Characterization of Gene Regulatory Networks in Shoot apex of *Arabidopsis thaliana*

Jayesh Kumar Sundaram

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali

April 2018

Certificate of Examination

This is to certify that the dissertation titled "**Characterization of Gene Regulatory Networks in Shoot Apex of *Arabidopsis thaliana***" submitted by Jayesh Kumar Sundaram (Reg.No. MS13054) for the partial fulfillment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Kuljeet Sandhu

Dr. Shashi Bhushan Pandit

Dr. Ram Yadav
(Supervisor)
Dated: 20th April, 2018

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Ram Yadav at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgment of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Jayesh Kumar Sundaram
(Candidate)
Dated: 20th April, 2018

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Ram Yadav
(Supervisor)

Acknowledgment

I would like to thank Indian Institute of Science Education and Research, Mohali for giving me a golden opportunity to be a part of the exciting BS-MS dual degree course. It really has boosted my confidence and has given me all the skills required to make a successfully career in the field of research. I really want to thank all the professors especially Dr. Samrat Ghosh, Dr. Mahak Sharma, Dr. Ramesh Ramachandran, Dr. K S Viswanathan, Dr. N G Prasad for inspiring me and also for being an excellent teacher. Special thanks to Dr. Lolitika Mandal and Dr. Sudip Mandal for believing my potential early on and also for training me actively. I also want to thank all the biology department faculties for their support throughout my stay at IISER, Mohali. Your support means a lot to me.

I want to thank my master thesis guide Dr. Ram Yadav for believing me and also for giving all the freedom to work in my own independent way. He was really supportive and motivational throughout my stay in the lab. I will take away with me all the experience and suggestions given by him for making a meaningful career in my life. I should agree that he has excellent management skills and can arrange anything the lab requires. I would like to thank Dr. Kuljeet Sandhu, Dr. Shashi Bhushan Pandit and Dr. Somdatta Sinha for the fruitful discussions. Special thanks to KEN, PhD student in Dr. Kuljeet's lab for helping me in crucial times. He was always a solution to my unsolvable problems. I also want to thank all my lab members with whom I have spent a lot of time in the last one year. I had a pleasant stay and the lab is organized better than many others labs. I am really happy that I was a part of it. I would like to thank Shivani for sharing her Protein-DNA interaction data and also for being very supportive. I would like to thank Prince for helping me a lot during early part of my stay in the lab. Thanks for helping me with the molecular biology techniques and also for sharing the Protein-Protein interaction network data. I would like to thank Sangram who helped me a lot with my thesis work. I learnt a lot of things and it was really helpful. He was a huge support for my thesis work and a person with an excellent character. Then I would like to thank Sonal from whom I learnt a lot of things. She is very professional, organized and I would like to have some of her qualities by the time I start my PhD. I believe she has an excellent working style which I would like to copy. Very good

at recording observations. The most influential person in the lab is Dr. Monika Mahajan, who is a mother of a kid. She is really balancing her social life and research life well which is something I always struggle to do. Thanks for sharing the cell type specific RNA-sequencing data. I would like to thank Harish Prateek for being so cool and also for considering me a worthy candidate to discuss his research problems. I would like to thank Anshul for being so friendly and for supporting me during my bad times. Your support means a lot to me. Thanks to Shalini for her support and guidance for all the molecular biology work I did. She was an excellent stress buster for me. I would like to thank Rimpy for her support and suggestion throughout my stay in the lab and also for discussing all the common problems we had together. The happy moments with all of you will be always remembered.

At the end I would to thank all my friends (Asish Moharana, Divita Gupta, Madhuvanthi G Athani, Saloni Rose, Sri Harsha Dantuluri, Akhil Bhardwaj, Neeraj Mann, V Praveen Kumar, Poonam Aggarwal, Harleen Kaur, Sathish Tiwari, Swathi Jayaram), family members (Appa, Amma and Pranesh) and all the supporting staffs (Baidnath Mandal and others) at IISER, Mohali for making this thesis possible. Special thanks to my Amma for helping me at bad times. Last but not the least I want to thank the excellent plant model organism *Arabidopsis thaliana*.

List of figures

List of tables

Notation

GRN: Gene Regulatory network.

eY1H: Enhanced yeast one hybrid.

Y2H: Yeast two hybrid.

PPI: Protein-Protein interaction network.

PDI: Protein-DNA interaction network.

TF: transcription factor.

SAM: Shoot Apical Meristem.

TFBS: Transcription factor binding site.

FIMO: Find Individual Motif Occurrences.

FACS: Fluorescence Activated Cell Sorting.

PSSM: Position specific scoring matrix.

AS: Alternative splicing.

Contents

Abstract

Cell types specific gene expression plays an important role in cell and tissue specialization in various developmental process. Specifically, TFs contribute a lot for the differential gene expression profiles observed among the cell types. The gene regulatory networks involved in the specification and maintenance of tissue layers in *Arabidopsis thaliana* SAM are not understood fully. Cell type specific gene expression profiling techniques have identified the genes that are differentially expressed or enriched including the TFs that regulate target genes. In this study, we have constructed PPI and PDI networks using yeast-two-hybrid (Y2H) and enhanced yeast-one-hybrid assays (eY1H), respectively, to understand how cell layer specific gene regulation is achieved in *Arabidopsis* SAM. Some of the interactions were validated by other experiments. The constructed PPI and PDI networks were characterized in the study by in-silico approaches. As the interactions were tested in the yeast model, the significance of the constructed network in-planta was analyzed using co-expression and co-occurrence properties of the interacting pairs. The nature of the transcription regulation in the PDI network were also predicted by co-expression of the TFs with its target genes and protein-protein interactions of the TFs with cofactors. The expression patterns of the interacting gene pairs were analyzed and the over-represented interaction types in the networks were identified. The eY1H assay was compared with other experimental methods available in the field, and we found that interactions captured by eY1H assay are both novel and reproducible in nature. The plant gene regulatory network is very complex and has many genetic redundancies. In order to identify redundant pairs, various association networks were created among TFs based on PPI similarity, target gene similarity, upstream regulator similarity and transcription factor binding site (TFBS) co-occurrence. SAM cell type specific PPI and PDI networks were constructed by in-silico approaches to understand

the formation of different cell layers and stem cells maintenance in the shoot apex. Finally, we also identified cell type specific annotated transcripts / isoforms in the SAM of *Arabidopsis thaliana* by RNA-sequencing.

Chapter – 1

Introduction

1.1 The model organism: *Arabidopsis thaliana*

Arabidopsis thaliana, a small flowering plant belongs to the Brassicaceae family. It is an excellent plant model system for several reasons. It has relatively short life cycle, generates huge number of seeds every generation, has relatively small genome size, and the genome is fully sequenced. In order to understand the role of different genes, genome wide insertional mutagenesis was performed in *Arabidopsis thaliana*. These mutants carry transfer DNA (T-DNA) at random locations within the genome. Several mutants with interesting phenotypes have helped us in understanding various biological process [1]. As opposed to animals, plants continuously form organs during post-embryonic development. Meristematic layers in the plant gives rise to various organs in the plant and keeps the plant growing. They are two types of meristem: Shoot Apical meristem (SAM), which gives rise to the shoot part of the plants like leaves and flowers and Root Apical Meristem (RAM), which gives rise to the root system of the plant [2, 3]. During embryonic development, both the meristems are placed in the opposite poles of the developing embryo. As the meristems are present at the tip, they are called apical meristems [2, 4]. Stem cells within these meristematic tissues grows indefinitely throughout the life of the plant. Stem cell niche maintains the stem cell population, integrates internal and external clues to regulate organogenesis. Niche cells secretes WUSCHEL (WUS) which is involved in the maintenance and specification of the stem cells in the shoot apex.

1.2 Organization of SAM in *Arabidopsis thaliana*

The SAM of *Arabidopsis thaliana* can be divided into different cell layers and zones. The first two layers are called epidermal (L1) and sub-epidermal (L2) cell layers, which undergoes mainly anticlinal pattern of cell division (Figure 1.1, F1) [5]. The inner layer (L3) undergoes both periclinal and anticlinal pattern of cell divisions. Stem cells are located at the tip of the shoot in central zone (CZ). They give rise to progenitors, which eventually differentiates to become the part of the organ primordia. Organizing center (OC) or the stem cell niche is located just below the CZ. In the organizing center (OC), *WUS* expresses the WUS protein which after being synthesized moves to CZ cell where it activates *CLV3* in the stem cells. The mature CLV3 ligand binds to CLV1 and CLV2-CRN receptor complex to repress the expression of *WUS* in OC. Hence, *CLV-WUS* feedback loop maintains loop maintains constant number of the stem cells number and size of SAM (Figure 1.1, F2) [5, 6, 7].

1.3 Gene regulatory networks in *Arabidopsis thaliana*

The nodes in the networks are genes and the edges through which they are connected called interactions between the genes. The interactions can be either Protein-Protein or Protein-DNA interactions. To understand any biological process, it is very essential to identify the gene regulatory network involved in the process. The common methods in the field to identify Protein-Protein interactions are yeast two-hybrid system (Y2H), affinity purification mass spectrometry (AP-MS), bimolecular fluorescence complementation (BiFC), and *in silico* prediction. Y2H is broadly used in the field to identify Protein-Protein interaction. It is easy and can be done relatively fast. But it has a high false positive rate. Y2H studies the interaction among the proteins in-vitro, while the other two experimental methods study the interactions in-

vivo. If two proteins have the proper interaction domain to interact but never co-express or co-localize together in plant then the Y2H assay will give false positive interaction. The common methods in the field to identify Protein-DNA interactions are DNaseI hypersensitive assay, ATAC-seq, Chromatin immuno-precipitation (ChIP) coupled with sequencing (ChIP-Seq), DAP-seq, Protein binding microarray, yeast-one-hybrid assays, etc [8]. The above-mentioned methods do not establish the type of regulation i.e. either positive or negative between the interacting nodes. Therefore, alternative experimental approaches are required to find the nature of regulation. Y1H assay has been used in the field to understand various biological process at systems level successfully [9].

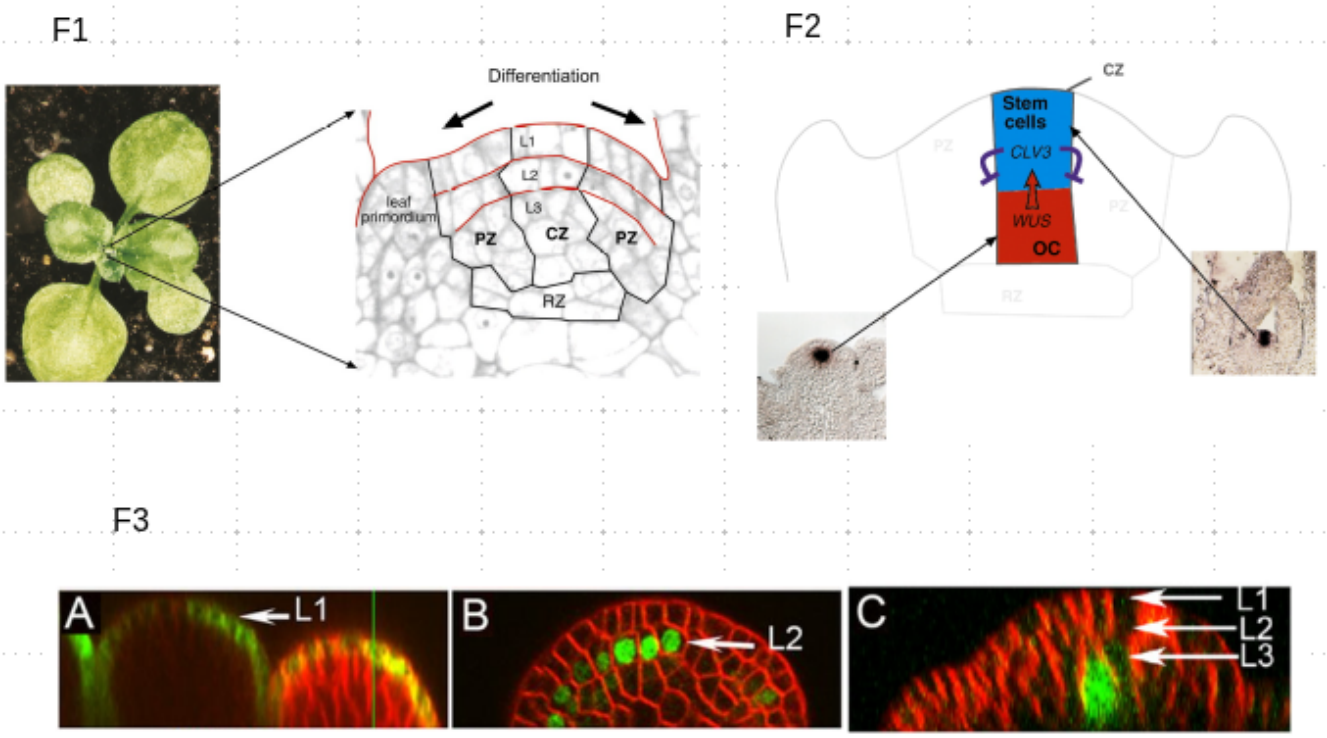

Figure 1.1: *Arabidopsis* SAM; F1: Organization of SAM in *Arabidopsis thaliana.* F2: Model describing the CLV-WUS feedback loop involved in stem cell maintenance and proliferation. F3: A) Side view of *pHMG::H2B-YFP* (yellow) and FM4–64 (red) in *ap1-1; cal1-1* shoot apex. *pHMG::H2B-YFP* (yellow) marks the L1 layer. B) Side view of *pHDG4:H2B-YFP* (yellow) and FM4-64 (red) in *ap1-1; cal1-1* shoot apex. *pHDG4:H2B-YFP* (yellow) marks the L2 layer of the

SAM. C) Side-view of *pWUS::mGFP5-ER* expression (green) and FM4–64 (red) in the wild type SAM. *pWUS::mGFP5-ER* (green) marks the L3 layer (arrows) and cell layers located beneath.

1.4 Objective

SAM of *Arabidopsis thaliana* is organized into three cell layers, namely L1, L2 and L3. The narrowly expressed genes play a very crucial role in the specification of these cell layers. Among the differentially expressed genes, TFs haves the ability to regulate gene expression. The objective of this study is to understand the regulation of TFs, which are specifically expressed in different cell layers. Cell type specific gene expression studies were conducted to know the identity of the genes that are expressed in these layers [10, 11]. Differentially expressed or enriched TFs in these layers were identified. Protein-protein interaction network among these TFs was constructed using Y2H assay. Protein-DNA network was constructed to identify the regulators regulating the transcription of differentially expressed or enriched TFs in these layers by using eY1H assay. A library of around 350 TFs, which are expressed in SAM of *Arabidopsis thaliana* was used to check their interaction with the upstream promoter sequence of the differentially expressed or enriched TFs. The main objective of the study is to characterize the Protein-DNA and Protein-Protein interaction network constructed based on Y1H and Y2H assay, respectively. As the interactions were checked in the yeast model system, it is very important to check whether it is relevant *in-planta*. It is also important to understand the nature of regulation i.e. either positive or negative. It is a tedious task to check or confirm all the interactions by experiments. The objective of the study is to understand the significance of the constructed network *in-planta* and also predict nature of the transcription regulation by in-silico approaches. We also want to characterize various network properties. Another objective is to understand the

expression pattern of interacting pairs in the network. We also want to compare the output of the eY1H assay with other experimental methods available. We want to know how similar these techniques are when compared to eY1H assay. Plant gene regulatory networks are highly complex and redundant. We want to construct association networks among TFs based on various properties, which will help us to identify the redundant gene pairs. We also wanted to construct shoot SAM cell type specific Protein-Protein interaction network and Protein-DNA interaction network by in-silico approaches to understand the formation of different cell layers and stem cells maintenance in the shoot apex. Finally, we wanted to identify cell type specific annotated transcripts / isoforms in the SAM of *Arabidopsis thaliana* by RNA-sequencing.

1.5 References

1) José M. Alonso et al., Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana;* Science 01 Aug2003:Vol.301, Issue5633, pp.653-657; DOI: 10.1126/science.1086391.

2) Dolan L., Cellular organisation of the Arabidopsis thaliana root; Development. 1993 Sep;119(1):71-84.

3) H. M. Ottoline Leyser, I. J. Furner, Characterisation of three shoot apical meristem mutants of Arabidopsis thaliana; Development 1992 116: 397-403.

4) M. K. Barton, R. S. Poethig., Formation of the shoot apical meristem in Arabidopsis thaliana: an analysis of development in the wild type and in the shoot meristemless mutant; Development 1993 119: 823-831.

5) Rita Groß-Hardt, Thomas Laux; Stem cell regulation in the shoot meristem; Journal of Cell Science 2003 116: 1659-1666; doi: 10.1242/jcs.00406

6) Ralf Müller et al., Dynamic and Compensatory Responses of *Arabidopsis* Shoot and Floral Meristems to *CLV3* Signaling; Plant Cell. 2006 May; 18(5): 1188–1198.

7) Reddy GV, Meyerowitz EM, Stem-cell homeostasis and growth dynamics can be uncoupled in the Arabidopsis shoot apex; Science. 2005 Oct 28;310(5748):663-7. Epub 2005 Oct 6.

8) Yixiang Zhang et al., Plant Protein-Protein Interaction Network and Interactome; Curr Genomics. 2010 Mar; 11(1): 40–46. doi: 10.2174/138920210790218016.

9) Brady SM et al., A stele-enriched gene regulatory network in the Arabidopsis root; Mol Syst Biol. 2011 Jan 18;7:459. doi: 10.1038/msb.2010.114.

10) Ram Kishor Yadav, Thomas Girke, Sumana Pasala, Mingtang Xie and G. Venugopala Reddy; Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche; PNAS March 24, 2009. 106 (12) 4941-4946

11) Ram Kishor Yadav, Montreh Tavakkoli, Mingtang Xie, Thomas Girke, G. Venugopala Reddy; A high-resolution gene expression map of the *Arabidopsis* shoot meristem stem cell niche; Development 2014 141: 2735-2744; doi: 10.1242/dev.106104

Chapter - 2

Characterization of Gene Regulatory Networks (GRN) in the SAM of *Arabidopsis thaliana*

2.1 Introduction

The shoot apex of *Arabidopsis thaliana* is divided into three layers, namely L1, L2 and L3 (Figure 1.1, F1) [1]. The first step in understanding the formation and specification of the three different layers is to identify the genes expressed in these layers. The promoter reporter constructs were made for three cell layers, and they were transformed into *ap1-1; cal1-1* genotype plant for cell sorting purposes. *pHMG::H2B-YFP*, *pHDG4:H2B-YFP* and *pWUS*::*mGFP5-ER* constructs were made to specifically mark L1, L2 and L3 layers (Figure 1.1, F3), respectively. By employing Fluorescence Activated Cell Sorting (FACS) method, cells were isolated from different layers, in separate experiments. Total RNA was isolated from these cell types and hybridized to ATH1 gene chip [2,3]. To understand the specification of these layers, it is important to know the differently expressed or enriched genes and their regulation in these three layers.

From the microarray data analysis, genes that are differentially enriched were identified. To understand the regulation of differentially enriched genes, eY1H and Y2H assays were carried out to construct the protein-DNA and protein-protein interaction networks, respectively. For eY1H, 3kb upstream promoter DNA sequence above the translation start site of differentially enriched TF genes was included in the bait list. Few interesting and broadly expressed TFs were also included in the bait list. The amplified promoter DNA fragments were cloned into Gateway compatible pDONOR vector, and later on sub-cloned into pMW2 destination vector. In total, 49

promoter fragments were rescued into pDONOR to make the baits. The prey TF library containing both narrowly and broadly expressed SAM TFs were used [4]. For Y2H, almost all the differentially enriched and a few broadly expressed interesting TFs were used as both baits and preys.

Complex protein-DNA and protein-protein interaction networks were constructed. On an average, one TF may regulate multiple TF genes, and vice versa, where multiple TF proteins bind one TF promoter. If a TF and its target gene in PDI network or the interacting pairs in the PPI network co-expresses together then their chances of interaction increase significantly. The TFs and their target genes in the PDI network and the interacting pairs in the PPI network were shown to co-express significantly. To compare eY1H assay with other experimental methods in the field, in-silico protein-DNA interaction network was constructed between the baits and preys present in the eY1H protein-DNA interaction network. For in-silico network, binding sites of the TF in the target gene promoter sequences were predicted using TF motif preference (known from other experiments) and 'Finding Individual Motif Occurrences' (FIMO) tool [5, 6, 7, 8, 9, 10]. Both eY1H and FIMO based networks had significant overlapping interactions. The interacting pairs in the PPI network are shown not to co-occur together in different species of the plant kingdom. The genes are really conserved in Brassicaceae family. The conservation of genes in the Monocots was poor [13]. Based on the various cell type specific expression data, the expression pattern of the interacting pairs of PPI and PDI networks were analyzed. In PDI network, the interaction pairs having higher overlapping patterns are shown to have higher positive Pearson correlation coefficient score indicating positive transcriptional regulation whereas the lesser overlapping patterns are shown to have higher negative Pearson correlation

coefficient score indicating negative transcriptional regulation. The nature of regulation of some the TFs in the PDI network were predicted based on its expression pattern with its target gene and its interaction with the co-activators or co-repressors (based on the known interactions) [11]. Plant regulatory networks have a lot of genetic redundancy. In order to find protein pairs which might play a redundant role, various association networks were made based on various properties [6]. Finally, in-silico protein-DNA and protein-protein networks of *Arabidopsis thaliana* SAM containing only TFs and co-factors were constructed based on the known and predicted interactions from various databases [12].

2.2 Results

2.2.1 Selection of baits and preys in eY1H and Y2H

The bait TFs for eY1H are chosen in such a way that they are enriched in the epidermal and sub-epidermal cell layers. Only a few interesting TF genes that plays important role in shoot and flower development but restricted in their expression pattern were added in this list as bait; e.g. *APETALA2, APETALA3, ATHB-15, MONOPOLE, PLETHORA7, SEPALLATA3, WRKY21,* and *WUSCHEL*. But for preys, genes which are enriched in a particular cell layer and also the genes which are broadly expressed in all layers were selected. The TFs for Y2H assay are selected in such a way that they are enriched only in a particular layer. But a few interesting and important genes which are broadly expressed were also considered.

Cell type specific microarray data of HMG (L1), HDG4 (L2) and WUS (L3) cells were taken [2, 3]. The data was normalized using MAS5 algorithm using 'affy' package in R. To find out genes that are expressed in a particular cell type, present call information of the non-parametric Wilcoxon signed rank test was computed. A gene is considered to be expressed in a particular

cell type if it has a 'present' call in all the replicates. The expression heatmap of both baits and preys selected for Y1H and Y2H assays were made (Figure 2.1, 2.2). A gene is said to be enriched in a particular cell type A in a group of A, B and C population if the gene is differentially expressed in A when compared to B and differentially expressed in A when compared to C. A gene is said to enriched in both A and B in a group of A, B and C population if the gene is differentially expressed in A when compared to C and differentially expressed in B when compared to C but not differentially expressed when A and B are compared. Thus, genes that are expressed in only one cell type are considered to be enriched in only one cell type. Genes which are expressed in two cell types are checked whether they are enriched in one cell type. If they are not enriched in one cell type then they are considered to be enriched in two cell types.

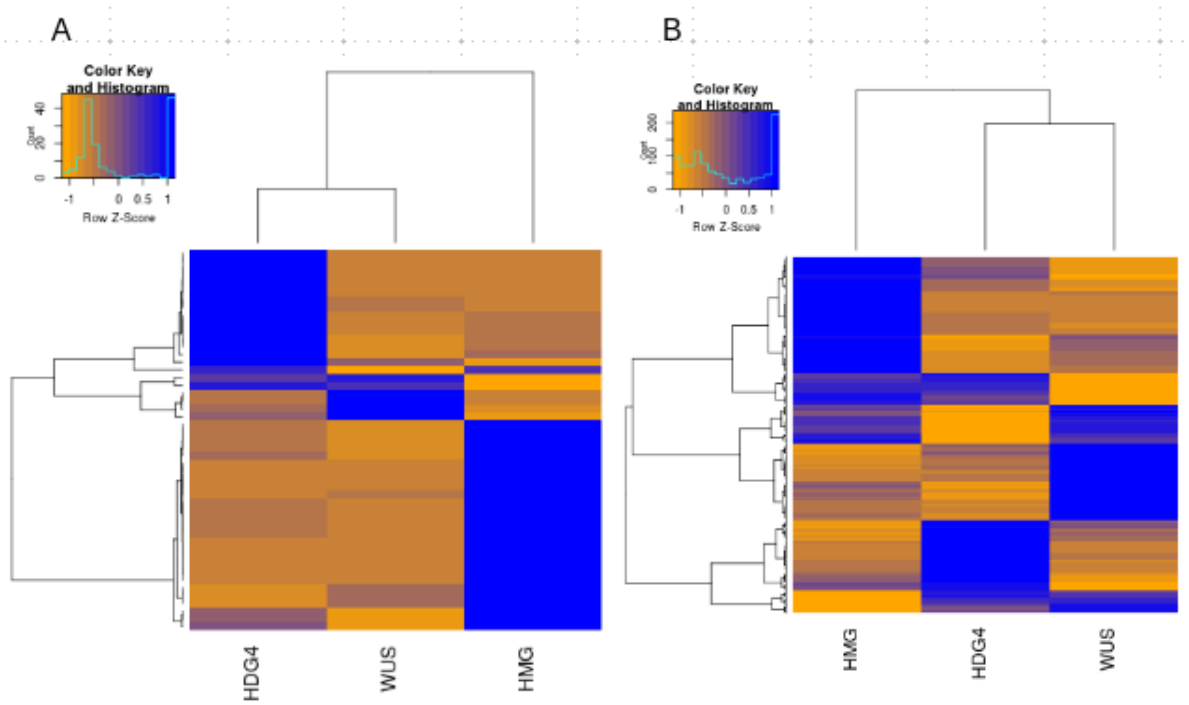

Figure 2.1: Gene expression patterns – PDI network; The expression patterns of the baits (A) and preys (B) chosen for the eY1H assay are shown in the form of heatmap. The data was normalized along the rows. Hierarchical clustering of genes was done using (1-correlation) matrix as the distance matrix.

Genes which are expressed in three cell types are checked whether they are enriched in one or two cell types. If they are not enriched in one or two cell types then they are considered to be enriched in all three cell types. The differential gene expression analysis was done using pairwise comparison function under library(simpleaffy) in R. t-test with a *P*-value of 0.05 and a fold change cut-off of ≥ 1.5 was used. Based on the above-mentioned criteria, enrichment degree of all the baits and preys chosen for the eY1H and Y2H assays were plotted (Figure 2.3). For eY1H, it is clear from the distribution that the baits selected for the study are narrowly enriched and the preys selected for the study are more broadly enriched. For Y2H assay, the genes chosen for the study are mostly narrowly enriched.

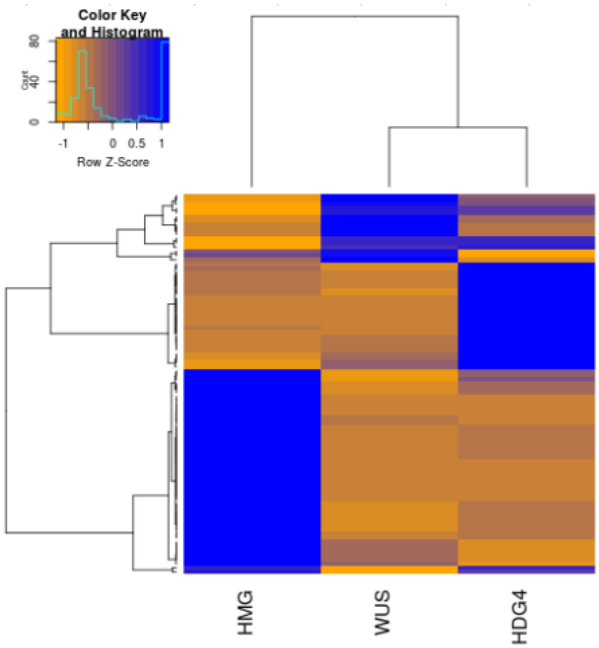

Figure 2.2: Gene expression patterns – PPI network; The expression patterns of the genes chosen for the Y2H assay are shown in the form of heatmap. The data was normalized along the rows. Hierarchical clustering of genes was done using (1-correlation) matrix as the distance matrix.

2.2.2 Network properties

The protein-DNA interaction network constructed by eY1H has 165 interactions (Shivani Bhatia and R. K. Yadav) (Figure 2.4, A). The network comprises of interactions between 37 baits and 53 preys. The total number of interactions on a bait promoter is called its in-degree, and the total number of interactions a prey has is known as its out-degree. The in-degree and out-degree distributions of bait and prey were made (Figure 2.5, A). It can be seen in the degree distribution that many baits and preys have less in-degree and out-degree, respectively. A very few baits and preys have really large in-degree and out-degree.

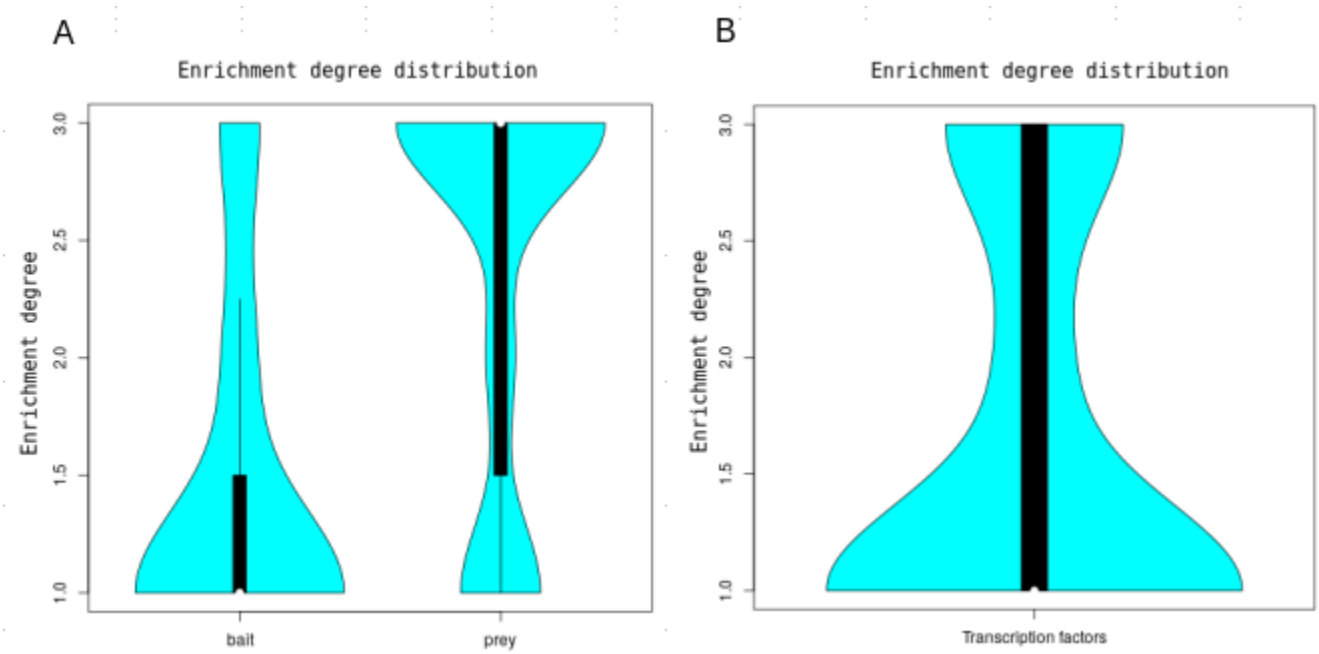

Figure 2.3: Enrichment degree; A) Enrichment degree of all the baits and preys chosen for the eY1H study. B) Enrichment degree of all the genes chosen for the Y2H study.

The protein-protein interaction network made by Y2H has 312 interactions, which corresponds to 291 unique interactions (Prince Saini and Ram Yadav) (Figure 2.4, B). It comprises of interaction between 33 baits and 36 preys. In total there are 52 unique TFs in the network. The total number of interactions which a gene has is its degree. The degree distribution of the genes

was made (Figure 2.5, B). It can be seen in the degree distribution that many genes have less degree and a very few genes have really large degree. Then Protein-DNA interaction density was calculated to know which kind of interactions are over-represented in the network. It is calculated as follows: Number of interactions observed upon number of interactions checked. The PDI density plot was made (Figure 2.4, C).

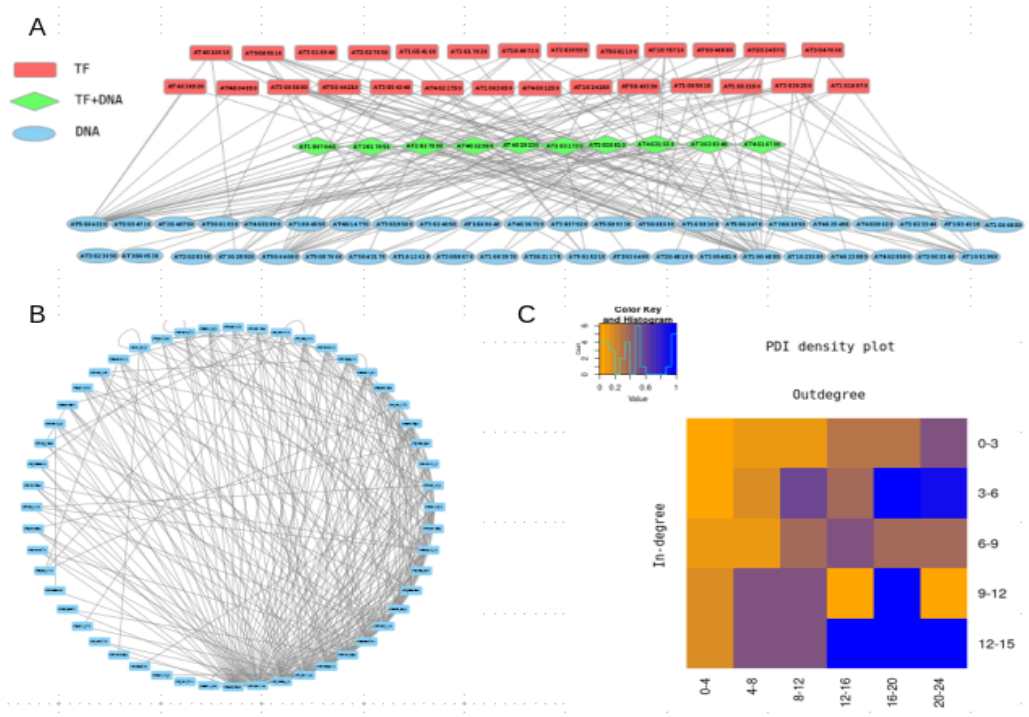

Figure 2.4: Interaction networks; A) eY1H based PDI network. B) Y2H based PPI network. C) PDI density plot based on in-degree and out-degree of the nodes in the network.

In PDI network, the following type of interactions are observed to be enriched: a) Interactions between preys having higher out-degree and baits having higher in-degree, and b) Interaction between preys having higher out-degree and baits having lower in-degree. If we consider the target genes, which have less in-degree (i.e. regulated by few TFs), the TF controlling them have mostly high out-degree (i.e. they are regulating many other target genes). It appears from this

analysis that not a single TF is dedicated to solely regulate one target gene. Evolution has utilized every single TFs for multiple purpose.

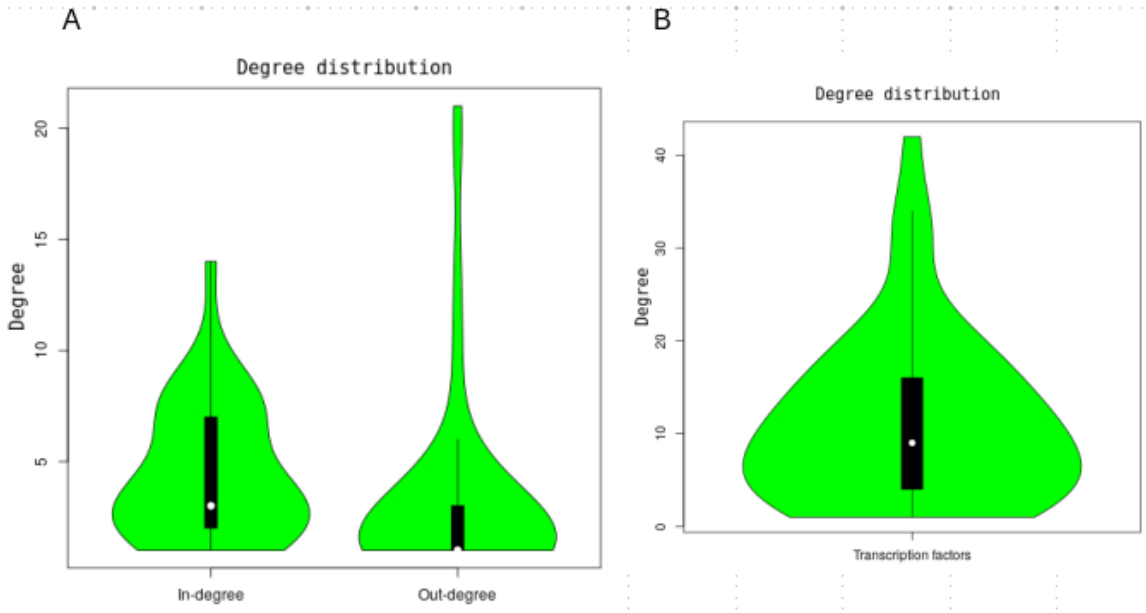

Figure 2.5: Degree distribution; A) In-degree and Out-degree distribution of the nodes in the PDI network. B) Degree distribution of the nodes in the PPI network.

2.2.3 In-vivo significance

The interactions were tested in yeast model. It is important to validate the interactions in *Arabidopsis thaliana*. The interactions observed in the eY1H assay were tested *in planta* to found the activators and repressor by qRT-PCR in the lab (Shivani Bhatia and Ram Yadav). Some of the interactions observed in Y2H assay were tested and verified by Bimolecular Fluorescence Complementation (BiFC) in the lab (Prince Saini). The interactions in the PDI and PPI are reproducible to a good extent. So, the interactions in the networks are reliable.

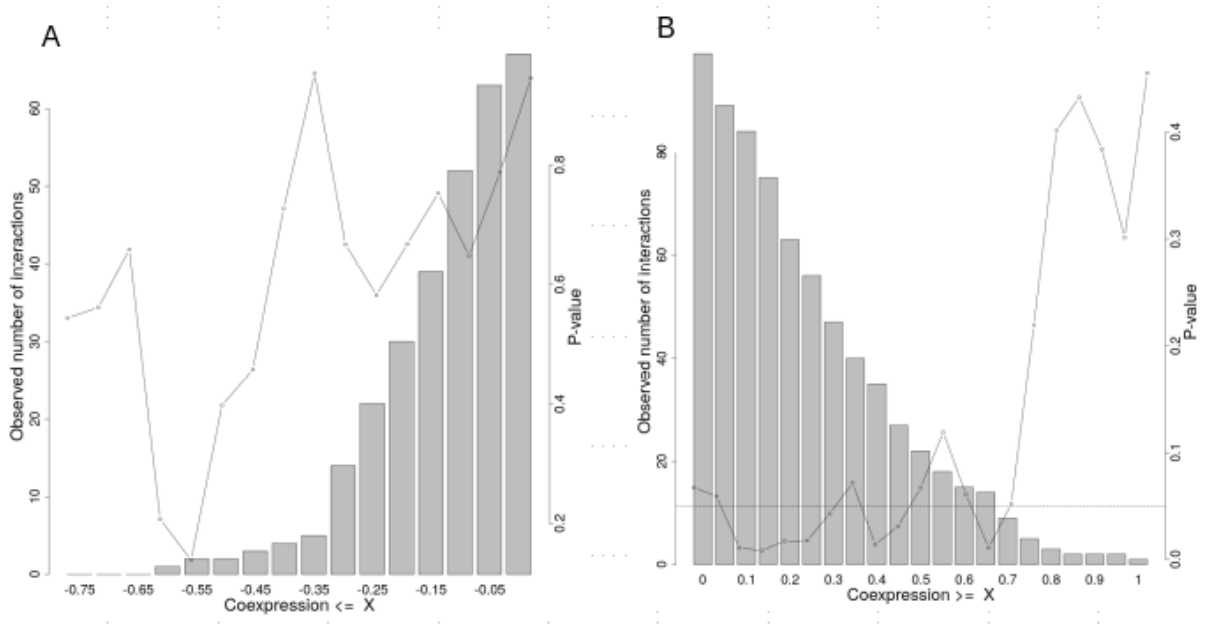

Figure 2.6: Co-expression score – PDI network; The bar plot represents the observed number of interaction given a particular Pearson correlation coefficient cut-off (A – Positive, B - Negative). The line plot represents the corresponding P-value for the observed number of interaction.

It is important to check whether the baits and preys, which are shown to have interactions express/enrich together or is there any correlation in their expression pattern. The expression data (Microarray data) is available for baits and preys in ten different cell populations that belong to the shoot apex of *Arabidopsis thaliana* under normal and other different experimental condition [2, 3]. To check whether they express together or not, Pearson correlation coefficient was calculated between all possible pairs of baits and preys.

The Pearson correlation coefficient value ranges from -1 to +1. A value close to +1 or -1 means that the expression of baits and preys are highly correlated. A value of 0 means no correlation. In case of +1, there is a positive correlation i.e. the bait and prey are enriched together. It also

means that the TF is positively regulating the target gene. In case of -1, there is a negative correlation i.e. the baits and the preys are not enriched together. It could also mean that the TF is negatively regulating the target gene.

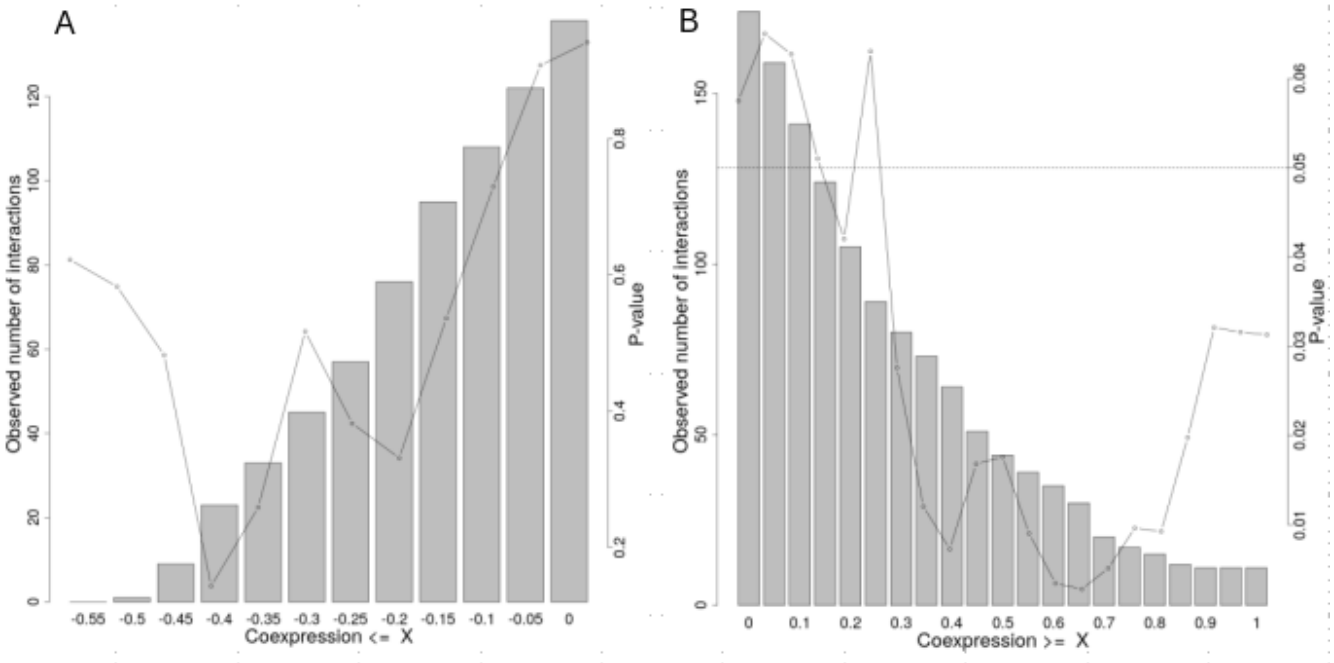

Figure 2.7: Co-expression score – PPI network; The bar plot represents the observed number of interaction given a particular Pearson correlation coefficient cut-off (A – Positive, B - Negative). The line plot represents the corresponding P-value for the observed number of interaction.

The number of interactions in both PDI and PPI networks having co-expression value greater/ lesser than different cut-offs were calculated. To check whether this number of interactions are significant or not, the networks were randomized 25000 times preserving the topology of the network i.e. keeping the in-degree and out-degree of the nodes same. Then the number of interactions having co-expression value greater/less than or equal to different cut-offs were calculated. The distributions of number of interactions were assumed to be normal distributions. P-value was calculated for the observed number of interactions (Figure 2.6, 2.7).

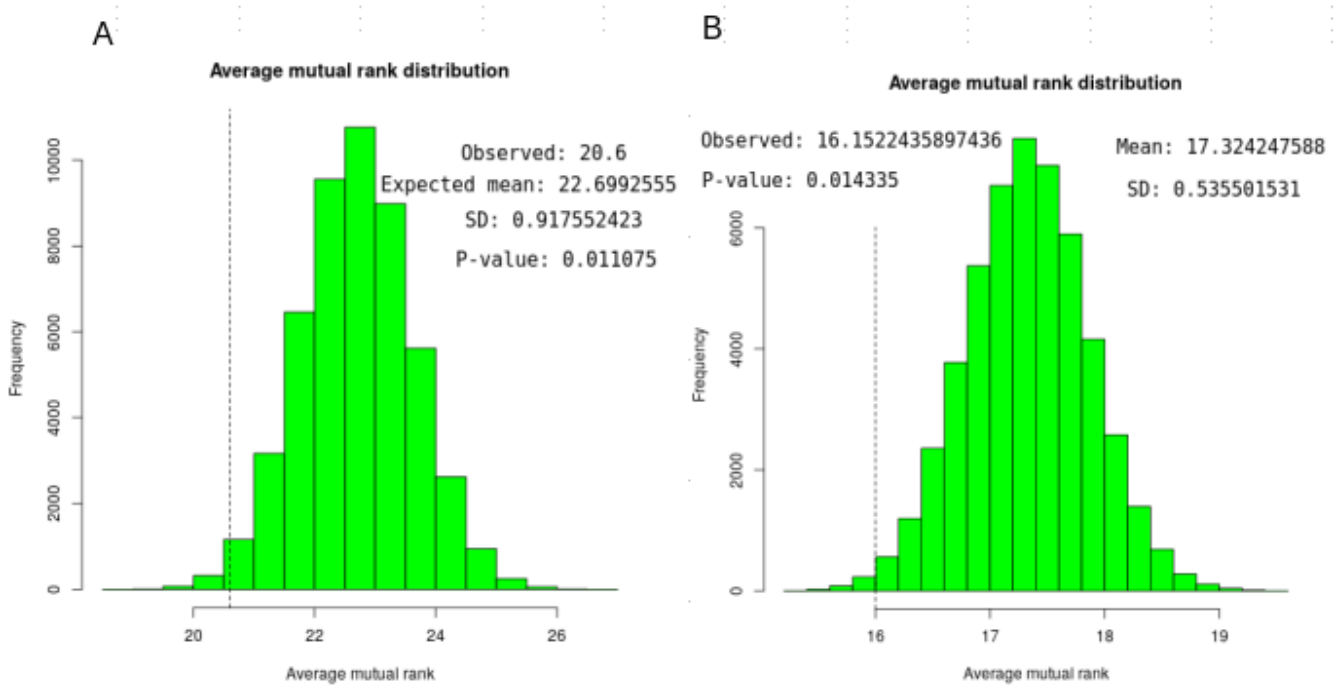

Figure 2.8: Average mutual rank – Interaction networks; Average mutual rank distribution of the 25,000 randomized networks generated from eY1H based PDI network (A) and Y2H based PPI network (B), preserving the topology of the network.

In both PDI and PPI network, one can clearly observe that the numbers of interactions are significant at different cut-offs in the positive direction but not significant in the negative direction.

To appreciate the co-expression of the interacting pairs in the interaction network better, rank based scoring was adopted. Every interacting pair was given three types of rank. The first one is the bait rank, which is the rank given to the prey with respect to bait. For ranking, the absolute value of the Pearson correlation coefficient was considered. The score was calculated for the bait of interest and for all the preys in our study. The preys having better correlation score with the bait of interest were given better rank. The second rank is the prey rank where the baits having better correlation score with the prey of interest were ranked better. Then finally the third rank is

the mutual or average rank which is the average of bait and prey rank. The observed average mutual rank of the PDI and PPI networks were significant. To check for significance, the networks were randomized 25000 times preserving the topology of the network i.e. keeping the in-degree and out-degree of the nodes same. Every time, average mutual rank was calculated. The distribution was assumed to be normal (Figure 2.8). The above observations clearly suggest that the interacting pairs expresses/enriched together or they have a very good correlation in their expression pattern.

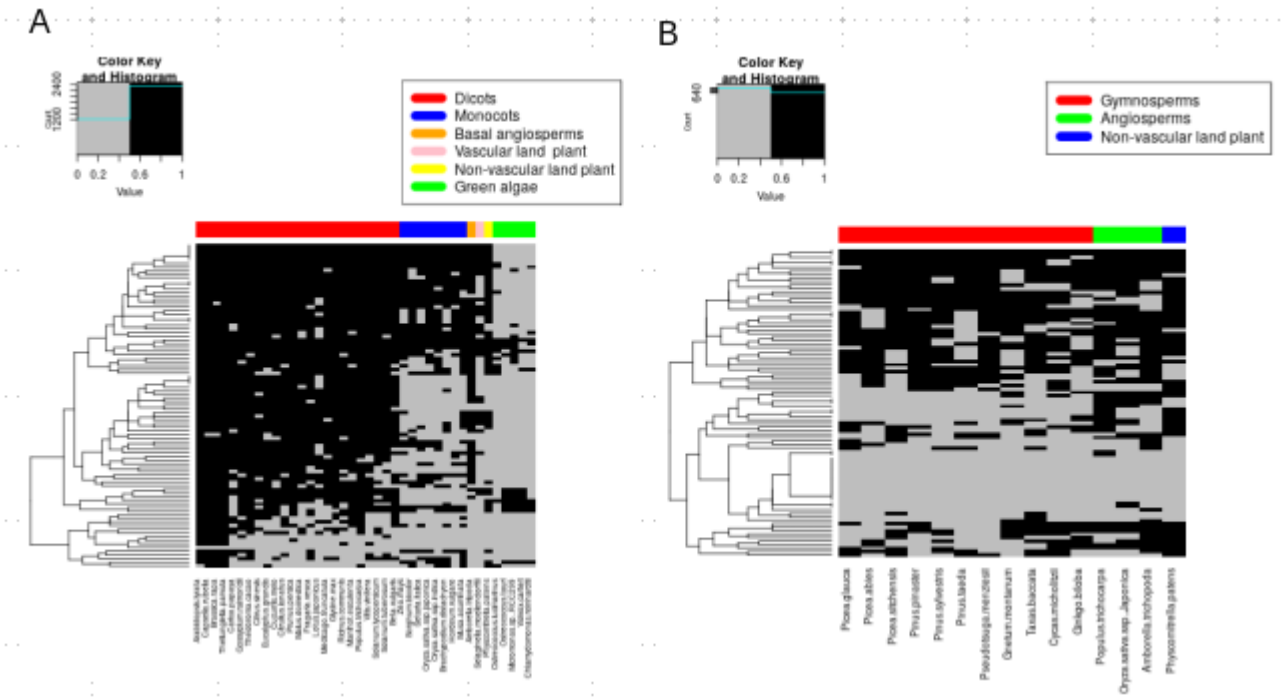

Figure 2.9: TF occurrence; Heatmap representing occurrence of genes chosen for the study in non-gymnosperm (A) and gymnosperms (B) species. Black color represents the presence of an ortholog and the grey color represents absence of an ortholog.

We wanted to check whether the TFs in our network were conserved in the plant kingdom and the interacting pairs occur together in different species of the plant kingdom. We collected the evidence of the presence of TF orthologs in different species from Plaza database [13]. The database had ortholog information predicted by four methods. They are: a) Orthologs detected

through best-BLAST hits; b) Orthologs detected through OrthoMCL clustering; c) Orthologs detected through phylogenetic trees; d) Orthologs detected through collinearity information.

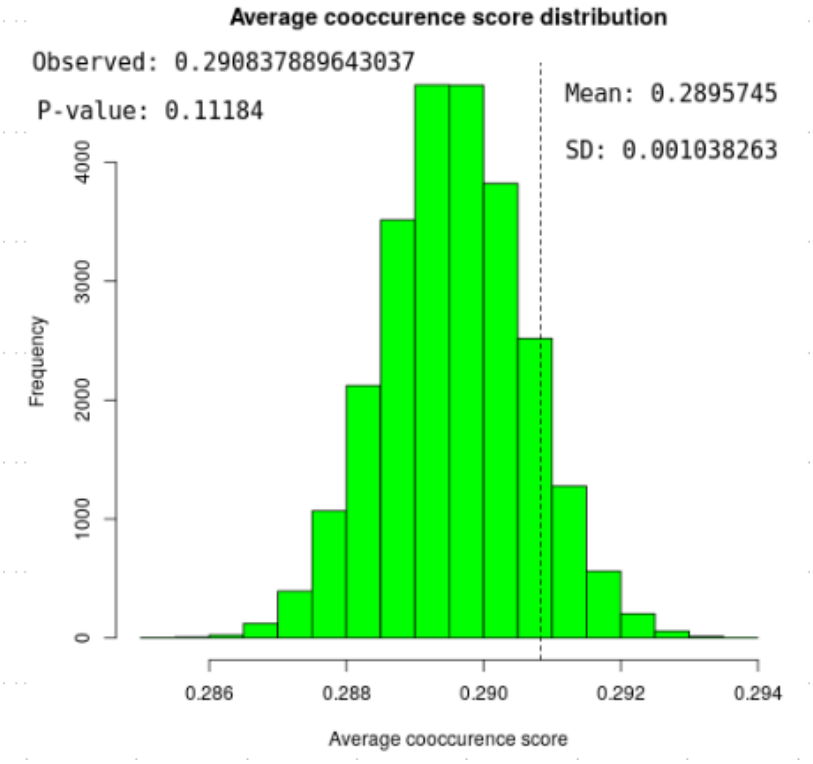

Figure 2.10: Average co-occurrence score; Average co-occurrence score distribution of the 25,000 randomized networks generated from Y2H based PPI network preserving the topology of the network.

In Angiosperms, the orthologs were predicted using all the four methods but in Gymnosperms the orthologs were predicted only by three methods but not based on collinearity information. A TF in *Arabidopsis thaliana* is assumed to have an ortholog in a species X from Angiosperms if at-least two out of four methods could predict the same gene in species X as an ortholog. A TF in *Arabidopsis thaliana* is assumed to an ortholog in a species Y from Gymnosperms if at-least two out of three methods could predict the same gene in species Y as an ortholog. Heatmap representing presence and absence of an ortholog in different species of the plant kingdom was made separately for Gymnosperms and non-gymnosperm species based on the data availability

(Figure 2.9). Based on the ortholog data, co-occurrence score was calculated for all the interacting pairs. The co-occurrence score for an interacting pair X and Y was calculated as follows,

$$\text{Cooccurence score of an interacting pair X and Y} = \sum_{i=1}^{N} (Ai*Bi) \;/\; N$$

where "Ai" is the number of evidence which predicts the presence of a particular ortholog of the protein X in the species "i" upon number of methods for which the data was available and "Bi" is the number of evidence which predicts the presence of a particular ortholog of the protein Y in the species "i" upon number of methods for which the data was available. N is the number of species considered.

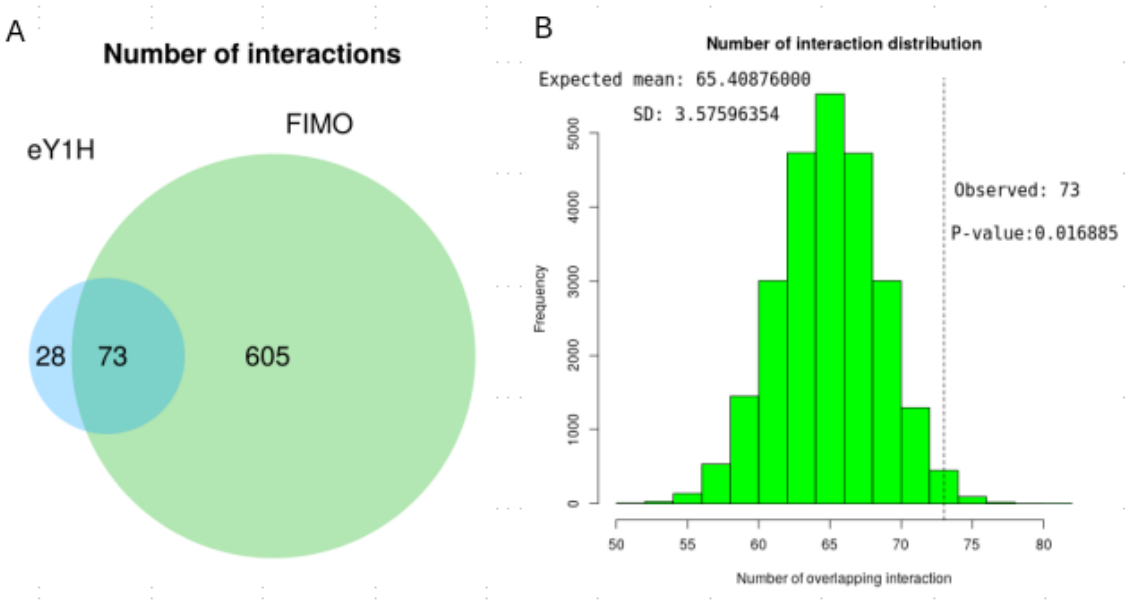

Figure 2.11: Comparison of eY1H with other methods; A) Blue portion represents the number of interactions found by eY1H assay and the green portion represents the number of interactions predicted by FIMO; between 37 baits and 29 preys for which the motif information is available. The intersection of blue and green portion represents the number of interactions which are present in both eY1H and FIMO based networks. B) Distribution of number of overlapping interaction between FIMO based network and 25000 times randomized network generated from eY1H based PDI network preserving the topology of the network.

The average co-occurrence score of all the interacting protein pairs in Y2H based Protein-Protein interaction network is 16.1522435897436. To check whether the observation is significant or not, he network was randomized 25000 times preserving the topology of the network i. e. keeping the degree of the baits and preys same. Every time the average co-occurrence score of the network was calculated. The distribution of average co-occurrence score was assumed to be normal (Figure 2.10).

The distribution had a mean value of 17.324247588 and standard deviation of 0.535501531. The observed average mutual rank of 16.1522435897436 (P-value: 0.014335) is significant. The above observations clearly suggest that the interacting pairs doesn't occur together in different species of the plant kingdom. It might be because the interactions responsible for the maintenance of SAM in *Arabidopsis thaliana* is unique. One can verify the same by considering more genes responsible for the maintenance of shoot apex and many species in the plant kingdom. Plant species usually undergo a lot of duplication events. Sometimes it also undergoes whole genome duplication as well. Duplication of genes always leads to formation of new genes with novel function. There is a good chance that the genes involved in the stem cell maintenance was obtained by the duplication event. From the ortholog heatmap, one can clearly see that the genes are conserved really highly in Brassicaceae family only. The conservation of genes in the Monocots appears to be poor.

2.2.4 Comparison of eY1H assay with other experimental methods

We wanted to compare the experimental output of eY1H assay with other different experiments which are used to find protein-DNA interactions. The motif or sequence preference for some of

the TFs are known by other techniques such as Chip-seq, DAP-seq, SELEX, etc. The position weight scoring matrix of different motifs preferred by different TFs are collected from different sources [6, 7, 8, 9, 10] (Sangram Sahu and Ram Yadav). Out of 53 TFs in the interaction network, sequence preference or motif information is known for 29 TFs. By using 'Finding Individual Motif Occurrences' (FIMO) tool, DNA sequence which are chosen as baits were scanned for the preferred motif sequence of different TFs [5].

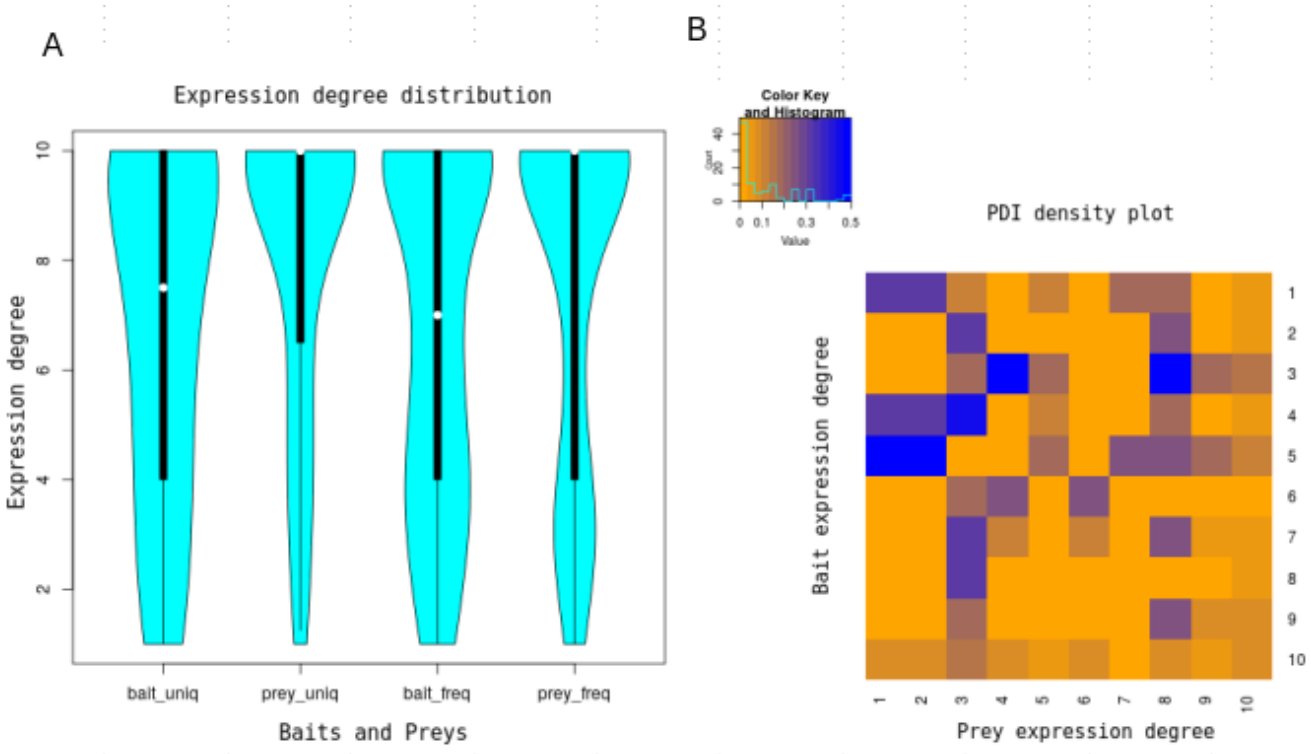

Figure 2.12: Expression pattern of the interacting pairs - A; A) The first two plots represent the expression degree distribution of all the unique baits and preys, which are part of the interaction network. The next two plots represent the expression degree distribution of all the baits and preys, which are part of the interaction network. B) Protein-DNA interaction density plot based on expression degree of the nodes in the interaction network.

The TF is said to regulate a target gene if it has at least one binding site in the DNA sequence chosen as bait with a p-value of <= 1e-4. Among 37 baits and 29 preys (motif information is available), 679 interactions were predicted. These 29 preys have 101 interactions in the eY1H

based PDI network. Out of 101 interactions, 73 interactions were matching with the FIMO based PDI network (Figure 2.11, A). To check whether the overlap is significant or not, FIMO based network was randomized 25000 times preserving the topology of the network i.e. keeping the in-degree and out-degree of the nodes same. Every time the number of interactions overlapping with the eY1H network was counted. The distribution of number of overlapping interactions was assumed to be normal. The distribution had mean of 65.40876000 and standard deviation of 3.57596354. The observed number of overlapping interaction i.e. 73 (P-value: 0.016885) is significant (Figure 2.11, B). The above observation suggests that the network obtained based on eY1H assay is not only very similar to the interactions obtained from other experimental method but also helpful in finding novel binding sites in the upstream sequence of various genes.

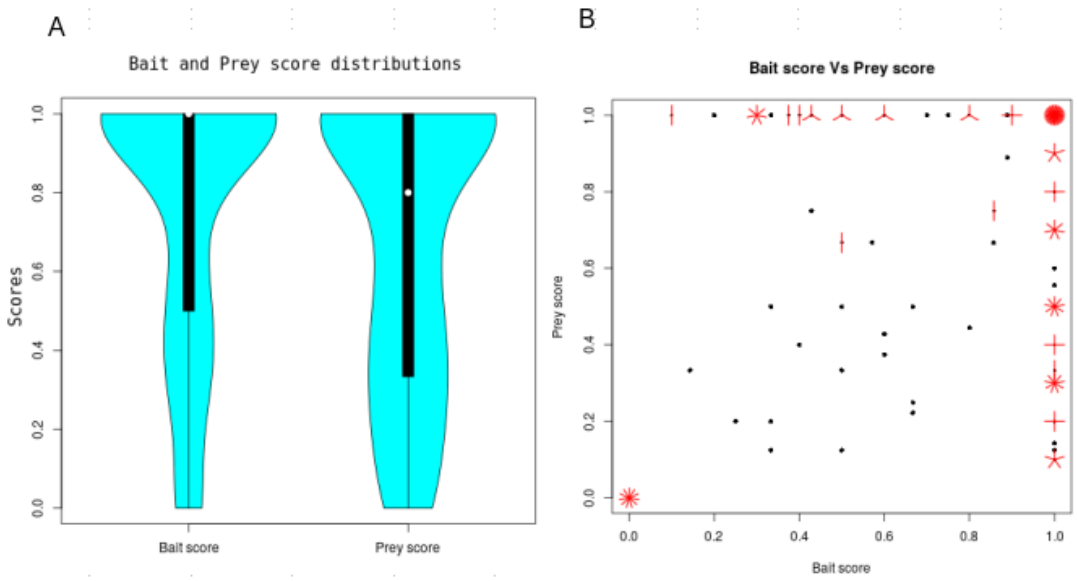

Figure 2.13: Expression pattern of the interacting pairs – B; A) Bait score and prey score distribution of all the interacting pairs in the eY1H based Protein-DNA interaction network. B) Sunflower plot of bait score against prey score of all the interacting pairs in the eY1H based Protein-DNA interaction network.

2.2.5 Expression patterns of the interacting TF and its target gene

The transcription regulation is very complex and a particular gene is regulated by multiple TFs. One can classify a particular transcription regulation into four types. They are: a) narrowly expressed TF regulating narrowly expressed target gene, b) narrowly expressed TF regulating broadly expressed target gene, c) broadly expressed TF regulating broadly expressed target gene and d) broadly expressed TF regulating narrowly expressed target gene. It will be interesting to know what type of interaction is really enriched in the interaction network.

The expression degree distribution was plotted for unique bait and prey list in the interaction network and also for the baits and preys in the interaction data (Figure 2.12, A). Mostly all the baits and preys are broadly expressed but preys are little bit more broadly expressed when compared to baits. Although the baits are narrowly enriched, they are broadly expressed. One of the reasons for the overall broad expression is that the cell populations considered are overlapping cell populations. The Protein-DNA interaction density based on expression degree was plotted (Figure 2.12, B). It can be noticed that the narrowly expressed TF regulating narrowly expressed target gene is over-represented in the network. So, the narrowly enriched target genes are more likely to be regulated by the narrowly enriched TFs.

To understand the spatial expression patterns of the interacting pairs, bait score and prey score were calculated. Bait score is defined as the number of the cells in which both the prey and bait expresses by the number of cells in which only bait expresses. Prey score is defined as the number of the cells in which both the prey and bait expresses by the number of cells in which

only prey expresses. Bait score and prey score ranges from 0 to 1. The distribution of bait score and prey score were made (Figure 2.13 A).

One can observe that the bait scores are closer to one. It is because baits are more narrowly expressed when compared to the preys. The bait score against prey score for all the interacting pairs were also plotted (Figure 2.13 B). One can observe a good number of interactions having either bait or prey score to be one. There are some interactions were both bait and prey scores are completely zero. It might be because of the negative regulation of TFs. When a TF has a positive influence on the target gene expression, one can expect a high overlap in the expression pattern. Similarly, when a TF has a negative influence on the target gene expression, one can expect a lesser overlap in the expression pattern. To confirm this, the co-expression score distribution of the interacting pairs having both bait and prey score >=0.5, and the co-expression score distribution of the interacting pairs having both bait and prey score <=0.5 were considered. The co-expression score distribution of the interacting pairs having both bait and prey score >=0.5 had a mean score of 0.176177586374959 and standard deviation of 0.317008853935565 (appears to be more). The co-expression score distribution of the interacting pairs having both bait and prey score <=0.5 had a mean value of -0.118397641589685 and standard deviation of 0.166799548547975. t-test was used to check whether there is any significant difference between these two distributions. The distributions were significantly different (P-value: 0.0001). As expected, on an average the interacting pairs with both bait and prey score >=0.5 had positive co-expression score and the interacting pairs with both bait and prey score <=0.5 had negative co-expression score (Figure 2.14, A).

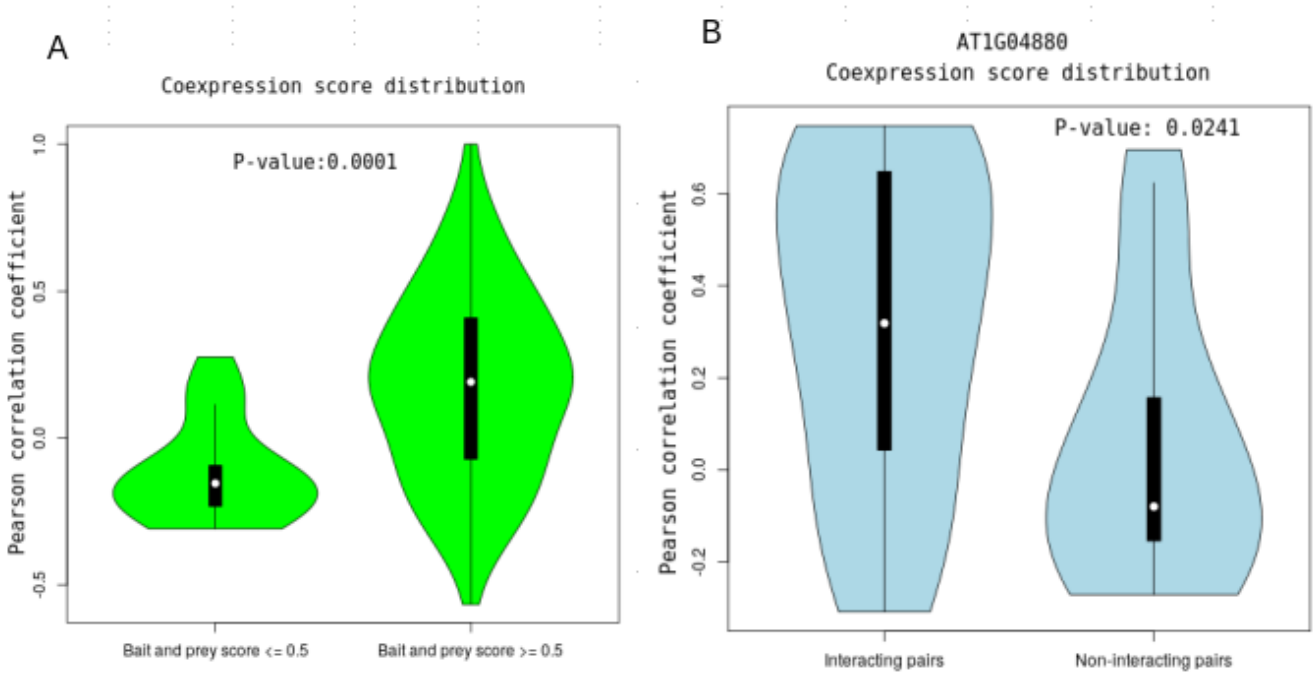

Figure 2.14: Co-expression score distribution; A) Co-expression score distribution of interaction pairs having lesser and higher overlaps, respectively. B) Co-expression score distribution of AT1G04880 with its targets and non-targets.

2.2.6 Predicting the nature of the transcription factor gene regulation

A TF can regulate the expression of the target gene both in a positive or negative way (sometimes bivalent). Based on gene ontology annotations from TAIR database [11], list of TFs which are known to be positive regulators and negative regulators were collected (Table 2.1). One can check the nature of regulation by qRT-PCR experimentally or can predict the nature of regulation by two ways: a) TF-cofactor interaction and b) TF-target gene co-expression values. If a TF interacts with cofactors which are known to be activators, then there is a good chance that the TF will also act as an activator and vice versa. The lists of all the TF-cofactor interactions were collected from Bar toronto, Biogrid and Inact databases [12].

Table 2.1: Interaction data; Interaction and annotation data collected from databases and TAIR ontology.

TF-Cofactor interaction data			Type of TF regulation	
TF	Co-factor	Cofactor type	TF	Literature information
At1g06850	At5g20900	Repressors	At1g04550	Repressors
At1g58100	At1g01160	Activators	At2g38340	Activators
At2g31730	At5g28640	Activators	At3g19580	Repressors
At2g36400	At1g01160	Activators	At3g21175	Activators
At2g36400	At4g00850	Activators	At3g24050	Activators
At2g36400	At5g28640	Activators	At3g50870	Activators
At2g38340	At5g28640	Activators	At3g54810	Activators
At4g14770	At3g48360	Not Known	At3g60530	Activators
At4g25490	At3g07740	Activators	At3g61850	Repressors
At4g25490	At4g16420	Activators	At4g16780	Repressors
At5g61590	At1g74950	Repressors	At4g25490	Activators
At5g61590	At3g07740	Activators	At4g28500	Activators
At5g61590	At5g28640	Activators	At4g32890	Activators
			At4g38620	Repressors
			At5g43170	Repressors
			At5g61590	Activators
			At5g64220	Activators

To predict the nature of TF regulation by co-expression values, we have considered only TFs which have high out-degree i.e. 6. For a TF, the co-expression value distribution for target and non-targets genes were made. t-test was used to check whether there is any significant difference between these two distributions. If the distributions are significantly different then based on the

target gene co-expression value distribution with respect to non-target gene co-expression value distribution, the nature of regulation of a TF is predicted. By this approach TF AT1G04880 was predicted to be a positive regulator. The distribution of target and non-target co-expression value distribution was plotted (P-value: 0.0241) (Figure 2.14 B).

2.2.7 Construction of in-silico association networks

The regulatory networks in *Arabidopsis thaliana* are very complex and has a lot of genetic redundancy. They are many genes which does the same function. Usually a robust phenotype is observed when more than one gene is deleted. Plants which had mutations in the genes chosen for our study didn't have a drastic robust phenotype. So, we have decided to create double mutants to observe robust phenotype. For choosing a pair of genes for creating double mutants, we created association networks. Associated networks connect two genes having very similar properties. We have created four association networks for understanding genetic redundancy based on the following properties. They are: a) Similar interacting partners, b) Binding site cooccurrence in the genome, c) Similar upstream regulating TFs and d) Similar downstream target genes (Figure 2.15, F1).

For the construction of association networks, data from plant transcription database and Y2H based PPI network were used. In plant transcription database, position specific scoring matrix (PSSM) or the motif preference of 27 out of 52 TFs were present [6]. A TF is assumed to regulate a target gene B if it has a binding site in the upstream 500 bp from the transcription start site of the target gene B. Overall the database had PSSM for around 600 plant TFs. Genome wide binding site information for all the TFs were also available.

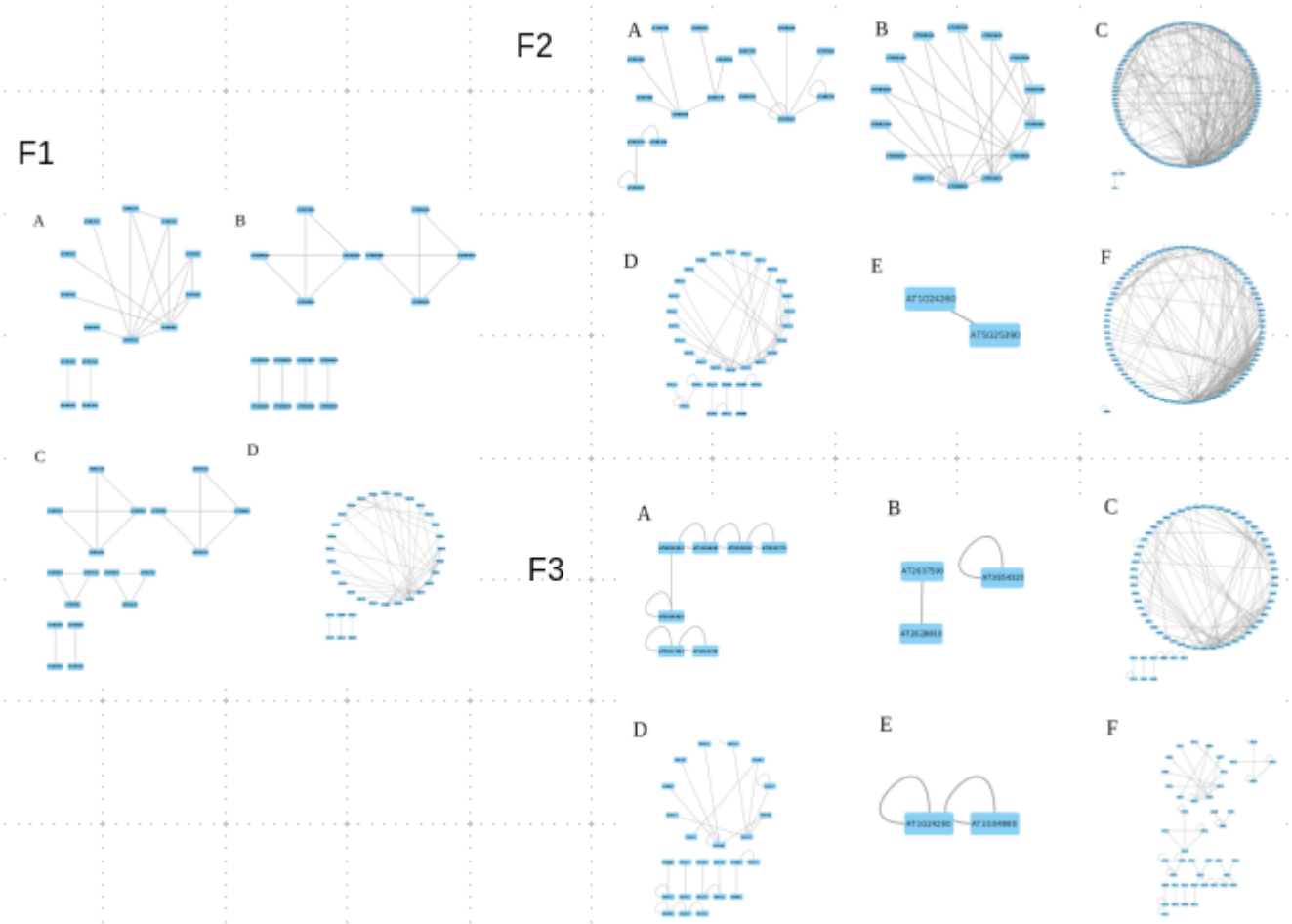

Figure 2.15: Association and in-silico PPI and PDI networks; F1: Association networks; A: Based on TF regulation similarity; B: Based on TF profile or target similarity; C: Based on TFBS cooccurrence; D: Based on PPI similarity based on Y2H based PPI network. F2 – PDI network, F3 – PPI network: A, B, C represents cell type specific network of HMG, HDG4 and WUS cells in HMG: HDG4: WUS group respectively. (D, E, F) represents cell type specific network of FIL, CLV and WUS cells in FIL: CLV: WUS group, respectively.

The networks were constructed only for the 52 TFs considered in our study. First the similarity score for all possible combinations of TFs were calculated. Finally, the pairs with the best scores were filtered for constructing the network. The similarity scores were calculated as follows:

$$\text{PPI similarity} = \frac{\text{Number of common interactions of protein A and protein B}}{\text{Number of Protein A interactions + Number of protein B interactions}} \quad \text{(Cut-off used: =>0.35)}$$

$$\text{TF target or profile similarity} = \frac{\text{Number of common targets of protein A and protein B}}{\text{Number of protein A targets + Number of protein B targets}} \quad \text{(Cut-off used: =>0.15)}$$

$$\text{TF regulation similarity} = \frac{\text{Number of common upstream regulators of protein A and protein B}}{\text{Number of upstream regulators of protein A + Number of upstream regulators of protein B}} \quad \text{(Cut-off used: =>0.3)}$$

$$\text{TFBS cooccurence} = \frac{\text{Number of cooccurence of binding sites of proteins A and B}}{\text{Number of Protein A binding sites X Number of Protein B binding sites}} \quad \text{(Cut-off used: =>1e-4)}$$

2.2.8 Cell type specific in-silico PDI and PPI networks

Interaction data was collected from Bar toronto, Biogrid and Inact [12]. The data includes both experimentally validated and prediction results. Only TFs and cofactors were considered for the construction of networks. To construct network specifically for the *Arabidopsis thaliana* SAM, co-expression mutual rank values for all the interacting pairs were calculated. Only the top 75% mutual rank pairs were considered for the tissue specific networks. Two groups were considered for the cell type specific network. The first group includes HMG, HDG4 and WUS which are non-overlapping and the other group includes WUS, FIL and CLV3 which are once again non-overlapping. The interacting pair is assumed to be specific to a particular cell type A if the interacting pairs are present only in cell type A but not in cell types B and C in the group of cell types A, B and C. Then finally cell type specific PPI and PDI networks were constructed. These networks will be useful in understanding the specification of different cell layers and the mechanism involved in stem cell maintenance and differentiation (Figure 2.15, F2, F3).

2.3 Discussion

A small-scale PDI and PPI networks in SAM of *Arabidopsis thaliana* containing genes enriched in different cell layers of the SAM was constructed using eY1H and Y2H assays, respectively. The Protein-DNA interaction network appears to be really complex with TFs having multiple targets and the target genes being regulated by multiple TFs, which might be a reason for the complexity in the Eukaryotic domain. It appears that there are no TFs which will be dedicated to regulate only one target gene. Evolution has utilized every single TFs for multiple purpose. The Protein-Protein interaction network appears to be really complex with TFs having multiple interacting partners. It might be a major reason for many broadly expressed TF regulating a narrowly expressed target gene. In PDI network, some of the regulations were validated using qRT- PCR in the lab. In PPI network, some of the interactions were tested and verified by Bimolecular Fluorescence Complementation (BiFC). So, the interactions in these networks are reliable. By in-silico approach, the interacting pairs in these networks were shown to co-express in SAM of *Arabidopsis thaliana.* It clearly suggests that the interactions are relevant in our tissue of interest. But one can also observe interactions where the Pearson correlation coefficient score is very close to zero. It might be possible that these interactions do not hold true in the SAM but might be true in some other different tissue. In the PPI network, the interacting TFs pairs doesn't seem to cooccur significantly in different species of the plant kingdom. It might be because the interactions responsible for the maintenance of SAM in *Arabidopsis thaliana* is unique. There is a good chance that the genes involved in the stem cell maintenance were obtained by the duplication event. TFs under study were shown to be conserved really highly in Brassicaceae family but have a poor conservation in monocots. The experimental output of eY1H and other experiments were compared. eY1H based PDI network not only has significant similarity with

in-silico network constructed by other experimental output but it also has its own novel experimental output. When the expression pattern of the interacting pairs was analyzed, it was observed that narrowly expressed TF regulating narrowly expressed target gene is over-represented in the network. So, the narrowly enriched target genes are more likely to be regulated by the narrowly enriched TFs. Interacting pairs having very less overlap is significantly shown to have negative regulation while pairs having very high overlap is significantly shown to have positive regulation. The nature of TF regulation (either positive or negative) was predicted based on the co-expression value of the TF with its target genes and its interactions with various coactivators and corepressors. The predicted results had to be validated experimentally. To identify TFs having redundant roles, association networks are made based on similarity scores of different properties. Double mutants of genes having high similarity scores had to be made to get a robust phenotype.

Finally, in-silico PDI and PPI networks of *Arabidopsis thaliana* SAM containing only transcription factor TFs and co-factors were constructed based on the known and predicted interactions from various databases. Based on the expression pattern of various TFs and co-factors, cell type specific PDI and PPI networks were also constructed. It can be used to understand the identity of various cell layers and stem cell maintenance in the SAM of *Arabidopsis thaliana.* The future plan is to increase the scale of the PDI network by increasing the bait and prey library so as to find the cis-regulatory modules specific to various cell layers. A large scale PDI and PPI networks will help us understand the gene regulation in different cell types better.

2.4 References

1) Rita Groß-Hardt, Thomas Laux; Stem cell regulation in the shoot meristem; Journal of Cell Science 2003 116: 1659-1666; doi: 10.1242/jcs.00406

2) Ram Kishor Yadav, Thomas Girke, Sumana Pasala, Mingtang Xie and G. Venugopala Reddy; Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche; PNAS March 24, 2009. 106 (12) 4941-4946

3) Ram Kishor Yadav, Montreh Tavakkoli, Mingtang Xie, Thomas Girke, G. Venugopala Reddy; A high-resolution gene expression map of the *Arabidopsis* shoot meristem stem cell niche; Development 2014 141: 2735-2744; doi: 10.1242/dev.106104

4) Siobhan M Brady, Lifang Zhang, Molly Megraw, Natalia J Martinez, Eric Jiang, Charles S Yi, Weilin Liu, Anna Zeng, Mallorie Taylor-Teeples, Dahae Kim, Sebastian Ahnert, Uwe Ohler, Doreen Ware, Albertha J M Walhout, Philip N Benfey; A stele-enriched gene regulatory network in the Arabidopsis root; DOI 10.1038/msb.2010.114 | Published online 18.01.2011; Molecular Systems Biology (2011) 7, 459

5) Charles E. Grant, Timothy L. Bailey and William Stafford Noble; FIMO: scanning for occurrences of a given motif; Bioinformatics. 2011 Apr 1; 27(7): 1017–1018.

6) Jin JP, Tian F, Yang DC, Meng YQ, Kong L, Luo JC and Gao G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, 45(D1):D1040-D1045.

7) Khan A, et al; JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework; Nucleic Acids Res. 2018.

8) José M. Franco-Zorrilla, Irene López-Vidriero, José L. Carrasco, Marta Godoy, Pablo Vera and Roberto Solano; DNA-binding specificities of plant transcription factors and their potential to define target genes; PNAS February 11, 2014. 111 (6) 2367-2372.

9) Weirauch et al., Determination and inference of eukaryotic transcription factor sequence specificity; Cell. 2014 Sep 11;158(6):1431-1443. doi: 10.1016/j.cell.2014.08.009.

10) O'Malley RC et al., Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape; Cell. 2016 May 19;165(5):1280-1292. doi: 10.1016/j.cell.2016.04.038.

11) Tanya Z. Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait and Eva Huala; The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome; genesis 2015 doi: 10.1002/dvg.22877.

12) Jane Geisler-Lee, Nicholas O'Toole, Ron Ammar, Nicholas J. Provart, A. Harvey Millar, Matt Geisler; A Predicted Interactome for Arabidopsis; Plant Physiology; October 2007.
DOI: https://doi.org/10.1104/pp.107.103465

13) Sebastian Proost et al., PLAZA 3.0: an access point for plant comparative genomics; Nucleic Acids Res. 2015 Jan 28; 43(Database issue): D974–D981.

Chapter - 3

Identification of different isoforms in the SAM of *Arabidopsis thaliana* by RNA-sequencing

3.1 Introduction

Alternative splicing (AS) is one of the most important post-transcriptional modification that plays a very big role in increasing the diversity of protein content within a cell. It was estimated that 60% of the intron containing genes in plants undergo alternative splicing. AS can have two major outcomes: regulating the transcript levels by changing the mRNA stability or regulation by changing the protein localization, stability or function [7]. In plants, AS was shown to play a crucial role in flowering, organogenesis, circadian rhythm and in biotic stress responses. Overall AS in plants is shown to be a novel means of regulating environmental fitness of the plant. AS in plants has contributed a lot to the complexity of the GRNs and will play a huge role in improving crop and plant phenotype [7, 8]. The environmental and developmental clues, which affect the gene expression at transcription and splicing level are poorly documented at single cell population resolution (Figure 3.1) [7]. Even in animals, AS is shown to play a major role in stem cell biology. AS being a part of the transcriptional and post-transcriptional networks controls pluripotency and differentiation of stem cells [6]. To appreciate the role of AS in *Arabidopsis thaliana* shoot development, we wanted to identify ensemble of cell type enriched annotated and novel isoforms. We identified the presence of various annotated isoforms of four different cell population in SAM by RNA-sequencing. The quality of the RNA-seq biological replicates of

different cell types were analyzed. Finally, microarray profile was compared with RNA-seq to check the similarity.

3.2 Results

3.2.1 Identification of all SAMs expressed isoforms

The promoter reporters of HMG, FIL, WUS and CLV genes were created in *ap1-1; cal1-1* genotype [1, 2]. By using FACS, these different populations were separated. Finally, RNA isolation was done followed by RNA-sequencing (Monika Mahajan and Ram Yadav). After RNA-sequencing, the raw data was analyzed. The quality of the data was checked using fastqc [3]. It was observed to be good. Then the adapter sequence in the reads were trimmed using Trimmomatic [4]. Finally, the reads were assembled and aligned to the *Arabidopsis thaliana* genome to get the transcript and isoform counts using Cufflinks [5] (By Sangram).

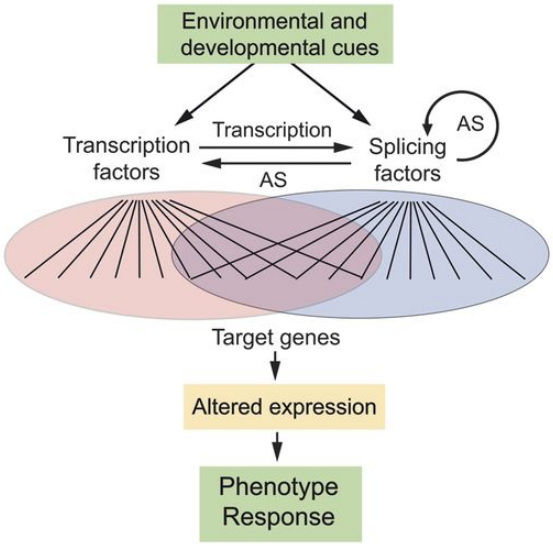

Figure 3.1: Alternative splicing; Dynamic regulation of RNA and protein levels.

The transcript or isoform is assumed to be present if it has a FPKM value of > 1. A transcript is assumed to be expressed in a cell type if it is present in all the replicates. Of 33656 annotated transcripts, 17635 transcripts were detected in at least one of the four cell types. Out of 54832 annotated isoforms, 24155 isoforms were detected in at least one of the four cell types. Then we wanted to compare it with the number of transcripts detected in Microarray. Cell type specific microarray data of these four cell types were taken. The data was normalized using MAS5 algorithm using 'affy' package in R. To find gene expressed in a particular cell type present call information of the non-parametric Wilcoxon signed rank test was computed. A gene is considered to be expressed in a particular cell type if it has a 'present' call in all the replicates. Of the 22591 gene probes used, 14187 genes were detected in at least one of the four cell types [1, 2]. It can be clearly seen that we were able to detect more genes and isoforms by RNA-sequencing.

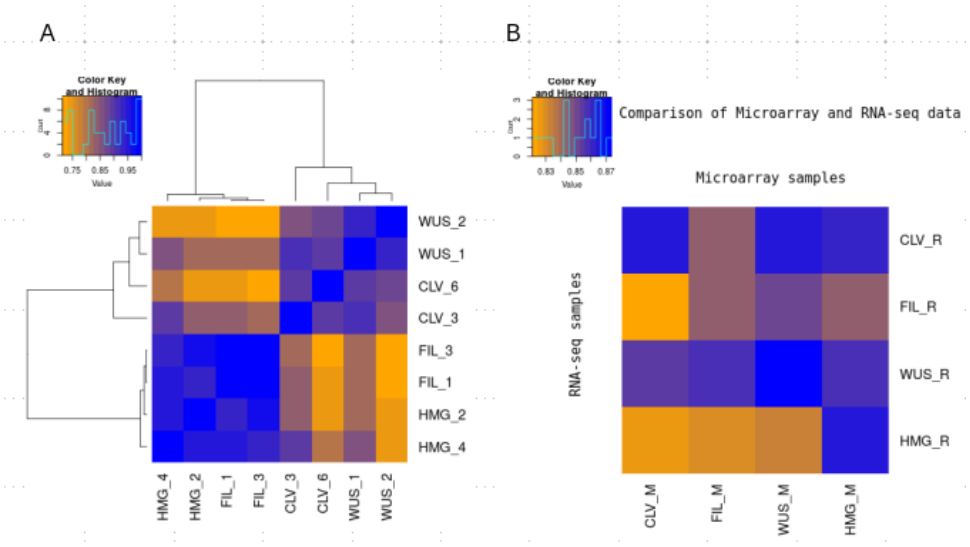

Figure 3.2: Comparing Microarray and RNA-seq data; A) The Pearson correlation coefficient score among the RNA-seq replicates is shown in the form of heatmap. Hierarchical clustering of genes is done using (1-correlation) matrix as the distance matrix. B) Spearman's rank correlation coefficient score among the RNA-seq replicates is shown in the form of heatmap.

3.2.2 Comparison of RNA-sequencing with Microarray data

To start with, similarity among the RNA-sequencing replicates was checked by Pearson correlation coefficient score. The score was represented as a heatmap and the hierarchical clustering was done using (1-correlation) matrix as the distance matrix (Figure 3.2, A). It can be seen that the replicates of the same cell type get clustered together. This shows that the data can be reliable and there is a difference in the signature of gene expression in these cell types.

Finally, the mean values of all the replicates of a particular cell type was considered to compare RNA-sequencing and microarray data. Spearman's rank correlation coefficient was calculated between all cell type specific RNA-sequencing and microarray data. The score was represented as a heatmap (Figure 3.2, B). It is observed that the RNA-sequencing samples were very close to the corresponding microarray data except FIL cell type specific data. It might be because of less number of replicates considered for the study. Overall the correlation of microarray and RNA-sequence data was good.

3.3 Discussion

To find all the transcripts and isoforms present in different cell types in SAM of *Arabidopsis thaliana*, RNA isolation from specific cell types was done using FACS followed by RNA-sequencing. By RNA-sequencing, we were able to identify the presence of huge number of annotated transcripts and isoforms when compared to microarray. When the similarity among different replicates were tested, replicates from the same cell types were observed to be similar. When the data from microarray and RNA-sequencing was compared, except FIL all the other cell types in RNA sequencing were very similar to their corresponding cell types in microarray.

Generating more replicates might solve this problem. In the future we have planned to identify novel transcripts and isoforms present in the shoot apex. In the future, we will identify the differentially expressed or enriched transcripts and isoforms in various cell types of the shoot apex. Finally, we want to identify or predict various annotated and novel long non-coding RNAs.

3.4 References

1) Ram Kishor Yadav, Thomas Girke, Sumana Pasala, Mingtang Xie and G. Venugopala Reddy; Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche; PNAS March 24, 2009. 106 (12) 4941-4946

2) Ram Kishor Yadav, Montreh Tavakkoli, Mingtang Xie, Thomas Girke, G. Venugopala Reddy; A high-resolution gene expression map of the *Arabidopsis* shoot meristem stem cell niche; Development 2014 141: 2735-2744; doi: 10.1242/dev.106104

3) Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data.

4) Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

5) Cole Trapnell et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks; Nat Protoc. 2012 Mar 1; 7(3): 562–578.

6) Kenian Chen, Xiaojing Dai, and Jiaqian Wu; Alternative splicing: An important mechanism in stem cell biology; World J Stem Cells. 2015 Jan 26; 7(1): 1–10. Published online 2015 Jan 26. doi: 10.4252/wjsc.v7.i1.1

7) Nancy A. Eckardt; *The Plant Cell* Reviews Alternative Splicing; Plant cell, Published October 2013. DOI: https://doi.org/10.1105/tpc.113.251013

8) Xudong Shang, Ying Cao, and Ligeng Ma; Alternative Splicing in Plant Genes: A Means of Regulating the Environmental Fitness of Plants; Int J Mol Sci. 2017 Feb; 18(2): 432. Published online 2017 Feb 20. doi: 10.3390/ijms18020432