsis Cvi by Dry after-Ripening, Low Temperature, Nitrate and Light Shows Common Quantitative Patterns of Gene Expression Directed by Environmentally Specific Sensing. The Plant Journal 51 (1), 60-78.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. 2004. Bioconductor: Open Software Development for Computational Biology and Bioinformatics. Genome Biology 5 (10), R80.

Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A. and Dudoit, S. 2005. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer Science+Bussines Media, China.

Gentleman R, C. V., Huber W, Irizarry RA and Hahne F. 2010. Genefilter: Methods for Filtering Genes from Microarray Experiments. Bioconductor.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J. and Horvath, S. 2006. Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. PLoS Genet 2 (8), e130.

Heber, S. and Sick, B. 2006. Quality Assessment of Affymetrix Genechip Data. OMICS: A Journal of Integrative Biology 10 (3), 358-368.

Heggie, L., and Halliday, K.J. 2005. The Highs and Lows of Plant Life: Temperature and Light Interactions in Development. Int J Dev Biol 49 675-678.

Holdsworth, M. J., Bentsink, L. and Soppe, W. J. J. 2008. Molecular Networks Regulating Arabidopsis Seed Maturation, after-Ripening, Dormancy and Germination. New Phytologist 179 (1), 33-54.

Holdsworth, M. J., Finch-Savage, W. E., Grappin, P. and Job, D. 2008. Post-Genomics Dissection of Seed Dormancy and Germination. Trends in Plant Science 13 (1), 7-13.

HOMMEL, G. 1988. A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. Biometrika 75 (2), 383-386.

Hu, Z., Mellor, J., Wu, J. and DeLisi, C. 2004. Visant: An Online Visualization and Analysis Tool for Biological Interaction Data. BMC Bioinformatics 5 (1), 17.

Hu, Z., Snitkin, E. S. and DeLisi, C. 2008. Visant: An Integrative Framework for Networks in Systems Biology. Brief Bioinform bbn020.

Irizarry R.A, G. L., Bolstad B.M., and Miller C. 2010. Affy: Methods for Affymetrix Oligonucleotide Arrays. Bioconductor.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. 2003. Summaries of Affymetrix Genechip Probe Level Data. Nucl. Acids Res. 31 (4), e15-.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. 2003. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostat 4 (2), 249-264.

Kozarewa, I., Cantliffe, D.J., Nagata, R.T. and Stoffella, P.J. . 2006. High Maturation Temperature of Lettuce Seeds During Development Increased Ethylene Production and Germination at Elevated Temperatures. J. Amer. Soc. Hort. Sci 131 (4), 564-570

Kucera, B., Cohn, M.A. and Leubner-Metzger, G. 2005. Plant Hormone Interactions During Seed Dormancy Release and Germination. Seed Science Research 15 (4), 281-307.

Langfelder, P., and Horvath, S. 2008. Wgcna: An R Package for Weighted Correlation Network Analysis. BMC Bioinformatics 9 (1), 559.

Langfelder, P. a. H., S. 2009. Wgcna: An R Package for Weifhted Correlation Network Analysis (On Line). Visited: 22/01/2010. Available in <http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/>.

Lefebvre, V., North, H., Frey, A., Sotta, B., Seo, M., Okamoto, M., Nambara, E. and Marion-Poll, A. 2006. Functional Analysis of Arabidopsis Nced6 and Nced9 Genes Indicates That Aba Synthesized in the Endosperm Is Involved in the Induction of Seed Dormancy. The Plant Journal 45 (3), 309-319.

Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. 2009. High Density Synthetic Oligonucleotide Arrays. Nat Genet 21 20-24.

Lockhart, D. J., and Winzeler, E.A. 2000. Genomics, Gene Expression and DNA Arrays. Nature 405 (6788), 827-836.

MacLennan, N. K., Dong, J., Aten, J. E., Horvath, S., Rahib, L., Ornelas, L., Dipple, K. M. and McCabe, E. R. B. 2009. Weighted Gene Co-Expression Network Analysis Identifies Biomarkers in Glycerol Kinase Deficient Mice. Molecular Genetics and Metabolism 98 (1-2), 203-214.

Nakashima, K., Fujita, Y., Kanamori, N., Katagiri, T., Umezawa, T., Kidokoro, S., Maruyama, K., Yoshida, T., Ishiyama, K., Kobayashi, M., Shinozaki, K. and Yamaguchi-Shinozaki, K. 2009. Three Arabidopsis Snrk2 Protein Kinases, Srk2d/Snrk2.2, Srk2e/Snrk2.6/Ost1 and Srk2i/Snrk2.3, Involved in Aba Signaling Are Essential for the Control of Seed Development and Dormancy. Plant Cell Physiol. 50 (7), 1345-1363.

Park, S.-Y., Fung, P., Nishimura, N., Jensen, D. R., Fujii, H., Zhao, Y., Lumba, S., Santiago, J., Rodrigues, A., Chow, T.-f. F., Alfred, S. E., Bonetta, D., Finkelstein, R., Provart, N. J., Desveaux, D.,

Rodriguez, P. L., McCourt, P., Zhu, J.-K., Schroeder, J. I., Volkman, B. F. and Cutler, S. R. 2009. Abscisic Acid Inhibits Type 2c Protein Phosphatases Via the Pyr/Pyl Family of Start Proteins. Science 324 (5930), 1068-1071.

Piskurewicz, U., Jikumaru, Y., Kinoshita, N., Nambara, E., Kamiya, Y. and Lopez-Molina, L. 2008. The Gibberellic Acid Signaling Repressor Rgl2 Inhibits Arabidopsis Seed Germination by Stimulating Abscisic Acid Synthesis and Abi5 Activity. Plant Cell 20 (10), 2729-2745.

Riefler, M., Novak, O., Strnad, M. and Schmulling, T. 2006. Arabidopsis Cytokinin Receptor Mutants Reveal Functions in Shoot Growth, Leaf Senescence, Seed Size, Germination, Root Development, and Cytokinin Metabolism. Plant Cell 18 (1), 40-54.

Saini, H. S., Consolacion, E. D., Bassi, P. K. and Spencer, M. S. 1986. Requirement for Ethylene Synthesis and Action During Relief of Thermoinhibition of Lettuce Seed Germination by Combinations of Gibberellic Acid, Kinetin, and Carbon Dioxide. Plant Physiol. 81 (4), 950-953.

Schulze, A. and Downward, J. 2001. Navigating Gene Expression Using Microarrays a Technology Review. Nat Cell Biol 3 (8), E190-E195.

Schwember, A. R., and Bradford, K.J. 2005. Drying Rates Following Priming Affect Temperature Sensitivity of Germination and Longevity of Lettuce Seeds. Hortscience 40 (3), 778-781.

Seed Biotechnology Center, U. 2010. Seed Biotechnology Center: Trait Integration (On Line). Visited: 01/22/2010. Available in <http://sbc.ucdavis.edu/research/trait_integration.htm>.

Seo, M., Hanada, A., Kuwahara, A., Endo, A., Okamoto, M., Yamauchi, Y., North, H., Marion-Poll, A., Sun, T., Koshiba, T., Kamiya, Y., Yamaguchi, S. and Nambara, E. 2006. Regulation of Hormone Metabolism in Arabidopsis Seeds: Phytochrome Regulation of Abscisic Acid Metabolism and Abscisic Acid Regulation of Gibberellin Metabolism. The Plant Journal 48 (3), 354-366.

Sheard, L. B. and Zheng, N. 2009. Plant Biology: Signal Advance for Abscisic Acid. Nature 462 (7273), 575-576.

SIMES, R. J. 1986. An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika 73 (3), 751-754.

Tamura, N., Yoshida, T., Tanaka, A., Sasaki, R., Bando, A., Toh, S., Lepiniec, L. and Kawakami, N. 2006. Isolation and

Characterization of High Temperature-Resistant Germination Mutants of Arabidopsis Thaliana. Plant Cell Physiol. 47 (8), 1081-1094.

Tan, B., Joseph, L.M., Deng, W., Liu, L., Li, Q., Cline, K. and McCarty, D.R. 2003. Molecular Characterization of the <I>Arabidopsis</I> 9<I>-Cis</I> Epoxycarotenoid Dioxygenase Gene Family. The Plant Journal 35 (1), 44-56.

Tang, H. and Therneau, T. M. 2009. Statistical Metrics for Quality Assessment of High-Density Tiling Array Data. Biometrics 9999 (9999),

Tarquis, A. M., and Bradford, K.. 1992. Prehydration and Priming Treatments That Advance Germination Also Increase the Rate of Deterioration of Lettuce Seeds. J. Exp. Bot. 43 (3), 307-317.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y. and Stitt, M. 2004. Mapman: A User-Driven Tool to Display Genomics Data Sets onto Diagrams of Metabolic Pathways and Other Biological Processes. The Plant Journal 37 (6), 914-939.

Toh, S., Imamura, A., Watanabe, A., Nakabayashi, K., Okamoto, M., Jikumaru, Y., Hanada, A., Aso, Y., Ishiyama, K., Tamura, N., Iuchi, S., Kobayashi, M., Yamaguchi, S., Kamiya, Y., Nambara, E. and Kawakami, N. 2008. High Temperature-Induced Abscisic Acid Biosynthesis and Its Role in the Inhibition of Gibberellin Action in Arabidopsis Seeds. Plant Physiol. 146 (3), 1368-1385.

Toh, S., Yoshida, S., Imamura, A., Nakabayashi, K., Okamoto, M., Hanada, A., Yamaguchi, S., Kamiya, Y., Nambara, E. and Kawakami, N. 2004. Expression of Aba and Ga Related Genes in Thermoinhibited Arabidopsis Seeds.

van Leeuwen, H., Stoffel, K., Mathrakott, S., Chen, F., Tan, X., Yao, J., Wong, G., Cui, X., Kozik, A., Truco, M.J., Michelmore, R. and Van Deynze, A.E. . 2008. Analysis of Single Feature Polymorphism Markers with a Lettuce Affymetrix Genechip for Massively Parallel Marker Discovery, Genotyping and Mapping.

van Leeuwen, H. a. A., H. 2009. Lettuce Sfp Chip Project (On Line). Visited: April 22, 2009. Available in <http://chiplett.ucdavis.edu/>.

Vlad, F., Rubio, S., Rodrigues, A., Sirichandra, C., Belin, C., Robert, N., Leung, J., Rodriguez, P. L., Lauriere, C. and Merlot, S. 2009. Protein Phosphatases 2c Regulate the Activation of the Snf1-Related Kinase Ost1 by Abscisic Acid in Arabidopsis. Plant Cell tpc.109.069179.

WU, H. 2010. Maanova: Tools for Analyzing Micro Array Experiments. Bioconductor.

Zhang, B. a. H., Steve 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4 (1),