Natural variation of gene regulatory networks in *Arabidopsis thaliana*

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

Ammarah Anwar
aus Islamabad, Pakistan

Würzburg, 2022

Eingereicht am: ……………………………………………..……..

Mitglieder der Promotionskommission:

Vorsitzender:

Gutachter: Dr. Arthur Korte

Gutachter: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums: ………………………..

Doktorurkunde ausgehändigt am: ……………………..…….

EIDESSTATTLICHE ERKLÄRUNGEN NACH §4, ABS. 3, SATZ 3a DER PROMOTIONSORDNUNG DER FAKULTÄT FÜR BIOLOGIE

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation: "Natürliche Variation von genregulatorischen Netzwerken in *Arabidopsis thaliana*„, eigenständig, d. h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen, als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben. Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Weiterhin erkläre ich, dass bei allen Abbildungen und Texten bei denen die Verwertungsrechte (Copyright) nicht bei mir liegen, diese von den Rechtsin- habern eingeholt wurden und die Textstellen bzw. Abbildungen entsprechend den rechtlichen Vorgaben gekennzeichnet sind sowie bei Abbbildungen, die dem Internet entnommen wurden, der entsprechende Hypertextlink angegeben wurde.

Affidavit

I hereby declare that my thesis entitled: "Natural variation of gene regulatory networks in *Arabidopsis thaliana*„ is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis. Furthermore, I verify that the thesis has not been submitted as part of another examination process neither in identical nor in similar form.

Besides I declare that if I do not hold the copyright for figures and paragraphs, I obtained it from the rights holder and that paragraphs and figures have been marked according to law or for figures taken from the internet the hyperlink has been added accordingly.

Würzburg, 2022

Ammarah Anwar

Abstract

Understanding the causal relationship between genotype and phenotype is a major objective in biology. The main interest is in understanding trait architecture and identifying loci contributing to the respective traits. Genome-wide association mapping (GWAS) is one tool to elucidate these relationships and has been successfully used in many different species. However, most studies concentrate on marginal marker effects and ignore epistatic and gene-environment interactions. These interactions are problematic to account for, but are likely to make major contributions to many phenotypes that are not regulated by independent genetic effects, but by more sophisticated gene-regulatory networks. Further complication arises from the fact that these networks vary in different natural accessions. However, understanding the differences of gene regulatory networks and gene-gene interactions is crucial to conceive trait architecture and predict phenotypes.

The basic subject of this study – using data from the Arabidopsis 1001 Genomes Project – is the analysis of pre-mature stop codons. These have been incurred in nearly one-third of the ~ 30k genes. A gene-gene interaction network of the co-occurrence of stop codons has been built and the over and under representation of different pairs has been statistically analyzed. To further classify the significant over and under-represented gene-gene interactions in terms of molecular function of the encoded proteins, gene ontology terms (GO-SLIM) have been applied. Furthermore, co-expression analysis specifies gene clusters that co-occur over different genetic and phenotypic backgrounds. To link these patterns to evolutionary constrains, spatial location of the respective alleles have been analyzed as well. The latter shows clear patterns for certain gene pairs that indicate differential selection.

Zusammenfassung

Das Verständnis des kausalen Zusammenhangs zwischen Genotyp und Phänotyp ist ein wichtiges Ziel in der Biologie. Das Hauptinteresse liegt darin, die Merkmalsarchitektur zu verstehen und Loci zu identifizieren, die zu den jeweiligen Merkmalen beitragen. Genome-wide association mapping (GWAS) ist ein Werkzeug, um diese Zusammenhänge aufzuklären und wurde erfolgreich in vielen verschiedenen Arten eingesetzt. Die meisten Studien konzentrieren sich jedoch auf marginale Markereffekte und ignorieren epistatische und Gen-Umwelt-Interaktionen. Diese Wechselwirkungen sind problematisch zu erklären, werden aber wahrscheinlich einen wichtigen Beitrag zu vielen Phänotypen leisten, die nicht durch unabhängige genetische Effekte, sondern durch ausgefeiltere genregulatorische Netzwerke reguliert werden. Eine weitere Komplikation ergibt sich aus der Tatsache, dass sich diese Netzwerke in verschiedenen natürlichen Akzessionen unterscheiden. Das Verständnis der Unterschiede zwischen genregulatorischen Netzwerken und Gen-Gen-Interaktionen ist jedoch entscheidend, um die Merkmalsarchitektur zu konzipieren und Phänotypen vorherzusagen.

Das grundlegende Thema dieser Studie – unter Verwendung von Daten aus dem Arabidopsis 1001 Genomes Project – ist die Analyse von vorzeitigen Stop-Codons. Diese sind in fast einem Drittel der ~ 30k-Gene aufgetreten. Ein Gen-Gen-Interaktionsnetzwerk des gleichzeitigen Auftretens von Stop-Codons wurde aufgebaut und die Über- und Unterrepräsentation verschiedener Paare wurde statistisch analysiert. Um die signifikante über- und unterrepräsentierte Gen-Gen-Interaktion in Bezug auf den biologischen Prozess der kodierten Proteine weiter zu klassifizieren, wurden genonkologische Begriffe (GO-SLIM) verwendet. Darüber hinaus spezifiziert die Koexpressionsanalyse Gencluster, die über verschiedene genetische und phänotypische Hintergründe hinweg gleichzeitig auftreten. Um diese Muster mit evolutionären Einschränkungen in Verbindung zu bringen, wurde auch die räumliche Lage der jeweiligen Allele analysiert. Letzteres zeigt klare Muster für bestimmte Genepaare, die auf eine differentielle Selektion hinweisen.

(Translated through google translate from original)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, all praise to the Almighty Allah, the most beneficent and the most merciful who has given me knowledge, and the ability to write, and has given me strength to accomplish this thesis. I am thankful for the countless blessings upon me.

I would like to pay my regards to my Supervisor Dr. Arthur Korte for hiring me and giving me the opportunity to pursue research in his group. I would like to express my heartfelt gratitude for his kindness, limitless support, and guidance at every step until the completion of this project. I thank him for his great ideas, optimism, self-motivation, and encouragement during the whole research period.

Moreover, I want to thank Dr. Arthur Korte and Prof. Dr. Thomas Dandekar for being members of thesis committee.

I would extend the acknowledgement by paying my deepest gratitude to my family. I am indebted to my parents, Mr. Anwar-ul-haq Alvi and Mrs. Hamida Anwar for their prayers, everlasting love, care, and encouragement. Without their support, both financially and morally I was not able to achieve success. I am heartedly thankful to my husband Dr. Asadullah whose limitless support, and encouragement throughout the processing and completion of this study truly kept me going. I also want to thank my siblings Hamnah, Dr. Yumnah and M. Huzaifa for being supportive during the study period.

Furthermore, I would like to acknowledge my friends and colleagues at CCTB Torsten Paul, William Lopez, Ludwig Leidinger and Markus Ankenbrand for their support and assistance during the study period. I would like to thank Patrick for the technical assistance.

Lastly, I would take the opportunity to thank everyone at CCTB for providing a great environment to work.

Ammarah Anwar

Chapter 1

1 Introduction: A model plant

Since the completion of the remarkable project '*Arabidopsis Genome Initiative*' in late 2000, several genome projects have been pivoted toward better and avant-garde understanding of *Arabidopsis thaliana* for deep perspicacity into plant genomics. With each passing year the whole Arabidopsis scientific community is inching towards new insights. The importance of *Arabidopsis thaliana* as a model to understand the adaptive mechanism in wild species is pronounced by the fact that how its competent genes maneuvered themselves through naturally occurring genetic variations in the hierarchy of genes in a vast gene pool; from an endemic specie to a completely adapted plant in diversified geographical distributions.

Section 1

1.1 History of Arabidopsis - Pre- and post-genomic era

In the first section there is introduction to the history of *Arabidopsis thaliana* and the causes and impacts of natural variations in its wide ranged geographical accessions.

1.1.1 Early discoveries and dawn of genomic era

Arabidopsis thaliana was the first flowering plant (considered as universal reference plant) that had been sequenced and the genomic sequence was initially published in 2000 (Somerville & Koornneef, 2002), (The Arabidopsis Genome Initiative, 2000), (Koornneef & Meinke, 2010). Genome sequencing was first inducted in 1996 by an international team of scientists (The Arabidopsis Genome Initiative, AGI). The chromosome 1, 3 and 5 were reported by (Tabata et al., 2000; Salanoubat et al., 2000; Theologis et al., 2000) respectively. Furthermore, the sequences of chromosome 2 and 4 were sequenced and published by (Mayer et al., 1999; Lin et al., 1999). Since the initiative there had been multiple studies conducted and latest findings in Arabidopsis Genome were published. TIGR (The Institute of Genomic Research) released genome versions 1-5 (Quackenbush et al., 2001) (Haas et al., 2005) until TAIR (The

Arabidopsis Information Resource, https://www.arabidopsis.org) took the initiative in 2005. Later, advanced genome versions of *Arabidopsis thaliana* were published yearly by genome annotation team of TAIR as TAIR 6-7 (Swarbreck et al., 2007), TAIR8 (Buisine et al., 2008) (Lister et al., 2008), TAIR9, TAIR10 (Lamesch et al., 2012) and Araport11 (Cheng et al., 2017) each with varied number of annotated genes.

The genome of *Arabidopsis thaliana* constitutes 5 diploid chromosomes with altogether 33602 genes of which 27,416 genes were reported to be protein coding genes according to TAIR10 data (Lamesch et al., 2012). The size of the genome of *Arabidopsis thaliana* ecotype Columbia (ref-0) was estimated to be 157 Mbp according to (Bennett et al., 2003) in comparison with TAIR10 (Lamesch et al., 2012) where it had been reported to be 125 Mbp which is the smallest genome known among all other flowering plants. (The Arabidopsis Genome Initiative, 2000). This characteristic feature had destined the possibility to sequence, assemble and annotate the genome. This annotated gene set most particularly the molecular functions had contributed towards multiple ecological and evolutionary studies (Atwell et al., 2010; Alonso-Blanco et al., 2009; Fournier-Level et al., 2011b; Bergelson & Roux, 2010; Tian et al., 2003; Mitchell-Olds & Schmitt, 2006) Rapid reproduction resulting in huge number of progenies is also one of the distinctive features of Arabidopsis that assisted scientific community to study its biological processes and helped classifying the genes.

Arabidopsis is native plant of Western Eurasia. Its natural habitat is stony shallow land, but it was also discovered in drastic, low-nutrients, poor, sandy and forest habitats (Mitchell-Olds & Schmitt, 2006). According to the ecological evidence and collected history, traces of glacial refugia of Arabidopsis were found in Mediterranean region (Brennan et al., 2014). The versatile ecotypes also migrated to North America, Africa and east Asia during past 100 years expanding its climatic and geographical range (Hohmann et al., 2014; M. Koch & Matschinger, 2007). This migration pattern unveils immense environmental changes and vast climatic fluctuations in the past that might had affected population size at different intervals (Beck et al., 2008). The relocation of different ecotypes in distinctive continents and survival in harsh climatic conditions also depicts the adaptive nature of the Arabidopsis genome. Figure I shows the Map of worldwide distribution of *Arabidopsis thaliana* cited in (Bresinsky & Strasburger, 2013).

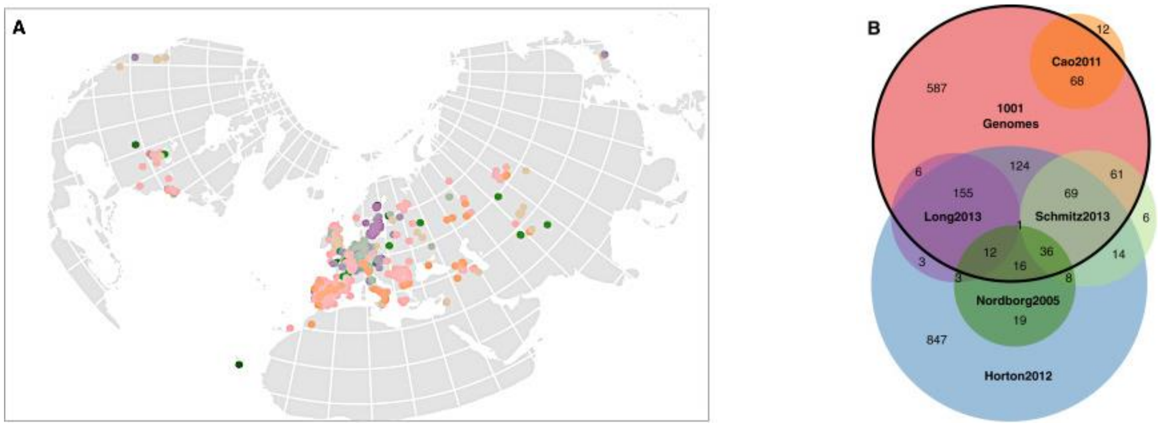


Figure I-A: Locations of the 1135 accessions in 1001 Genomes on the basis of diversity set. The colored dots represent the geographical location of 1135 accessions all over the world. B: Comparison of 1001 genomes with other Arabidopsis thaliana studies (Cao et al., 2011; Long et al., 2013; Horton et al., 2012; Nordborg et al., 2005; Schmitz et al., 2013)

Image Credit: (Alonso-Blanco et al. 2016; 1001 Genomes)

1.1.2 Natural variation and Local adaptation based on geographical location in Arabidopsis

Due to the broad geographical range of *Arabidopsis thaliana* (M. A. Koch, 2019) there is an exceptionally wide-ranging natural phenotypic variation among the accessions. Every investigated phenotypic trait so far has reportedly shown significant natural variations (Koornneef et al., 2004). Within the whole accession range of Arabidopsis, climate differs extensively (Hoffmann, 2002) where environmental gradients like precipitation and temperature play a significant role in natural variation and local adaptation (Ågren & Schemske, 2012; Fournier-Level et al., 2011b; Hancock et al., 2011a; Lasky et al., 2012).

The adverse effect of climatic changes on many plant species can already be seen in the form of changes in their latitudinal and altitudinal distribution (Parmesan & Yohe, 2003). Niche models based on the current climatic changes infer upcoming variations in distribution as relocations or local extinctions (Jezkova & Wiens, 2016; Thuiller et al., 2005). However, the adaptive capabilities of the species are overlooked in such methods i.e., an adequate number of variations in the genes can lead to local adaptations. In the following 50 to 100 years, drastic climate changes leading to extreme drought periods (Siepielski et al., 2017), are predicted to be prevalent among the Arabidopsis accessions in Eurasia (Dai, 2013). Therefore, to accurately predict the responses to future climatic changes and selective pressures, it is very crucial to understand the modes of adaptations by plants and their genetic modifications in present climatic fluctuations. In the world-wide geographical distribution of Arabidopsis, the challenges for the accessions include early spring or late Autumn extremely low temperatures in cold climatic regions and high temperatures in the regions with less rainfall (Hoffmann, 2002). These fluctuations in growth conditions and environmental stress has been overcome through phenotypic adaptations among the different accessions and is underpinned by genetic variations in different accessions adapted to specific environmental conditions. Potentially varied traits reveal substantial variations in response to biotic stresses e.g., fungi, bacteria, viruses, and insects; tolerance to abiotic stimuli including high temperature, freezing low temperature, drought, carbon dioxide, salt stress, and water deprivation (Langridge & Griffing, 1959; Murphy & Taiz, 1995; Rao & Davis, 1999; Sharma et al., 1979; J. Zhang & Lechowicz, 1995), as well as changes in physiological traits like: phosphate uptake, calcium ion signaling, seed dormancy and water usage efficiency (P. Krannitz et al., 1991; Nienhuis et al., 1994; Ratcliffe, 1976) and biochemical traits including glucosinolate, epicuticular wax composition and several enzymatic activities (Magrath et al., 1994; Mitchell-Olds & Pedersen, 1998; Mithen et al., 1995; Rashotte et al., 1997).

Apart from these traits many other phenotypes depict variation including seed size, vernalization reaction and flowering period, which are all important life history traits. (Alonso-Blanco et al., 1999, p.; Candela et al., 1999; Koornneef et al., 1998; P. G. Krannitz et al., 1991; Mitchell-Olds, 1996) .

A considerable deviation in flowering pattern of multiple accessions has been observed which associates with latitude of the sample collecting location. With focus on flowering time of *Arabidopsis thaliana*, although variation in *FRI* has already a key role in flowering time adaptations, yet some other flowering-time genes have reportedly exhibited natural variations where they mostly schedule an early flowering-time cycle (Roux et al., 2006). Some studies even found early flowering as an adaptation to be more beneficial in some particular climatic conditions like high temperatures in summer and drought, as it prevents the plant from stressful conditions which set a negative effect on the seed production (Pigliucci et al., 2003; Callahan & Pigliucci, 2002; Mckay et al., 2003). Some accessions also exhibit considerable adaptations in responses to the varying drought constraints in their native environments which result in diverse genetic modifications among them in traits that are significant in climatic adaptations to drought e.g., regulation of guard cells, stomatal closure, and flowering time (Mckay et al., 2003). (Exposito-Alonso et al., 2018) tested the prospective ability of *Arabidopsis thaliana* to adaptation in severe drought periods by treating the plants under extreme drought conditions mocking the future climate change scenarios and accordingly predicted the genetic changes in populations. They found that the pool of genetic variations in the southern European accessions are capable to give the individuals immunity against extremely harsh drought conditions in future (Hampe & Petit, 2005; Lee-Yaw et al., 2016). They also revealed that not only southern regions, but the northern range of the species is also likely be adaptive to extreme drought periods due to the great variety of drought survival genes in both populations. In general, three strategies used by plants to manage drought stress were described by (Ludlow, 1989) e.g., 1) tolerance to dehydration which involves survival in internal water deficiency in dry environment (Scott, 2000), 2). preventing dehydration that refers to maintaining internal water levels in unfavorable conditions, and lastly 3) escaping drought which is achieved through a shorter life span by starting early reproduction (Mckay et al., 2003; Sherrard & Maherali, 2006). So far, research on natural variations based on local adaptations suggests that they have enhanced the overall performance of the plants respective to their environmental gradients, moreover, it is also evident that the reproductive success has a direct relationship with local environmental gradients.

Over the geographic range of a species natural variation is partly retained by local adaptations where various alleles are favored by natural selection under diverse climatical conditions (Hereford, 2009; Kawecki & Ebert, 2004). Local adaptations eventually lead to evolutionary history in population of a species due to their response to the variable geographical selection and are considered predominantly central in determining diversity among species (Alonso-Blanco & Koornneef, 2000). The contribution of local adaptations in maintaining genetic diversity is also remarkable and taken as a stepping stone towards ecological speciation and facilitate in the expansion of species range (Tiffin et al., 2014). Already various studies conducted in plants focusing on local adaptations (A. A. Hoffmann & Sgrò, 2011; Kronholm et al., 2012; Montesinos-Navarro et al., 2012; Méndez-Vigo et al., 2011) have highlighted how environment effects the expression of genes in a synchronized way (Barbujani, 2000). Since Arabidopsis is adapted to diverse habitats, the discovered genomics loci and known geographic regions lead towards identification of diversely modified and locally adapted traits which are also biologically relevant (Banta et al., 2007; Bouchabke et al., 2008; Shindo et al., 2007). Genome wide study of SNPs by (Fournier-Level et al., 2011b) and (Hancock et al., 2011b) reported genomic loci that relate with local adaptations due to climatic conditions. Despite the history of widespread gene flow, the local adaptations in overall range of accessions have been documented (Ågren & Schemske, 2012; Exposito-Alonso et al., 2018; Hancock et al., 2011; Fournier-Level et al., 2011; Weigel & Nordborg, 2015), and several genomic loci have been identified with their association in climate related local adaptations based on correlation studies and numerous field experiments. The association studies on correlation of SNP mutations with climatic adaptations have revealed the enrichment of non-synonymous SNPs among other SNPs triggering environmental variations (Lasky et al., 2014; Hancock et al., 2011), however, very less is known about the novel genes contributing to the natural phenotypic variations. Therefore, identification of the specific loci leading to phenotypic variations and isolating the responsible genes remains one of the main goals in the field of study.

1.1.3 An introduction to genetic mutations and Single Nucleotide Polymorphisms (SNPs) in Arabidopsis genome

Studies on genetic mutations have played a significant role in identification of the phenotypic adaptations among all 1135 accessions. It is a widely researched topic in order to discover novel alleles and genes particularly involved in adaptive responses to climatic and environmental abiotic or biotic stimuli (Alonso-Blanco & Koornneef, 2000; Pigliucci, 1998; Shindo et al., 2007). Genetic mutations are explained as naturally occurring differences found in the genome of the individuals of same species (Koornneef et al., 2004) and they might be randomly pass on to the next generation. According to previous studies, mutations are the ultimate source of variation in a certain population (Suzuki et al., 2000). It is believed that some mutations proliferate chances of survival and enhance reproduction by shifting biotic pressures e.g.

(pathogens, insects, or environmental fluctuations) and thus become common in the given population. Other mutations might be destructive and lead to functional loss and they are more likely to be lost in the population (Kerwin et al. 2015; Glazebrook 2005; Jones and Dangl 2006; Holub 2007; Holub 2001). Mutations change the DNA sequence effecting genotype of the species, by either deletion of a single or multiple nucleotide base pairs or addition, duplication, inversion, or translocation. According to Hunt et al., a mutation is called SNP when there is a germline addition or deletion of a single nucleotide at a given location, and it is existent in at least 1% of the population (Hunt et al., 2009; Single-Nucleotide Polymorphism, 2015). As reported in literature, SNPs are detected in both coding and non-coding regions in the genome sequence. Non-coding region SNPs can emerge in the intronic and intergenic regions, and other non-coding regions like transcription factor binding sites or promoters (Hunt et al., 2009) whereas coding regions SNPs develop in the exonic regions (Zhan et al., 2015). These SNPs in coding regions are further categorized into synonymous: also known as silent SNPs as they do not affect the encoding amino acid sequence despite affecting the codon, henceforth, primary sequence of proteins remains unchanged therefore they have LOW impact on the genome (Hunt et al., 2009); and non-synonymous SNPs which either lead to codon change altering amino acid sequence and known to be as missense SNPs or Non-sense SNPs which consequently result in an incomplete protein structure due to occurrence of premature stop codon in the gene, therefore, they hold MEDIUM and HIGH impact on the DNA sequence respectively (Filichkin et al., 2010; Lodish H et al., 2000).

SNPs are currently considered as the major source responsible for phenotypic variations in many diverse species (Kruglyak, 1997; Kwok et al., 1996). These markers could be utilized to carry out linkage disequilibrium and association mapping studies, thus help in the identification of polymorphic variants triggering phenotypic changes in the population or species (Mähler et al., 2017). From previous studies, it has been found that different accessions of *Arabidopsis thaliana* show common sequence polymorphisms equally in non-coding and coding regions (Schwarte et al., 2013). In humans 90% of sequence variation is due to the occurrence of single nucleotide polymorphisms (Collins et al., 1998). Similarly, in *Arabidopsis thaliana* single nucleotide polymorphisms are the most prevalent in all types of genetic variations compared to InDels (insersion-deletions) (Initiative, 2000). According to TAIR10 genome release 10,709,466 SNPs were discovered in Arabidopsis genome of which 1,854,599 SNPs were found in coding regions (Alonso-Blanco et al., 2016; Lamesch et al., 2012). In total, 28,148 SNPs resulted in occurrence of premature stop codon in the genome which estimates the density as 1 SNP per 3.3 kb compared to 14,570 InDels ranged between 2 bp – 38 Kbps with an average space of 6.1 kb (Cao et al., 2011; Initiative, 2000; Weigel & Mott, 2009).

From previous studies a strong relationship of non-synonymous mutations and climatic variations has been observed, in continuation our main emphasis is the role of premature stop codons (due to occurrence of SNPs) in initiating natural variations among 1135 accessions of *Arabidopsis thaliana*. Actual stop codons are distinguished from premature stop codons on the basis of their position of occurrence in genome sequence. It is said that in a protein coding gene if the stop codons i.e., (TAA, TGA, and TAG) occur in the second or third reading frame, they are labeled as premature stop codons (here: PSCs) (Wong et al., 2008). A single nucleotide mutation in a triplet codon resulting in any of the three stop codons and it is presumed that premature stop codons have a strong impact on the gene function (Y.-F. Chang et al., 2007) therefore, their occurrence lead to an altered gene function (Savas et al., 2006) due to variations in gene expression. The most vulnerable amino acid codons are reported to be Trp, Tyr, Cys, Glu, Gln, Ser, Leu, Arg, and Gly (Lueck et al., 2019).

In the next topic there is a brief introduction to the gene expression and the codon changes leading to a differed expression which are responsible for local adaptations in certain populations.

1.1.4 Differential gene expression and effects of premature stop gain SNPs on regulation of gene expression

Gene expression is a process in which gene information in the form of genetic code is translated into a functional protein (Alberts B et al., 2002). Studies on gene expression variation provide data of genomic and environmental effects on gene function. While populations in a species differ significantly in genome sequence and gene expression across their geographical range it is expected that any prolific variation in genome sequence due to genetic mutation either impacts their gene function or initiates a differential gene expression.

Differentially expressed genes are identified either by statistical analysis of RNA-seq data or through DNA microarray experiments (Storey & Tibshirani, 2003). They are known to be statistically significant genes with varying expression levels or read counts in two diverse experimental conditions (Anjum et al., 2016). Under varied environmental conditions the differentially expressed genes regulate modified cellular functions. This alteration in gene expression level is stated as either up regulation or down regulation, where up regulation refers to an increased expression of a single gene or multiple genes resulting in higher production of the encoded proteins, whereas down regulation of expression is the decreased expression of gene resulting in less amount of protein (Anjum et al., 2016). Several studies on the effects of over and under expression of gene expression in *Arabidopsis thaliana* were published in the past decade, i.e., delay in flowering time due to down regulation of flowering genes (Agliassa et al., 2018), down regulation of catalytic subunit NatA caused growth delays (Linster et al., 2015), impact of AtCCR1 down-regulation on phenotype, lignins, and cell wall degradation (Goujon et al., 2003), and Down regulation of CLPR2 affects chloroplast size and delays plant development (Rudella et al., 2006).

The term regulation of gene expression points toward several complex mechanisms that control the production of proteins and RNA. These mechanisms play vital role in several developmental pathways, responses on environmental fluctuations, tolerance to biotic or abiotic stresses, cell damage, in turn provide flexibility, strength, and increase adaptability of the organism (Bell et al., 2011). It has been found previously that a certain number of phenotypic variations occur due to abnormal factors in gene regulatory network that profoundly affect gene expression like, flowering time fluctuations (Johanson et al., 2000; Schwartz et al., 2010), semi-dwarfism (Barboza et al., 2013), and variations in flotation of seeds (Saez-Aguayo et al., 2014). In general, regulation of gene expression is controlled at several steps starting from transcription until post translation, however, it may be affected by several factors such as environmental stimuli and mutations (Initiative, 2000).

Among natural accessions of Arabidopsis differential gene expression is prevalent however the eventual causes instigating variations are still being explored. In the molecular function prediction studies, gene functions can be postulated from the expression polymorphism (Kesari et al., 2012; Rockman, 2008). Additionally, differentially expressed genes can be assembled and categorized together according to their functional domains and contribute to Gene Ontology annotations (Kvam et al., 2012). Since they are considered to be potential source of phenotypic heterogeneity therefore studying the gene expression variations due to variations in coding regions in diverse geographical locations are believed to be responsible for local adaptation to build tolerance against abiotic stresses. This is an ongoing field of research and one of the main objectives in our study.

Section 2

1.2 Introduction to the methods applied in research

This section explains the approaches applied to the data for the respective analyses.

1.2.1 Co-occurrence analysis and construction of gene co-expression network based on co-expression coefficient

Co-occurring genes are defined as the genes which are expressed in common accessions holding similar mutations. It is hypothesized by (P.-J. Kim & Price, 2011; H. Müller & Mancuso, 2008) that co-occurring genes are functionally connected. Co-occurrence analysis is also named as functional enrichment analysis (P.-J. Kim & Price, 2011). Generally, the functional product of genes are proteins, and they are always dependent on other proteins to work either as enzymatic catalysts or as a binding pair in metabolic pathways (Marcotte, Pellegrini, Thompson, et al., 1999; Tatusov et al., 2001). To develop profound understanding of functional interactions of genes, several studies about co-occurring genes have been conducted in the past (Huynen & Bork, 1998; Pellegrini et al., 1999; Von Mering et al., 2007). Basically co-occurrence analysis is carried out to determine common unpredicted functional relations between gene pairs (King et al., 2003; Mostafavi & Morris, 2010). In our study, by co-occurrence analysis we intend to identify the statistically significant co-occurring gene pairs which are knocked out in common accessions of *Arabidopsis thaliana*, that lead us towards the assumption that their co-occurrence in shared accessions is functionally related. To validate the function of premature stop codons we are using gene expression data. We hypothesize that if the accessions are randomly picked up together which genes significantly co-occur together. This co-occurrence analysis helps in finding the possible functional connection between the two genes that might be involved in protein inhibition or activation processes. It also contributes to categorization of significant functionally linked genes in their corresponding GO-Slim terms emphasizing on over and under-represented gene pairs (Huang et al., 2007).

It has been reported earlier that incorporation of biological networks with omics data is a useful method for interpretation of differential expression of genes in fluctuating conditions and identification of unknown cellular mechanisms and subnetworks

(Pierrelée et al., 2021). Therefore, we found the best approach to analyze our co-occurrence matrix exhibiting gene-to-gene correlations by building a co-expression network. Analysis of co-expression data using a biological network is an extremely powerful method over the traditional methods that provides a thorough understanding of fundamental molecular functions and shared cellular processes. The approach has several advantages such as the analyses based on the networks are more data driven and less controlled by the current annotation limitations. Due to this reason, the network-based co-expression analyses are also less inclined toward the specific known regulatory pathways but comprise greater coverage over known genes (Charitou et al., 2016). It also indicates that the biological progressions are not primarily driven by a single protein or distinct, directly independent pathways, rather, it consists of way too complex network of cellular interactions. Understanding the relationships of genes or proteins in a co-occurrence network is highly important in the identification of significant nodes and other topological features, which also play fundamental role in regulating the network.

Numerous networking techniques have been developed so far of which most common include: neural networks (Wahde & Hertz, 2000), Bayesian networks (Friedman et al., 2000; Hartemink et al., 2000), Boolean networks (Akutsu et al., 2000; Ideker et al., 2001). These inter-related networks can portray protein-protein, protein-DNA and gene co-expression interactions (Mitra et al., 2013).

Networks which are derived from expression data of the genes are known as gene co-expression networks (GCNs). These interactions are the graphical representation of the relationships between inter-linked genes in the same organism and built with evidence of functional relationships based on co-occurring genes present in gene databases or published in research articles (Jenssen et al., 2001; Schuemie et al., 2004; Stapley & Benoit, 1999). In these undirected connections each entity (gene) is characterized by a node and two nodes are attached by a link or edge indicating a functional gene-gene correlation (P.-J. Kim & Price, 2011) constructing a scale-free network that possesses few highly interconnected nodes (hub genes) and several nodes with very few connections (outliers). Network topology in terms of arrangement of nodes with respective attributes is responsible to describe the correlation between genes (Barabasi & Oltvai, 2004; Mitra et al., 2013). In a co-occurrence network gene names can be used as an entity and differentially regulating gene candidates can be derived from the resultant network (H. Müller & Mancuso, 2008), e.g., a highly interconnected subnetwork of genes indicates involvement in a certain pathway or formation of a protein complex (Stuart et al., 2003) also known as a ‘module’. Through modules phenotypic traits of unknown genes can be predicted based on strong association with known genes that are already linked to those traits. This approach is named as Guilt-by-Association and is now a widely used prediction method already applied in vast range of species including *Arabidopsis thaliana* (Atias et al., 2009; I. Lee et al., 2010; Mao et al., 2009)*, Oryza sativa* (Ficklin et al., 2010; Ficklin & Feltus, 2011; T.-H. Lee et al., 2009), *Zea mays* (Ficklin & Feltus, 2011), *Homo sapiens* (H. K. Lee et al., 2004) and *Mus musculus* (MacLennan et al., 2009).

Overall, the analysis of co-occurrence through a GCN assists in identification of the genes being regulated in the functionally correlated similar transcriptional pathway or their gene product is involved in a shared biological process. Practically their common functional information is integrated through Gene Ontology enrichment explained in next section.

1.2.2 Gene Ontology Enrichment

Gene ontology is a well-ordered vocabulary of terms that define the function of a gene (http://geneontology.org). This process of gathering information related to the gene’s biological activity is known as functional annotation of gene. There is a specific code along with detailed description and its reference attached to each annotation. These GO terms have been implemented in multiple research projects and considered as a standard process for functional enrichment of genes (Camon et al., 2003; Dwight et al., 2002; Hazbun et al., 2003; Kanapin et al., 2003; Menges et al., 2003).

The annotation involves broad functional categories including biological process, molecular function, and cellular location (Ashburner et al., 2000). The biological process terms refer to a chain of processes involving multiple molecular functions. Furthermore, the GO terms categorized into molecular function elucidate the genes biochemical activity. The terms listed in category cellular component describe the location of the expression of gene in subcellular parts. These categories are supposedly non-overlapping and intended to broadly describe functions, biological activities, and location of the gene products (Berardini et al., 2004). There are several genes that have been assigned an unknown molecular or cellular function since they have been manually examined without any evidence from computational prediction or any findings in literature (Ashburner et al., 2000). In the Arabidopsis genome initiative, it has been reported that 69% genes were categorized in different GO terms according to sequence homology in all organisms, moreover, 9% of the genes were classified based on experiments (The Arabidopsis Genome Initiative, 2000).

In this study, we were specifically interested in finding the molecular function of the genes which contained a functional premature stop codon. Our focus was on molecular functions listed in TAIR GO slim enrichment list publicly available at https://www.arabidopsis.org/download_files/GO_and_PO_Annotations/Gene_Ontology_Annotations/TAIR_GO_slim_categories.txt. These molecular functions are associated with a unique GO identifier such as Kinase activity (GO:0016301), DNA or RNA binding (GO:0003677) (GO:0003723), hydrolase activity (GO:0016787), receptor binding or activity (GO:0005102) (GO:0004872), nucleotide binding (GO:0000166), nucleic acid binding (GO:0003676), structural molecule activity (GO:0005198), protein binding (GO:0005515), transferase activity (GO:0016740), transporter activity (GO:0005215), transcription factor activity (GO:0003700), other enzyme activity (GO:0003824), other binding (GO:0005488), unknown molecular functions (unknown function for these genes) (GO:0005554) and other molecular functions (0003674). GO terms are structured hierarchically i.e., from parent to child (Glass & Girvan, 2014).

Section 3

1.3 Statistical techniques, tools, and data visualization methods based in R

As major proportion of our study was based on statistical analysis and visualization of complex data, therefore, powerful tools and packages were required to handle large scale high throughput data which in our case is Genome data incorporating SNPs encoding premature stop codons and RNA-Seq data. For this purpose, R is considered one of the best programming languages as it provides a wide range of packages for statistical analysis along with the basic functions that include processing, cleaning, and exploration of data. Moreover, it possesses diverse features to visualize data. In comparative genomics studies, visualization is considered as the best strategy to explore, analyze, and present the data and ensures the accuracy of results. R provides a wide variety of core as well as user contributed packages for basic plotting i.e., histograms, box plots and heatmaps.

1.3.1 Statistical testing methods

1.3.1.1 P-value and Multiple testing

In comparative genomics, there are numerous approaches and statistical methods for an estimation of the levels of significance among co-occurring genes which help in interpretation of results and drawing conclusions. Almost all empirical and statistical inferences are interrelated with significance level α as a cut-off commonly stated as $P<0.05$. This p-value cut-off reveals the compatibility of data set with null hypothesis or statistical model (Andrade, 2019).

P-value is a widely applied method because of its computational simplicity and direct results. However, when multiple hypotheses are tested simultaneously in a large population the probability distribution is based on a huge sample size which results in outwardly statistically significant values that are basically false positives and only appear by chance. This phenomenon is known as multiple testing problem or look-elsewhere effect (Bender & Lange, 2001).

Several methods are established to neutralize the number of interpretations. The most common method is to constrict the significant threshold value (HOCHBERG, 1988). Bonferroni correction method is considered to be the simplest multiple testing procedure (Bender & Lange, 2001). The confidence intervals corrected by Bonferroni method are calculated by dividing cut-off threshold with the total number of statistical inferences(Olejnik et al., 1997).

When $\alpha = 0.05$, according to Bonferroni method, in N number of statistical tests inferences which are recognized as statistically significant have:

$$P < \alpha/N$$

Implementation of multiple testing correction procedure to the data set exhibits significant results after application of controlled error rate.

1.3.1.2 Tau score

A pairwise similarity score also known as correlation coefficient is necessary to build a gene co-occurrence network. Tau score is widely used method that is independent of cut-off in the formula. It indicates the strength of relationship among two variables [3]. The value of Tau lies between +1 and -1. +1 and -1 signify a strong positive and negative correlation respectively whereas 0 designates no correlation. It is assumed from the negative value of tau that the elements are inversely or indirectly linked, which implies that with the incline of one variable, there is decline in the other. Instead, with positive tau score values it is presumed that both variables are directly associated with each other.

Tau score is symbolized as "τ" and is calculated by using following formula:

$$\tau = (\text{con-discon}) / (\text{con+discon})$$

where, "con" indicates the concordant pairs and "discon" denotes number of discordant pairs.

To calculate the Tau Score, count the total number of concordant pairs and discordant pairs from the data. The results can be obtained through R. If the pairs are more than

10, Tau usually follows normal distribution. To calculate the z-score for Tau, the following formula is used:

$$z = 3\tau^{*}\sqrt{num(num-1)} / \sqrt{2(2num+5)}$$

Where, "τ" indicates the calculated tau and "num" indicates the total number of pairs.

Tau Score in R

To calculate the Tau score for two vectors in R, the kendall.tau() function is used. This function is present in the VGAM library. The generic syntax of this function is: kendall.tau(x,y) where "x" and "y" are the two vectors having equal length.

1.3.1.3 Peacock test

For comparison of clustering in two samples, Two-sample Kolmogorov D statistic test is applied. This test precisely measures the difference in shape and locality of analytically collective distribution functions of the two samples.

The outcome of two-dimensional Peacock test is D-statistic value, which is built on the hypothesis of genetic association among mutated and reference alleles. Moreover, their correctness is assessed by evaluating specific coexistences of all, which is elaborated as the foremost alteration among the theoretical and empirical values.

1.3.1.4 Haversine Formula

The haversine formula is mainly used to calculate the geographical distance on earth. To calculate the great-circle distance, which is generally the shortest distance from one point to another, latitudinal and longitudinal values are needed. The word "Haversine" was introduced by Professor James Inman in the year 1835. Haversine function is widely and more frequently used while developing the Geographic Information System (GIS) applications.

1.3.1.5 What is the Kolmogorov D statistic?

In Kolmogorov D statistic, the letter "D" represents the "Distance". Technically, D calculates the total vertical or perpendicular distance among the Empirical Cumulative Distribution Function (ECDF) of specific data set. Moreover, the Cumulative

Distribution Function (CDF) is also measured. The calculation of point D could be divided into two parts:

- The value of statistic D is calculated as the highest difference among the ECDF and distribution of reference when the reference distribution is beyond or above the Empirical Cumulative Distribution Function. As the distribution reference is growing monotonically, the highest value of D always appears at the right side of Empirical Cumulative Distribution Function.
- The value of statistic D is calculated as the highest difference among the ECDF and distribution of reference when the reference distribution is below the Empirical Cumulative Distribution Function. The lowest value of D always appears at the left side of Empirical Cumulative Distribution Function.

We reject the null hypothesis if the value of statistic is larger compared to the critical value.

1.3.2 R libraries

In the below section there is introduction to the R packages and their specific functions which were applied to carry out the analysis.

1.3.2.1 Dplyr

Dplyr (Wickham et al., 2015) package provides the grammar of data manipulation with all the functions that work as verbs in order to carry out a task. It has three main objectives:

- Dplyr identifies the most critical verbs of data manipulation and ease them to be utilized in R.
- By writing key chunks in C++, it offers intensely fast performance for large data.
- It utilizes the same display to work with data without needing an extra storage.

This package is designed primarily for data frame operations. It resembles the basic functionality in the purr package and can usually be used in the similar way purr package is used by using %>% operator. This way the series of actions can be stringed together in a pipeline. It provides multiple highly useful functions making data handling and analysis much simpler.

Few examples include:

- *Select*(): for selecting the columns.
- *rename*(): specific columns can be renamed by keeping the remaining columns.
- *matches*(): is considered as the most dominant basic function and allows user to select the columns through regular expressions.
- *filter*(): select rows based on a condition.
- *distinct*(): provides all the distinctive rows from designated columns.
- *arrange*(): is used to sort the rows of table by values in the targeted columns.
- *desc*(): allows users to sort the rows in descending order.
- *mutate*(): allows user to add the columns into data frame that are relied on expressions.
- *summarise*(): reduces a data set to a single entity basically keeps only one copy of the variable.

Practically, all these functions can be combined with '*group_by*()' in order to perform an operation group-wise.

1.3.2.2 Plyr

Plyr package in R language is used to split apart or segregate the data, perform certain task or activity on that data and join that data together again. This step is very usual and commonly used during data manipulation. Furthermore, this package makes it easy to regulate the input data format as well as output from constant group of functions. Although it is possible to use with already existing methods split and the apply functions, but plyr makes the processing convenient.

Plyr package is composed of the built-in apply functions that provide the users to have control upon the input as well as output formats of data. Moreover, it keeps the syntax of all the variants consistent. In addition, it offers more functions such as parallel processing, processing of errors, and display progress bars for better visualization. A widely used function *ddply*() works as: target the data frame, segregate that up, perform certain task, and send back that data frame.

A reason for when plyr is not appropriate to use would be when there are huge datasets to deal with and that involve too much sub-setting which makes processing slower.

1.3.2.3 Tidyr

Tidyr (Wickham & Wickham, 2017) package provides the tools to make the data clean and tidy, where every column represents the variable, observation is represented by each row and every cell consists of a single value.

It provides the tools for pivoting (altering the shape) and hierarchy (also known as nesting and 'unnesting') of targeted set of data. It converts the intensely nested lists to the data frames in rectangular shape (also known as "rectangling"). Moreover, allows the extraction of values out of columns containing string. It also facilitates the users to work with missing values that are both explicit and implicit in nature.

To preserve the breadth of data frame, chopping and un-chopping function is used. *Chop*() function is used to make the data frame smaller by using the conversion of rows of every group into list-columns. However, *unchop*() function is used to make the data frame lengthier by escalating the list-columns in such a way that each and every entity of list-column acquires its individual row as output.

In general, *unchop*() function is more advantageous than the *chop*() function as it makes the complex data simpler. This function keeps the record of tracks of the type of elements, as *unchop*() function is capable of reconstitution of precise and accurate vector type even in the situation of void list-columns.

1.3.2.4 data.table

data.table() (Dowle et al., 2019) is new class of R, which is basically the extension of data.frame object. The syntax of data.table package is similar to the SQL syntax. To enhance the speed, the code of this package is written in C language. Due to this reason, this package works much faster with huge data sets and computes the major operations within seconds.

To read the data from table as data.frame object, *read.table*() function is used. Likewise, the *fread*() function allows to read the data as a data.table object. Datdt is to be defined to read the data in the format of function data.table. colClasses argument is used in the statement to minimize the time duration in reading the data. Hence, the time duration is reduced which then R utilizes to identify the classes of variables. The na.strings argument is used in the data.frame situation, as the missing values are encoded as "-".

1.3.2.5 Reshape

reshape() (Wickham, 2011) package provides the framework for multiple types of reshaping and aggregation of data. The model of "melting" and "casting" is used in this package, where the targeted dataset is "melted" to the differentiable shape that is being measured and identified by the variables. Afterwards, they "cast" them into the new form, whether it would be a list, an array of high dimension or a data frame.

In R language, there are multiple functions that offer the aggregation of data such as tapply. Likewise, for reshaping the data, reshape function is used. Any of the functions offered by R, tends to be dealt very well by in one or more than one situation. However, these functions need different input arguments. Reshape function offers the support for unique data structures that includes high multi-dimensional arrays and list of matrices.

1.3.2.6 Splitstackshape

Different tools that are being used to collect the data such as Google Forms usually disseminates multiple-response queries with all the data concatenated in cells. The *cSplit*() family of functions (concat.split) ruptures such kind of data into segregated cells. Splitstackshape package contains different functions to stack different groups of columns. It also reshapes the extensive data, even in the cases where the targeted data is "unbalanced". This package was introduced by Ananda Mahto in 2019.

1.3.2.7 ggplot2

ggplot2 uses multiple programming methods that are put together by the Grammar of Graphics introduced by (Wickham, 2011a). The second version of this package includes comprehensive description of new geometrics and includes more significant alterations.

Complex plots can be created from data formatted in the form of data frames by using ggplot2. The package provides a set of programmatic commands that specify the variables which can be plotted, and their general graphical properties can be defined. It is also a more practical tool when using different types of plots while only marginal modifications are required if one plot is changed to another, examples include, *barplot*(), *scatterplot*(), *geompoint*(), *geomline*() etc. Moreover, the plots are also high quality and easy to manipulate.

The best structured layout to work on *ggplot*() is the data in long format. *Melt*() function from dplyr often helps achieve the desired data format. Another feature of graphics in ggplot2 is the layering. After designing the main structure, attributes are added layer by layer which offers possibilities for plot customization and gives extra flexibility.

1.3.2.8 ggmap

ggmap (Kahle & Wickham, 2013) package allows spatial picturing by collecting the geographic data of stationary maps originated from OpenStreetMap, Google Maps or Stamen Maps by utilizing graphics execution of ggplot2. It visualizes the data on the maps that have been downloaded from service provided by the google maps.

The main idea behind *ggmap*() is to download the image containing map, using ggplot2 function. In ggmap, the earth map is downloaded in the form of an image. After downloading, with the help of *get_map*() function, the image can be formatted. Get_map is more specifically a package for underlying functions such as *get_openstreetmap*(), *get_stamenmap*() etc which offer an extensive range of arguments.

The most important argument of *get_map*() function is the location input. Basically, the location consists of pair of longitude and latitude values which specifies the center position of the map. This value and specific location is accompanied by the zoom function that ranges from 3 to 20. The zoom function specifies spatial range around the center, with 3 as the continent level, whereas 20 as the roughly building level.

1.3.2.9 Geosphere

The Geosphere package implies the spherical trigonometry explicitly for geographic applications. There are variety of functions that calculate the distance and direction along the circular path and lines of continuous bearing. Many other functions include calculation of location of an entity at specific distance and direction. Moreover, it calculates the given perimeter and area of the spherical polygon.

The prerequisites are the geographical locations (latitude and longitude) that should be provided in degrees. Degrees are basically the decimal numbers. For example: 14 degrees, 12 minutes, 30 seconds = 14 + 12/60 + 30/3600 = 14.208 degrees.

1.3.3 Visualization tool for Co-occurrence data

Despite availability of libraries for basic plotting data in R, there are specific packages related to genomics of which most unique are circular plots such as circos and ideograms which demonstrate the characteristics over whole genome. Since our co-occurrence data is based on inter-linked genes in all 5 chromosomes. So, in order to demonstrate their connection where the information is distributed on multiple tracks, however, the main object is their connection, so instead of a linear layout, the efficient and comprehensive approach is to visualize information in a circular form. For this purpose, we have used Rcircos. In the below section the features of Rcircos are explained.

1.3.3.1 Rcircos

Circos (Krzywinski et al., 2009) is a distinctive package that visualizes genetic data by producing round ideograms which consists of “2D tracks” of line plots. More specifically, Circos is useful for comparison of genetic data among the entities, populations, or other species. It is an open-source tool and considered as one of the most powerful and effective software to showcase highest quality multi-dimensional images. In genomics research, there is a wide usage of Circos plot, especially in next generation sequencing to discover genomic relocations and gene correlations (Hagège et al., 2007). It is highly beneficial in distributing the information for explicit data points.

Initially Circos was developed in perl language and operated through command-line, however, there were certain shortcoming, first, the complexity in installation, i.e., to use the Circos, too many perl packages needed to be installed and secondly, those packages were sensitive to the versions of operating system being used. Moreover, it also needed the users to have high computational skills and hence remained a challenge for genomics data analysis (Diaz-Garcia et al., 2017).

The recently developed RCircos (H. Zhang et al., 2013) is a R language version of Circos. The Rcircos package offers package of graphical functions by drawing basic Circos 2D ideogram for visualization of genome structure and the relationship between position and structure of genomic intervals. R graphics is used to implement the Circos tool which needs R base to be installed. Hence the usage complexity has been significantly reduced and flexibility is increased by incorporating other R pipelines of data processing (H. Zhang, 2016).

1.3.4 Visualization tool for genetic networks

To represent the functional connectivity of co-occurring genes in common accessions we build a gene-gene co-expression network by using Cytoscape. Below is a brief introduction to Cytoscape.

1.3.4.1 Cytoscape

Cytoscape is an open-source tool developed for designing, analyzing, and displaying genetic and molecular interaction networks. It is also used for visualization of co-related data. In a network graph built by Cytoscape nodes represent species or genes or proteins and between edges nodes represent intermolecular interactions (Shannon et al., 2003) .

The steps to create a genetic co-expression network with Cytoscape include:

1. Collection of respective up or down regulated expression data
2. Data integration based on node and edge attributes
3. Parsing GO data into annotations or attributes
4. Data visualization in different available layouts and exporting network.

To integrate the biomolecular relation networks with high quantity of data expression and the molecular states into a cohesive conceptual framework, Cytoscape is used.

Cytoscape is freely available open-source software, which is considered as the most powerful tool when used in combination with the huge databases of protein-DNA, protein-protein and other genetic relations that are progressively present for humans and model organisms. Cytoscape provides the functionality to design and explore the network; to integrate the network visually by using the expression profiles, and phenotypes etc.; and to connect the network with databases for functional interpretations. The Cystoscope's Core is extendable through direct plug-in architecture, which allows the users for speedy progress in additional computational analyses and extra features (Shannon et al., 2003). There are different plug-ins in Cytoscape, which basically are the extensions that could be downloaded to the main software. They provide functions such as collecting network data from public sources and network analysis topology to discover new and interesting biological patterns.

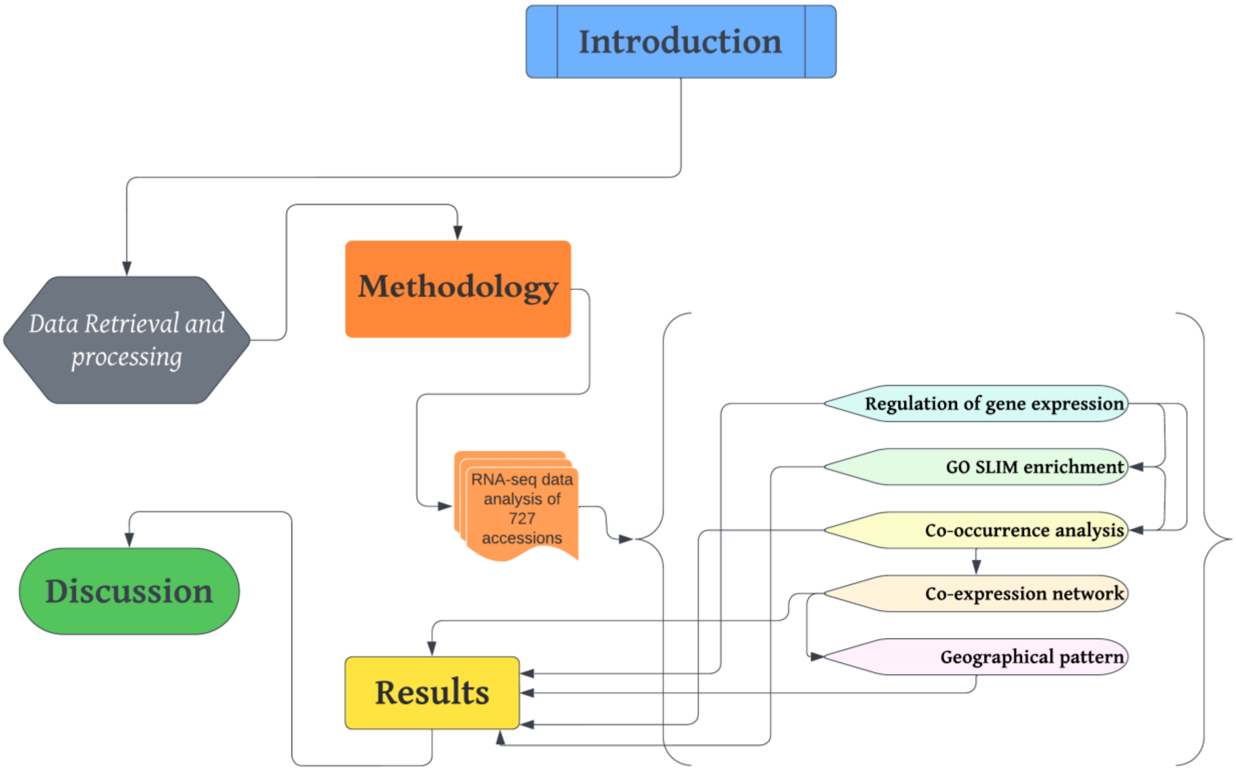
Following are the required resources for using Cytoscape:

Hardware: A computer having CPU power 1GHz or greater, high quality graphics card, hard disk space of at least 60 MB or greater, free physical RAM space of at least 512 MB or greater, and minimum screen resolution of 1024 × 768. In actual, the requirements solely depend upon the size of network that is to be imported and to be analyzed. To maintain the network data from online databases, high speed internet connectivity is also required. However, internet connectivity is not required to generate or visualize data from file located locally.

Software: Operating system that allows Cytoscape to operate are Windows, Linux, Mac OS X or any other operating platform that allows Java, any Standard Edition, having version 5.0 or greater. To download the network files, compatible browsers are Mozilla Firefox, Apple Safari and standard Microsoft Internet Explorer (Yeung et al., 2008) [8].

Overview

Flowchart 1. Overview of complete analysis



CHAPTER 2

2 Research Methodology

The methodology has been carried out using different libraries of R to analyze the data retrieved from public repositories mentioned in next section. The whole research is divided into two sections, first: development of scripts; and second: their implementation on the data sets to carry out the results. The data sources and procedures followed are explained in this chapter.

a. Reproducibility

For data processing of large matrices high computational power was required, in order to regulate computational power and time and maintain stability and reproducibility containers using singularity platform (Kurtzer et al., 2017) were built and ran on Julia cluster of the University of Würzburg which primarily operates in a batch mode by using SLURM workload manager (Yoo et al., 2003).

b. Technical Sources

In our study we have used Rstudio (RStudio Team, 2019) with integrated R (R Core Team, 2021) in order to formulate data frames, matrix, and tables at each step to extract, process and visualize information from the data. Various R libraries were used including plyr (Wickham, 2011b), dplyr (Wickham et al., 2015), tidyr (Wickham 2017), splitstackshape (Mahto, 2018), ggplot2 (Wickham, 2011a), data.table (Dowle et al., 2019), reshape (Wickham 2015), stringr (Wickham 2019). Shiny (W. Chang et al., 2015) was used to visualize graphics; ggmap (Kahle & Wickham, 2013) was used for spatial mapping and R Markdown (Baumer et al., 2014) was used for presentations.

SECTION A

2.1 Data Retrieval

With reference to the 1001 genome project (Alonso-Blanco et al., 2016) 10,707,430 biallelic SNPs were discovered in the Arabidopsis genome. This characterizes that on average at every 10 bp there is one variant, showing the densest variant map compared to any other species, that also includes the latest release of 1000 Human Genome project. From 10,709,466 SNPs altogether 1,854,599 SNPs were incurred in the coding regions. Being focused on the non-synonymous SNPs; 28,148 encode for a premature stop codon (PSCs) in the genome sequence that affected 9,999 genes which is one third of the total genes (~30k) in Arabidopsis.

2.2 Detailed Analysis of pre-mat Stop codon

2.2.1 Preparation of data tables

To create our pre-mat stop gained data file, we took gene and SNPs information from publicly available *Arabidopsis thaliana* annotation file (attached in the soft copy). We integrated accession information from snpeff vcf file. The start and stop position of the genes were extracted from Tair10 available at https://www.arabidopsis.org. To extract allele frequency information from vcf.snpeff.gz vcftools version v4.1 (Danecek et al., 2011) was used.

The final data table including all SNPs that acquired premature stop codons in Arabidopsis genome is available in soft copy as STOP_GAIN_TABLE. It has multiple columns with SNPs information including genes, chromosome number, start position of SNP, stop position of SNP, No. of Alleles, Allele Frequency, reference start position, reference stop position and accession IDs.

2.2.2 Data filtration

Data frame 'STOP_GAIN_TABLE' based on SNPs was filtered considering certain factors:

1. We filtered the table for SNPs knocked out in less than 2 accession IDs.

2. Next, we minimized the rest of the data set by removing SNPs with more than 1 alt allele instead only biallelic genes were selected for the study.
3. Accession Ids with missing genotype ('./.') were also removed from the data table.

To get insights about the statistics of PSCs gained on gene and accession ids level we formulated some questions:

2.3 What is the relative position of SNPs on genome?

We calculated the relative position of the occurrence of premature stop codon in each gene in order to find out if they were more inclined towards start or stop of the gene. We retrieved start and stop position of the gene from STOP_GAIN_TABLE. The following formula was used to calculate the relative position:

Total distance = Stop position – Start position

Relative position = Position on genome / Total distance

Where position on genome was calculated by using start and stop position of the gene and position of SNP on the gene respective to the '+' or '-' orientation as given in tair10 file.

2.4 How many premature stops were gained by each gene?

At first, having insights about the data and an overall review, we were interested to find out the frequency of occurrence of premature stop codons in each gene in which they were knocked out.

We calculated the number of occurrences of PSCs in each of all 9,999 genes. The script first generated a table from our main STOP_GAIN_TABLE and extracted SNP and gene information. Table function from data.table (Dowle et al., 2019) package was used to count total number of occurrences of PSCs in each gene.

2.5 How many premature stops were gained by each accession ID?

The gene centric table provided the statistics only about PSCs gained in each gene, however it was still unclear that how many knock-outs were found in each accession Id of *Arabidopsis thaliana*. Moreover, it was also important to know which geo location had acquired maximum number of premature stop codons.

To calculate the number of PSCs gained by each accession Id, we took gene and accession information from STOP_GAIN_TABLE, listed all accessions together with their corresponding genes in which they were knocked out by using the functions 'summarize' and 'collapse' from dplyr library (Wickham et al., 2015). 'cSplit' function again from dplyr (Wickham et al., 2015) library was used to split all accession Ids in the list in each row and shaped the table in long direction resulting in a table with 1 accession Id per row with its corresponding gene. Altogether 9999 rows (total number of genes) were expanded to 716199 rows. Duplicate rows were removed from the data frame by using 'unique' function. The table was then collapsed again respective to the total number of accession Ids subsequently creating 1135 rows. Occurrences of number of PSCs in each accession were calculated by using 'table' function from data.table library (Dowle et al., 2019). List of genes which got premature stop codon in each particular accession Id were also collapsed into each of 1135 rows. Lastly, accession centric table was generated including 1135 accession Ids, total number of premature stop codons in each accession Id and the genes which gained those stop codons.

SECTION B

Analysis of Differential Expression data and Co-occurrence of high confidence down regulated genes

2.6 RNA seq data of 727 accessions

To interrogate the functionality of premature stop codons, we accessed RNA-seq expression data from (Kawakatsu et al., 2016) which was available publicly for 727 accessions out of 1135 Arabidopsis accessions. The number of accessions were further reduced to 664 when gene data from (1001 Genomes) was incorporated. (There were 62 genes in RNA-seq data from (Kawakatsu et al., 2016) that were absent in Tair10 (Lamesch et al., 2012) data, hence we removed those genes).

We eliminated SNPs with low minor allele frequency (MAF) hence the data set was further reduced. Furthermore, we removed SNPs with missing genotype data in the snpeff annotation and it resulted in the reduction of data set to 4 times approximately.

2.7 Extracting differentially expressed genes from RNA-seq data

We filtered RNA-seq expression data set in 2 steps: 1). First, the matrix was filtered for the genes which overlap with the tair10 data. 2). Next, the accessions which were absent in tair10 data were removed. Proceeding further with formation of expression data table, we obtained SNP, Gene, and accession Id information from STOP_GAIN_TABLE. Then we extracted the gene list from the table and selected unique set of genes from the list by using the *unique()* function of base R (R Core Team, 2021). We then transposed the expression data matrix in order to make a smaller subset of the 727-expression data set by using *subset()* function of library dplyr (Wickham et al., 2015). To filter the accession IDs common in tair10 data and transcription data we used the table expression data table and separated all accession Ids knocked out in one SNP to each in a single row by applying *separate_rows()*

function from dplyr (Wickham et al., 2015). As a result, a data-frame of 312,109 rows was formed each with one accession Id. This step assisted in extracting accessions which had an overlap between both data sets. After filtering the rows, we again collapsed all accessions separated by ‘,’ snp-wise by using *summarize()* and *collapse()* from dplyr (Wickham et al., 2015) and merged the filtered accessions column in expression data table. We calculated the mean expression of all knock-out and Wild-type accessions by using *mean()* as R base function (R Core Team, 2021) to collectively quantify the variation of expression.

2.7.1 Independent two-sample t-test

To find out the possible significant difference among the calculated means of knock out and wild type genes, we performed an independent two-sample t-test to obtain the p-value.

t.test (x1, x2, paired = FALSE) $p.value

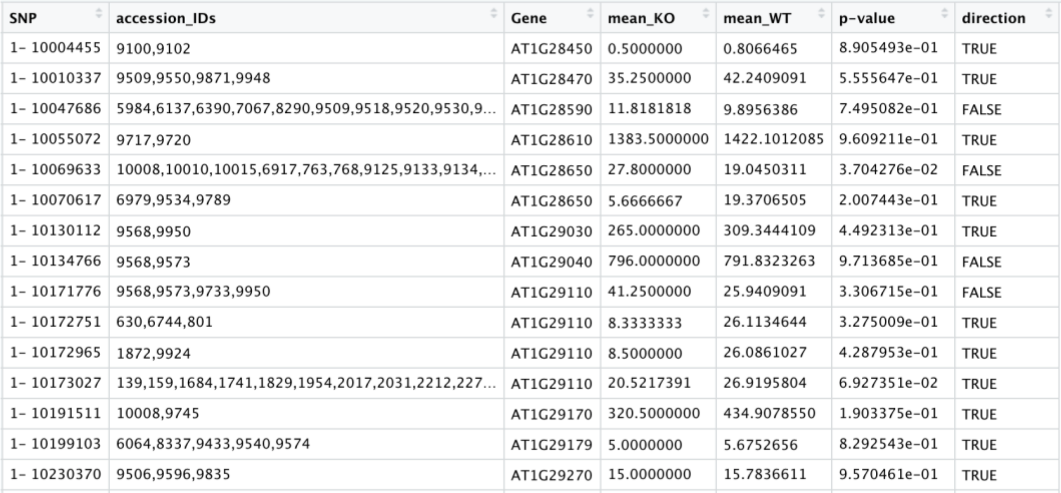
For an unpaired t.test R assumed that the variance of the groups of samples: i.e., mean of knock out and wild type accessions being compared are different. The independent t-test tested null hypothesis that: “There is no difference in the mean expression of knock-out and wild-type accessions”. The significance level of t-test was calculated in the form of p-value. With the p-value obtained less than 0.05, we rejected the null hypothesis, and the result was concluded as statistically significant and therefore the data set was named as significant at Threshold-1 (T1). In addition to setting confidence interval at 0.05 we corrected threshold value 0.05 with Bonferroni multiple testing equation and named it as Threshold-2 (T2). Explanatory Image 2-I shows the expression data table with all essential columns.

2.8 Regulation of Gene expression

We figured the changes in gene expression levels by computing difference in size between mean expression of wild type and knock out accessions. *abs()* function from R Math was applied to return the positive absolute value of expression data. It was assumed that if the mean expression of accessions with wild type allele appeared to be greater than mean expression of knocked out genes, we specified them as downregulated. Simultaneously, if the mean expression of accessions with knocked out genes was higher than in wild type, the SNPs were categorized as upregulated. We indicated the results in our table as TRUE/ FALSE for up and downregulated SNPs respectively.

Processing further with our analysis, we selected the downregulated differentially expressed data set for co-occurrence analysis since downregulated genes are reportedly involved in suppression of gene transcription or in translation of protein leading to either truncated protein or complete loss of function.

Explanatory Image 2-I: Image of differential expression table showing the table format along with mean expression values of Wild type and Knockout accessions.



SNP	accession_IDs	Gene	mean_KO	mean_WT	p-value	direction
1- 10004455	9100,9102	AT1G28450	0.5000000	0.8066465	8.905493e-01	TRUE
1- 10010337	9509,9550,9871,9948	AT1G28470	35.2500000	42.2409091	5.555647e-01	TRUE
1- 10047686	5984,6137,6390,7067,8290,9509,9518,9520,9530,9...	AT1G28590	11.8181818	9.8956386	7.495082e-01	FALSE
1- 10055072	9717,9720	AT1G28610	1383.5000000	1422.1012085	9.609211e-01	TRUE
1- 10069633	10008,10010,10015,6917,763,768,9125,9133,9134,...	AT1G28650	27.8000000	19.0450311	3.704276e-02	FALSE
1- 10070617	6979,9534,9789	AT1G28650	5.6666667	19.3706505	2.007443e-01	TRUE
1- 10130112	9568,9950	AT1G29030	265.0000000	309.3444109	4.492313e-01	TRUE
1- 10134766	9568,9573	AT1G29040	796.0000000	791.8323263	9.713685e-01	FALSE
1- 10171776	9568,9573,9733,9950	AT1G29110	41.2500000	25.9409091	3.306715e-01	FALSE
1- 10172751	630,6744,801	AT1G29110	8.3333333	26.1134644	3.275009e-01	TRUE
1- 10172965	1872,9924	AT1G29110	8.5000000	26.0861027	4.287953e-01	TRUE
1- 10173027	139,159,1684,1741,1829,1954,2017,2031,2212,227...	AT1G29110	20.5217391	26.9195804	6.927351e-02	TRUE
1- 10191511	10008,9745	AT1G29170	320.5000000	434.9078550	1.903375e-01	TRUE
1- 10199103	6064,8337,9433,9540,9574	AT1G29179	5.0000000	5.6752656	8.292543e-01	TRUE
1- 10230370	9506,9596,9835	AT1G29270	15.0000000	15.7836611	9.570461e-01	TRUE

2.9 GOSLIM enrichment of significant up and down regulated genes

We looked for the prospective gene ontology terms of significantly up and downregulated highly expressed genes which encode for premature stop codons. To carry out further analysis, the up-to-date version of ATH_GO annotation text file available publicly at tair https://www.arabidopsis.org was used. GO annotations in the file are based on research literature and electronic sources from Tair, UNIPROT, GO Consortium, IntAct and TIGR for Arabidopsis genes.

ATH_GO file (Berardini et al., 2004) refers to Gene Ontologies of *Arabidopsis thaliana*. The file consists of 15 different columns that explain several attributes associated with each of the GO term in which Arabidopsis genes are categorized. The locus name denotes gene names in standard AGI convention names format. We filtered the ATH_GO file for Arabidopsis genes which gained premature stop codon, GO ID representing them and their corresponding GO Slim terms. Subsequently, we created a subset table for GO slim terms elucidating molecular functions listed in chapter 1 Section 1.2.2.

We performed GO enrichment analysis on both up and downregulated data sets at T1 and T2 in two sequential steps. 1) We calculated total number of genes in whole genome involved in each of the GO slim molecular functions categorized in ATH GO file. 2) The first data set i.e., all the genes in downregulated T1 were categorized according to their annotation in terms of molecular functions. To check the probability of success at confidence level 0.95, we stated the null hypothesis as: "The number of genes categorized in GO Slim terms in both whole genome and highly significant downregulated at T1 are same according to their proportion in population". To experiment this, we implemented binomial test in R and implied resulting p-value at alpha level of significance to reject null hypothesis. The below binomial equation was used:

binom.test (x, y) $p.value

where, x represents number of successes i.e., genes in whole genome,

and y illustrates number of trials.

The resulting p-value showed the significance level of characterization of genes in different GO slim categories in comparison to whole genome. Next, we tabularized GO slim terms, Gene Frequency in each GO slim term and their respective p-values. The same process was repeated simultaneously for highly significant downregulated genes at T2, upregulated genes at T1 and T2.

Table 3: Number of up and down regulated genes in each GO slim term in comparison to Genome.

GO SLIM TERMS	GENOME	DOWN-REG T1	DOWN-REG T2	UP-REG T1	UP-REG T2
CARBOHYDRATE BINDING	186	5	1	10	3
CATALYTIC ACTIVITY	7986	252	62	146	48
CHROMATIN BINDING	201	6	1	3	0
DNA BINDING	2441	30	5	45	16
DNA-BINDING TRANSCRIPTION FACTOR ACTIVITY	1970	14	3	23	7
ENZYME REGULATOR ACTIVITY	530	5	3	7	3
HYDROLASE ACTIVITY	3895	131	39	79	30
KINASE ACTIVITY	2529	87	23	65	23
LIPID BINDING	350	4	0	5	3
MOTOR ACTIVITY	69	2	1	0	0
NUCLEASE ACTIVITY	329	15	3	6	3
NUCLEIC ACID BINDING	1339	20	5	25	7
NUCLEOTIDE BINDING	1216	60	11	53	16
OTHER BINDING	4045	116	32	111	37
OTHER MOLECULAR FUNCTIONS	282	10	5	4	0
OXYGEN BINDING	7	1	1	0	0
PROTEIN BINDING	6895	111	32	90	25
RNA BINDING	2551	41	12	12	2
SIGNALING RECEPTOR ACTIVITY	210	3	2	16	9
SIGNALING RECEPTOR BINDING	124	0	0	1	0
STRUCTURAL MOLECULE ACTIVITY	422	5	1	0	0
TRANSCRIPTION REGULATOR ACTIVITY	186	5	0	2	0
TRANSFERASE ACTIVITY	5842	166	37	108	28
TRANSLATION FACTOR ACTIVITY, RNA BINDING	110	3	1	2	1
TRANSLATION REGULATOR ACTIVITY	10	0	0	0	0
TRANSPORTER ACTIVITY	2444	48	13	23	9
UNKNOWN MOLECULAR FUNCTIONS	9722	217	54	226	86

2.10 Co-occurrence analysis of down regulated genes

To predict functional pairing between the genes we looked if the co-occurrence of these genes deviates from mendelian law of segregation.

We first created a table with all accession Ids knocked out in each of 646 genes. We used *ddply()* function of dplyr library (Wickham et al., 2015) to summarize the list of genes and combined all correlated accession Ids separated by ',' in one cell as a list. The table is attached in soft copy as GENEWISE.Rda.

To find out the co-occurrence of all the genes we created a matrix of 646x646 having number of rows and columns equal to total number of genes in gene-wise table. Column names and row names of matrix were set according to the downregulated gene names saved in a separate list. As the nested script loops through each row and each column of the matrix, it obtains the gene name from the row at designated index; refers back to the gene-wise table and splits the list of corresponding accession Ids using *strsplit()* function of library splitstackshape (Mahto, 2014). Consecutively, it takes the gene name from the respective column, again iterates through the gene-wise table, splits the accession Ids separated by ',' knocked out in that specific gene and simultaneously counts common accession Ids knocked out in both genes.

The matrix with connection values was then reshaped to create a co-occurrence network table. In order to create network table, data from the matrix was selected from upper triangle of the matrix with values in diagonal set as false. We used 2D linear indexing operation from MATLAB in R to extract the values from co-occurrence matrix and converted into a table. To filter the most accurate and statistically significant connections, we removed all the gene pairs within 100kb distance on the same chromosome and also on different chromosomes to rule out the fact that the potential connection might be due to the closest proximity of both genes. Later, tair10 data was merged including chromosome number, start and stop position corresponding to both genes. We used R library matrixstats (Bengtsson, 2014) to apply functions *rowMaxs()* and *rowMins()* for calculating the actual distance between co-occurring gene pairs according to their start and stop position. Later we filtered the table and excluded the genes within 100kb on same and different chromosomes on the basis of calculated distance. Over and under-represented p-values were calculated with deviation from higher or lower observed frequency. To investigate the significance of co-occurrence of genes, we applied the hypergeometric distribution method (Johnson et al., 1992) including *qhyper()* and *dhyper()* functions from R Hypergeometric{stats} package to calculate the p-value. The following formula was applied:

qhyper(p, x, n, y)

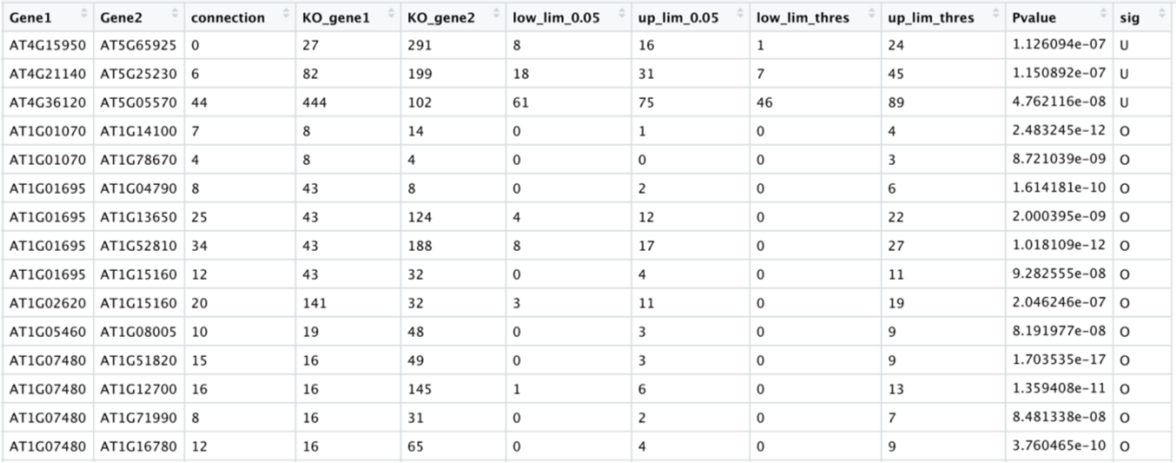
qhyper(threshold, x, n, y, lower.tail=FALSE)

dhyper(connection, x ,n ,y)

where p is threshold 0.05, x is total number of accession Ids knocked out in gene 1, n is (total number of accessions) – (knocked out in gene 1), y is total number of accession Ids knocked out in gene 2, and set lower.tail to FALSE.

We calculated lower limit and upper limit as 0.05 and multiple testing threshold (0.05/nrow(df)) respectively by *qhyper ()* and then acquired significance of the connection in the form of p-value by *dhyper ()* function. With this p-value we determined over and under-represented connections between the gene pairs from the expected values. Highly significant under-represented and over-represented connections were filtered with threshold based on Bonferroni correction/multiple testing (HOCHBERG, 1988) to the alpha level of 0.05 to rule out the possibility of predominantly false positive significant p-value results due to large population size. The image of final co-occurrence table is shown in explanatory image 2-II.

Explanatory Image 2-II: Screenshot of co-occurrence table showing highly significant over and underrepresented co-expressed gene pairs.



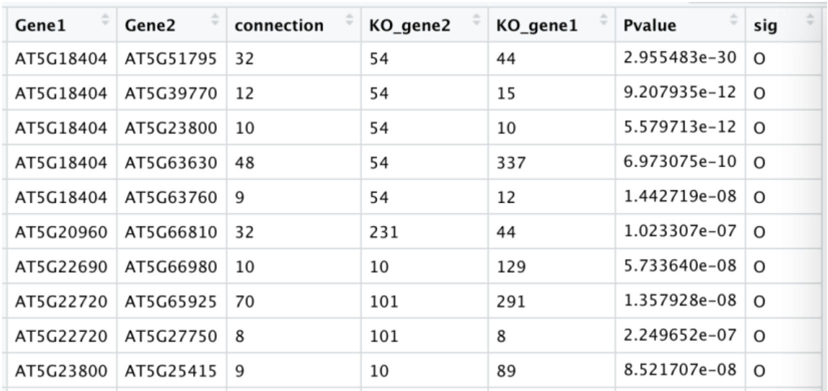
Gene1	Gene2	connection	KO_gene1	KO_gene2	low_lim_0.05	up_lim_0.05	low_lim_thres	up_lim_thres	Pvalue	sig
AT4G15950	AT5G65925	0	27	291	8	16	1	24	1.126094e-07	U
AT4G21140	AT5G25230	6	82	199	18	31	7	45	1.150892e-07	U
AT4G36120	AT5G05570	44	444	102	61	75	46	89	4.762116e-08	U
AT1G01070	AT1G14100	7	8	14	0	1	0	4	2.483245e-12	O
AT1G01070	AT1G78670	4	8	4	0	0	0	3	8.721039e-09	O
AT1G01695	AT1G04790	8	43	8	0	2	0	6	1.614181e-10	O
AT1G01695	AT1G13650	25	43	124	4	12	0	22	2.000395e-09	O
AT1G01695	AT1G52810	34	43	188	8	17	0	27	1.018109e-12	O
AT1G01695	AT1G15160	12	43	32	0	4	0	11	9.282555e-08	O
AT1G02620	AT1G15160	20	141	32	3	11	0	19	2.046246e-07	O
AT1G05460	AT1G08005	10	19	48	0	3	0	9	8.191977e-08	O
AT1G07480	AT1G51820	15	16	49	0	3	0	9	1.703535e-17	O
AT1G07480	AT1G12700	16	16	145	1	6	0	13	1.359408e-11	O
AT1G07480	AT1G71990	8	16	31	0	2	0	7	8.481338e-08	O
AT1G07480	AT1G16780	12	16	65	0	4	0	9	3.760465e-10	O

2.11 Calculation of gene co-expression coefficient and building a Gene- gene co-occurrence network

To measure the strength of the co-occurrence we calculated the co-expression coefficient from Kendall Tau method in R. Several steps were involved where each one is explained in detail below.

We first filtered the essential columns from co-occurrence table i.e., information of both data sets (gene names), number of connections, number of accessions where gene 1 was knocked out and number of gene 2 knock-out accessions, their significance of co-occurrence (p-value) and the information about over or under-representation as presented in Explanatory Image 2-III.

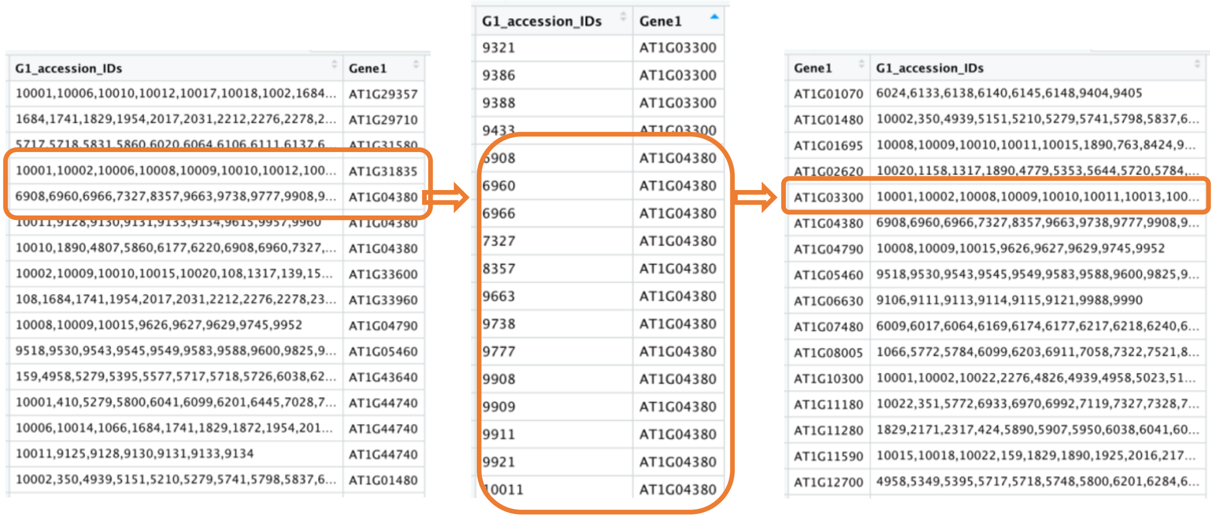
Explanatory Image 2-III: Filtered co-occurrence table.



Gene1	Gene2	connection	KO_gene2	KO_gene1	Pvalue	sig
AT5G18404	AT5G51795	32	54	44	2.955483e-30	O
AT5G18404	AT5G39770	12	54	15	9.207935e-12	O
AT5G18404	AT5G23800	10	54	10	5.579713e-12	O
AT5G18404	AT5G63630	48	54	337	6.973075e-10	O
AT5G18404	AT5G63760	9	54	12	1.442719e-08	O
AT5G20960	AT5G66810	32	231	44	1.023307e-07	O
AT5G22690	AT5G66980	10	10	129	5.733640e-08	O
AT5G22720	AT5G65925	70	101	291	1.357928e-08	O
AT5G22720	AT5G27750	8	101	8	2.249652e-07	O
AT5G23800	AT5G25415	9	10	89	8.521707e-08	O

Due to the gain of more than one SNP in single gene and since accession Ids were listed SNP-wise in differentially expressed downregulated data table (as shown previously in explanatory image 2-I), therefore, to convert the data structure in gene-wise format another table with only accession and gene information was required. For that, first, unique gene list 1 with (332 genes) and list 2 (331 genes) was obtained from co-occurrence table and stored in the form of vector in two different variables. With these lists the accession information was extracted from downregulated differential expression table by using *filter()* from dplyr (Wickham et al., 2015) (explanatory image 2-IV-a). The aim was to combine fragmented accession Ids in multiple rows along with the genes into single unique list. So, all the accessions in each row were split using *separate_rows()* function also from library dplyr (Wickham et al., 2015) which resulted in a long table with each row having one accession ID along with the respective knocked-out gene (explanatory image 2-IV-b), than *ddply()* function was applied where *summarize()* and *collapse()* functions joined all the unique accessions into a single list again (explanatory image 2-IV-c).

Explanatory Image 2-IV: Different form of data structures during processing of common accession Ids data.



(a)

G1_accession_IDs	Gene1
10001,10006,10010,10012,10017,10018,1002,1684...	AT1G29357
1684,1741,1829,1954,2017,2031,2212,2276,2278,2...	AT1G29710
5717,5718,5831,5860,6020,6064,6106,6111,6137,6...	AT1G31580
10001,10002,10006,10008,10009,10010,10012,100...	AT1G31835
6908,6960,6966,7327,8357,9663,9738,9777,9908,9...	AT1G04380
10011,9128,9130,9131,9133,9134,9615,9957,9960	AT1G04380
10010,1890,4807,5860,6177,6220,6908,6960,7327,...	AT1G04380
10002,10009,10010,10015,10020,108,1317,139,15...	AT1G33600
108,1684,1741,1954,2017,2031,2212,2276,2278,23...	AT1G33960
10008,10009,10015,9626,9627,9629,9745,9952	AT1G04790
9518,9530,9543,9545,9549,9583,9588,9600,9825,9...	AT1G05460
159,4958,5279,5395,5577,5717,5718,5726,6038,62...	AT1G43640
10001,410,5279,5800,6041,6099,6201,6445,7028,7...	AT1G44740
10006,10014,1066,1684,1741,1829,1872,1954,201...	AT1G44740
10011,9125,9128,9130,9131,9133,9134	AT1G44740
10002,350,4939,5151,5210,5279,5741,5798,5837,6...	AT1G01480

(b)

G1_accession_IDs	Gene1
9321	AT1G03300
9386	AT1G03300
9388	AT1G03300
9433	AT1G03300
6908	AT1G04380
6960	AT1G04380
6966	AT1G04380
7327	AT1G04380
8357	AT1G04380
9663	AT1G04380
9738	AT1G04380
9777	AT1G04380
9908	AT1G04380
9909	AT1G04380
9911	AT1G04380
9921	AT1G04380
10011	AT1G04380

(c)

Gene1	G1_accession_IDs
AT1G01070	6024,6133,6138,6140,6145,6148,9404,9405
AT1G01480	10002,350,4939,5151,5210,5279,5741,5798,5837,6...
AT1G01695	10008,10009,10010,10011,10015,1890,763,8424,9...
AT1G02620	10020,1158,1317,1890,4779,5353,5644,5720,5784,...
AT1G03300	10001,10002,10008,10009,10010,10011,10013,100...
AT1G04380	6908,6960,6966,7327,8357,9663,9738,9777,9908,9...
AT1G04790	10008,10009,10015,9626,9627,9629,9745,9952
AT1G05460	9518,9530,9543,9545,9549,9583,9588,9600,9825,9...
AT1G06630	9106,9111,9113,9114,9115,9121,9988,9990
AT1G07480	6009,6017,6064,6169,6174,6177,6217,6218,6240,6...
AT1G08005	1066,5772,5784,6099,6203,6911,7058,7322,7521,8...
AT1G10300	10001,10002,10022,2276,4826,4939,4958,5023,51...
AT1G11180	10022,351,5772,6933,6970,6992,7119,7327,7328,7...
AT1G11280	1829,2171,2317,424,5890,5907,5950,6038,6041,60...
AT1G11590	10015,10018,10022,159,1829,1890,1925,2016,217...
AT1G12700	4958,5349,5395,5717,5718,5748,5800,6201,6284,6...

This process was simultaneously repeated for gene 2. After getting both lists of accession Ids we wanted to list down the common accession Ids where both genes were knocked out. By using a for loop each row of both columns (G1 accession IDs) and (G2 accession IDs) was accessed and simultaneously unlisted, split and stored in 2 different variables. *Intersect()* function from base R was applied and obtained mutual accession Ids were stored in a third variable for each row. The loop resulted in a collective table with a separate column of combined accession Ids.

Next step was to convert each row of the data frame into a separate table to distinguish concordant and discordant pairs on the basis of which Kendall's rank coefficient was calculated. We applied binary method (0,1); 1 for knock out and 0 for wild type for both gene1 and gene2 data. Sorting of accession Ids in ascending order in the list was a crucial step in order to get same results in each row for both datasets (gene1 and gene2). The concept behind concordant and discordant pairs is explained in the form of ranks, e.g., we call a pair Concordant when the value of the subject (accession Id) is high in both variables and the pair is discordant if the value of the subject is higher in one variable and lower in the other.

The formula to calculate *Kendall's Tau* (τ) is:

$$\tau = (\text{Con - Discon}) / (\text{Con + Discon})$$

where:

Con is the total number of concordant pairs and

Discon represents the number of discordant pairs

To calculate *tau* score in R the formula applied was:

cor.test (x, y, method = 'kendall')

where:

x represents gene1 data set and

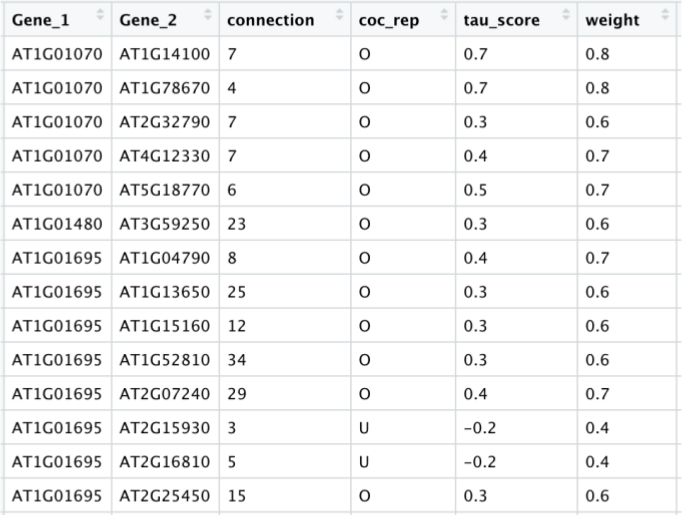
y represents gene2 data set.

cor.test() function provided results in the form of a *list* including p-value and z score which were converted into *data.frame* using *do.call()* function where all rows were combined using rbind. After calculation of tau score, weight of co-expression was calculated to generate a signed network. Signed network showed the direction of the relationship either on the positive or the negative side. To calculate we used the formula as proposed by:

$$0.5+0.5 * (\text{Con - Discon / Con + Discon})$$

Next, after computing all required values, gene co-expression network was evaluated and visualized in Cytoscape (Shannon et al., 2003). The data set was loaded as "network from table". In the data-frame each row had 2 nodes and the edge attributes i.e., their connection strength (*tau* score) and relationship as over or underrepresented as shown in explanatory image 2-V below.

Explanatory Image 2-V: Network table loaded in Cytoscape



Gene_1	Gene_2	connection	coc_rep	tau_score	weight
AT1G01070	AT1G14100	7	O	0.7	0.8
AT1G01070	AT1G78670	4	O	0.7	0.8
AT1G01070	AT2G32790	7	O	0.3	0.6
AT1G01070	AT4G12330	7	O	0.4	0.7
AT1G01070	AT5G18770	6	O	0.5	0.7
AT1G01480	AT3G59250	23	O	0.3	0.6
AT1G01695	AT1G04790	8	O	0.4	0.7
AT1G01695	AT1G13650	25	O	0.3	0.6
AT1G01695	AT1G15160	12	O	0.3	0.6
AT1G01695	AT1G52810	34	O	0.3	0.6
AT1G01695	AT2G07240	29	O	0.4	0.7
AT1G01695	AT2G15930	3	U	−0.2	0.4
AT1G01695	AT2G16810	5	U	−0.2	0.4
AT1G01695	AT2G25450	15	O	0.3	0.6

The genes were defined as unique nodes (altogether 333 nodes) and their connection strength was based on *tau* score. Furthermore, their direction of relationship positive or negative was defined by the signed weight of the edges (below 0.5 was considered as negative). Size of the nodes in all modules was based on the degree of the node where degree was defined as total number of connections the gene was involved in. Different visualization styles available in Cytoscape were used to explore the network and analytics tools were used to perform the clustering analysis e.g., MCODE was used to identify the central clusters in the network.

MCODE algorithm splits the processing in three steps:

I. Weight determination: scoring is based on the number of nodes interconnected where highest score is given to the most interconnected node.

II. Cluster prediction: the algorithm starts with hub gene (highest weighing score) recursively going out while adding more connected nodes with weight score about the set threshold.

III. Post processing: in this step filter i.e., haircut or fluffy is applied to improve the quality.

Next, we wanted to investigate the geographical pattern of highly significant co-occurring gene pairs, to find out how geographic location and genetic variability correlate with each other.

Geographical coordinates of all 664 accessions in the form of latitude (lat) and longitude (long) were downloaded from (1001 Genomes). Despite availability of the geographical coordinates for 1135 accessions all over the world, we selected 664 accessions in continuation to the highly significant cooccurrence data set. Due to random, and unequal distribution of accessions outside of Europe, only accessions from Europe were selected for further analysis, thus, the total number was reduced to 553 geo-localized accessions.

The highly significant co-occurring gene pairs were selected for the analysis to find out their physical location, origin, the pattern of co-occurrence and their correlation with genome wide association studies (GWAS). We applied imputation method on our input dataset to create a matrix using binary numbers where the imputation method is described as: ‘A statistical inference method that replaces the missing value with an attributable value which retains information and the overall structure of experiment’. It was applied on the data set in the form of 1 and 0 for co-occurring genes knocked out in set of common accessions or the wild type respectively. Table 4 shows the section of matrix with binary values and accessions.

Table 4: Section of gene-gene matrix with corresponding knocked out and wildtype accessions. Rows represent the accession IDs while columns constitute co-occurring gene pairs.

	AT1G1270_AT1G13650	AT1G13650_AT1G21210	AT1G12700_AT1G51820	AT1G20350_AT1G51820	AT1G12700_AT1G60070
108	0	0	0	0	0
265	1	0	0	0	1
997	0	0	0	0	0
1002	0	0	0	0	0
1066	1	1	0	0	1
1317	0	0	1	1	0
4779	1	0	1	0	1
4807	0	0	0	0	0
4939	0	0	0	0	1
4958	0	0	0	0	1
5023	1	0	1	0	1
5104	0	0	0	0	0

After formation of matrix, the analysis entails 4 sequential steps:

1. Minor allele count (MAC) was calculated aiming to pre-filter the dataset for gene-gene pairs having common MAC more than 1. This step reduced the data set to 1114 gene pairs.

2. Next step was calculation of mean latitude and longitude of all knock out and wild type accessions in filtered dataset and computing distance between them. R function *tapply()* (Becker et al., 1988) was applied on the data table to calculate the mean latitude and longitude and to compute distance, *distm()* function of library geosphere was applied.

   Distm (c (x1, y1), c (x0, y0), fun = distHaversine)

   where, x represents the longitude value while y acts for the latitude of knock out and wild type accession as 1 and 0 respectively. Moreover, *distHaverstine()* is the R function that compute shortest distances between two points also known as great circle distances by assuming earth in spherical trigonometry. It works in R behind the formula:

   haverstine (p1, p2, r)

   p1 is the longitude/latitude in the form of vector of two numbers, (first as longitude and second as latitude of gene-pairs with knock out accessions), similarly p2 for wild type accessions. r is the radius of earth (default value = 6378137 m)

3. To check whether the statistically significant clustering of co-occurring gene pairs at given geographical locations occurs and if the data is normally distributed or not, as a 3rd step in the analysis Peacock test (Massey Jr, 1951; Peacock, 1983) was applied. It is a Kolmogrov-Smirnov test for two-dimensional space wherein test statistics are calculated based on Monte Carlo method. Peacock2 (Xiao, 2017) package is developed in R in which two-sample test statistic method in multidimensional space given by Peacock in 1983 is implemented. The function *Peacock2()* required longitude and latitude values

of both knock out and wild type accessions in the form of a matrix as a prerequisite. The function iterated through each row of co-occurring gene-gene pairs, grouped the latitude and longitude values of wild type {0} and knock out {1} accessions and simultaneously stored the values in matrix x and y respectively. The resulting D-statistics value was stored in a list. The formula below was used to calculate D statistics.

Peacock2(x,y)

where ***x = m[df[[cols[i]]] == 0,]*** and ***y = m[df[[cols[i]]] == 1,]***

m is the matrix constituting latitude and longitude values; df is the data table; cols represent the column names of data table representing co-occurring genes and i is the loop index.

As the for loop was finished and in an attempt to unlist the resulting values along with their co-related gene pair, list was converted to a *data.frame* using *sapply()* function from base R that converted the *list* to *matrix* and *data.frame* converted *matrix* to a *dataframe*. c was used to concatenate the results.

data.frame(sapply(results,c))

4. In the end, to standardize the values and to assess the significance of D-statistics, z-score for two-sample test was calculated using the formula below:

$$Z = \sqrt{}\ ((n1 * n2) / (n1+ n2)) * D$$

n1 = difference between population - sample (n-m)

where, n is total population of accession Ids i.e., population size.

m is MAC.

n2 is sample size of mutated alleles,

and D is the d-statistic value derived from peacock test.

After computing z-scores describing the variance from the mean value in terms of standard deviation, p-value was calculated intending to explicitly accept or reject the null hypothesis. To calculate p-value from z-score a built-in R function *pnorm()* was used which determines the 'cumulative density function' also known as CDF. The standard formula for (normal distribution) *pnorm()* in R is:

pnorm (x, mean = 0, sd = 1)

However, to compute p-values from z-scores we used the formula below:

2*pnorm(q)

Here, q in the equation denotes z-score. In *pnorm()* function, standardized mean μ = 0 and standard deviation σ = 1 are default parameters and since z-scores are already standardized normally distributed values, therefore, only q as z-score was given as input.

The flow chart below explains the step-by-step process in spatial location analysis of high confidence co-occurring gene pairs.

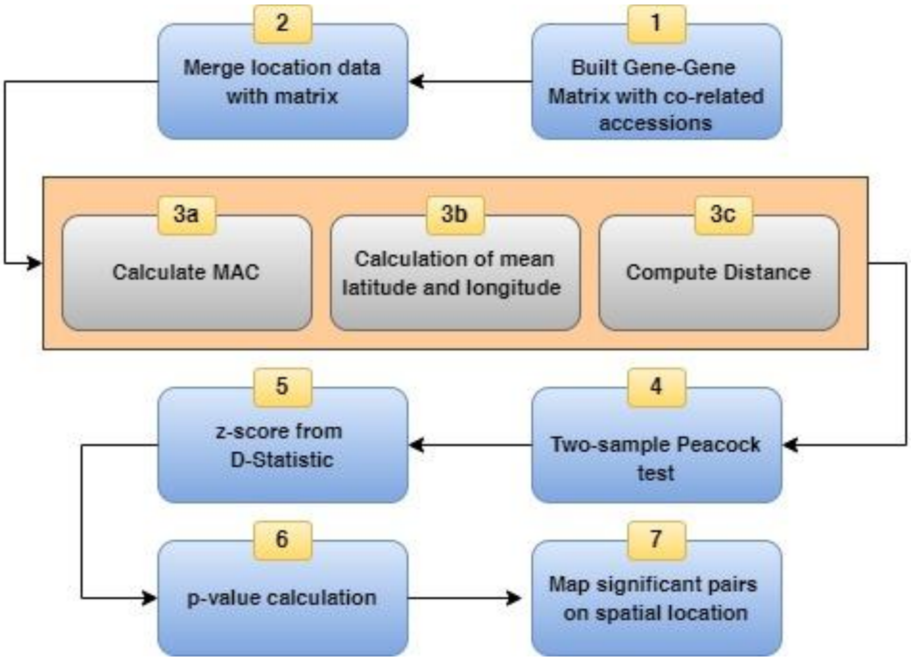

1
Built Gene-Gene Matrix with co-related accessions
2
Merge location data with matrix
3a
Calculate MAC
3b
Calculation of mean latitude and longitude
3c
Compute Distance
4
Two-sample Peacock test
5
z-score from D-Statistic
6
p-value calculation
7
Map significant pairs on spatial location

Table 5 below shows segment of the data.frame with calculated MAC, mean latitude, mean longitude of gene pairs with both wild type and knock out accessions, their distance, D-statistics, z-scores and p-values.

GENE1-GENE2	MAC	MEAN LAT 0	MEAN LAT 1	MEAN LONG 0	MEAN LONG 1	DISTANCE	D STATS	Z SCORE	P_VALUE
AT1G03300_AT3G58050	211	47,46	46,3	9,02	3,5	439419,819	0,37	4,23	2,35E-05
AT1G20350_AT2G32790	17	46,55	60,75	6,5	17,1	1723690,65	0,93	3,79	0,00015094
AT1G20350_AT2G40050	16	46,53	62,11	6,57	15,27	1819808,27	0,96	3,77	0,00016203
AT1G43640_AT5G66630	50	47,2	45,08	7,66	-1,09	714184,178	0,62	4,19	2,84E-05
AT1G43640_AT2G29710	44	47,28	43,89	7,58	-1,48	800176,923	0,65	4,14	3,45E-05
AT2G15930_AT2G27340	23	47,04	46,18	7	3,23	303589,093	0,34	1,59	0,11126483
AT2G15930_AT2G29710	132	47,58	45,24	8,51	1,76	579818,098	0,43	4,27	1,94E-05
AT2G15930_AT4G34460	72	46,92	47,51	7,8	0,69	541121,882	0,51	4,06	4,98E-05
AT2G16365_AT2G41710	2	47,01	45,97	6,87	-1,53	653878,286	0,87	1,22	0,22139829
AT2G18320_AT2G36815	2	46,97	55,74	6,81	14,13	1099982,01	0,93	1,32	0,18733056
AT2G18700_AT4G15310	163	48,05	44,63	8,53	2,98	571140,326	0,43	4,53	5,84E-06
AT2G18700_AT5G08030	41	47,47	41,34	7,46	-0,7	941572,487	0,76	4,69	2,75E-06
AT3G08947_AT5G59510	18	46,49	61,62	6,46	17,51	1826564,67	0,97	4,02	5,71E-05
AT3G23610_AT4G09490	41	47,46	41,51	7,5	-1,17	953821,126	0,75	4,61	3,95E-06
AT3G23610_AT5G08030	35	47,4	41,3	7,39	-1,11	958405,122	0,76	4,32	1,58E-05
AT3G23610_AT5G10800	50	47,45	42,71	7,56	-0,17	803182,44	0,68	4,56	5,06E-06
AT3G27330_AT3G27700	31	47,28	42,56	7,24	0,24	761142,334	0,71	3,84	0,00012378
AT3G28610_AT4G09490	49	47,3	44,07	7,64	-1,08	766637,386	0,69	4,6	4,13E-06
AT4G09012_AT5G66980	75	45,86	53,99	5,98	12,06	1003937,95	0,63	5,08	3,70E-07
AT4G09490_AT4G11521	30	47,07	45,82	7,33	-1,37	681209,899	0,67	3,54	0,00039932
AT4G09490_AT4G18840	7	47,06	42,93	6,9	1,89	605764,93	0,71	1,88	0,06036947
AT4G10200_AT4G34460	8	46,96	49,84	6,84	6,51	321308,065	0,56	1,58	0,11425823
AT4G10200_AT5G51795	2	46,96	57,42	6,85	4,02	1180187,28	0,98	1,38	0,16687456
AT4G13810_AT5G25910	3	46,98	51,04	6,88	-1,68	770511,592	0,92	1,59	0,11103501
AT4G36650_AT5G45820	2	47,01	45,01	6,85	3,18	360233,596	0,72	1,01	0,31114967
AT4G37460_AT5G09700	50	46,67	50,23	6,72	7,98	407822,622	0,45	3,03	0,00248391
AT4G37460_AT5G66810	9	46,98	48,52	6,71	13,99	570592,736	0,63	1,87	0,0610278
AT5G01330_AT5G10800	34	47,36	41,76	7,27	0,45	824404,152	0,79	4,48	7,56E-06

CHAPTER 3

3 Results

This chapter presents all the results obtained at different sequential steps during the whole analysis.

3.1 Basic understanding of Arabidopsis genomic data

The publicly available Arabidopsis genome dataset from 1001 genome project revealed that 28,148 non-synonymous SNPs instigated premature stop codons (PSCs) in the genome sequence that affected 9,999 genes which, on average estimate for one third of the total genes in *Arabidopsis thaliana* (~30,000). The density map with SNPs distributed over distance of 100 kbs shows the uniform distribution of PSCs across the genome in comparison to whole genome within all 5 chromosomes as shown in figure 1. No specific pattern is observed in this step except there are less SNPs in centromeric region which is also expected.

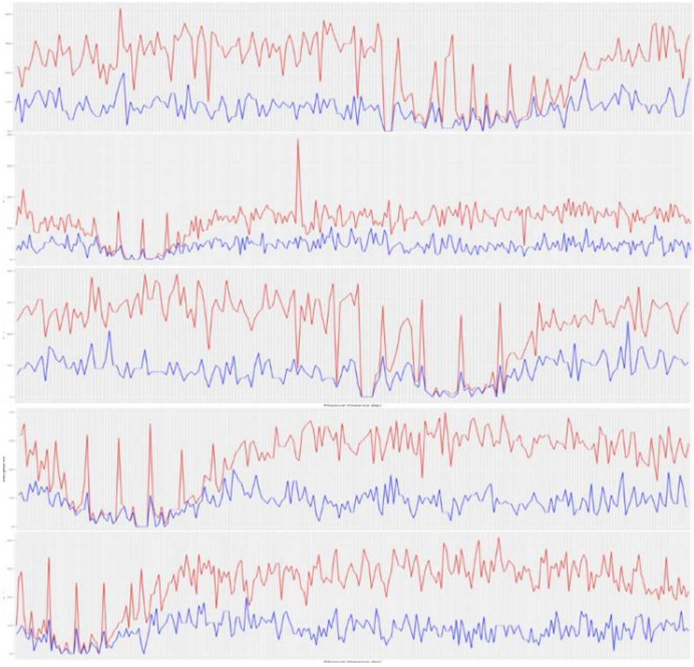


Figure 1: Density map of occurrence of pre-mat stop codon in all 5 chromosomes of Arabidopsis genome at 100 kb distance. Blue lines represent pre-mat codon while red lines indicate the whole genome for comparison.

3.2 How are the PSCs distributed in the Genome according to their relative position?

The distribution of premature stop codons (PSCs) respective to start and stop position in the gene across the genome displayed a W-shaped normalized uniform pattern with equally high number of PSCs gained in both start and stop position and less in the center also indicating centromeric region. We presumed that PSCs originated close to the start position in gene had the potency to produce a non-functional protein due to an incomplete truncated gene sequence. On contrary the genes that had gained PSCs close to the stop position possess less chances of functional loss. Figure 2 shows the distribution of PSCs relative to their start and stop position in gene across the whole genome.

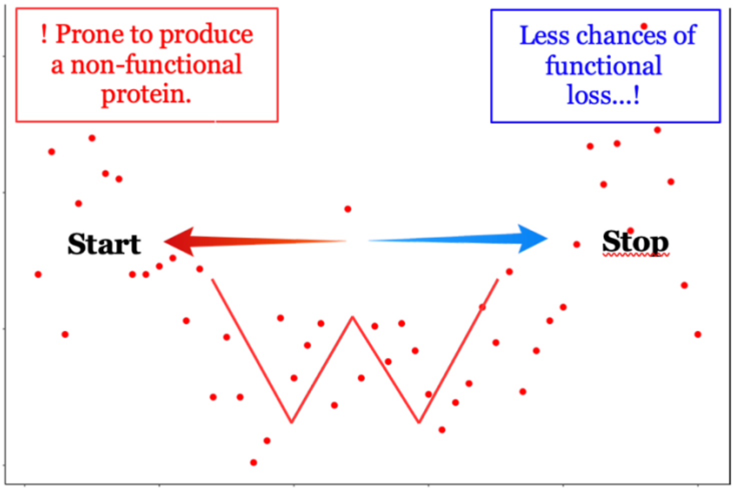


Figure 2: Distribution of pre-mat stop codons (PSCs) in the genes relative to their start and stop position.

3.3 How many premature stop codons were gained by each gene?

We observed that 4,741 out of 9,999 (~ 50%) genes have gained at least 1 PSC, and the other 50% genes have gained multiple PSCs. Interestingly, we found out that a single gene ID AT2G10440 (MED15_2) has gained 59 PSCs at multiple locations which are found to be the highest number of PSCs gained in any Arabidopsis gene. According to functional annotation of Tair10, this gene acts as a mediator of RNA polymerase II transcription subunit. It enables protein binding, chromatin DNA binding and transcription coactivator activity and locates in the nucleus of cell (M. J. Kim et al., 2016). Sequentially, AT3G42723 (ATP binding / aminoacyl-tRNA ligase/ nucleotide binding protein) has gained 54 PSCs, located in plasmodesma (Fernandez-Calvino et al., 2011) however, the biological process is unknown and hence molecular function is also unknown (Ashburner et al., 2000). AT3G43148 (a myosin heavy chain like protein coding gene) has gained 49 premature stop codons. The whole statistics of the number of PSCs gained in each gene are shown in Supplementary Table 1. Figure 3 shows the plot of gene centric summary.

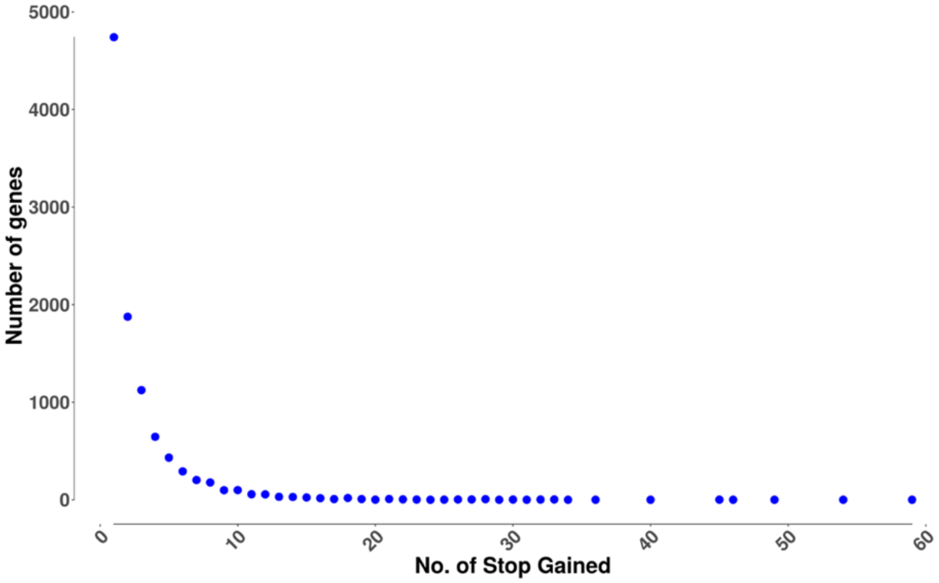


Figure 3: Premature stop codons gained by each gene. On the x-axis the numbers show the total PSCs gained by number the genes on y axis.

3.4 How many premature stop codons were knocked-out in each accession Id?

In all 1135 accessions and 9,999 genes, 790 PSCs were the maximum that had been observed in the accession ID 9533 (IP-Cem-0) (Alonso-Blanco et al., 2016) from Spain suggestive of dry environment gradient and drought conditions. Whereas the lowest number of occurrences of PSCs were 170, observed in accession ID 7208 (Lan-0) from

UK. The zero premature stops in Columbia (ref-0) can also be observed in figure 4. We estimated that on average each accession ID had 630 different genes in which at least one pre-mat stop codon was knocked out also shown in Figure 4. The ACCESSION_CENTRIC.csv table is attached in soft copy.

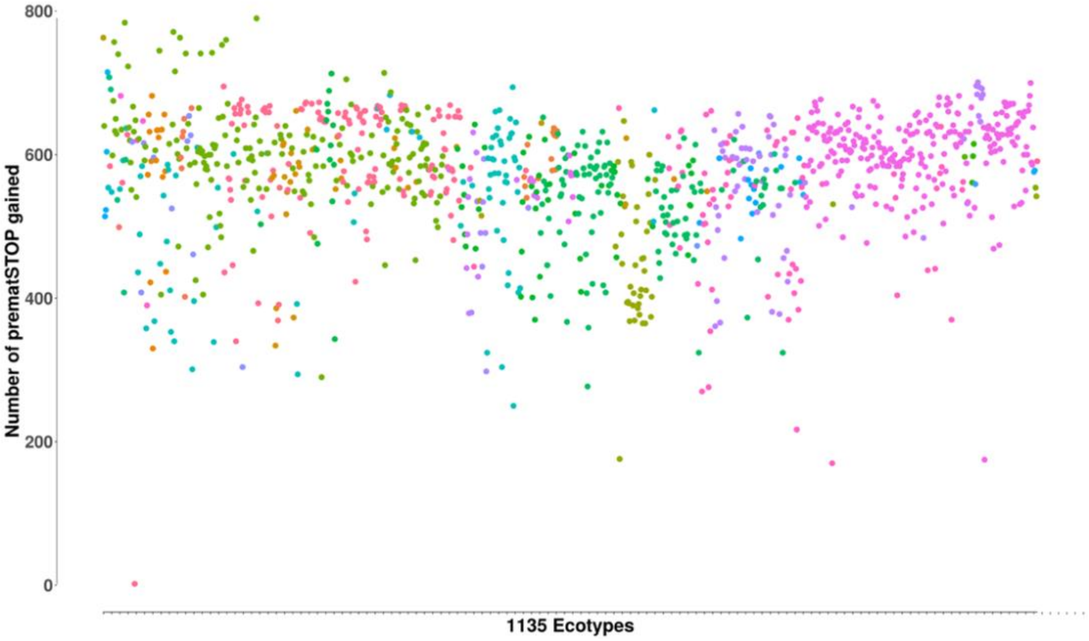


Figure 4: pre-mature stop codons gained by each accession ID. The color palette is used to represent different countries where these accessions were knocked out.

3.5 Filtering results of RNA seq data

This section elucidates the results of all the processes explained in section B chapter 2 after performing detailed analysis of the RNA seq data.

As the tabular results below show, premature stop gained data was filtered from 10 million SNPs and altogether 28,148 SNP markers were filtered out. Furthermore, the number was reduced to 25,063 after keeping only biallelic SNPs in all 1135 accessions. After incorporating transcriptome data from (Kawakatsu et al., 2016) with 727 accessions, 13440 SNPs were left affecting 6363 genes. Moreover, the number of accessions was reduced to 664. Additionally, the SNPs with missing genotype data were excluded which in turn reduced the data set to nearly 4 times leaving behind 7411

SNPs, 3985 genes. Aiming to optimize the data further for statistical analysis we also excluded singletons and remained with 6359 SNPs and 3570 genes.

Table 1: Tabular results of RNA seq data in terms of Differential expression and Regulation of gene expression

	STOP SNPS	GENES	ACCESSIONS	DATASET
PRE-MAT STOP GAIN DATA FROM 1001 GENOME	28148	9999	1135	Comp
AFTER KEEPING ONLY BIALLELIC SNPS	25063	9726	1135	F1
AFTER REMOVAL OF LOW MAF SNPS	13440	6363	664	F2
REMOVAL OF ACCESSIONS WITH MISSING GENOTYPE	7411	3985	664	F3
FILTERED SINGLETONS	6359	3570	664	F4
664 SIG EXPRESSED T1	1407	1141	664	Sig-T1
664 SIG T1 UP	647	562	664	Sig-T1-up
664 SIG T1 DOWN	760	646	664	Sig-T1-down
664 SIG EXPRESSED T2	412	366	664	Sig_T2
664 SIG T2 UP	227	205	664	Sig-T2-up
664 SIG T2 DOWN	185	166	664	Sig-T2-down

The density graph in figure 5 shows the occurrence of PSCs according to the relative distance of 100kb on genome. The bin size was set at 0.01 as a relative distance. It was observed that there were slightly high number of PSCs gained in the start of gene sequence and this pattern remained uniform retaining the peak structure at each filtering step.

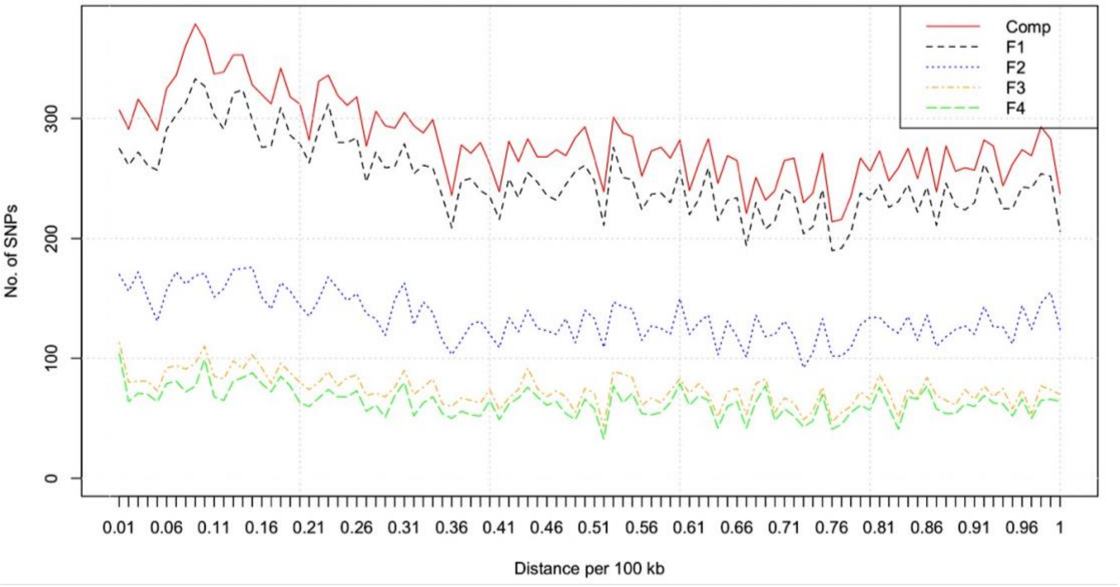
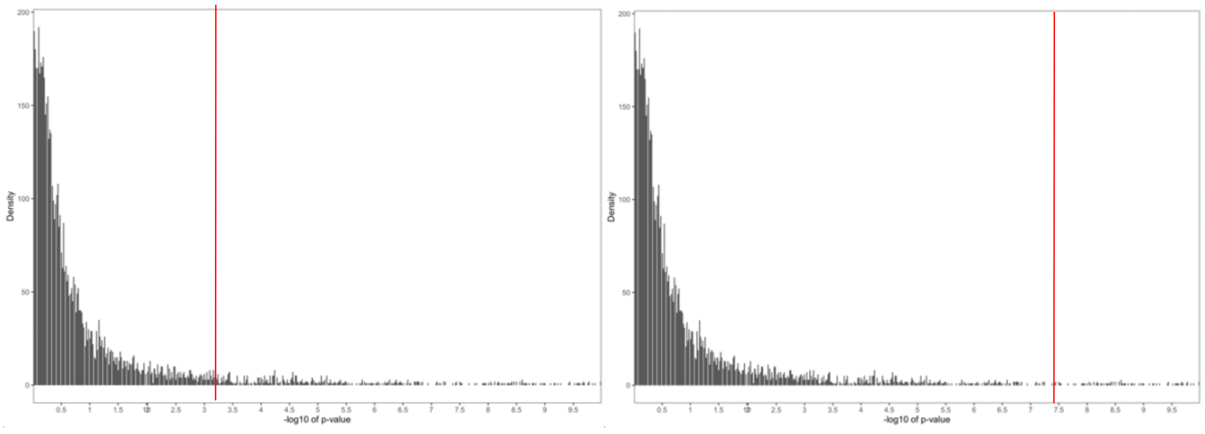


Figure 5: Density graph of occurrence of premature stop codons over the genome according to relative position in all the filtering steps. (Comp: Red line graph represents the total number of premature stops gained in the genome; F1: black dashed line for only biallelic SNPs; F2: Blue dotted line for high MAF SNPs; F3 brown dashed line represents data set with known genotype; and F4: dataset after filtering singletons shown in light green dashed line.

3.6 Regulation of gene expression analysis in terms of up and downregulation

Overall results of RNA expression data in 664 accessions revealed that 22.13% of genes encoding premature stop codon were significantly up and down regulated whereas 77.87% of the expression data showed non-significant results according to T1 (threshold 1) at α level of significance. Bonferroni correction of α significance level indicated 6.4% (412 SNPs in 366 genes) of data as highly significant to be instigating the potential differential expression in the affected genes according to T2 (threshold 2) p-value < $7.86x10^{-6}$. The significant data according to T1 included 1407 SNPs and 1141 genes (Sig_T1). Regulation of the gene expression showed that 54% premature stop codons at T1 (760 SNPs in 646 genes) were found downregulated (Sig_T1_down) whereas 46% (647 SNPs in 562 genes) were upregulated (Sig_T1_up). According to T2 by multiple testing 45.3% premature stop codons (185 SNPs in 166 genes) had led to a reduction in gene expression (Sig_T2_down) whereas 54.7% (227 SNPs in 205 genes) led to a higher expression (Sig_T2_up). Table 1 shows the respective numbers along with their data set symbols.



(a) (b)

Figure 6: P-value plot of SNPs in RNA-seq expression data of 664 accessions plotted against -log10(p-value).

The distribution of -log10 of p-values at T1 and T2 are shown in figure 6-a and 6-b respectively. The red line indicates -log10 of marked threshold value which is $5x10^{-2}$ in (a) and multiple testing Bonferroni correction of in (b) (HOCHBERG, 1988). The plot displayed a skewed curve with substantial difference in the number of SNPs inducing differential expression at each threshold level. It is also observable that only a small portion of p-values are significant at T2 however they are generally lower thus highly significant as compared to T1.

The regulation of gene expression analysis resulted in significantly high number of down regulated SNPs compared to the up regulated at T1. Nevertheless, despite the high number of down regulated genes, it was observed that up regulated genes were knocked out in considerably high number in the start of the genome sequence and the numbers were reduced significantly until the end however in case of genes which were down regulated, the occurrence of PSCs was observed uniform throughout the genome. According to the Bonferroni correction threshold (T2) similar pattern was retained in high confidence up and down regulated genes although the number of SNPs was reduced to almost 1/4th in case of T2_down and 1/3rd in case of T2_up.

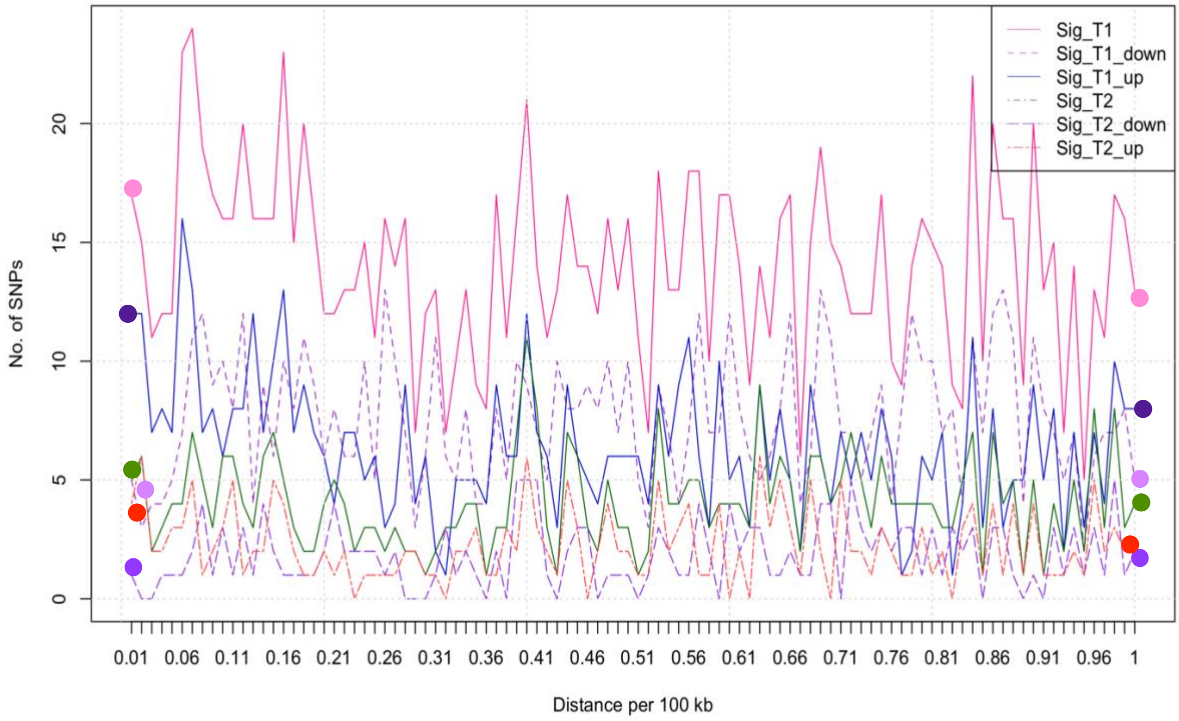


Figure 7: Density graph of occurrence of premature stop codons which are significantly up and downregulated at Threshold 1 (T1) and Threshold 2 (T2) according to relative position on genome.

(Sig_T1: Pink straight line shows significantly expressed SNPs at T1; Sig_T1_down: small, dashed line in lavender represents downregulated SNPs at T1; Sig_T1_up: dark purple straight line illustrates upregulated SNPs at T1; Sig_T2: Green straight line shows significantly expressed SNPs at T2; Sig_T2_down: Large dashed line in violet signifies downregulated SNPs at T2; Sig_T2_up: red dashed line denotes T2 significant upregulated SNPs.

Table 2: Tabular form of Expression data set with calculated mean and p-values. Mean KO indicates average mean expression of all Knock out accessions. Mean WT refers to mean expression of wild type alleles. P-value tells the significance level of expression whereas Direction signals the regulation of differentially expressed genes as either up or downregulated.

SNPs	Accession IDs	Gene	Mean KO	Mean WT	p-value	Direction
1- 10004455	9100,9102	AT1G28450	0.50	0.81	8.91e-01	TRUE
1- 10010337	9509,9550,9871,9948	AT1G28470	35.25	42.24	5.56e-01	TRUE
1- 10055072	9717,972	AT1G28610	1383.50	1422.10	9.61e-01	TRUE
1- 10070617	6979,9534,9789	AT1G28650	5.67	19.37	2.01e-01	TRUE
1- 10130112	9568,995	AT1G29030	265.00	309.34	4.49e-01	TRUE
1- 10134766	9568,957	AT1G29040	796.00	791.83	9.71e-01	FALSE
1- 10171776	9568,9573,9733,9950	AT1G29110	41.25	25.94	3.31e-01	FALSE
1- 10172751	630,6744,801	AT1G29110	8.33	26.11	3.28e-01	TRUE
1- 10172965	1872,992	AT1G29110	8.50	26.09	4.29e-01	TRUE
1- 10191511	10008,97	AT1G29170	320.50	434.91	1.90e-01	TRUE
1- 10199103	6064,8337,9433,9540,9574	AT1G29179	5.00	5.68	8.29e-01	TRUE
1- 10230370	9506,9596,9835	AT1G29270	15.00	15.78	9.57e-01	TRUE
1- 1024866	8337,9653,9659,9964,9981	AT1G03990	224.20	231.02	8.78e-01	TRUE

3.7 What are these significant downregulated genes doing?

GO-Slim enrichment analysis revealed clear differences between the knock out genes in comparison with whole genome. Numerous genes were accompanied with multiple GO slim terms from each of the parent subcategories during the analysis demonstrating their association in different cellular pathways.

In comparison to the whole genome, downregulated genes revealed significant involvement in catalytic activity, hydrolase activity, unknown molecular functions, and other binding whereas there were significantly smaller number of genes involved in protein binding, lipid binding, DNA and RNA binding and DNA-binding transcription factor activity. Similarly, the GO enrichment analysis of upregulated genes exhibited that strikingly increased number of genes were involved in carbohydrate binding, other binding while less genes were involved in protein binding, DNA and RNA binding and transporter activity according to the p-value threshold. It is also noticeable that a significant number of both up and downregulated genes were participating in Kinase activity (Down:13.5% and Up:11.6%) and nucleotide binding (Down:9.2% and Up:9.4%) in comparison to only 8.9% and 4.2% respectively in overall genome. The graphical representation of GO slim enrichment in the form of bubble chart is shown in figures 8 (a and b) and 9 (a and b) for both down and upregulated genes at T1 and T2 respectively. The overall pattern of bubbles (number of genes) in the plot remained persistent at T1 and T2 despite the cutback of dataset to almost 67% due to Bonferroni correction with multiple testing. Evidently, almost 35% of the gene data was annotated into unknown molecular functions. Furthermore, it was also found out that none of the downregulated genes were involved in signaling receptor binding and translation regulator activity. Contrary to this, none of the upregulated genes were found active in oxygen binding, motor activity and structural molecule activity. Only 0.02% of protein coding genes were found functioning in oxygen binding therefore it was witnessed that oxygen binding is non-significant at the lowest number as indicated by red square in both figures 8 and 9. The highest percentage of downregulated genes i.e., 39% were functioning in catalytic activity whilst 40% of upregulated genes had unknown molecular functions. Table 3 below displays all the percentage values of number of genes in each category of GO slim term.

Table 3: GO Enrichment summary table of statistically significant up and downregulated genes at T1 and T2.

GO SLIM TERMS	GENOME	DOWN REG T1	DOWN REG T2	UP REG T1	UP REG T2
CARBOHYDRATE BINDING	0.65	0.77	0.60	1.78	1.46
CATALYTIC ACTIVITY	28.02	39.01	37.35	25.98	23.41
CHROMATIN BINDING	0.71	0.93	0.60	0.53	0.00
DNA BINDING	8.56	4.64	3.01	8.01	7.80
DNA-BINDING TRANSCRIPTION FACTOR ACTIVITY	6.91	2.17	1.81	4.09	3.41
ENZYME REGULATOR ACTIVITY	1.86	0.77	1.81	1.25	1.46
HYDROLASE ACTIVITY	13.67	20.28	23.49	14.06	14.63
KINASE ACTIVITY	8.87	13.47	13.86	11.57	11.22
LIPID BINDING	1.23	0.62	0.00	0.89	1.46
MOTOR ACTIVITY	0.24	0.31	0.60	0.00	0.00
NUCLEASE ACTIVITY	1.15	2.32	1.81	1.07	1.46
NUCLEIC ACID BINDING	4.70	3.10	3.01	4.45	3.41
NUCLEOTIDE BINDING	4.27	9.29	6.63	9.43	7.80
OTHER BINDING	14.19	17.96	19.28	19.75	18.05
OTHER MOLECULAR FUNCTIONS	0.99	1.55	3.01	0.71	0.00
OXYGEN BINDING	0.02	0.15	0.60	0.00	0.00
PROTEIN BINDING	24.19	17.18	19.28	16.01	12.20
RNA BINDING	8.95	6.35	7.23	2.14	0.98
SIGNALING RECEPTOR ACTIVITY	0.74	0.46	1.20	2.85	4.39
SIGNALING RECEPTOR BINDING	0.44	0.00	0.00	0.18	0.00
STRUCTURAL MOLECULE ACTIVITY	1.48	0.77	0.60	0.00	0.00
TRANSCRIPTION REGULATOR ACTIVITY	0.65	0.77	0.00	0.36	0.00
TRANSFERASE ACTIVITY	20.50	25.70	22.29	19.22	13.66
TRANSLATION FACTOR ACTIVITY, RNA BINDING	0.39	0.46	0.60	0.36	0.49
TRANSLATION REGULATOR ACTIVITY	0.04	0.00	0.00	0.00	0.00
TRANSPORTER ACTIVITY	8.58	7.43	7.83	4.09	4.39
UNKNOWN MOLECULAR FUNCTIONS	34.11	33.59	32.53	40.21	41.95

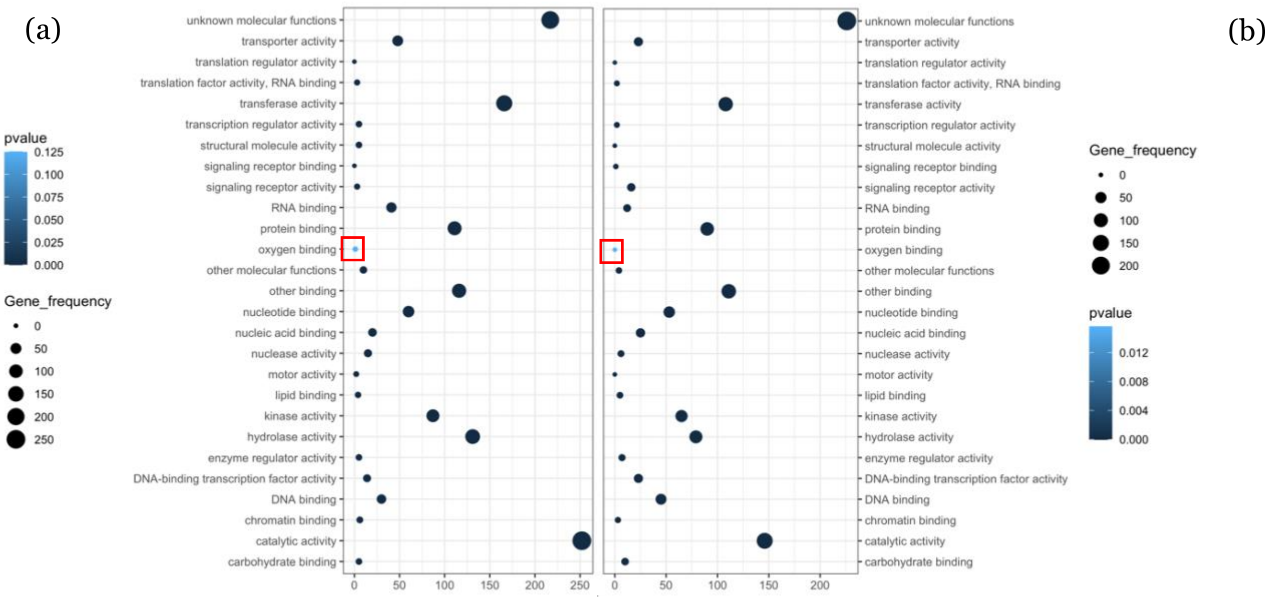


*Figure 8: GO Slim enrichment of differentially expressed genes at Threshold-1 (T1) **

a). Down regulated b). Up regulated

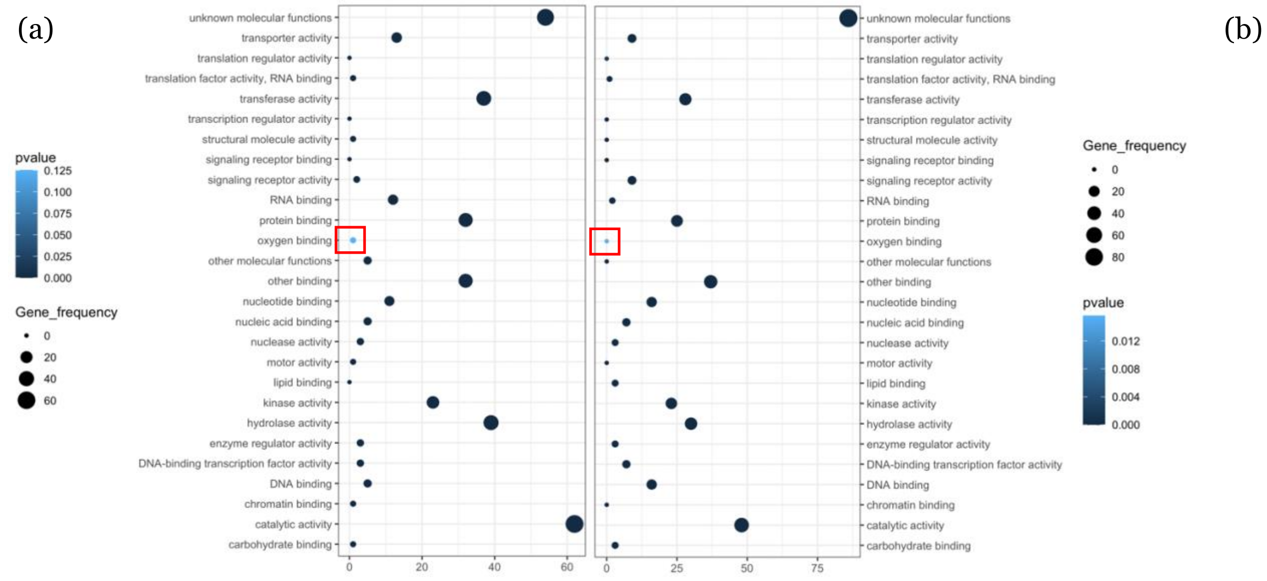


*Figure 9: GO Slim enrichment of differentially expressed genes at Threshold-2 (T2) **

a). Down regulated b). Up regulated

(Legend Fig 8 and 9: Bubble size depends on the total number of genes on x-axis whereas their color gradient represents the p-value).

3.8 Co-occurrence analysis of the downregulated genes

The idea for computing co-occurrence statistics originated with the hypothesis that a mutated allele and an associated change in function are more likely to appear together rather than a random co-existence. In this work, the theory behind the statistical analysis on the co-occurrence data was to estimate the association strength and the functional connection in between significantly down regulated gene pairs by doing comparison of their observed and expected frequency in all accessions. To uncover the concept, Mendel's law of segregation was applied. Due to plausible loss of function in down regulated genes, we chose significantly low expressed genes at T1 for co-occurrence analysis.

The matrix of 646 genes created 208012 gene pairs. We removed 88 gene pairs lying within the 100 kb distance on same and different chromosomes and altogether 207,924 gene-pairs were left for further analysis. To obtain the significant number of pairs which co-occurred together more or less often than expected, correction for multiple-testing of the p-values was performed according to Bonferroni correction. Connections were sorted at significance level 2.40x10^{-7}. Only 0.82% of gene pairs were found to be statistically significantly co-occurring together. We found 1,696 highly significant gene pairs in which 139 pairs were found to be co-occurred less frequent referring to unrelated genes having greater distance within 87 unique negatively inter-connecting genes and therefore named as under-represented. However, 1,557 gene pairs were found to co-occur statistically significantly more often than expected hence labeled as over-represented. The total number of unique over-represented genes were found to be 325.

Figure 10 shows the p-value histogram of co-occurring gene pairs binned at relative distance of 0.01 with the red line pointing towards multiple testing threshold. It can be observed that very few gene pairs were found to be significant at this threshold however with highly significant low p-values.

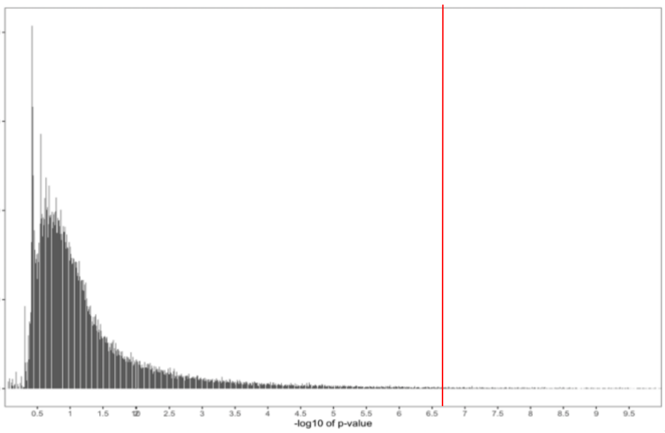


Figure 10: P-values of co-occurrence of gene pairs plotted against -log10.

3.8.1 GO slim enrichment of under and over-represented co-expressed downregulated gene pairs

To uncover the relevant molecular functions of highly significant co-occurred gene pairs, a comparison of GO slim enrichment of all premature stop gained genes (PSCs) and downregulated over and under-represented co-expressed gene pairs with whole genome was conducted. The most striking results were observed in significantly reduced number of genes involved in protein binding (GO:0005515), DNA binding (GO:0003677), DNA-binding transcription factor activity (GO:0003700), and lipid binding (GO:0008289). However, a higher number of genes were involved in hydrolase activity (GO:0016787), kinase activity (GO:0016301), nucleotide binding (GO:0000166) and nuclease activity (GO:0004518). Overall, the results remained consistent with GO enrichment of downregulated genes in both over and under-represented category however, few differences were observed between over-represented co-occurring pairs where interestingly 38.15% of genes were involved in catalytic activity (GO:0003824) compared to under-represented where the number was significantly reduced to 32% which was however still higher in comparison to genome and all other PSCs gained genes. Contrarily, much greater number of under-represented genes were working in enzyme regulator activity (GO:0030234), carbohydrate binding (GO:0030246) and transcription regulator activity (GO:0140110) in comparison with gene pairs which were co-occurred more often than expected. It was also noticed that ~34.5% of over-represented genes and ~38% of under-represented genes were categorized into unknown molecular functions (GO:0003674). The term is described as: “A molecular process that can be carried out by the action of a single macromolecular machine, usually via direct physical interactions with other molecular entities".

Almost no activity was observed in oxygen binding (GO:0019825), translation regulator activity (GO:0005198) and chromatin binding (GO:0003682) in over and under-represented gene pairs to a little activity in all stop gained and whole genome. The graphical representation of GO slim enrichment of co-occurring over and under-represented genes is shown in figure 14 in the form of a sunburst plot. The percentage values of GO enrichment in whole genome, all stop gained, and down regulated over and under-represented genes are shown in table 6.

Table 4: Tabular Results of GO Slim enrichment in Co-occurring down regulated over and under-represented gene pairs in comparison to whole genome and all stop gained.

GO SLIM TERMS	GENOME	STOP_GAINED	COC_OVER	COC_UNDER
HYDROLASE ACTIVITY	12.98	14.00	21.54	22.99
KINASE ACTIVITY	8.43	9.82	12.92	13.79
TRANSFERASE ACTIVITY	19.47	19.46	25.54	21.84
DNA-BINDING TRANSCRIPTION FACTOR ACTIVITY	6.57	6.22	1.85	3.45
DNA BINDING	8.14	7.04	4.31	5.75
NUCLEIC ACID BINDING	4.46	4.48	3.38	5.75
PROTEIN BINDING	22.98	17.05	18.15	20.69
CATALYTIC ACTIVITY	26.62	27.03	38.15	32.18
RNA BINDING	8.50	4.34	7.38	6.90
OTHER BINDING	13.48	13.40	16.00	11.49
TRANSPORTER ACTIVITY	8.15	8.04	7.69	5.75
ENZYME REGULATOR ACTIVITY	1.77	1.68	0.92	2.30
STRUCTURAL MOLECULE ACTIVITY	1.41	0.65	0.92	1.15
TRANSCRIPTION REGULATOR ACTIVITY	0.62	0.57	0.31	1.15
SIGNALING RECEPTOR ACTIVITY	0.70	0.93	0.00	0.00
LIPID BINDING	1.17	1.13	0.62	0.00
NUCLEOTIDE BINDING	4.05	4.85	7.38	6.90
MOTOR ACTIVITY	0.23	0.30	0.31	0.00
OTHER MOLECULAR FUNCTIONS	0.94	0.67	2.15	3.45
CHROMATIN BINDING	0.67	0.68	0.31	0.00
SIGNALING RECEPTOR BINDING	0.41	0.30	0.00	0.00
TRANSLATION FACTOR ACTIVITY, RNA BINDING	0.37	0.29	0.62	1.15
CARBOHYDRATE BINDING	0.62	0.85	0.31	1.15
OXYGEN BINDING	0.02	0.03	0.00	0.00
TRANSLATION REGULATOR ACTIVITY	0.03	0.01	0.00	0.00
UNKNOWN MOLECULAR FUNCTIONS	32.41	38.60	34.46	37.93
NUCLEASE ACTIVITY	1.10	1.46	3.38	4.60

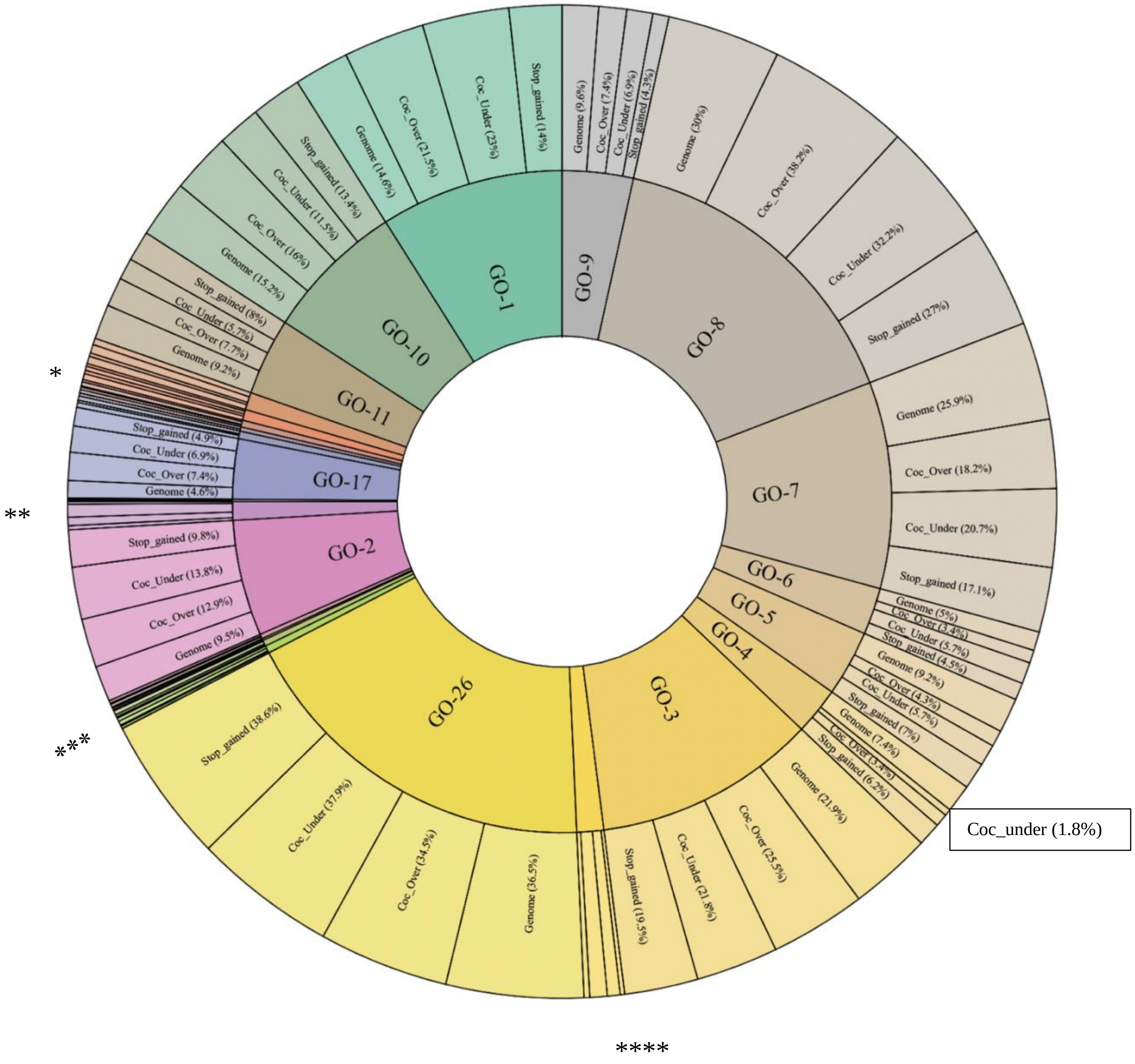
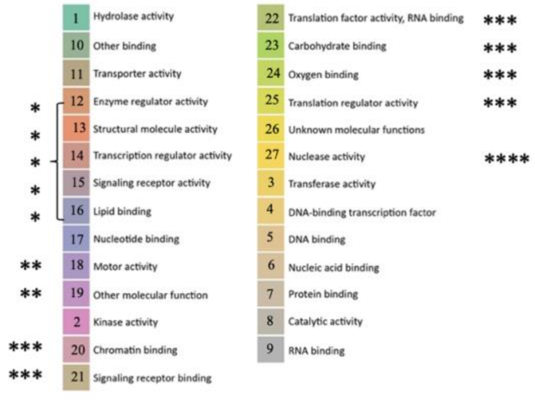


Figure 11: Sunburst plot showing functional enrichment categorization according to molecular function. A comparison between whole genome, all stop gained, under and overrepresented co-occurring gene pairs. Go terms are numbered in the innermost circle with their names associated in the legend below.



*(*Asterisks are given at the categories where almost little to no activity was observed, therefore no space was allocated in the plot. Their percentage values can be observed in table 6).*

3.8.2 Visualization of co-occurred over and under-represented gene pairs

Both under and over-represented co-occurring gene pairs were visualized on a circos plot using RCircos (H. Zhang et al., 2013) library for a better image of self-loops and gene-gene co-occurrences at chromosomes level. Roughly 40.5% of connections were found to be in between genes on the same chromosome forming loops. Interestingly, the highest number of loops were observed in between genes on chromosome 5. Their highly significant p-values yielded very well interpretable context in the form of their molecular function. The GO enrichment of these genes revealed that ~26.6% were involved in protein binding and nearly 16.6% in hydrolase activity and catalytic activity as parent GO slim terms. However, 37% of genes had unknown molecular functions as shown in figure 11. Furthermore, it was also observed that clusters of genes originating in chromosomes 2 and 5 were noticeably much more as compared to chromosome 1,3 and 4.

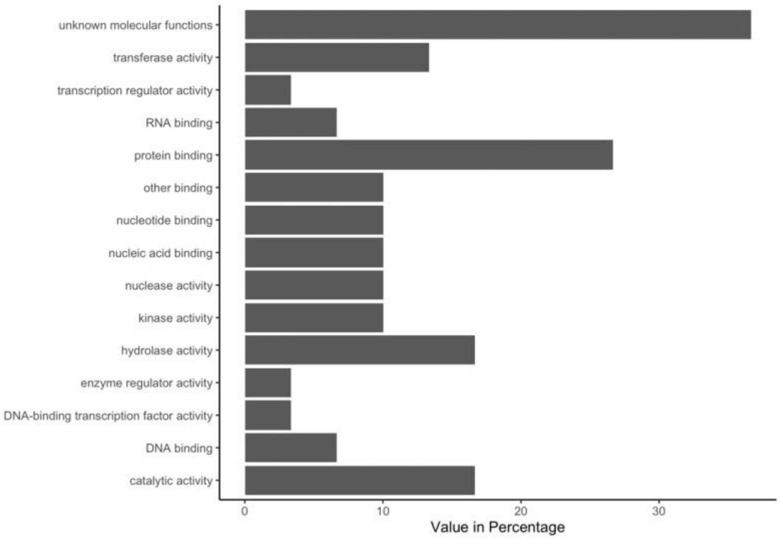


Figure 12: GO enrichment of highly significant genes in chromosome 5

Some genes were acting as a hub and connected with numerous other genes in different chromosomes. For example: AT2G15390 (FUT4 belonging to FUCOSYLTRANSFERASE 2-RELATED family) was knocked out in 311 accessions. Due to considerably high number of accessions where FUT4 is knocked out it was also acting as a hub gene in the co-expression network. However, zero co-occurrences were found with AT5G16330 knocked out (KO) in 66 accessions yielding highly significant under-represented connection at the lowest p-value 1.27x10^{-18}. Likewise, 0 connections were also observed with AT5G18404 (KO: 54) and AT2G25450 (KO: 40) and only 1 common accession with AT2G36815 (KO: 60). Interestingly, 83 common accessions were found with AT4G09012 (Mitochondrial ribosomal protein L27) with significant p-value 1.45x10^{-09}. According to the latest annotation of molecular function by (Ashburner et al., 2000), this gene enables structural constituent of ribosome (Gaudet et al., 2010). Another highly significant example was co-occurrence pattern between AT5G09700 a pseudogene of glycosyl hydrolase family 3 protein, (involved in arabinan and xylan catabolic process) shared 24 common knocked accessions with AT1G12700 (KO: 145) and AT5G65925 (KO: 291) with a highly significant p-value of 5.20x10^{-15}. A section of co-occurrence table with gene pairs, their connections, lower and upper limits of connections at T1 and T2 along with their significant p-values is shown in table 3. Figure 12 below represents the graphical representation of co-occurring gene pairs at multiple testing threshold and illustrates the clusters of connected pairs in all 5 chromosomes.

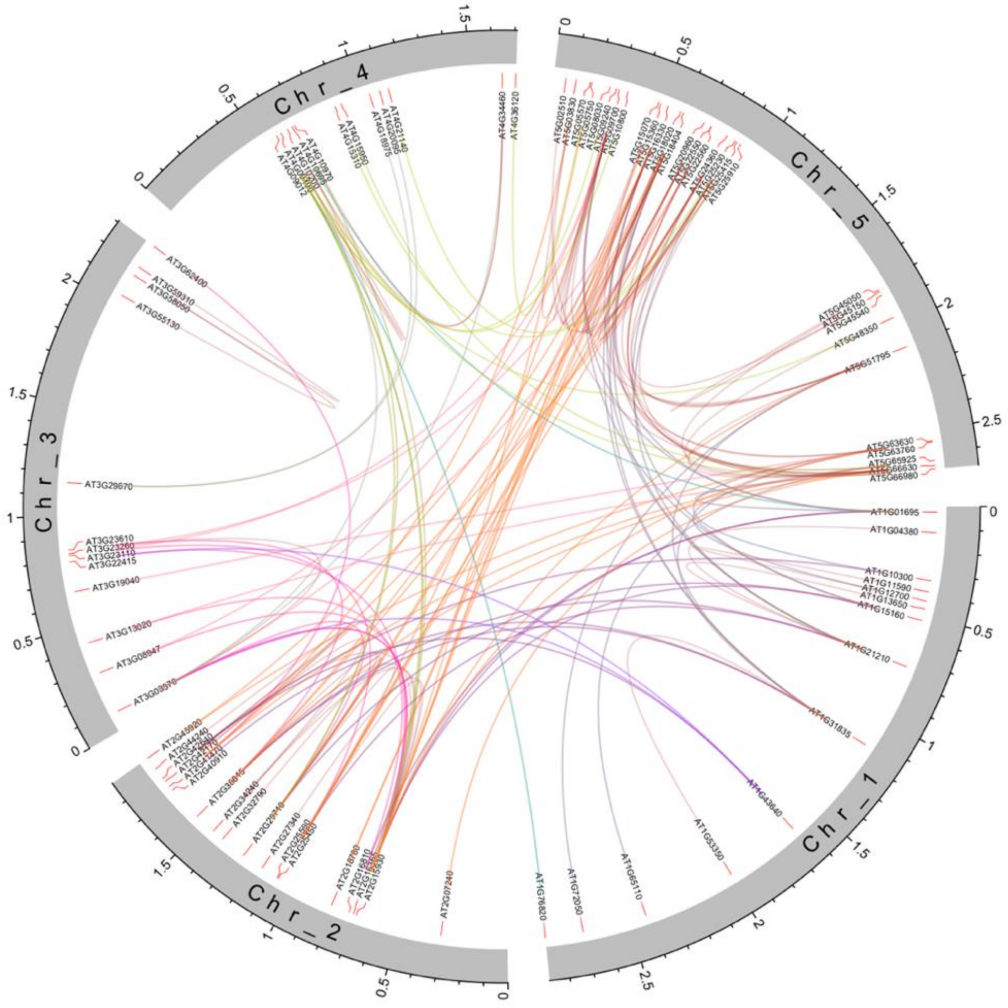


Figure 13: Co-occurrence pattern between under-represented gene pairs. The first (outer) circle in the form of an ideogram indicates all 5 chromosomes, and distance mapped as relative position. Red markers in the second circle indicate the position of knocked out gene on the chromosome. The third track shows the gene names at the marked position. The Lines in the center represent connecting pairs of genes with p-value ≤2.40x10^{-7}. Brown line indicates self-loops within same chromosomes, while blue, red, magenta, green and orange lines distinguish connections originating from each of chromosome 1-5 respectively

Table 5: Co-occurrence table of overrepresented gene pairs exhibiting all connections, the total number of knock out accessions of each gene and their calculated p-values based on correction of multiple testing threshold. (The rows are randomly chosen from a table of 1557 highly significant co-occurring pairs.)

Gene1	Gene2	Connection	Ko_gene1	Ko_gene2	Low_lim_0.05	Up_lim_0.05	Low_lim_thres	Up_lim_thres	Pvalue
AT3G60440	AT3G60966	17	39	44	0	5	0	12	6,22E-12
AT2G16380	AT4G14385	14	21	41	0	3	0	9	9,39E-14
AT1G11180	AT1G33960	25	107	43	3	11	0	20	5,63E-11
AT1G11180	AT1G29710	37	107	73	7	17	0	28	1,16E-13
AT5G16330	AT5G18404	46	61	54	2	8	0	17	2,47E-49
AT3G08947	AT5G25415	30	60	89	4	12	0	23	2,93E-13
AT1G65110	AT1G67270	29	34	205	6	15	0	24	2,00E-11
AT2G41710	AT4G18840	20	44	39	0	5	0	12	6,83E-16
AT4G08480	AT4G36650	20	100	36	2	9	0	18	5,34E-09
AT1G76740	AT2G25450	11	18	40	0	3	0	8	2,09E-10
AT1G20350	AT2G32790	17	38	48	0	5	0	13	1,97E-11
AT3G13020	AT5G18404	19	23	54	0	4	0	10	4,90E-19
AT3G27325	AT5G10800	17	27	80	1	6	0	14	1,40E-10
AT2G34240	AT4G09012	46	63	256	18	30	7	43	4,69E-09
AT2G16380	AT3G60440	16	21	39	0	3	0	9	1,12E-17
AT1G07480	AT1G51820	15	16	49	0	3	0	9	1,70E-17
AT2G34240	AT5G15360	31	63	56	2	9	0	18	7,42E-21
AT4G09012	AT5G59510	34	256	44	12	22	3	33	6,22E-08
AT2G34240	AT5G18404	33	63	54	2	9	0	18	2,27E-24
AT2G25450	AT5G25415	31	40	89	2	9	0	18	5,33E-22
AT5G16330	AT5G51795	35	61	44	1	7	0	15	2,53E-33
AT5G16330	AT5G45050	20	61	28	0	5	0	12	1,33E-16
AT5G15360	AT5G51795	26	56	44	1	7	0	15	7,90E-20
AT3G23610	AT3G32920	57	207	88	21	34	9	49	2,13E-12
AT4G07400	AT4G07825	151	309	223	94	114	73	134	3,49E-15
AT1G29710	AT4G18840	24	73	39	1	8	0	16	1,55E-15
AT1G60500	AT3G24360	18	40	21	0	3	0	9	1,74E-21
AT2G07240	AT2G34240	33	98	63	5	14	0	25	5,76E-14
AT2G40910	AT3G13020	14	16	23	0	2	0	6	2,94E-21
AT1G65110	AT4G00970	21	34	50	0	5	0	12	1,30E-17
AT2G18320	AT5G63630	88	115	337	50	66	34	83	3,87E-10
AT3G59250	AT5G02630	13	45	22	0	4	0	10	3,30E-11
AT4G13810	AT4G18840	18	46	39	0	5	0	13	8,76E-13
AT1G15160	AT5G63760	9	32	12	0	2	0	6	8,46E-11
AT1G60070	AT2G04280	59	181	101	21	34	9	49	4,20E-13
AT1G12700	AT5G48320	14	145	14	1	6	0	12	3,36E-10
AT1G01695	AT5G16330	33	43	61	1	7	0	15	3,35E-30

A

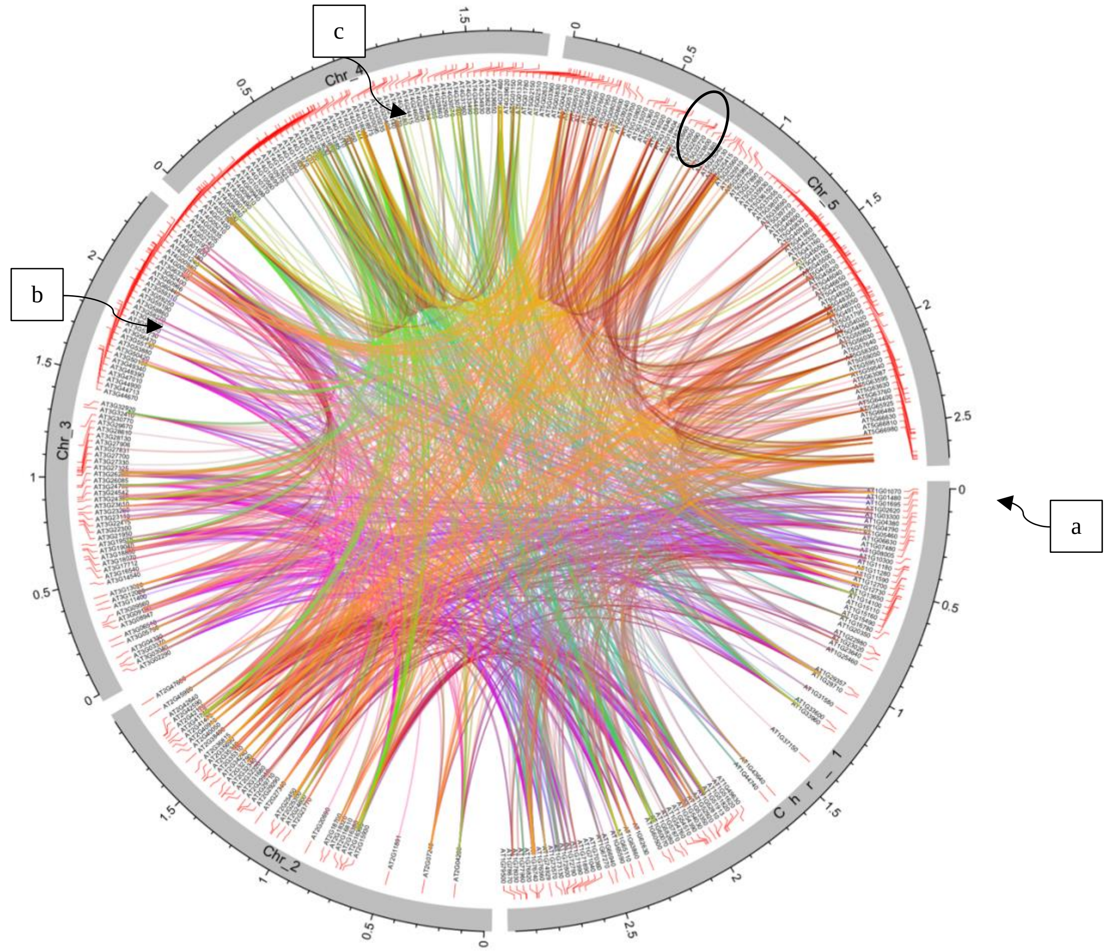


Figure 14: A: Circos representation of highly significant overrepresented gene-gene co-occurring pairs. The first (outer) circle in the form of an ideogram indicates all 5 chromosomes of Arabidopsis thaliana, their length along with the distance mapped as relative position (a). Red markers in the second circle indicate the position of knocked out gene on the chromosome (b). The third track show the gene names at the marked position. The lines in the center represent connecting pairs of genes with p-value ≤2.40x10^{-7}. Brown line indicates self-loops within same chromosomes, while blue, red, magenta, green and orange lines distinguish connections originating from each of chromosome 1-5 respectively.

In figures 13 and 14, RCircos library was used to generate the circular plot in the form of an ideogram. Three tracks were built on which distance was mapped as relative position. Figure 14 shows significant mutated genes overly co-expressed in multiple accessions represented by colored links and illustrates their connection pattern. It was observed that genes in all 5 chromosomes were highly connected to each other, but no specific clustering pattern was detectable from the circular plot also showed in figure 14: -B, -C, -D, -E and -F.

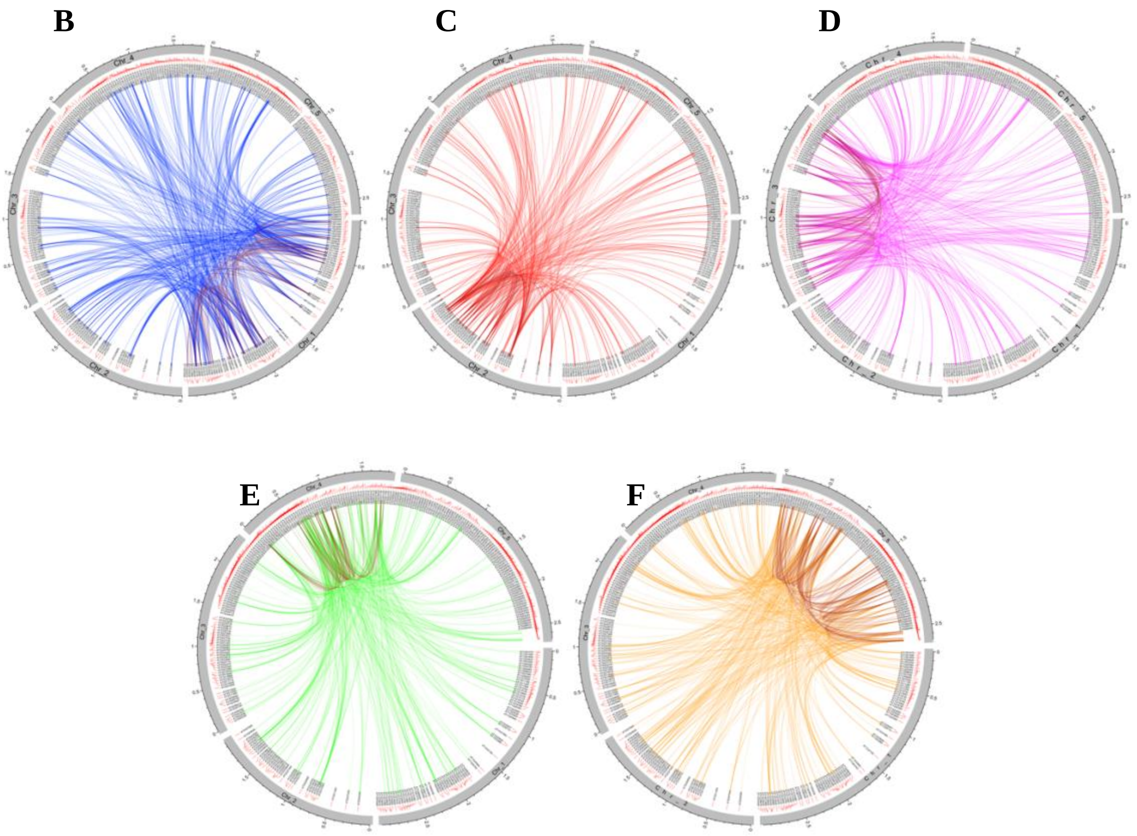


Figure 14 (extension): Each circular plot from B-F shows connections between genes in a chromosome to each of four other chromosomes. Brown lines show self-loops. B: blue lines connecting between genes on chr1 with other 4 chr; C: similarly red lines for gene-gene connections between chr 2 and rest; D: magenta lines indicate connections of chr 3; E: green lines connecting with chr 4 and F: Orange lines representing connections between chr 5 and chr 1,2,3 and 4 respectively.

Nevertheless, it gave an overall depiction of connections between genes in all 5 chromosomes. Yet, by looking at the highly significant p-values derived from hypergeometric distribution method in table 4, we identified certain gene clusters where a set of genes were strongly connected with many other genes. Few of them were outliers with not many connections while others had crucial roles in diverse molecular functions, hence were acting as a central connecting point, therefore we labelled them as hub genes. For example: the highly significant co-occurrence with lowest p-value of $2.47x10^{-49}$ in 46 accessions was found between AT5G16330 (KO: 61) (NC domain-containing protein-like protein) and AT5G18404 (KO: 54) (showed evidence of transcription activity). Moreover, AT5G16330 (KO: 61) was also co-expressed in 35 and 33 accessions with AT5G51795 (KO: 44) (DNA/RNA-binding protein Kin17) p-value: $2.53x10^{-33}$ and AT1G01695 (43) (TON1 RECRUITING MOTIF 33, TRM33) with p-value: $3.35x10^{-30}$, respectively. Similarly, gene AT5G51795 (KO: 44) (DNA/RNA-binding protein Kin17) was knocked out in 26 common accessions with AT5G15360 (KO: 56) (a transmembrane protein whose molecular function is unknown) p-value $7.90x10^{-20}$. Likewise, AT2G40910 (KO: 16) 14 common accessions with AT3G13020 (KO: 23) at p-value $2.94x10^{-21}$. Another clustering gene, AT5G18404 (KO: 54) besides having strongest association with AT5G16330, was found to be significantly co-occurred in 19 and 33 accessions at a p-value of $4.90x10^{-19}$ with AT3G13020 (KO: 23) (hAT transposon superfamily protein) and AT2G34240 (KO: 63) (a protein with domains of unknown function DUF627 and DUF632) having p-value $2.27x10^{-24}$. The gene clusters of AT5G16330 and AT5G18404 in chr-5 are marked in black circle as shown in figure 12-A.

3.8.3 Quantification of co-expression by Kendall Tau co-expression coefficient

Irrespective of the fact that the number of over-represented connections were almost 10 times more as compared to the underrepresented, the other reason behind no visual clustering or perceivable patterns was a lot of noise data included in this plot meaning that although the co-occurrence of the genes was highly significant, nevertheless the strength of their co-expression specific to the accessions was still unknown. Therefore, *Kendal's Tau* co-expression coefficient τ (KENDALL, 1938) was calculated by Kendall Tau method to quantify the association between co-occurring genes and to get insights about more robust co-expression. It also helped to get a better visualization of the connections for extraction of meaningful biological modules from the co-occurrence.

After calculation of *tau* score the co-occurred gene pairs were then divided into 3 categories: $\tau \geq 0.7$: characterized as highly significant strong co-expression. τ score between 0.4 and 0.6: termed as moderate co-expression while a τ score of 0.3 and less: perceived to be a very weak connection however $\tau < 0$ was perceived as negatively correlated. We discovered 37 gene pairs with a τ score 0.7 and above and simultaneously they were also the highly significant gene pairs exhibiting the lowest p-values. Moreover, 509 pairs exhibited moderate but significant co-expression with τ score between 0.4 and 0.6, and 1011 pairs had τ score of 0.3 and less, thus considered very weak connections.

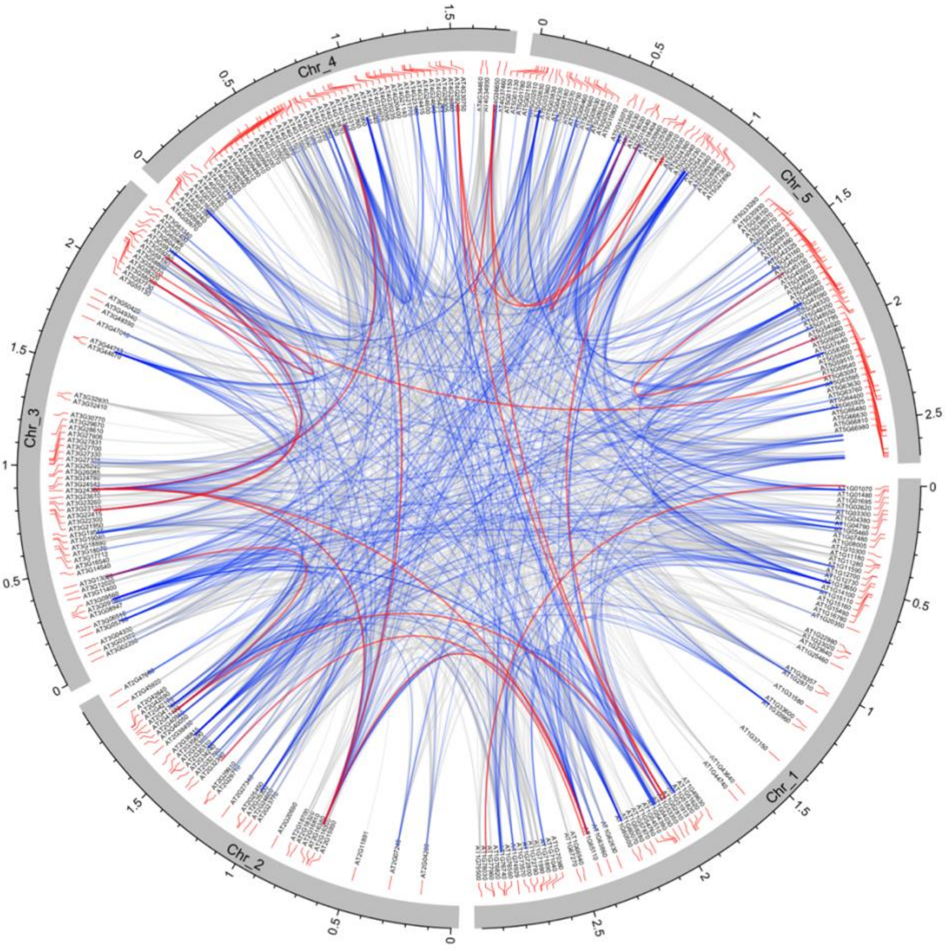


Figure 15: Co-occurrence circular plot based on co-expression coefficient tau score values. Outer most circle indicates the chromosome in the form of an ideogram. Red markers in 2nd track indicate the location of allele on the chromosome whereas the 3rd circle shows knocked out gene names. The top-most layer comprising red lines embody highly significant connections with correlation coefficient τ = 0.7 and above. Blue lines represent medium connection strength with tau score in between 0.4 and 0.6 and the inner most grey lines represent weak connections having Tau score (τ) 0.3 and less.

Next, we extracted the co-occurring gene pairs possessing strong co-expression coefficient ($\tau \geq 0.7$) to comprehend the pattern of co-occurrence of gene pairs and to filter significant modules with a biological implication. Overall, the maximum τ score of 1.0 was observed in 3 different gene pairs:

1) AT4G31360 (KO: 5) (selenium binding protein, involved in gene expression regulation, negative regulation of cellular process, epigenetic and seed development) surprisingly co-expressed in all 5 accessions from Italy (9679, 9680, 9681, 9682, 9683) with p-value of 3.33×10^{-146} together with AT4G36280 (KO: 5) (MORC2, involved in defense response to virus and bacterium, hypersensitive response, positive regulation of systemic acquired resistance, regulation of DNA repair, enables ATP hydrolysis activity, DNA binding, RNA binding, endonuclease activity, protein binding). Both genes were located in nucleus and expressed in vascular leaf, stem, cauline leaf, carpel, flower, flower pedicel, collective leaf structure, guard cell, inflorescence meristem, leaf apex , hypocotyl, leaf lamina base, plant embryo, petiole, petal, seed, shoot apex, shoot system, root, and sepal.
2) AT3G56470 (KO: 3) (F-box family protein, involved in cellular response to organic substance and signal transduction, enables unknown molecular function) connected in all 3 accessions from Sweden (6255, 6268, 7517) p-value: 3.33×10^{-146} with AT5G54880 (KO: 3) (DTWD2B, active in cellular component, involved in tRNA modification and enables tRNA-uridine amino-carboxy-propyl-transferase activity). They are expressed in guard cell.
3) AT5G40830 (ICA, encodes a SAM-dependent methyltransferase superfamily protein, acts upstream of or within phloem or xylem histogenesis) co-occurred in all 4 knocked out accessions (9128, 9130, 9133, 9134) from Armenia at p-value score 3.33×10^{-146} with AT5G49710 (RING finger protein, involved in response to inorganic substances and enables unknown molecular function) are expressed in vascular leaf, sepal, cauline leaf, hypocotyl, plant sperm cell, flower pedicel, pollen, petiole, plant embryo, inflorescence meristem, shoot apex, leaf apex, stamen, cotyledon, seed, carpel, guard cell, petal, stem, collective leaf structure, flower, leaf lamina base, root, shoot system.

Furthermore, 2 gene pairs were found with τ score of 0.9 and 6 pairs with co-expression co-efficient score (τ = 0.8) depicting strongest connections. We looked into the common accession Ids where they were co-occurred together and remarkably five assorted clusters were identified from Sweden, US, Spain, Italy, and Armenia where majority of the gene pairs were expressed in related accessions. Numerous other gene pairs were discovered which knock out together from Russia, Georgia, Uzbekistan, Afghanistan, Kazakhstan with unexpected frequency of highest significant connections and since it was also difficult to scrutinize each significant co-expression and correlate them, a better approach was to build a gene co-expression network (GCN) for both over and under-represented pairs. Highly significant gene pairs with strongest tau score (τ) coefficient are shown in table 6 below.

Table 6: Tabular results of co-expression analysis of downregulated gene pairs with strongest co-expression coefficient value, i.e., ($\tau \geq 0.7$).

Gene1	Gene2	connection	KO gene1	KO gene2	Pvalue	Tau score	Tau pvalue
AT1G01070	AT1G14100	7	8	14	2,48E-12	0.7	4.42e-64
AT1G01070	AT1G78670	4	8	4	8,72E-09	0.7	1.24e-73
AT1G52920	AT4G31360	5	7	5	1,98E-11	0.8	1.10e-104
AT1G52920	AT4G36280	5	7	5	1,98E-11	0.8	1.10e-104
AT1G52920	AT5G22690	7	7	10	1,10E-14	0.8	1.78e-102
AT4G31360	AT4G36280	5	5	5	9,44E-13	1.0	3.33e-146
AT4G31360	AT5G22690	5	5	10	2,38E-10	0.7	1.60e-73
AT4G36280	AT5G22690	5	5	10	2,38E-10	0.7	1.60e-73
AT1G65110	AT2G16380	20	34	21	3,35E-28	0.7	1.06e-80
AT1G65110	AT3G24360	20	34	21	3,35E-28	0.7	1.06e-80
AT1G65110	AT4G15950	21	34	27	9,27E-26	0.7	2.24e-68
AT4G15950	AT4G18840	22	27	39	4,70E-26	0.7	3.72e-65
AT2G16380	AT3G24360	18	21	21	2,24E-29	0.9	8.61e-107
AT2G16380	AT4G15950	18	21	27	7,67E-26	0.7	1.84e-82
AT2G47680	AT3G44900	4	9	4	1,57E-08	0.7	1.47e-65
AT2G47680	AT5G40830	4	9	4	1,57E-08	0.7	1.47e-65
AT2G47680	AT5G49710	4	9	4	1,57E-08	0.7	1.47e-65
AT5G40830	AT5G49710	4	4	4	1,25E-10	1.0	3.33e-146
AT3G09560	AT5G25560	13	17	21	6,62E-19	0.7	1.90e-68
AT1G54510	AT4G30570	7	13	8	1,24E-12	0.7	5.31e-69
AT3G21950	AT3G55130	11	17	13	3,72E-18	0.7	1.07e-79
AT3G21950	AT3G59310	12	17	14	3,99E-20	0.8	4.52e-88
AT3G21950	AT5G02630	13	17	22	1,60E-18	0.7	2.95e-65
AT3G59310	AT5G02630	12	14	22	4,11E-18	0.7	8.62e-68
AT3G55130	AT3G59310	12	13	14	8,50E-23	0.9	1.64e-115
AT3G56470	AT5G54880	3	3	3	2,06E-08	1.0	3.33e-146
AT4G03935	AT5G27750	5	7	8	1,10E-09	0.7	1.31e-65
AT3G28130	AT5G55960	5	5	11	4,36E-10	0.7	6.58e-67
AT1G65590	AT2G31680	3	3	5	2,06E-07	0.8	3.03e-88
AT5G16330	AT5G18404	46	61	54	2,48E-49	0.8	2.31e-90
AT2G36815	AT5G18404	40	60	54	3,01E-37	0.7	1.19e-67
AT1G01695	AT5G02510	27	43	36	2,34E-29	0.7	4.80e-66
AT2G40910	AT3G13020	14	16	23	2,95E-21	0.7	3.60e-77
AT1G06630	AT5G05180	7	8	13	1,24E-12	0.7	5.31e-69

3.9 Co-occurrence Network of co-expressed genes

Network represents relationships in large datasets; however, the idea is very generic. Our focus was to build a gene-oriented network exhibiting regulatory interactions between co-occurring genes. The concept behind the story was, if there is a phenotype associated with mutation A and a phenotype associated to mutation B and they both co-occur in same accession; more is expected than a random co-expression. Therefore, after determining pairwise correlation between the highly significant co-occurring gene pairs, we represented them in the form of a gene-gene network and defined it as a Gene Co-occurrence Network (GCN). In the overall network analysis, total number of nodes and edges, average number of neighbors, clustering coefficient, network density, multi-edge node pairs, and self-loops were checked. No self-loops were kept in the network.

We interrogated all the identified modules and the network for identification of hub genes, common regulatory pathways, and also looked into the functional enrichment of co-occurring genes. Moreover, we also identified the common knocked out geospatial locations specifically for the modules on European map.

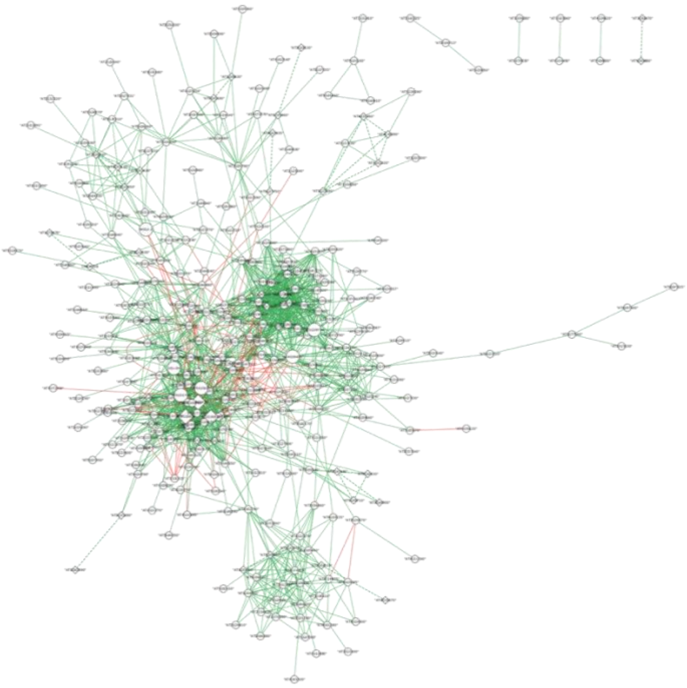


Figure 16: Gene Co-occurrence Network (GCN) including positive and negative correlation. Degree is represented by node size. Green edges represent positive correlation and red edges indicate a negative correlation.

Interestingly, we observed that negative correlations were highly interconnected with hub genes in over-represented network. To acquire a better understanding of the underlying pattern of both over and under-represented network, we analyzed them separately.

3.9.1 Negative correlation in co-expression network due to negative coefficient score

The gene pairs which were less often co-occurred together than expected had a negative co-expression coefficient hence were presented as negatively correlated. All these under-represented genes with a negative tau value i.e., τ < 0 were inversely connected with many over-represented hub genes. It also relates to the activation and inhibition of molecular functions of the occurring gene pair. These connections are represented by red edges in figure 17.

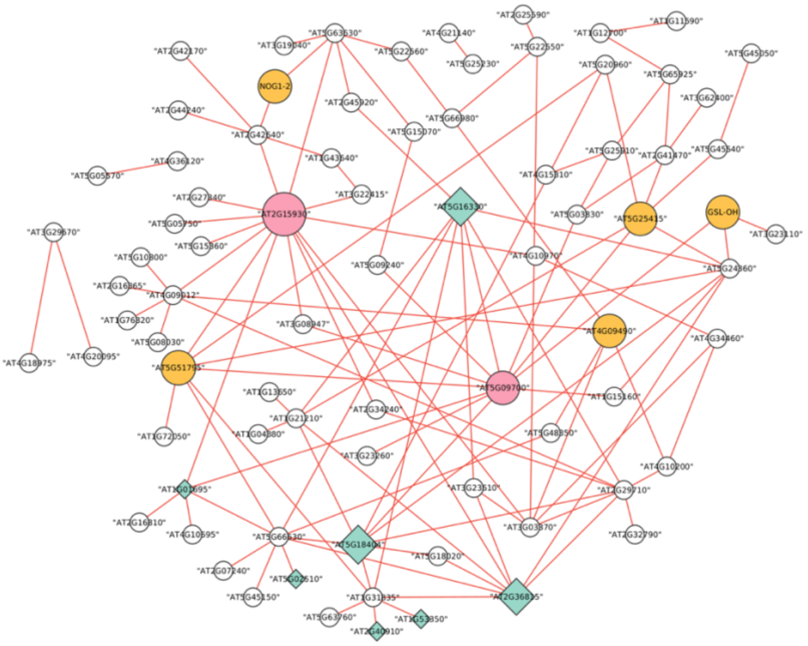


Figure 17: Gene Co-occurrence network of under-represented genes represented as circles and diamonds. Red edges indicate the negative correlation. Increased node size indicates hub genes where different size denotes degree of the node. Orange nodes represent both over (positive correlated) and under-represented (negative correlated) hub genes. Pink nodes are the hub genes in negative correlations only. Lastly, green diamonds indicate genes which have tau τ >= 0.7 in over-represented network and have negative correlations too.

Interestingly, several genes were acting as hub genes that were knocked out together with multiple other genes in numerous accessions. We presumed that their co-occurrence was not random, rather there was a biological connection between the co-occurring gene pairs. Several examples include: AT2G15930 (KO:311) (a putative uncharacterized protein located in mitochondrion and enables unknown molecular function) and AT5G09700 a (protein coding pseudogene of glycosyl hydrolase family 3 involved in arabinan catabolic process and expressed in rosette leaf) (Aryal et al., 2014). They were acting as hub genes having multiple correlations but less often connected with other deeply interconnected co-occurring genes despite being knocked out in sufficient number of accessions. Furthermore, there was a relatively strong negative correlation (Kendall tau τ = -0.3, p-value = $2.04\text{x}10^{-13}$) between AT2G36815 (involved in mRNA cis splicing, located in mitochondrion, and enables unknown molecular function) and AT2G15930 (also located in mitochondrion and enables unknown molecular function) where both genes were acting as core hub genes in the network and co-expressed in accession ID 6201 from Sweden. Another core hub gene AT5G09700 (pseudogene of glycosyl hydrolase family 3 protein, located in cytosol, reportedly involved in arabinan catabolic process, and showed expression in rosette leaf) is negatively correlated (Kendall tau τ = -0.3, p-value = $2.04\text{x}10^{-13}$) with; 1) AT5G16330 (NC domain-containing protein-like protein, active in cellular component, enables unknown molecular function), 2) AT5G18404 (involved in light stimulus response and pigment biosynthetic processing, located in mitochondrion and showed expression in guard cell and enables unknown molecular function), AT5G51795 (DNA/RNA-binding protein Kin17, enables double stranded DNA binding, located in nucleus) AT5G03830 (CDK inhibitor P21 binding protein, active in nucleus, located in mitochondrion, expressed in guard cell and enables unknown molecular function), AT1G01695 (TON1 RECRUITING MOTIF 33, TRM33, enables unknown molecular function), AT5G25415 (DUF239, located in cellular component, enables unknown molecular function), AT2G25450 (involved in glucosinolate biosynthetic process, regulation of glucosinolate biosynthetic process, located in cytosol, expressed in rosette leaf), AT1G15160 (MATE efflux family protein involved in transmembrane export, located in plasma membrane and expressed in guard cell) AT3G08947 (ARM repeat superfamily protein; located in cytosol and cytoplasm, enables nuclear localization sequence binding, expressed in guard cell and rosette leaf), AT3G23260 (F-box and associated interaction domains-containing protein, enables unknown molecular function, expressed in guard cell).

Overall, we identified 8 genes involved in only negatively correlated connections i.e., AT1G21210 (WAK4, WALL ASSOCIATED KINASE 4) enables calcium ion binding and polysaccharide binding, expressed in root and trichome, AT1G31835 (expressed in guard cell), AT1G72050 (TFIIIA, TRANSCRIPTION FACTOR IIIA), AT2G25590 (Plant Tudor-like protein), AT4G36120 (filament-like protein (DUF869), AT5G22560 (transmembrane protein, putative (DUF247)), AT2G44240 (NEP-interacting protein (DUF239)), and AT5G45540 (transmembrane protein, putative (DUF594)). Their GO slim analysis revealed that except AT1G21210 involved in carbohydrate binding (GO:0030247) and AT1G72050 involved in DNA binding (GO:0080084), DNA-binding transcription factor activity (GO:0003700) and RNA binding (GO:0008097; GO:0003723) all other genes were involved in unknown molecular functions.

3.9.2Positive correlation in co-expression network for over-represented genes

Based on correlation scores, shared knocked out accessions and mutual information, we first defined individual relationships between each gene pair (Butte & Kohane, 1999; Steuer et al., 2002). The resemblance between expression pattern of gene pairs in all accessions was then interpreted through these relationships. An extraordinarily related information in a connection signified that the co-occurring genes were purposefully correlated; therefore, we hypothesized that the two were biologically associated. We observed five large clusters detectable in the positively correlated network besides some small highly significant clusters. We identified six intramodular genes connecting with hub genes in two different modules and named them "intra-modular connectors". Moreover, there were other genes which co-expressed with all other hub and peripheral genes, they were named "inter-modular hub genes".

3.9.2.1 Intra-modular connectors

A very unexpected connectivity pattern was exhibited by the intra-modular connectors shown as purple squares in figure 17. They were observed to be co-expressed with numerous other inter-modular hub genes in both connecting modules. Since we presumed that each knocked out cluster had a different expression pattern and it was also anticipated that each cluster is distinct from the other either in molecular function or their pattern of expression or possibly both, therefore, these connectors also had possible roles in both modules, and as multi-functional genes they were regulating several cellular and sub-cellular pathways. With all these presumptions it was essential to investigate their annotations. Hence, it was found out:

i. AT3G18070 (BGLU43, BETA GLUCOSIDASE 43), was located in chloroplast, extracellular region, involved in carbohydrate metabolic process, defense response to bacterium, macromolecule catabolic process, negative regulation of gene expression and tissue development (Depuydt & Vandepoele, 2021), and enables hydrolase activity, beta-glucosidase activity, and hydrolyzing O-glycosyl compounds, and is expressed in guard cell (Obulareddy et al., 2013).

ii. AT5G08030 (GDPD6, GLYCEROPHOSPHODIESTER PHOSPHODIESTERASE 6), is involved in lipid metabolic process, and glycerophosphodiester phosphodiesterase activity and expressed in flower, sepal, and collective leaf structure (Schmid et al., 2005)).

iii. AT1G12700 (RPF1, RNA PROCESSING FACTOR 1), a pentatricopeptide repeat (PPR) protein that acts within mitochondrial mRNA modification (Hölzle et al., 2011), enables mitochondrial mRNA 5'-end processing (Schleicher & Binder, 2021), located in chloroplast and mitochondrion, and expressed in guard cell and plant embryo).

iv. AT3G47010 a glycosyl hydrolase family protein, that is involved in regulation of defense response to other organisms, regulation of defense response (Depuydt & Vandepoele, 2021), response to organonitrogen compound, carbohydrate metabolic process, glucan catabolic process. Its located in cytoplasm and chloroplast, enables beta-glucosidase activity, scopolin beta-glucosidase activity and expressed in vascular leaf, collective leaf structure, plant sperm cell, seed, inflorescence meristem, endosperm, flower, petiole, root, hypocotyl, sepal, plant embryo, leaf lamina base, shoot apex, cauline leaf, leaf apex, flower pedicel, shoot system, cotyledon, stem, guard cell (Schmid et al., 2005).

v. AT3G02290 belongs to RING/U-box superfamily protein, that's located in nucleus, involved in protein ubiquitination, and enables ubiquitin-protein transferase activity. The expression was observed in cauline leaf, carpel, leaf apex, vascular leaf, plant embryo, collective leaf structure, stamen, shoot apex, flower, root, shoot system, petal, seed, flower pedicel, sepal, hypocotyl, pollen, inflorescence meristem, cotyledon, guard cell, leaf lamina base, petiole, stem by (Schmid et al., 2005).

vi. AT4G09490 a Polynucleotidyl transferase, ribonuclease H-like superfamily protein, has an unknown biological process, and enables nucleic acid binding and RNA-DNA hybrid ribonuclease activity (Tair), is expressed in cauline leaf, cotyledon, vascular leaf, petal, carpel, flower, leaf lamina base, petiole, stamen, root, collective leaf structure, guard cell, flower pedicel, plant embryo, sepal, seed, inflorescence meristem, hypocotyl, shoot system, shoot apex, leaf apex, stem, (Schmid et al., 2005).

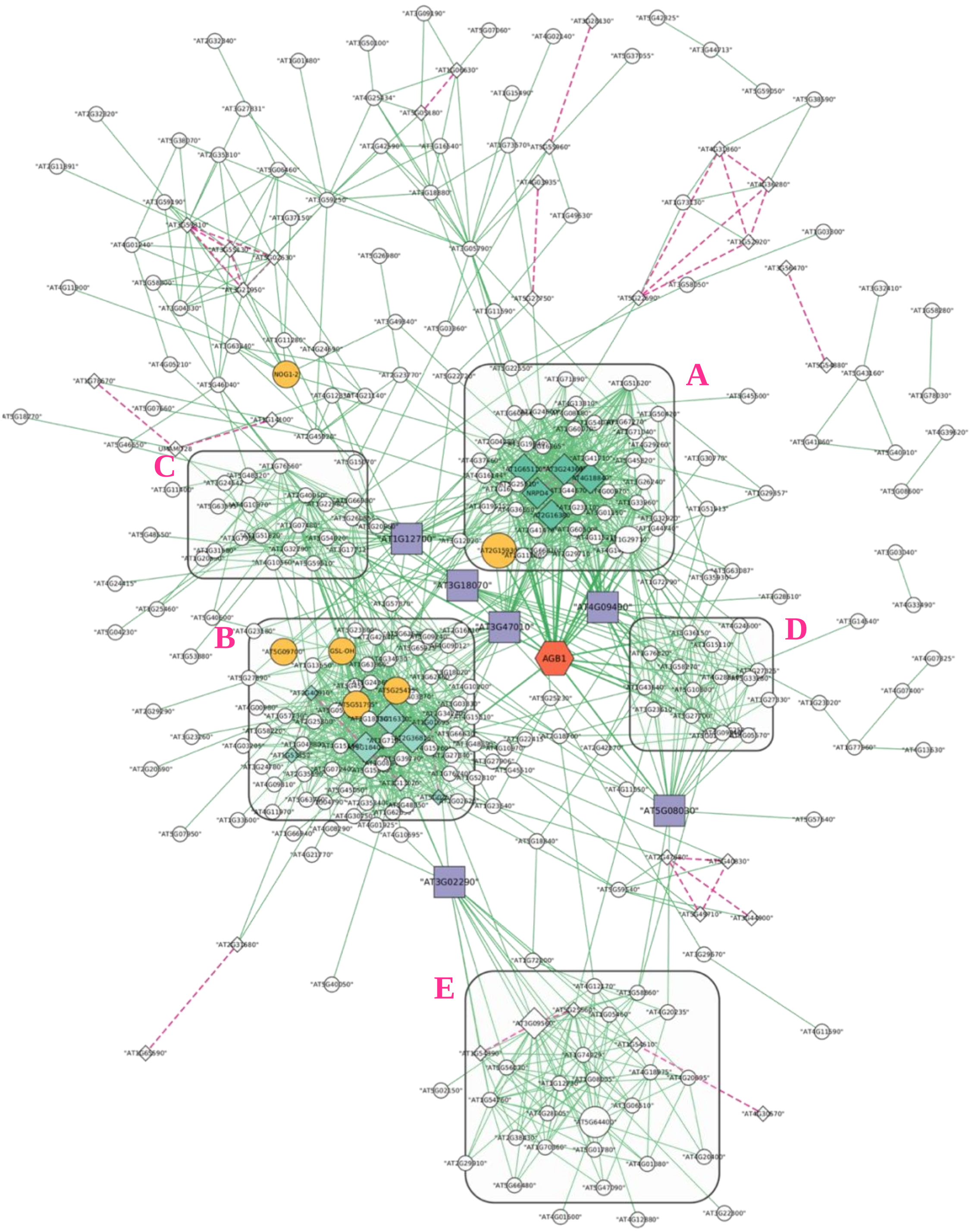


Figure 18: Co-occurrence network of over-represented genes. Green edges indicate the positive correlation. Pink dashed edges denote topmost strong positive τ ≥0.7 to perfect positive relationship τ = 1. Increased node size indicates hub genes where different size denotes degree of the node. Orange nodes represent hub genes in both over (positive correlated) and under-represented (negative correlated) connections. Green diamonds indicate inter-modular hub genes with tau τ >= 0.7. Purple squares indicate intra-modular connector between two distinct modules while red hexagon is the connecting gene between four adjacent modules.

3.9.2.2 Inter-modular hub genes

By looking at the network in figure 17, it was clearly noticeable that highly significant inter-modular hub genes (presented as sea green diamonds) possessing edge weight in the range of 0.7 to 1.0 had higher number of connecting nodes compared to the ones with weight in the lowest range i.e., less than 0.4. Due to their surprisingly high number of connections, there was high curiosity to know who these inter-modular hubs were and where were they expressed? Therefore, few hub genes were selected from each module to examine their molecular functions.

Module A was clustered with the genes mostly expressed in different parts of leaf as well as root and guard cell. Some hub and peripheral genes were involved in signal transduction and cellular transport e.g., AT2G16380 (Sec14p-like phosphatidylinositol transfer family protein), AT1G51620 (Protein kinase superfamily protein) and AT1G60070 (AP1G1, AP-1 COMPLEX GAMMA SUBUNIT 1) and also involved in immune homeostasis e.g., AT1G65110 (Ubiquitin carboxyl-terminal hydrolase-related protein); AT4G15950 (NRPD4, RNA-DIRECTED DNA METHYLATION 2) and auto immune response e.g., AT3G44670 (DANGEROUS MIX2H, DM2H) There were few other genes expressed in mitochondrion and were likely involved in seed development, Cell growth and metabolism e.g., AT3G24360 (ATP-dependent protease/crotonase family protein, involved in valine catabolic process). Furthermore, the GO molecular function terms revealed enrichment in hydrolase activity, protein binding, catalytic activity, nucleotide binding and unknown molecular functions for most of the genes.

Genes clustered in module B showed major involvement in regulation of gene expression e.g., AT2G36815, AT5G51795 (DNA/RNA-binding protein Kin17) and AT5G18404 (evidently involved in transcription) and were also acting in stress response to light or other environmental stimuli e.g., AT2G25450 (GSL-OH, GLUCOSINOLATE HYDROXYLASE) and AT5G18404 (acts in response to light stimulus, pigment biosynthetic process). Moreover, there were some other genes also involved in cellular detoxification and transport e.g., AT1G15160 (MATE efflux family protein), AT2G34240 (ubiquitin carboxyl-terminal hydrolase-like protein). Interestingly AT5G45050 (WRKY16, encodes a member of the WRKY Transcription Factor family)(tair) was found to be active in several diverse regulation pathways e.g., cellular defense response to other microcellular organisms, immune response and activation of signal transduction, and transcription regulation. The Go enrichment analysis identified multiple genes including hub genes with unknown molecular functions e.g., AT5G16330 (NC domain-containing protein-like protein). It was also discovered that most of the genes in the cluster were expressed in flower and guard cell and located in cytoplasm, mitochondrion, and nucleus, whereas few in plasma membrane.

Module C was a relatively smaller cluster of genes mostly expressed in guard cell, which exhibited their functions in regulation of growth and development in response to certain stimuli including anoxia, heat, Abscisic acid, Jasmonic acid and activation of defense response to other micro-organisms e.g., AT1G76560 (CP12-3, CP12 DOMAIN-CONTAINING PROTEIN 3), AT5G59510 (RTFL5, ROTUNDIFOLIA LIKE 5, DEVIL 18, DVL18). Another hub gene in the cluster was involved in aging which may be preceded as growth maturation, e.g., AT2G40050 (Cysteine/Histidine-rich C1 domain family protein). Some other genes were regulating gene expression including AT1G07480 (Transcription factor IIA, alpha/beta subunit).

In Module D most of the genes were involved in cellular regulation of seed germination, and flowering time with adaptations to stress i.e., long-day photoperiodism e.g., AT3G27700 (zinc finger (CCCH-type) family protein), AT3G23610 (DSPTP1, DUAL SPECIFICITY PROTEIN PHOSPHATASE 1) and AT5G10800 (RNA recognition motif (RRM)-containing protein). Genes in the cluster were expressed in flower, flower pedicel, guard cell, plant embryo and sepal. The GO slim analysis showed the functions in RNA binding, protein binding and some genes had unknown molecular functions.

Module E exhibited one of the distinguished cluster of genes with their functions in drought tolerance and response to abiotic stimuli e.g., AT5G64400 (AT12CYS-1, CHCH domain protein) and AT4G18975 (Pentatricopeptide repeat (PPR) superfamily protein); heat resistance, water deprivation, salt excess e.g., AT5G56030 (HSP81-2, HEAT SHOCK PROTEIN 81-2); response to temperature stimulus e.g., AT1G12730 (GPI transamidase subunit PIG-U). Some were also involved in regulation of lipid synthesis and cellular transport e.g., AT3G09560 (PAH1, PHOSPHATIDIC ACID

PHOSPHOHYDROLASE 1, lipin family protein). The genes were co-expressed in guard cells and regulating stomatal closure to prevent loss of water.

Molecular functions and expressions of major inter-modular hub genes and intra-modular connectors were explored, and several connecting factors were identified. In addition to cell growth and development, the most shared molecular functions included cellular transport, regulation of defense response, lipid biosynthesis, signal transduction, and regulation of gene expression.

3.9.2.3 Intra-modular AGB1 cluster

AGB1 (GTP BINDING PROTEIN BETA 1) part of Cul4-RING E3 ubiquitin ligase complex acts an intra-modular hub and connects with four modules labelled as A, B, C and D. By looking into the biological processes, it is a multifunctional gene involved in various biological process in multiple cellular locations e.g., response to ethylene, regulation of root development, root development (Pandey et al., 2008), lateral root development, seed germination, (Trusov et al., 2008), reactive oxygen species metabolic process (Joo et al., 2005), fruit development, plant organ morphogenesis (Lease et al., 2001), jasmonic acid mediated signaling pathway, defense response to fungus (Llorente et al., 2005; Trusov et al., 2006), response to extracellular stimulus (Tanaka et al., 2010), stomatal movement (Yu et al., 2018), endoplasmic reticulum unfolded protein (S. Wang et al., 2007). It has shown involvement in regulation of root development regulation and enables GTPase activity (Lease et al., 2001), signaling receptor complex adaptor activity and protein binding (Chang et al., 2009; Fan Liu-Min et al., 2008; Heo et al., 2012; Jones et al., 2014; Lee et al., 2008; Mason et al., 2000; Mudgil et al., 2009; Urano et al., 2012; Wang et al., 2008; Yu et al., 2016) . Also expressed in various cellular locations i.e., hypocotyl, cotyledon, cauline leaf, pollen, flower pedicel, carpel, flower, sepal, leaf apex, petiole, shoot system, guard cell, seed, vascular leaf, petal, portion of vascular tissue, root, base, stamen, inflorescence meristem, hydathode, plant embryo, collective leaf structure, leaf lamina rosette leaf, shoot apex and stem.

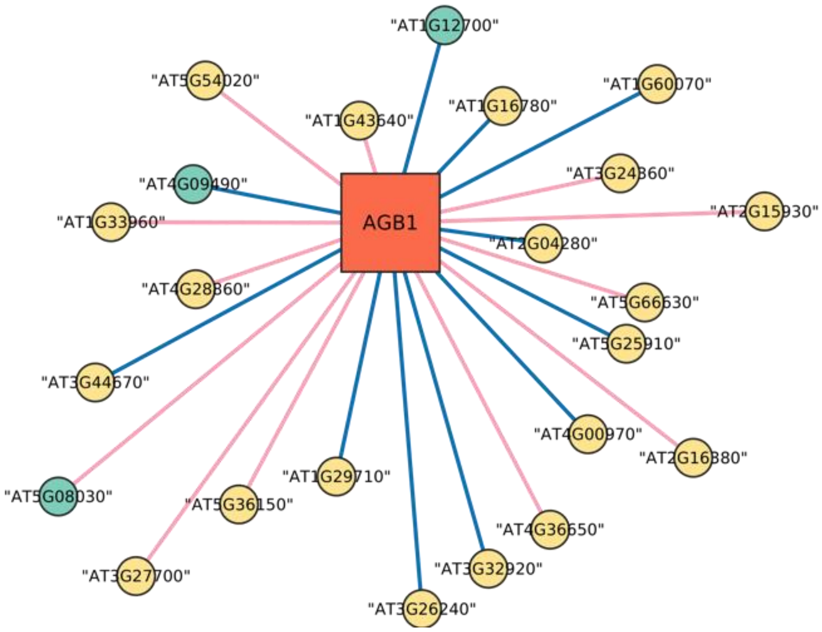


Figure 19: Common connections of AGB1 with first direct neighbors. Gene represented in green circle nodes are inter-connecting genes between 2 modules in over-represented network. Yellow color nodes denote peripheral genes. Blue edges show moderate (τ = 0.6 – 0.4) interconnection while pink edges show weak (τ < 0.3) connections.

We discovered that AGB1 was collectively co-expressed in 14 common accessions IDs from US including: (1684) Haz-10, (1741) KBS-Mac-74, (2017) MNF-Pin-40, (2031) Map 8, (2212) Pent-46, (2370) Yng-4, (687) LI-EF-018, (728) LI-SET-019, (7515) RRS-10, (506) BRR60, (7033) Buckhorn Pass, (7377) Tul-0, (2278) SLSP-35 and (8233) Dem-4 highly significantly (p-value: $2.5x10^{-23}$) with AT4G09490 (Polynucleotidyl transferase, ribonuclease H-like superfamily protein, enables RNA-DNA hybrid ribonuclease activity, nucleic acid binding); at (p-value: $6.42x10^{-18}$) with AT1G12700 (RPF1, RNA PROCESSING FACTOR 1, a pentatricopeptide repeat (PPR) protein acts upstream of or within mitochondrial mRNA modification, located in chloroplast and mitochondrion); and at (p-value: $1.2x10^{-8}$) with AT5G08030 (GDPD6, GLYCEROPHOSPHODIESTER PHOSPHODIESTERASE 6), involved in lipid metabolic process and enables glycerophosphodiester phosphodiesterase activity).

Other fairly strong co-expressions were discovered with:

i) AT1G16780 (VHP2;2, involved in proton transmembrane transport, enables pyrophosphate hydrolysis-driven proton transmembrane transporter activity and inorganic diphosphatase activity) at p-value $2.08x10^{-15}$; ii) AT1G60070 (AP1G1, AP-1 COMPLEX GAMMA SUBUNIT 1, involved in Golgi to vacuole transport, enables clathrin adaptor activity) (p-value: $1.18x10^{-13}$); iii) AT2G04280 (Calcium ion-binding protein acts upstream of or within cell cycle, cell growth, cell morphogenesis (Depuydt & Vandepoele, 2021), enables glycosyltransferase activity) (Nikolovski et al., 2012) with co-expression coefficient (p-value: $3.31x10^{-11}$); iv) AT5G25910 (RLP52, RECEPTOR LIKE PROTEIN 52, recognized disease resistance protein, showed involvement in defense response to chitin and fungus (Ramonell et al., 2005) signal transduction (Kobe & Kajava, 2001), enables kinase activity) shared 32 common accessions at (p-value: $2.06x10^{-12}$); v) AT4G00970 (CRK41, CYSTEINE-RICH RLK(RECEPTOR-LIKE PROTEIN KINASE) 41, encodes a cysteine-rich receptor-like protein kinase, located in chloroplast) having connection significance with (p-value: $9.96x10^{-17}$); vi) At p-value: $4.26x10^{-13}$ with AT3G32920 (P-loop containing nucleoside triphosphate hydrolases superfamily protein, involved in DNA repair, located in chloroplast, enables single-stranded DNA binding); vii). AT3G26240 (Cysteine/Histidine-rich C1 domain family protein, acts upstream of or within response to light stimulus (Depuydt & Vandepoele, 2021), involved in unknown molecular function, had 44 connections at (p-value: $3.20x10^{-12}$); viii) AT1G29710 (DYW4, DYW DOMAIN PROTEIN 4, Tetratricopeptide repeat (TPR)-like superfamily protein, shared 41 common accessions (p-value: $2.48x10^{-14}$); ix) AT3G44670 (DM2H, DANGEROUS MIX2H, one of a series of RPP1-like, tandemly duplicated Toll-Interleukin1-Receptor- related NLR receptors within the DANGEROUS MIX2 cluster, involved in defense response, defense response to bacterium (Botella et al., 1998), plant-type hypersensitive response (Stuttmann et al., 2016), signal transduction, defense response to oomycetes, active in nucleus (Ordon et al., 2021), enables ADP binding) significantly co-related with (p-value: $5.61x10^{-14}$).

Upon further analysis, it was revealed that all the co-occurring genes were mutually expressed in guard cell (Obulareddy et al., 2013) and plant embryo (that co-relates with the growth and development function of AGB1 directly). Moreover, more than half of the genes in the cluster were equally expressed in stamen, cauline leaf, seed, carpel, petiole, cotyledon, leaf lamina base, inflorescence meristem, vascular leaf, shoot apex, root, shoot system, collective leaf structure, leaf apex, flower pedicel, flower, hypocotyl, petal, sepal, and stem (Schmid et al., 2005). Overall, genes in the whole cluster were involved in morphogenesis, cellular transport, and defense response to various pathogens.

3.9.3Custer formation extracted through MCODE

We identified several modules in GCN where a whole set of gene pairs were knocked out together in several geographical locations e.g., Southern Europe including Spain, Italy, and a smaller region in Portugal, Sweden in Northern Europe, Georgia and Armenia from Caucasus, North America and southern Russia that indicated a common environmental gradient possessed by the genes and their expression in common geographical locations. These highly interconnected regions in the over-represented network were determined through MCODE in Cytoscape tools. We performed the analysis by looking at the hub genes (degree), their connection strength (co-expression coefficient Kendall tau) common accessions (connection), expression in the cell and GO enrichment. At first, we extracted the modules by using clustering analysis of MCODe in Cytoscape than each of these modules with their biological relevance and interpretation is discussed in the section below.

3.9.3.1 Module 1

This module was shaped with highest score of 19.758 with 34 nodes and 326 edges extracted where down regulated overrepresented co-occurring genes were strongly co-expressed in different parts of leaf including: collective leaf structure, cotyledon, leaf apex, petiole, stamen, vascular leaf (Schmid et al., 2005) and guard cell (Obulareddy et al., 2013). Overall, there were five interconnecting genes which paired highly significantly in multiple accessions, subsequently we explored their occurrences and molecular functions in the cell. As an example a hub gene: AT4G15950 (NRPD4, RNA-DIRECTED DNA METHYLATION 2, function in gene silencing, and part of RNA polymerase IV and V complex (Ream et al., 2009), located in nucleus (He et al., 2009), enables nucleotide binding) was strongly connected ($\tau \geq 0.7$) with 4 other hub genes: AT4G18840 (Pentatricopeptide repeat (PPR-like) superfamily protein, located in mitochondrion, have unknown molecular function) co-expressed with (τ = 0.7, p-value: 3.72x10^{-65}) in 22 common accessions; AT3G24360 (ATP-dependent caseinolytic (Clp) protease/crotonase family protein, involved in valine catabolic process, located in chloroplast and mitochondrion, enables hydrolase activity) co-occurred (τ = 0.8, p-value: 2.39x10^{-102}) in 20 common accessions; AT2G16380 (Sec14p-like phosphatidylinositol transfer family protein, acts in regulation of signal transduction, response to light intensity (Depuydt & Vandepoele, 2021), organic cyclic compound, located in Golgi apparatus and involved in transporter activity) co-expressed with (τ = 0.7, p-value: 1.84x10^{-82}) in 18 shared accessions; AT1G65110 (Ubiquitin carboxyl-terminal hydrolase-related protein, located in nucleus, enables cysteine-type deubiquitinase activity, involved in hydrolase activity) with (τ = 0.7 and p-value: 2.24x10^{-68}) shared 21 mutual accessions. Furthermore, AT2G16380 and AT3G24360 showed a robust co-expression at (τ = 0.9, p-value: 8.61x10^{-107}) in 18 common accessions. Moreover, they were also found to be strongly co-expressed with AT1G65110 at (τ = 0.7, p-value: 1.04x10^{-80}) in 20 common accessions. Beside these strongest connections, two other hub genes: AT3G60440 (Phosphoglycerate mutase family protein located in chloroplast (Zybailov et al., 2008), enables unknown molecular function) and AT3G44670 (DM2H, DANGEROUS MIX2H, one of a series of RPP1-like, involved in defense response to bacterium (Botella et al., 1998), plant-type hypersensitive response (Stuttmann et al., 2016), signal transduction, defense response to oomycetes, active in nucleus (Ordon et al., 2021), enables ADP binding) were connected with NRPD4 having slightly less coefficient score compared to the latter discussed above nevertheless very strong at (τ = 0.6, p-value: 2.36x10^{-53}) and (τ = 0.6, p-value: 5.26x10^{-54}) respectively.

Furthermore, all the hub genes i.e., AT4G18440, AT3G60440, AT3G24360 and AT3G16380 apart from connections with each of the other hub genes, were also strongly interconnected with peripheral genes with co-expression coefficient tau score between 0.4 and 0.6 in the cluster and also showed connections with outliers (weak correlation). Peripheral genes were mostly found in strong connection with hub genes and comparatively less connected with outliers. In this context outliers are also significant in the network but only their co-expression coefficient is low compared to peripheral genes. We identified AT1G60500 (DRP4C, DYNAMIN RELATED PROTEIN 4C, located in cytoplasm, enables microtubule binding, expressed in plant egg cell)(Wuest et al., 2010); AT2G41700 (ABCA1, ATP-BINDING CASSETTE A1, located in mitochondrion, enables amino acid transmembrane transporter activity, ATPase-coupled transmembrane transporter activity)(Ward, John, 2002); AT1G51620 (Protein kinase superfamily protein, involvement in phosphorylation of protein, located in nucleus, enables ATP binding, protein kinase activity); AT4G00970 (CRK41, CYSTEINE-RICH RLK (RECEPTOR-LIKE PROTEIN KINASE) 41, located in chloroplast); AT1G33960 (AIG1, AVRRPT2-INDUCED GENE 1, involved in response to bacterium (Reuber & Ausubel, 1996), located in chloroplast and mitochondrion); AT5G66810 (Ran-binding protein in the microtubule-organizing center protein, located in nucleus, involved in catalytic activity and enables unknown molecular function) as peripheral genes.

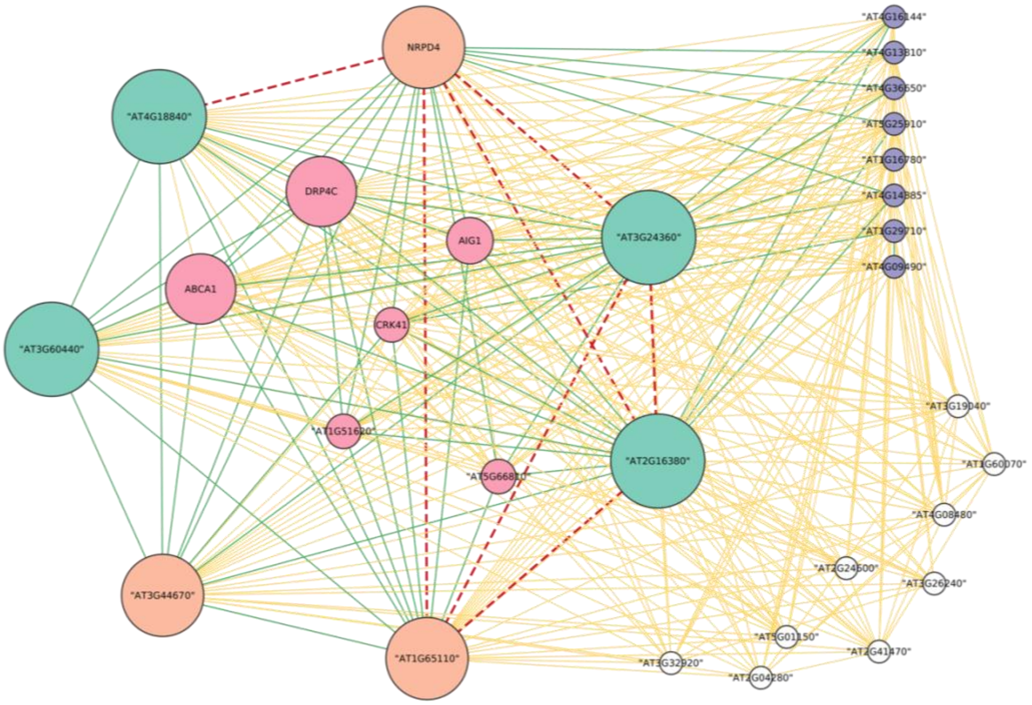


Figure 20: US Module in overrepresented network. with common expression in leaf. Red dashed edges indicate strongest correlation (tau >= 0.7) between connecting nodes. Green edges represent moderate strength connections (tau = 0.4-0.6) while yellow edges show weak connections (tau<=0.3). Green nodes are the hub genes (connected with each member of the network), Orange nodes are less connected as compared to hub genes. Pink nodes denote peripheral genes having strong connection with each of the hub genes while purples nodes are peripheral genes which are strongly correlated with either 1 or 2 hub genes. Nodes without color are outliers but still important in the network topology.

By looking further into the precise expression profiles of the clustered genes it was revealed that they were significantly knocked out together in 18 geographical locations with accessions from the US (listed in table below) Moreover, the GO enrichment of these genes revealed their involvement in protein binding, hydrolase activity, kinase activity and nucleotide binding including roles in DNA and RNA binding. Few genes were observed to be involved in unknown molecular functions irrespective of their coefficient scoring.

Table 7: Geographical locations of significantly knock out accession Ids from US

Accession Ids	Name	Latitude	Longitude
506	BRR60	40.8313	-87.735
687	LI-EF-018	40.9064	-73.1493
728	LI-SET-019	40.9352	-73.114
870	MIC-31	41.8266	-86.4366
1684	Haz-10	41.879	-86.607
1741	KBS-Mac-74	42.405	-85.398
1872	MNF-Pot-75	43.595	-86.2657
2017	MNF-Pin-40	43.5356	-86.1788
2031	Map-8	42.166	-86.412
2212	Pent-46	43.7623	-86.3929
2370	Yng-4	41.865	-86.646
7033	Buckhorn Pass	41.3599	-122.755
7248	Mv-0	41.3923	-70.6652
7377	Tul-0	43.2708	-85.2563
7515	RRS-10	41.5609	-86.4251
8077	PT2.21	41.3423	-86.7368
8132	RMX3.22	42.036	-86.511
8233	Dem-4	41.1876	-87.1923

3.9.3.2 Module 2

The second extracted module with a score of 15.4 was made of 21 nodes and 154 edges where five genes were co-expressed with highly significant strongest connections. Interestingly not all these genes were acting as hub genes, rather their connections were limited to the genes with highest tau scores meaning reduction of noise data. Only two genes i.e., AT1G15160 (MATE efflux family protein, located in plasma membrane, involved in transmembrane export) and AT5G51795 (DNA/RNA-binding protein Kin17, located in cytoplasm, chloroplast, and nucleus, enables double stranded DNA binding) were identified as hub genes which were significantly and strongly co-expressed with peripheral genes and outliers.

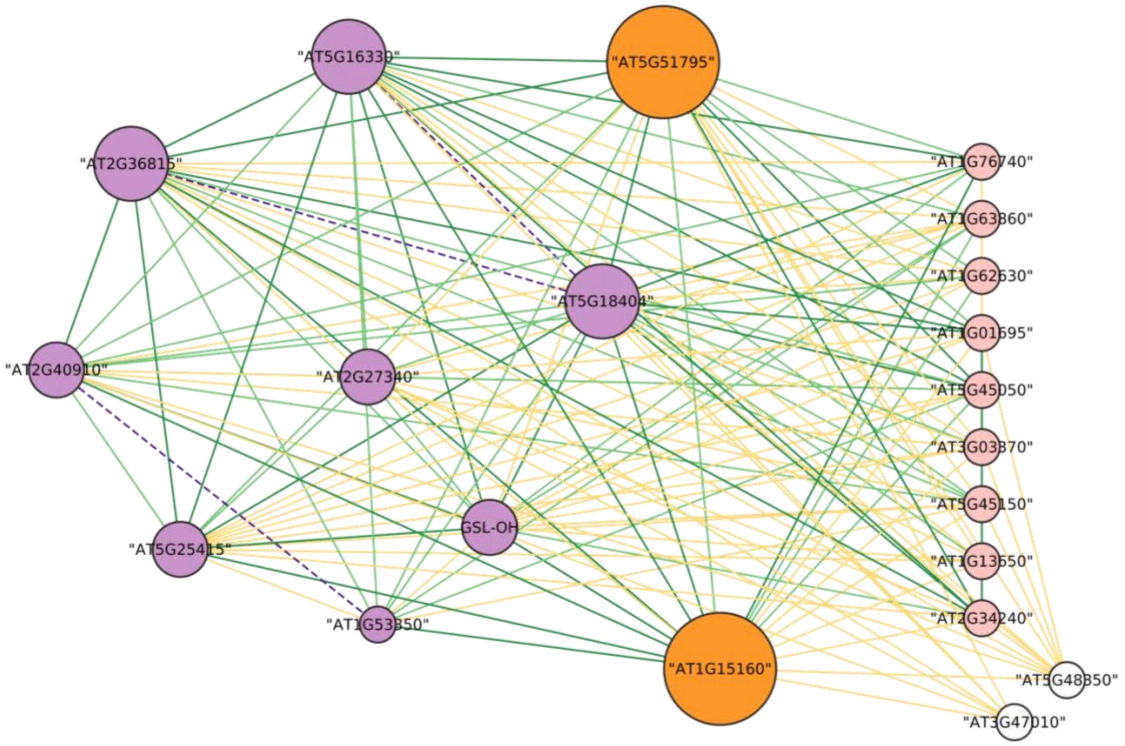


Figure 21: Russian module with common expression pattern in guard cells. Size of the nodes indicate degree of the node. Purple dashed edges indicate strongest correlation (tau >= 0.7) between connecting nodes. Green edges represent connections with moderate co-expression coefficient (tau = 0.4-0.6) while yellow edges show weak connections (tau<=0.3). Orange large nodes are the hub genes (connected with each member of the network), purple nodes are categorized as peripheral genes (less connected as compared to hub genes). Pink nodes denote peripheral genes having strong connection with hub genes (orange) as well as other peripheral genes (purple). Nodes without colour are outliers.

Examining the connections with strong co-expression coefficient, AT5G18404 (involved in transcription, acts in response to light stimulus (Depuydt & Vandepoele, 2021), pigment biosynthetic process and enables unknown molecular function) was co-occurred in 46 common accessions with AT5G16330 (NC domain-containing protein-like protein, enables unknown molecular function) at (τ = 0.8, p-value: 2.31x10^{-90}) and with AT2G36815 (involved in mRNA cis splicing, via spliceosome, enables unknown molecular function) at (τ = 0.7, p-value: 1.19x10^{-67}) in 40 shared accessions. A closer look at their origin revealed that few accessions were knocked out from Uzbekistan, Kazakhstan, and Kyrgyzstan while most of them had origin from Russia. Furthermore, AT2G40910 (enables unknown molecular function) was interconnecting with AT1G53350 (Disease resistance protein (CC-NBS-LRR class) family, enables ADP binding), at (τ = 0.7, p-value: 6.62x10^{-74}) in 9 accessions and interestingly they were also knocked out from Russia (9619, 9620, 9621, 9631, 9634, 9636, 9638, 9640, 9642). To get better understanding of the module structure and functional relevance of co-occurring genes, the moderately strong connections with tau >= 0.5 were also reviewed, few of the examples include: AT5G51795 co-expressed in 35 accessions with AT5G16330 at (τ = 0.6, p-value: 1.13x10^{-62}); and AT5G18404 in 32 accessions at (τ = 0.6, p-value: 4.27x10^{-59}). Furthermore, AT5G16330 co-occurred with AT2G36815 in 41 accessions at (τ = 0.6, p-value: 5.13x10^{-62}); AT1G01695 (TRM33, TON1 RECRUITING MOTIF 33, located in nucleus and enables unknown molecular function) at (τ = 0.6, p-value: 1.46x10^{-56}) in 33 accessions and with AT2G34240 (ubiquitin carboxyl-terminal hydrolase-like protein, located in nucleus and enables unknown molecular function) in 36 accessions with (τ = 0.5, p-value: 1.44x10^{-43}). Furthermore, AT2G40910 and AT1G15160 were co-expressed in 14 common accessions at (τ = 0.6, p-value: 5.36x10^{-55}) and they were all knocked out in several geospatial places across Russia.

Additionally, the analysis of some dominant peripheral genes revealed that gene AT5G18404 and AT5G45050 (WRKY16, encodes a member of the WRKY Transcription Factor family, involved in bacterial defense response, intrinsic immune response-activating signal transduction, regulation of transcription regulation, DNA-templated, signal transduction, located in plant-type vacuole, enables DNA binding transcription factor activity and protein binding) were co-expressed in 19 accessions from Russia at (τ = 0.5, p-value: 3.68x10^{-32}). Moreover, AT5G18404 was also co-expressed in 27 accessions at (τ = 0.5, p-value: 1.63x10^{-45}) with AT2G25450 (GSL-OH, GLUCOSINOLATE HYDROXYLASE, 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein, involvement in regulation of glucosinolate biosynthetic process (Hansen et al., 2008), hydrolase activity and enables 1-

aminocyclopropane-1-carboxylate oxidase activity) (Trentmann & Kende, 1995) and AT5G25415 (hypothetical protein (DUF239), enables unknown molecular function) in 41 common accessions at (τ = 0.5, p-value: 6.76x10^{-45}).

A very surprising fact from expression profiles of the mutated genes revealed that all the clustered genes were commonly expressed in guard cell (Obulareddy et al., 2013) though few were also expressed in other locations and some had missing expression information. Moreover, the cluster was commonly expressed in the listed knock out accessions at numerous locations throughout Russia in table 8. Most of the genes in the cluster were enriched with protein binding, DNA binding, nucleotide binding and transporter activity however several others were found to have unknown molecular functions.

Table 8: Geographical locations of significantly knock out accession Ids from Russia.

Accession Ids	Name	Latitude	Longitude
9607	Panik-1	53.05	52.15
9611	Lesno-1	53.04	51.9
9615	Parti-1	52.99	52.16
9616	Krazo-1	53.06	51.96
9619	Basta-1	51.84	79.48
9620	Basta-2	51.82	79.48
9621	Basta-3	51.84	79.46
9626	Kolyv-3	51.36	82.59
9627	Kolyv-5	51.32	82.55
9628	Kolyv-6	51.33	82.54
9629	K-oze-1	51.35	82.18
9631	Lebja-1	51.65	80.79
9632	Lebja-2	51.67	80.82
9633	Lebja-4	51.63	80.83
9634	Masl-1	54.13	81.31
9636	Noveg-1	51.75	80.82
9639	Panke-1	53.82	80.31
9640	Rakit-1	51.87	80.06
9642	Rakit-3	51.84	80.06
9951	Kly-1	51.3333	82.5667
9953	Koz-2	51.33	82.19

3.9.3.3 Module 3

The module was formed with 11 nodes and 42 edges with a collective score of 6.18 with the gene pairs connected with strong co-expression coefficients shown in figure 21. We identified two functional hub genes in the cluster i.e., AT12CYS-1 and HSP81-2 wherein AT5G64400 (AT12CYS-1, CHCH domain protein; involved in mechano-transduction and drought tolerance; acts in response to abiotic stimulus (Y. Wang et al., 2016); located in nucleus and mitochondrion; enables unknown molecular function, and provides cell-to-cell mobile RNA) (Thieme, et al., 2015) co-expressed with all other hub, peripheral and outlier genes with moderately strong co-expression coefficient between 0.5 and 0.6 and highly significant p-values and AT5G56030 (HSP81-2, HEAT SHOCK PROTEIN 81-2 and ERD8, EARLY-RESPONSIVE TO DEHYDRATION 8; involved in defense response, protein stabilization (Hubert et al., 2003), heat acclimation (Lim et al., 2006), , response to heat (Takahashi et al., 1992), salt stress, water deprivation (H. Song et al., 2009), and regulates stomatal closure (Clément et al., 2011); located in Golgi apparatus, cytoplasm, mitochondrion, plant-type cell wall, plasmodesma and plastid; enables ATP binding, ATP hydrolysis activity, mRNA binding and protein binding and produces cell-to-cell mobile RNA) strongly co-expressed with central hub genes including PAH1, AT12CYS-1AT5G25560, AGL85, AT1G74929 with edge weight ranging between 0.7 and 0.8 depicting strong interconnections in regulation pathways. Due to the fundamental importance in the module performing as central regulators and their support function of timely adaptations to the ongoing variations in the microenvironment, they have been named functional hub genes.

Moreover, there were four central core genes identified i.e., AT5G25560, AT1G74929, PAH1 and AGL85 where three of them possessed connections with strongest co-expression coefficient scores i.e., AT5G25560 (CHY-type/CTCHY-type/RING-type Zinc finger protein, located in nucleus, provides cell-to-cell mobile RNA as gene product) (Thieme, et al, 2015) co-expressed with AT3G09560 (PAH1, PHOSPHATIDIC ACID PHOSPHOHYDROLASE 1, lipin family protein, located in nucleus Golgi membrane and protein storage vacuole membrane, encodes a phosphatidate phosphohydrolase, involved in galactolipid biosynthetic process, phospholipid biosynthetic process (L. Wang et al., 2014), lipid metabolic processes, cellular response to phosphate starvation (Nakamura et al., 2009), intracellular protein transport) (Shen et al., 2011) in 13 common accessions at (τ = 0.7, p-value: 1.91x10^{-68}); AT1G54390 (ING2, INHIBITOR OF GROWTH 2, located in nucleus, enables methylated histone binding) (W. Y. Lee et al., 2009) in 10 accessions at (τ = 0.7, p-value: 4.39x10^{-53}) and AT1G54760 (AGL85, AGAMOUS-LIKE 85, located in nucleus, involved in transcription regulation, enables DNA-binding transcription factor activity), (Parˇenicová et al., 2003; Riechmann. et al., 2000) in 13 accessions at (τ = 0.5, p-value: 4.11x10^{-33}). Additionally, ING2 co-occurred in 10 shared accessions with PAH1 at (τ = 0.7, p-value: 8.70x10^{-66}), moreover, both genes also strongly co-expressed with AGL85 at (τ = 0.7, p-value: 2.72x10^{-55}) in 13 accessions and 14 accessions with coefficient score (τ = 0.7, p-value: 1.95x10^{-48}) respectively. Investigation of connections with peripheral genes showed that PAH1 connected at a moderate co-expression coefficient score of (τ = 0.5, p-value: 1.13x10^{-41}) in 12 accessions with AT2G38430 and (τ = 0.5, p-value: 4.85x10^{-43}) in 9 accessions with AT1G74929 (a hypothetical protein involved in response to light intensity, and involved in unknown molecular function), and AT1G70360 (F-box family protein, located in Mitochondrion and enables unknown molecular function) respectively. Additionally, AT1G12730 (GPI transamidase subunit PIG-U, located in mitochondrion, acts in response to temperature stimulus) another peripheral gene of great importance in the cluster co-occurred with PAH1, AT1G74929, AT1G70360, AGL85, AT5G25560 at moderate co-expression coefficient score of 0.5 and 0.6 with highly significant p-values. Upon investigation of the hub genes along with their connections with peripherals and outliers it was revealed that the gene pairs were knocked out together in accession IDs from Spain listed in table 9 below.

Table 9: Geographical locations of significantly knock out accession Ids from Spain.

Accession Ids	Name	Latitude	Longitude
9533	IP-Cem-0	41.15	-4.32
9542	IP-Fun-0	40.79	-4.05
9598	IP-Vim-0	41.88	-6.51
9545	IP-Her-12	39.4	-5.78
9871	IP-Nac-0	40.75	-3.99
9879	IP-Per-0	37.6	-1.12
9947	Ped-0	40.74	-3.9
9543	IP-Gra-0	36.77	-5.39
9583	IP-Sne-0	37.09	-3.38
9554	IP-Lso-0	38.86	-3.16

Furthermore, we also noticed that the whole gene cluster was commonly located in nucleus and expressed in guard cell (Obulareddy et al., 2013), nevertheless, most of them were also reported to be expressed in cauline leaf, leaf lamina base, plant embryo, stamen, flower pedicel, leaf apex, petal, seed, vascular leaf, carpel, cotyledon, shoot apex, collective leaf structure, petiole, hypocotyl, flower, inflorescence meristem, pollen, root, sepal, shoot system, stem (Schmid et al., 2005). GO enrichment of the clustered genes showed involvement in multiple and variable molecular functions with single genes performing several functions including protein binding, DNA binding transcription factor activity, hydrolase activity, DNA binding, nucleic acid binding, catalytic activity and few were found to have unknown molecular function.

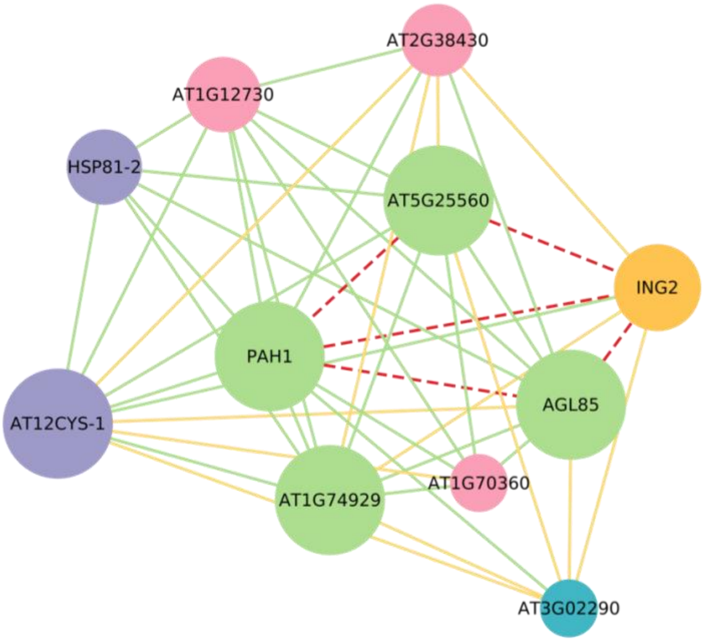


Figure 22: Spanish module I. The size of the node differs according to the node degree. Red dashed edges denote strongest correlation (tau >= 0.7). Green edges indicate moderately strong connections (tau = 0.4-0.6) while yellow edges represent weak connections (tau<=0.3). Purple nodes represent functional hub of the cluster whereas other hub genes are represented with large green nodes. Orange node indicate core gene having strong connection with hub genes (green). Pink node represents peripheral genes while blue node shows single outlier.

3.9.3.4 Module 4

We extracted another relatively small module formed of 6 nodes and 14 edges where all knocked out accession IDs were originated from Spain. We reviewed each connection in the cluster and incredible results were achieved. E.g., AT3G21950 (SABATH family methyltransferase, involved in methylation, located in nucleus) co-expressed with AT3G55130 (WBC19, WHITE-BROWN COMPLEX HOMOLOG 19, located in mitochondrion and vacuolar lumen, enables ATPase-coupled transmembrane transporter activity, involved in vacuolar transport) (Mentewab & Stewart, 2005) at (τ = 0.7, p-value: 1.07x10^{-79}); AT3G59310 (solute carrier family 35 protein (DUF914), involved in transmembrane transport, enables transmembrane transporter activity) at (τ = 0.8, p-value: 4.52x10^{-88}); and with AT5G02630 (7TM3, CAND6, CANDIDATE G-PROTEIN COUPLED RECEPTOR 6, lung seven transmembrane receptor family protein, involved in transport, G protein-coupled receptor signaling pathway (Gookin et al., 2008), located in chloroplast and active in membrane) at (τ = 0.7, p-value: 2.95x10^{-65}). Moreover, WBC19 co-occurred in 12 accessions with AT3G59310 at strongest coefficient score of (τ = 0.9, p-value: 1.64x10^{-115}) marking it highly significant in the cluster. Furthermore, AT3G59310 and 7TM3 were also found out to be expressed together in 12 accession IDs at a score of (τ = 0.7, p-value: 8.62x10^{-68}). Another important gene AT1G10300 (NOG1-2, involved in positive regulation of defense response to bacterium, stomatal movement (S. Lee et al., 2017), enables GTP binding, RNA binding, GTPase activity, located in nucleus) was discovered moderately interconnecting in the cluster with hub and central core genes like WBC19 and AT3G59310 in 13 accession IDs at (τ = 0.4, p-value: 6.16x10^{--22}) and 14 IDs with a score of (τ = 0.4, p-value:1.56x10^{-23}) respectively. Moreover, AT3G04330 was co-expressed with AT3G21950, AT3G59310 and WBC19 at moderate score of (τ = 0.4, p-value:1.50x10^{-20}), (τ = 0.4, p-value:2.21x10^{-22}) and (τ = 0.4, p-value:7.21x10^{-21}).

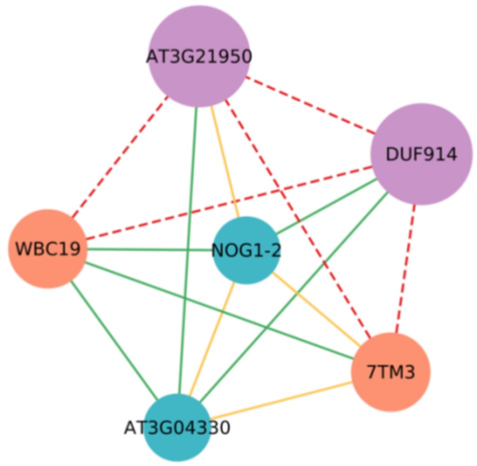


Figure 23: Spanish module II. Dashed edges in red indicate connections with strong coefficient score (tau >= 0.7), Green edges are moderate connections while yellow lines denote week connections. Purple nodes are the hub genes and simultaneously carry the strongest co-expression scores while also co-occurring with other genes. Orange nodes are distinguished in the form of central genes connecting highly strongly with hub genes. Blue nodes are peripheral genes.

The detailed analysis resulted in the following list of common accessions from Spain where the knock out genes were clustered at geographical locations listed in table 10 below.

Table 10: Clustered geographical locations in module 4 from Spain

Accession Ids	Name	Latitude	Longitude
9514	IP-Adm-0	39.15	-4.54
9515	IP-Ala-0	39.72	-6.89
9522	IP-Bea-0	36.52	-5.27
9537	IP-Cum-1	38.07	-6.66
9541	IP-Fue-2	38.26	-5.42
9560	IP-Mot-0	38.19	-6.24
9873	IP-Ndc-0	37.94	-5.45
9900	IP-Tri-0	37.38	-6.01
9943	Cdm-0	39.73	-5.74
9946	Mer-6	38.92	-6.34
9509	IP-Reg-0	39.29	-7.4
9511	IP-Vav-0	38.53	-8.02

Moreover, there were two accession IDs from Portugal i.e., IP-Reg-0 and IP-Vav-0 however, their latitudinal and longitudinal 2D coordinates lie in virtually similar geographical region. Furthermore, genes in the module were commonly expressed in guard cells although few genes were also found to be expressed in flower. It was also interesting to notice that few were also commonly located in nucleus while others in mitochondrion and cytoplasm. Additionally, studying their GO function, WBC19, 7TM3 and AT3G59310 were enriched with Transporter activity, AT3G21950 involved in transferase activity, whereas, NOG1-2 was involved in multiple processes i.e., hydrolase activity, nucleotide binding and other binding, moreover, AT3G04330 was found to have unknown molecular function (Ashburner et al., 2000).

3.9.3.5 Module 5

Given the size of the module with only 5 nodes and 10 edges, and because of the strong correlations between the genes it had comparatively a higher MCODE score = 5. Interestingly each of the gene was highly strongly connected to every other gene in the cluster and co-expressed in accession IDs from Italy. A profound analysis of the cluster showed promising results where AT1G52920 (GCR2, G-PROTEIN COUPLED RECEPTOR 2, is active in plasma membrane, located in nucleus and plasma membrane, enables abscisic acid binding and protein binding) (Liu et al., 2014) co-occurred with highly strong co-expression coefficient score ($\tau = 0.8$, p-value:1.10x10^{-104}), ($\tau = 0.8$, p-value:1.78x10^{-102}) and ($\tau = 0.8$, p-value:1.10x10^{-104}) with AT4G36280 (CRH1, CRT1 HOMOLOGUE 1, located in nucleus, involved in DNA repair regulation, positive defense response regulation to bacterium and virus (Kang et al., 2010, 2012), enables ATP hydrolysis activity, DNA binding, RNA binding, endonuclease activity, protein binding)(Moissiard et al., 2014); AT5G22690 (Disease resistance protein (TIR-NBS-LRR class) family, located in cytoplasm, involved in defense response, signal transduction and cellular response to oxygen-containing compound, hormone-mediated signaling pathway (Depuydt & Vandepoele, 2021), has gene product cell to cell mobile RNA (Thieme,et al, 2015), enables ADP binding) and AT4G31360 (selenium binding protein, located in nucleus, involved in regulation of gene expression, epigenetic, seed development, negative regulation of cellular process (Depuydt & Vandepoele, 2021), enables unknown molecular function) respectively. Furthermore, CRH1 co-expressed at the maximum coefficient score ($\tau = 1.0$, p-value:3.33x10^{-146}) with AT4G31360 depicting strong correlation in terms of function or expression, and with AT5G22690 (τ = 0.7, p-value:1.60x10^{-73}), whereas AT4G31360 and AT5G22690 were also observed to be strongly connected with each other at a score of (τ = 0.7, p-value:1.60x10^{-73}).

Further going down in the analysis, it was observed that AT1G73130 (ATI3C) had a significant role in the cluster due to its rare function and surprisingly it only co-occurred with genes in this specific cluster in the whole co-occurrence network which make its presence more reasonable. The thorough analysis exhibited (role in heat acclimation (Zhou et al., 2018), and response to stress, active in autophagosome and phagophore and located in the nucleus, enables protein binding, has gene product cell-to-cell mobile RNA) (Thieme, et al., 2015). The co-occurrence pattern showed moderate to fairly strong correlations with AT4G31360 at score (τ = 0.6, p-value:2.15x10^{-61}); CRH1 (τ = 0.6, p-value:2.15x10^{-61}), GCR2 (τ = 0.5, p-value:7.23x10^{-44}), and AT5G22690 (τ = 0.5, p-value:1.06x10^{-30}).

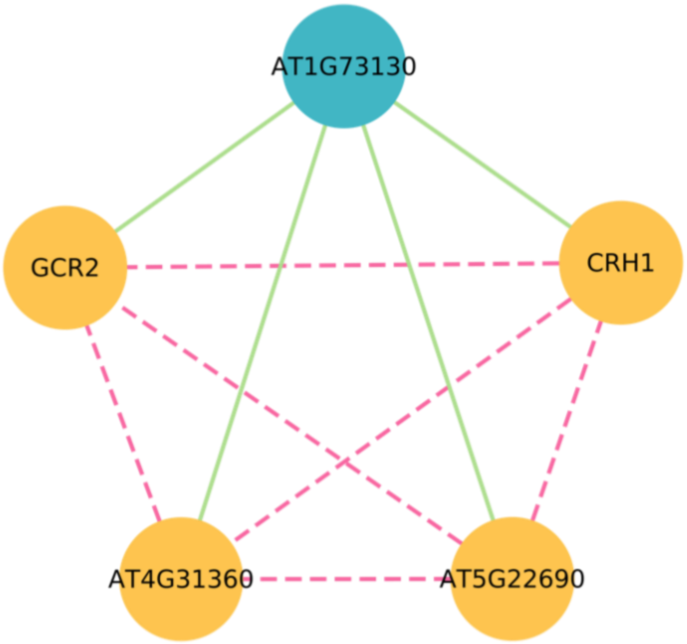


Figure 24: Italian module. Pink dashed edges represent strong correlations while green lined edges indicate moderate connections. Orange nodes are the strongly connected genes. Blue node has moderate connection with other genes. Size of the nodes is neutral due to same degree.

Co-occurrence of the whole cluster was observed in a series of accession Ids knocked out from Italy listed in table 11 below:

Table 11: Geographical locations of significantly co-expressed cluster from Italy

Accession Ids	Name	Latitude	Longitude
9679	Castelfed-1-195	46.34	11.29
9680	Castelfed-1-196	46.34	11.29
9681	Castelfed-1-197	46.34	11.29
9682	Castelfed-1-198	46.34	11.29
9683	Castelfed-1-199	46.34	11.29

The regulation pathways and molecular functions of the co-occurred genes indicated their mutual involvement in response to abiotic stress and defense mechanism. Gene expression pattern showed their presence in collective leaf structure, flower, plant embryo, shoot apex, seed, stem (Schmid et al., 2005) and guard cells (Obulareddy et al., 2013). GCR2, ATI3C, and CRH1 were enriched for protein binding however CRH1 was also found to be involved in nuclease activity, DNA and RNA binding. Moreover, AT5G22690 was enriched for GO terms including hydrolase activity, nucleotide binding and other binding while AT4G31360 was reported to have unknown molecular function.

3.9.3.6 Module 6

We discovered this small but highly significant and strongly connected module with 5 nodes and 10 edges (figure 24) where genes were paired in accession IDs knocked out from Georgia (Caucasus ecotypes). The co-expression analysis showed that AT1G06630 (F-box/RNI-like superfamily protein, located in nucleus, enables unknown molecular function) was strongly connected with AT5G05180 (myosin heavy chain, striated protein, located in cytoplasm, involved in cytoskeleton-dependent cytokinesis, cellular developmental process and enables unknown molecular functions) at coefficient score of (τ = 0.7, p-value:$1.60\text{x}10^{-73}$) whereas mildly strong co-expressions were observed with AT4G25434 (NUDT10, NUDIX HYDROLASE HOMOLOG 10, located in chloroplast, cytoplasm, cytosol and nucleus, enables ADP-ribose diphosphatase activity, NAD binding and NADH pyrophosphatase activity) at

(τ = 0.6, p-value: 4.26x10^{-51}) and AT3G05790 (LON4, LON PROTEASE 4, located in cytoplasm and mitochondrion, involved in quality control of misfolded proteins and chaperone-mediated protein complex assembly, enables single-stranded DNA binding) with (τ = 0.5, p-value: 8.67x10^{-35}). Furthermore, AT5G05180 co-occurred at (τ = 0.6, p-value: 1.51x10^{-47}), (τ = 0.6, p-value: 2.74x10^{-61}) with LON4 and NUDT10 respectively, whereas both NUDT10 and LON4 were strongly interconnected with score (τ = 0.5, p-value: 6.67x10^{-42}). Additionally, for a thorough understanding of the downstream mechanism of function, we looked for the association between rest of the connections. It was observed that AT3G18880 (Nucleic acid-binding, OB-fold-like protein, involved in translation) co-occurred with AT1G06630 at (τ = 0.6, p-value: 5.16x10^{-46}); AT5G05180 at (τ = 0.6, p-value: 4.08x10^{-55}); NUDT10 at (τ = 0.6, p-value: 1.17x10^{-58}) and LON4 at (τ = 0.6, p-value: 3.49x10^{-62}).

The whole cluster was evenly knocked-out in the geo locations from Georgia. The details are listed in table 12 below. It was also noteworthy finding out that the pairs were commonly expressed in flower, root (Schmid et al., 2005) and guard cell (Obulareddy et al., 2013).

Table 12: Geographical locations of significantly co-expressed cluster from Georgia

Accession Ids	Name	Latitude	Longitude
9106	Lag1-8	41.8296	46.2831
9111	Lag2-4	41.8296	46.2831
9113	Lag2-6	41.8296	46.2831
9114	Lag2-7	41.8296	46.2831
9115	Lag2-10	41.8296	46.2831
9988	Bak-2	41.7942	43.4767

The GO enrichment of the genes showed the joint role of LON4 and NUDT10 in hydrolase activity, however, LON4 was also involved in DNA binding and catalytic activity while NUDT10 in nucleotide binding. Moreover, AT3G18880 was reportedly involved in structural molecule activity whereas AT1G06630 and AT5G05180 had role in unknown molecular functions.

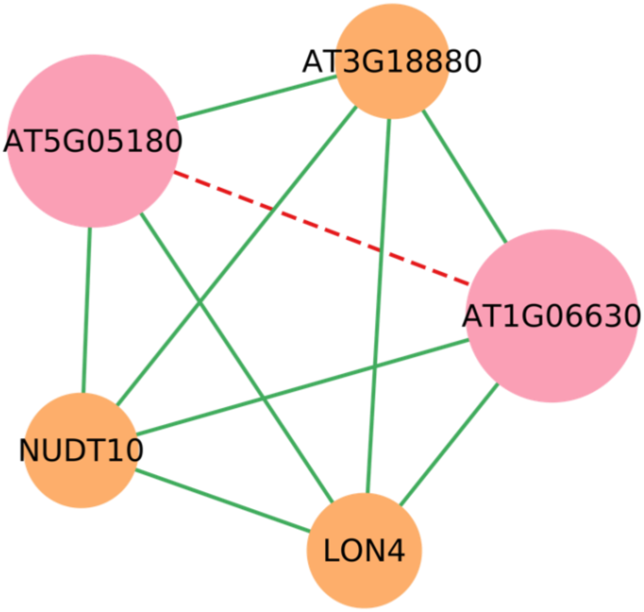


Figure 25: Georgian module with common expression pattern in flower. Red dashed edge indicates strongest connection, green edges represent moderately strong connection according to co-expression coefficient score. Pink nodes are strongly connected genes while orange nodes represent their connection with moderate strength.

3.9.3.7 Module 7

This module comprised six strongly connected edges and highly significant mutated gene pairs forming 4 nodes, was also detected from Eastern Europe with accession IDs from Armenia which shares the same geographical region as Georgia. We were interested to know which diverse climatic conditions in the region remained the cause of local adaptations in these accessions and which underlying biological processes and molecular functions were affected. Our results showed highly strong co-expression at coefficient score of (τ = 1.0, p-value: $3.33\text{x}10^{-146}$) between AT5G40830 (ICA, INCREASED CAMBIAL ACTIVITY, encodes an SAM-dependent methyltransferase superfamily protein, involved in phloem or xylem histogenesis (H. Kim et al., 2016)) and AT5G49710 (RING finger protein, acts in response to inorganic substance, located in nucleus and cytoplasm, enables unknown molecular function). Moreover, both AT5G40830 and AT5G49710 were observed to have strong connection with

AT2G47680 (zinc finger (CCCH type) helicase family protein, enables mRNA binding (Reichel et al., 2016; Bach-Pages et al., 2020), located in nucleus) with an identical score of (τ = 0.7, p-value: 1.47x10^{-65}) and moderately strong correlation with AT5G59540 (2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein, located in cytoplasm, acts in response to osmotic stress, negative regulation of signal transduction, defense response regulation, defense response to fungus, bacterium, inorganic substance and abscisic acid and enables unknown molecular function) was noticed at (τ = 0.6, p-value: 8.21x10^{-54}). Furthermore, AT5G59540 was also found to be co-expressed with AT2G47680 at (τ = 0.6, p-value: 2.49x10^{-53}).

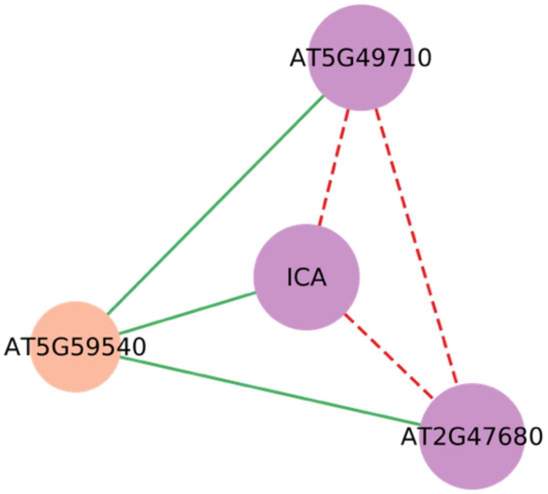


Figure 26: Armenian Module. Red dashed edges with purple nodes represent highly strong correlations whereas orange node with green edges shows moderately strong co-expression.

It was perceived from the above information that the gene cluster was apparently involved in plant's response to abiotic stress. We noticed that the knocked out genes were mutually expressed in: inflorescence meristem, petiole, hypocotyl, collective leaf structure, leaf lamina base, flower, leaf apex, vascular leaf, cauline leaf, sepal, shoot system, shoot apex, stamen, stem (Schmid et al., 2005) and guard cell (Obulareddy et al., 2013) and were jointly located in cytoplasm and nucleus in the cellular component of the cell. This expression pattern led us to the assumption that the adaptations were ensued in various parts of the plant. The shared accession Ids between the clustered gene pairs are listed below in table 13:

Table 13: Geographical location of accession Ids clustered from Armenia.

Accession Ids	Name	Latitude	Longitude
9128	Yeg-2	39.8692	45.3622
9130	Yeg-4	39.8692	45.3622
9133	Yeg-7	39.8692	45.3622
9134	Yeg-8	39.8692	45.3622

Further looking into Gene Ontology of the genes we found that AT2G47680 was enriched for multiple diverse processes including hydrolase activity, catalytic activity, and RNA binding, whereas AT5G49710 and AT5G59540 had unknown molecular functions.

3.9.3.8 Module 8

Lastly, this extremely tiny but highly strong cluster of three interconnected genes was identified in the highly significant data set with co-expression in accession IDs from Sweden.

Interestingly all three genes were found to be fully annotated with known functions. The co-expression statistics showed strong correlation of AT1G01070 (UMAMIT28, USUALLY MULTIPLE ACIDS MOVE IN AND OUT TRANSPORTERS 28, Nodulin MtN21 /EamA-like transporter family protein, located in mitochondrion and plasma membrane, involvement in amino acid export to the developing seed, involved in seed development (B. Müller et al., 2015), enables L-glutamine transmembrane transporter activity) at tau score (τ = 0.7, p-value: $4.42\text{x}10^{-64}$) with AT1G14100 (FUT8, FUCOSYLTRANSFERASE 8, member of Glycosyltransferase Family- 37, located in Golgi apparatus, involved in xyloglucan biosynthetic process and protein glycosylation, enables galactoside 2-alpha-L-fucosyltransferase activity and fucosyltransferase activity) and at a score of (τ = 0.7, p-value: $1.24\text{x}10^{-73}$) with AT1G78670 (GGH3, GAMMA-GLUTAMYL HYDROLASE 3, located in extracellular region, plant-type vacuole, involved in tetrahydrofolylpolyglutamate metabolic process, has cell to cell mobile RNA as gene product (Thieme et al., 2015)). Moreover, FUT8 and GGH3 were also interconnected with a moderately strong coefficient score (τ = 0.5, p-value: 1.79x10^{-42}).

Table 14: Geographical locations of the genes co-expressed in Sweden

Accession Ids	Name	Latitude	Longitude
6024	Fly2-2	55.7509	13.3712
6140	T880	55.9392	13.5539
6145	T930	55.9497	13.5533
9405	HolA-1 2	55.7491	13.399

The genes were found to be mutually expressed in collective leaf structure, flower, and sepal (Schmid et al., 2005). Moreover, the molecular functions of the knocked-out genes show their involvement in transport system, flowering time and related to the amino acid regulatory pathway. Their GO slim terms revealed involvement of UMAMIT28 in transporter activity, FUT8 was enriched with transferase activity while GGH3 was found to be involved in catalytic activity and hydrolase activity.

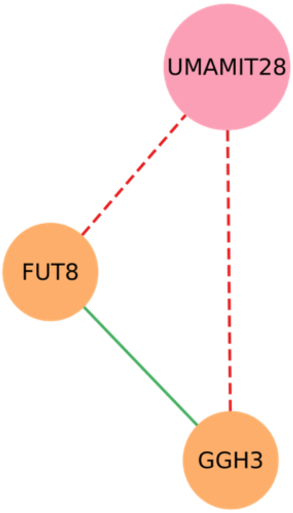


Figure 27: Swedish module. Red dashed edges represent strong co-expression according to the coefficient score. Green edge shows moderately strong co-occurrence. Pink large node distinct the orange node by connecting with only strong coefficient scores while orange node represents genes with both strong and moderate connections.

3.10 How the geographical pattern of over-represented significant co-occurred pairs looks on the European map?

The spatial disposition of highly significant allele pairs was examined. The location of all the accessions in the form of latitudinal and longitudinal geographical coordinates was available at https://arapheno.1001genomes.org. We obtained the clustering data from co-occurrence network table of 1,674 gene pairs out of 1,696 (22 pairs had significantly less co-occurrence with 0 connections so they were removed) with altogether 664 accession IDs from the worldwide accessions of *Arabidopsis thaliana*. After filtration to only European accessions (we chose this for a uniform sample collection), 533 IDs were left. The minimum cut-off value of MAC was set to 2 and initial data set was further filtered by keeping only the pairs with MAC >= 2, therefore we were left with 1114 pairs. 759 pairs were found to be significantly clustered at their geospatial locations at T1 (peacock p-value < 0.05) whereas the Bonferroni correction of p-value: $4.49\text{x}10^{-5}$ showed 79 highly significant clustered pairs.

We selected two gene pairs one with high and the other with low peacock p-value to demonstrate their hypothetical clustering on the European map. High value of MAC was chosen for a better visual segregation of the knock out and wild type accessions. We observed that in the high peacock p-value clusters the accession IDs were arbitrarily scattered in several geographical locations without portraying any distinct pattern of expression, contrarily gene pairs with significantly low peacock p-value showed distinguished co-expression regions from rest of the knock out and wild type accessions. We mapped selected gene pairs to show the clustering on European map considering their low and high peacock p-value and a high MAC count. Figures 28 and 29 show the mapping of gene pairs.

On the map with high peacock p-value, random distribution of co-expressed knock out accessions was observed, comparatively, mapping of low peacock p-value cluster on their spatial locations resulted in clear distinction of co-occurred knock out accession IDs from wild type. Additionally, the gravitational center of both wild type and knock out accession Ids was lying in the close vicinity in high p-value gene pairs which evidenced that there were no significant differences in both sets, on the contrary, it was observed to be quite distant from each other in low peacock p-value gene pair as shown in figures 28 and 29 (marked in white circles). Moreover, a high p-value was observed for gene-pairs with low MAC in general.

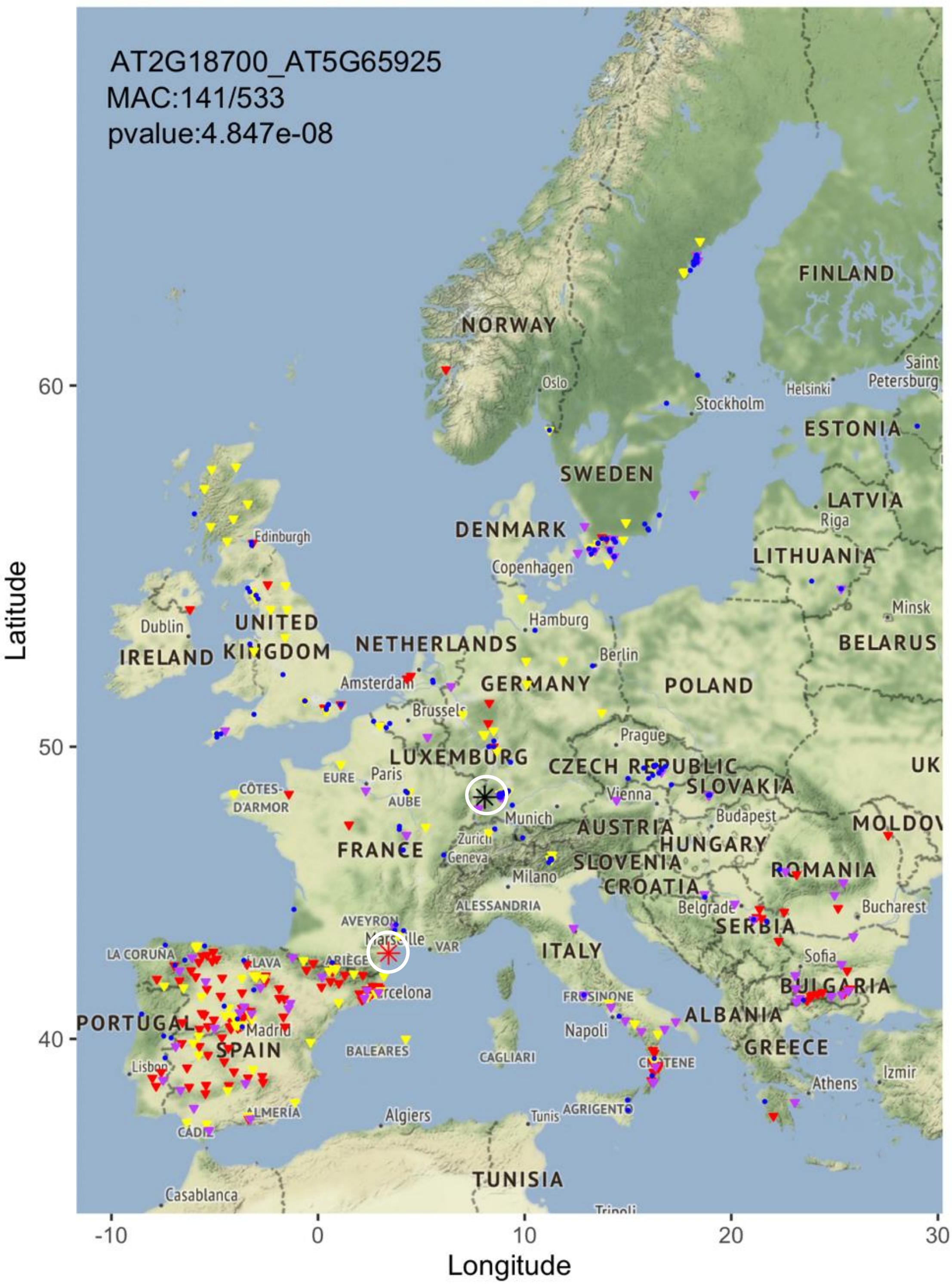


Figure 28: Mapped spatial distribution of co-expressed gene pair with low p-value. Red inverted triangles point out the knock-out accessions common in both genes (1). Blue dots represent wild-type accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of knock out and wildtype accessions respectively.

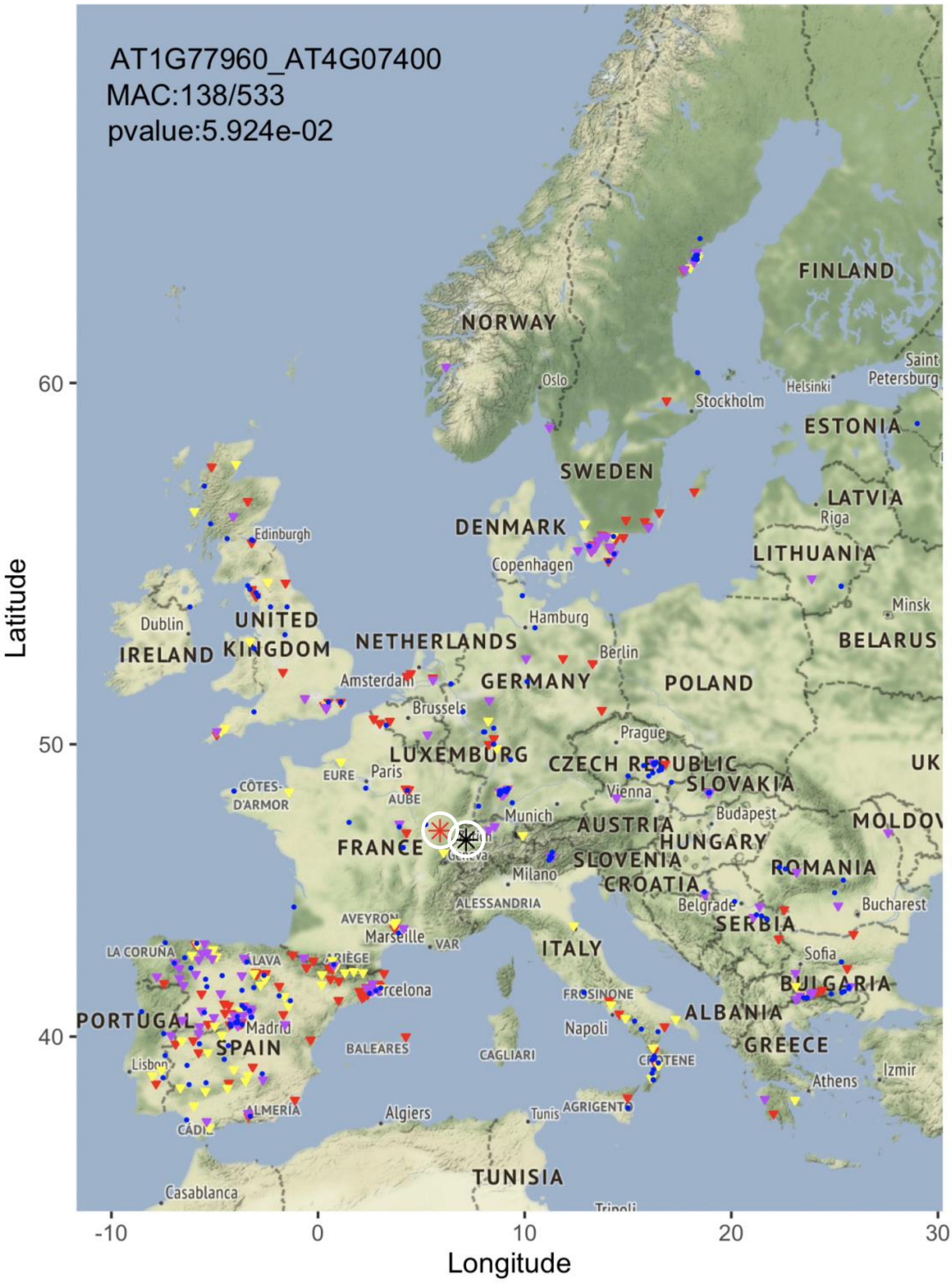


Figure 29: Mapped spatial distribution of gene pair with high p-value. Red inverted triangles point out the knock-out accessions common in both genes (1). Blue dots represent wild-type accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of knock out and wildtype accessions respectively.

We also mapped the highly significant modules where strong co-expression was identified in diverse regions in Europe as explained in section 3.9. Cluster formation in the particular location was evident in all regions and the distance between central geographic location of knocked out and wild type accessions also confirmed module formation. It was also noticed that the higher the mean distance between the knocked out and wild type accessions goes, more the co-occurrence is clustered. Furthermore, z-score of the D-statistic was also high in strongly connected allelic pairs suggesting of high deviation from standardized mean. The clusters of the highly significant knocked out accessions can be seen in the figures 30 -33 in several geo spatial locations across Europe.

By looking at figure 30, cluster formation of shared accession IDs in AT3G21950 (gene-1) and AT5G02630 (gene-2) indicated through red markers was distinctively visible. The were some other knock outs expressed in the same region in gene-1 and in Eastern Europe in gene-2 represented by yellow and orchid markers respectively. On the other hand, wild type accessions showed a uniform distribution.

Furthermore, shared, and non-shared accession IDs of AT3G09560 (gene-1) and AT5G25560 (gene-2) from Spanish module II in section 3.9 were mapped on their geographical locations. Similar pattern of co-expression in the south of Spain was visible with formation of clearly discrete clusters of knock out vs wildtype. Moreover, significantly higher mean distance of the center of gravity of both data sets can be seen in figure 31 (marked in white circle).

In figure 32 and 33, gene pairs AT1G52920 (gene-1) and AT5G22690 (gene-2) from Italian module and AT1G01070 (gene-1) and AT1G14100 (gene-2) from Swedish module (as explained in section 3.9) were mapped. Due to low Mac in both gene pairs the knock out cluster seemed very small on the map. Since there were no markers found for non-shared knock out accessions of both gene-1 and gene-2, it was also assumed that both genes were explicitly co-expressed in that geographical location which somehow connected to their molecular function in that certain environmental gradient i.e., Italy being dry region with mild drought conditions and Sweden with overall cold conditions affecting flowering time and seed development etc.

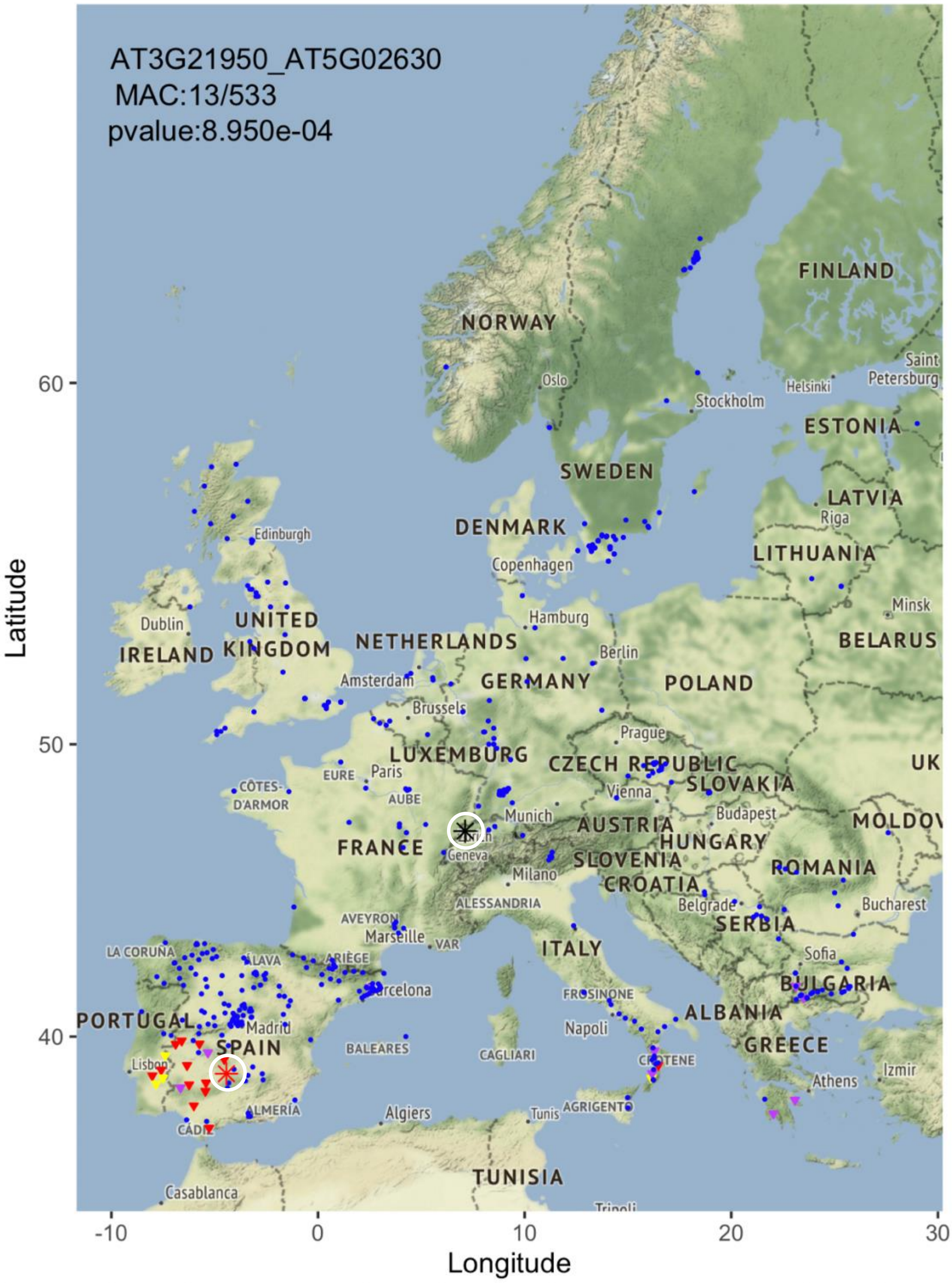


Figure 30: Spatial distribution of significantly co-expressed gene pair from Spanish Module-I on European map. Red inverted triangles point out the knock-out accessions common in both genes (1). Blue dots represent wild type accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of mutated and wild type accessions respectively.

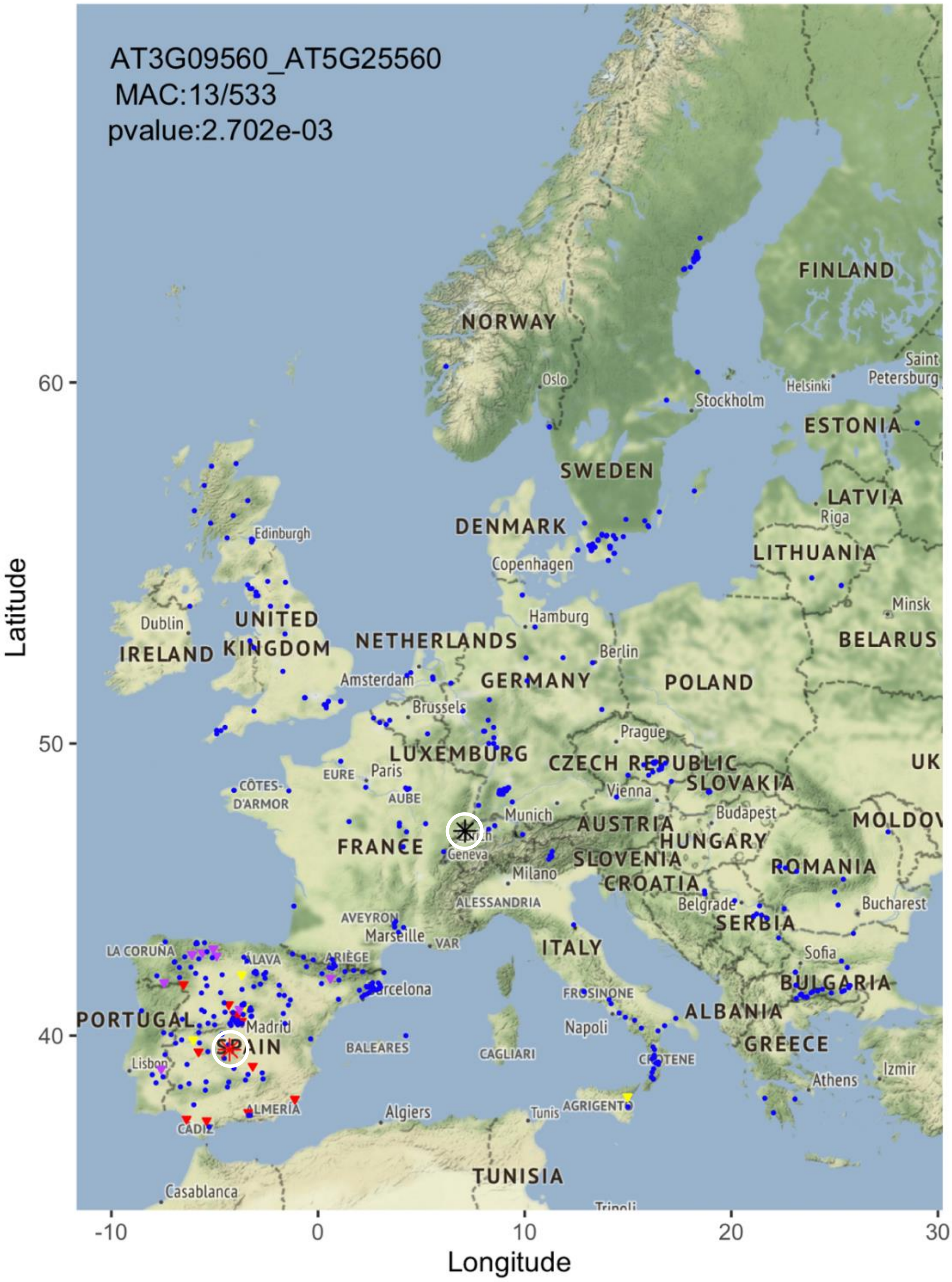


Figure 31: Spatial distribution of significantly co-expressed gene pair from Spanish Module-II on European map. Red inverted triangles point out the knock-out accessions common in both genes (1). Blue dots represent wild-type accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of knock out and wild type accessions respectively.

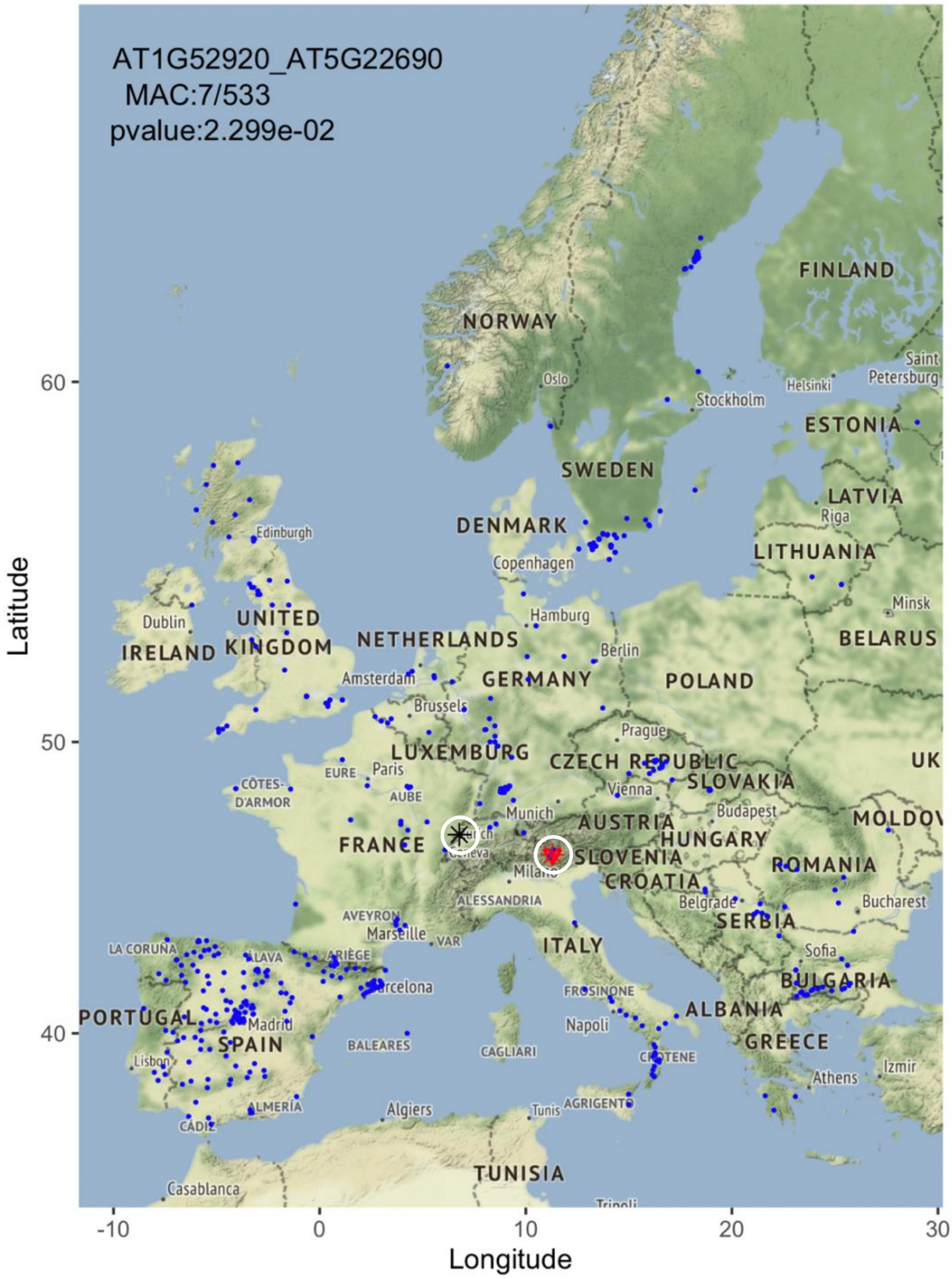


Figure 32: Spatial distribution of significantly co-expressed gene pair from Italian Module on European map. Red inverted triangles point out the knock out accessions common in both genes (1). Blue dots represent wildtype accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of knock out and wildtype accessions respectively.

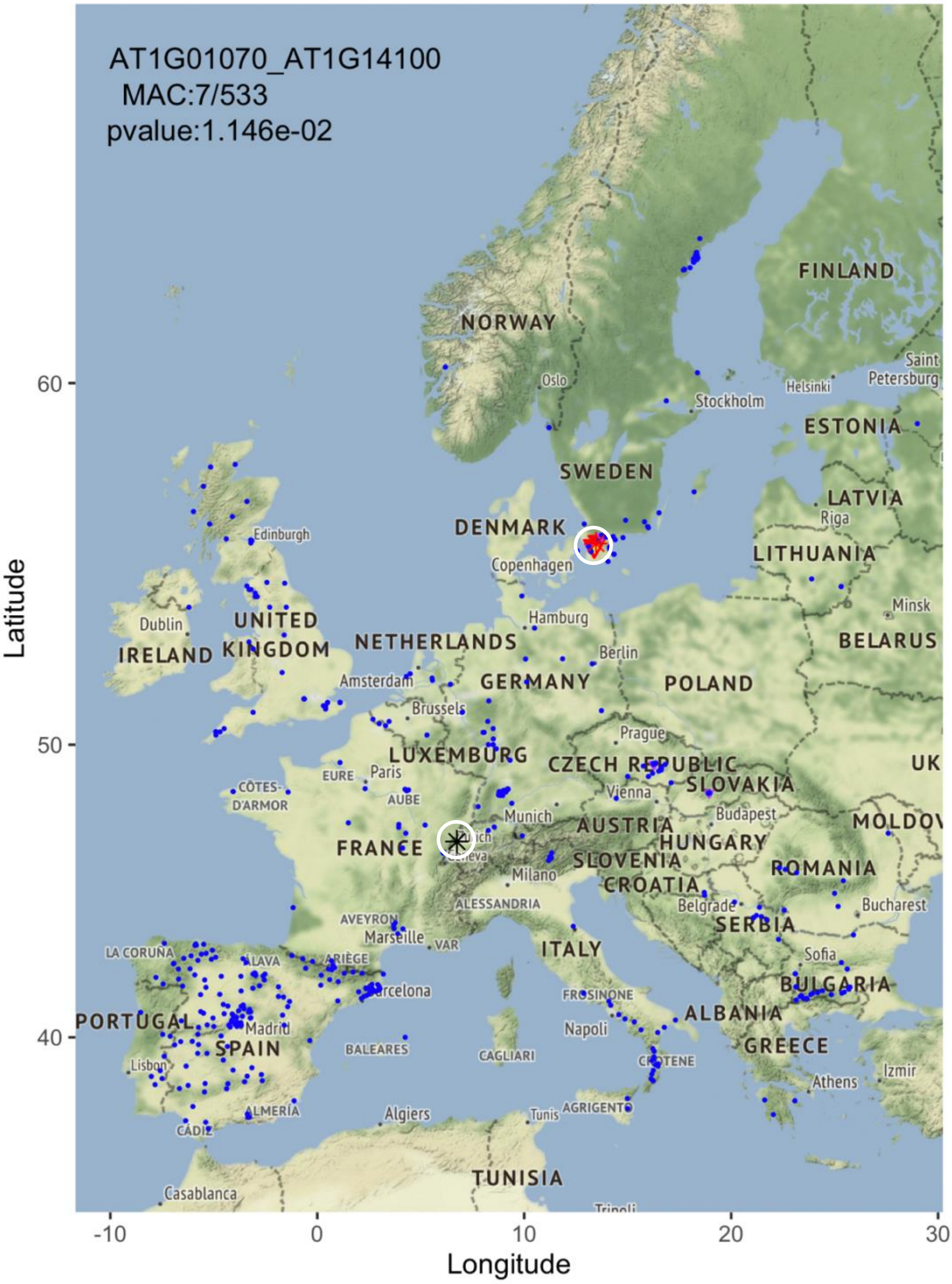


Figure 33: Spatial distribution of significantly co-expressed gene pair from Swedish Module on European map. Red inverted triangles point out the knock-out accessions common in both genes (1). Blue dots represent wild-type accessions (0). Yellow color shows gene-1 knock outs while orchid color indicates knock outs of gene-2. Red and black stars represent earth central gravity of knock out and wildtype accessions respectively.

CHAPTER 4

4 Discussion

Variations in genome sequence of *Arabidopsis thaliana* have been widely studied (Kerwin et al., 2015; Pignatta et al., 2014; Schmitz & Ecker, 2012; Vaughn et al., 2007; Weigel, 2012), however the study of natural adaptations due to advent of non-sense mutations (pre-mature stop codons, PTCs) in *Arabidopsis thaliana* worldwide accessions is a new subject. It is believed that PTCs directly disrupt either transcription process or result in incomplete truncated proteins (Byers, 2002; Gregersen et al., 2000) which might act negatively in cellular functions (Schell et al., 2002). However, it is also evident that PTCs may lead to either non-functional proteins or provide added stability and altering functions adaptive for the environmental selection (Savas et al., 2006).

The key purpose of the study was to determine the natural variations associated with differential gene expression due to PSCs gain and to identify and explain their correlation in a gene co-expression network (GCN) in a worldwide collection of *Arabidopsis thaliana* accessions. Another major objective was to correlate the adapted genotypic traits with their spatial location of occurrence in terms of environmental gradients. Exploring the overall frequency of occurrence of PSCs and their density based on relative position on the genome in all 1135 accessions, we found a uniform distribution. While no differences were observed, therefore it was presumed that the location of stop gains (in start or stop of the genome) does not relate to their possible change of function. Overall, we observed highest number of mutations in accessions from Spain which co-relates to the dry climatic conditions and water deprivation where one crucial cause could be global warming. While every SNP was typically linked to one gene, there were many genes knocked out on multiple locations, their association with more than one SNP indicated that several loci could impact the gene expression either collectively or distinctly. Nearly 20% significantly down regulated genes selected using hypergeometric distribution method were assumed to be potentially controlling the regulation of gene expression. This method was chosen over binomial distribution because of finite, without replacement, dependent trials needed to be applied on the expression data of knock out and wild type accessions. More genes were down regulated as compared to up regulated at significant thresholds, which, we also expected (due to the premature stop codon gain which typically led to loss of function) e.g., in case of cellular response to increased stimuli and abiotic conditions. Furthermore, our co-expression matrix with nearly 200,000 connections presented only a tiny fraction ~1% of gene pairs with highly significant (Bonferroni correction of p-value) over and under-represented connections. This shedding of insignificant genes removed plenty of noise data that simplified and expediated the next steps. Based on the previous studies, it was found out that (King et al., 2003; Mostafavi & Morris, 2010) co-occurrence analysis helps determine common unpredicted functional relations between gene pairs. Addressing our key objectives, this down regulated RNA expression data set was therefore used to build a robust co-expression network. Kendall tau correlation coefficient method was used to quantify the association between co-expressed gene pairs. This method was selected over Pearson and Spearman's rho because it is more intuitive and is considered as more robust, efficient and quantifies a monotonic relationship. Moreover, it provides a direct interpretation in terms of probability for concordant and discordant pairs. Kendall tau is also a preferred method when there are outliers in the data which are usually prevalent in the free-scale co-expression network. On the other hand, Pearson correlation required a continuous data in linear format. Moreover, Spearman's rho method is believed to be almost alike Kendall tau except Spearman being more sensitive to errors and instigating discrepancies with small sample sizes. From the results, we observed a standard pattern in which tau coefficient score gets stronger when p-value gets closer to zero, thus, their inverse relation makes the correlation highly strong and significant.

Talking about the co-expression scores, while comparing to the networks in system biology studies where the samples are collected from individual tissues of identical genotype e.g., (Klepikova et al., 2016), our gene-gene co-expression coefficients showed low numbers. Anyway, the co-occurrence still exhibited characteristic features of scale-free network including hubs, edges, and distinctive modules. In general, significant number of genes in the network showed less connectivity meaning majority of the arbitrary mutations also did not reflect in co-relatedness of the module. However, agreeing with some previous studies e.g., (Mähler et al., 2017) few of our inter-modular hub genes were acting as the central regulators in the cluster and the expression of these genes shared related adaptions/variations in the other co-expressed genes. Thus, it can be stated that each of our modules were sharing almost similar and interconnected expression pattern within themselves and even if all the co-expressed genes shared same evolutionary history, the many-to-one feature of our scale-free network had still been accurate; meaning nodes were likely be congregating mutational induced variations in the network unless there is intervention of natural selection in its topology e.g., prevention of build-up. We also deduced from the network analysis that the distribution of mutations in the clustered genes is not random rather they are responsible for the adaptations in various cellular mechanisms e.g., in response to environmental stimuli or plant pathogen immune response. Another aspect of GCN analysis is the identification of putative loss of function or diseased genes through their association with multiple other faulty genes (Gillis & Pavlidis, 2012). However, our network is more focused towards a positive gene regulation i.e., adaptations in the genome for a response to foreign stimuli. Since GCNs encode functional information and it was also previously reported that 'gene co-expression networks have been lately another source of inferring function of unknown genes in broad spectrum based on their connectivity with other hub genes' (Depuydt & Vandepoele, 2021). The similar concept was applied for the genes with unknown molecular function in our network modules. Based on their co-expression with annotated genes of known function, we assumed their function to be in the correlated cellular pathways, however, the prediction is considered highly localized with limited abilities, but it still hints towards a direction. This method is also helpful in a gene regulation network when unannotated or under-annotated genes are encountered.

It is reported that the quantification of co-expression usually describes the strength of relationship among the co-occurring genes, however, the mutual information holds utmost importance in pointing towards extraction of biologically meaningful gene clusters (L. Song et al., 2012). We observed the similar pattern in our modules where targeted regulatory pathways were being controlled by few functional hub genes e.g., AGB1 cluster (GTP BINDING PROTEIN BETA 1) part of Cul4-RING E3 ubiquitin ligase complex and involved in G-protein coupled receptor signaling pathway. According to Various studies based on AGB1 its involvement in signaling pathways of Jasmonic acid is known (Trusov et al., 2006), Abscisic acid and Brassinosteroid (Tsugama et al., 2013). Similarly, AGB1 is also involved in positive regulation of cell elongation by affecting phosphorylation and transcriptional activities (T. Zhang et al., 2018). As the overall cluster depicted involvement in distinct processes like morphogenesis, cellular transport, and defense response to various pathogens, we could relate the co-expression of AGB1 with multiple genes from diverse protein families of known domains where the co-relatedness was certain in the underlying mechanisms e.g., DYW4 (DYW DOMAIN PROTEIN 4) from Tetratricopeptide repeat (TPR)-like superfamily, proteins belonging to TTR family are reportedly essential elements in signal transduction. Moreover, their involvement in response to abscisic acid and tolerance for osmotic stress is also proved by (Schapire et al., 2006); CRK41 (CYSTEINE-RICH RLK(RECEPTOR-LIKE PROTEIN KINASE) from CRK gene family play several vital roles including stress adaptation and plant development. Transcriptional induction of CRKs are reportedly involved in abiotic stress response conditions for example, excess salt, drought, salicylic acid, and UV light (Bourdais et al., 2015; K. Chen et al., 2003, 2004; Yeh et al., 2015). Furthermore, 41VHP2;2 (Vacuolar H+-PyroPhosphatase) is an inorganic pyrophosphate donor and energy source according to (Segami et al., 2018). Numerous studies have reported improvement in overall plant growth, salt and drought stress tolerance due to overexpression of H^+-PPase (Arif et al., 2013; Pizzio et al., 2015; Vercruyssen et al., 2011) (Gaxiola et al., 2001). Many genes indicated co-relation in growth and regulation of gene expression e.g., AT4G09490 (Polynucleotidyl transferase from Ribonuclease H-like (RNHL) superfamily. Proteins in the family play roles in replication, DNA repair, and in general nucleic acid metabolism. GDPD6 a glycerophosphoryl diester phosphodiesterase family protein with roles in glycerol and lipid metabolic pathway, under phosphate limiting conditions, mutants show defects in root growth; RPF1 (RNA PROCESSING FACTOR 1) from pentatricopeptide repeat (PPR) protein family mediates in RNA splicing, editing, and RNA translation. AT3G32920 encodes proteins from P-loop containing nucleoside triphosphate hydrolases superfamily and involved in DNA repair. Moreover, since AGB1 has a major role in plant defense to pathogens and specifically fungi, we found co-expression with RLP52 (RECEPTOR LIKE PROTEIN 52) a chitin responsive gene with specific role in disease resistance (G. Wang et al., 2008) from fungal pathogens e.g., powdery mildew (Ramonell et al., 2005). Other defense regulating genes include RPP1-like Probable disease resistance protein RPP1; DM2H, DANGEROUS MIX2H increase resistance together with ENHANCED DISEASE SUSCEPTIBILITY1 (EDS1) in the form of complex (Stuttmann et al., 2016) and Cysteine/Histidine-rich C1 domain family protein encoded by AT3G26240. It is also reported to be involved in growth and development of plant and reacts in response to different stresses (Hwang et al., 2014). Additionally, we presumed connected roles of AGB1 and AT2G04280 in Ca2+ signaling pathway as reported by (Tanaka et al., 2010). In general, Ca2+ act like phosphate ions and are crucial factors in signal transduction (Grzybowska, 2018). On the other hand, they also act as cell membrane transporters (Yáñez et al., 2012).

Overall, the whole set of genes co-expressed in a cluster illustrate a chain of different interlinked biological processes where one process stimulates the other which also means that the whole set of co-expressed genes in a single module are functionally connected for a given trait affected by an external stimuli e.g., defense response to microorganisms i.e., in terms of simplified reverse-genetics technique, bacterium or oomycetes stimulate immune response of the plant which further activates signal transduction, cellular transport and eventually leads to differential regulation of gene expression. Our findings relate with the "Omnigenic" model proposed by (Boyle et al., 2017) which proposes that a few biologically relevant 'core' genes are partially responsible for any given trait along with their regulators and associated pathways. Moreover, it is also said that 'Peripheral' genes despite not being part of the key pathways surpass core genes in contributing to the heritability of a certain trait.

Due to rapid continuous environmental fluctuations local adaptations are reportedly very essential in the survival of a species (Fournier-Level et al., 2011a). Because clusters in our network are accession dependent, therefore, most of the co-occurring genes share similar biological processes in specific geographical zones which also relate to the changing environmental gradients of the region e.g., climatic fluctuations like dry, hot, or extreme cold weather, light intensity, water availability, mineral concentration that led to local adaptation or modification in the genome.

In northern hemisphere, a shorter day length with less sunlight overlaps with increased precipitation and cold temperatures. This excessive water and cold stress with less sunlight effects plant growth severely. The expression of the co-occurring genes in flower, leaf, guard cells and seed points towards the mutations led changes in flowering time and seed germination. This observation agrees with the infinitesimal model presented by (Fisher, 1918) which states that mutations in one part of the genome potentially affect other phenotypes indirectly. In a former study it is stated that day length and temperature are the essential factors incorporated in plants in order to optimize their flowering time (Lutz et al., 2015). These seasonal changes of both temperature and sunlight all-around the year are sensed by the plants to adjust their flowering time accordingly. It is also reported that the differential incorporation of day length and temperature information in plants help them adjust their flowering time between different ecotypes in *Arabidopsis thaliana* and even other species. It also needs to be noted that leaves measure the day length that also aligns with our findings in module A from US and few accessions from Caucasus region. Moreover, we have identified co-occurrence of disease resistance genes along with abiotic stress responsive genes in the cold climatic regions e.g., Atlantic North America, Sweden in Scandinavia and accessions from Caucasus regions including southern Russia, Georgia, and Armenia where temperatures go as low as -10°C on average in winters. It was found out from a previous study that pathogenic related genes show an enhanced disease resistance when exposed to low temperatures as they perceive cold signals to pathogens, however their mechanism and signaling pathway is mostly unknown (Seo et al., 2010).

Involvement of some key genes like AP1G1, NRPD4, DM2H, DRP4C, ABCA1, CRK41, AIG1, GSL-OH, WRKY16 TRM33, NUDT10, LON4, ICA, UMAMIT28 and FUT8 in diverse biological processes e.g., regulation of gene expression, stress response to light, cellular detoxification and transport, cellular defense response, auto immune response in seed development, Cell growth and metabolism, activation of signal transduction, and transcription regulation explain the underlying mechanisms. Protein families like ATP-dependent caseinolytic (Clp) protease/crotonase family, Sec14p-like phosphatidylinositol transfer family, Protein kinase superfamily, Pentatricopeptide repeat (PPR-like) superfamily, Phosphoglycerate mutase family were dominantly involved in the key processes where few were acting as regulators.

Going further, due to increasing effects of global warming, temperatures are rising (Y. Chen et al., 2021). This effect can be observed in the accessions knocked out in southern Europe where our knock-out alleles have developed phytohormone signaling pathways based on Abscisic acid and Jasmonic acid in response to abiotic stresses e.g., heat (high temperature) and water deprivation (drought or dehydration) e.g., in GCR2 and NOG1-2. The climate gradient in southern part of Europe i.e., in Spain, Portugal and Italy is mostly dry and hot however, temperatures rise up-to 37°c in summer which induce mild drought conditions in the plant when water availability is limited. As plants adapt with the changing temperatures and environmental conditions, it is postulated that due to increasing dry conditions and water scarcity, the variations are altering the regulation of guard cells and closure of stomata which in turn prevents water loss also stated by (Tuteja, 2007) (Kostaki et al., 2020) e.g., downregulated overexpressed NOG1-2 revealed its role in stomatal opening and stimulation of defense responses in Spanish module. The mechanism demonstrates its role in prevention of water loss by forming silica body around stomata (Pant et al., 2022). It also reportedly functions in guard cell signaling by regulating Jasmonic acid and Abscisic acid pathways in response to abiotic and biotic signals (S. Lee et al., 2017). ERD 8 (HSP-81.2) reportedly showed its roles in heat tolerance (Lim et al., 2006), stomatal closure (Clément et al., 2011) and water deprivation (H. Song et al., 2009). PAH1 reported to be strong stimulus in the stress signaling pathway (Kuhlmann et al., 2020). Furthermore, role of AT12CYS-1 in mechano-transduction leading to increased drought tolerance is another example of stress induced variations. http://bar.utoronto.ca. From this we presume that co-expression of downregulated over-represented genes in the guard cell in our Spanish module is more than a co-incidence.

Key protein families e.g., Cytosolic chaperones of the HSP90 family, CHCH domain, Pentatricopeptide repeat (PPR) superfamily, lipin family, F-box family, SABATH family, ABC transporter White- Brown Complex (WBC) family, lung seven transmembrane receptor family and Disease resistance protein (TIR-NBS-LRR class) family involved in regulation of drought tolerance, stomatal closure, lipid synthesis, response to abiotic stimuli, heat resistance, water deprivation, salt excess, temperature stimulus, and cellular transport played roles in local adaptation in the genome sequence at cellular level and in some cases also affected the phenotype.

Beside investigating molecular functions and co-expression patterns it was also very interesting to map the clustered genes on their spatial locations. Earlier discoveries revealed that genetically alike Arabidopsis accessions originated from more closely linked geographical locations which suggested the strong pattern of geographical clustering in different regions (Anastasio et al., 2011; Horton et al., 2012). Our distance matrices from knocked out geographical locations and their genotype retain the relationship information between all pairs and also represent the densely clustered structure on the map. The clustering method is based on peacock D-statistic whose accuracy is assessed by specific co-occurrences of genes among the groups based on the probability of a genetic relationship comprising four populations. Like every other statistical analysis our null hypothesis stated, “there is no well-defined clustering rather uniform distribution of genes on the map”. As very low p-values associated with extremely high/low z-scores are lying in the tail of normal distribution. Our clustering analysis yielded low p-values with mostly high and few low z-scores which indicated that it is highly unlikely that our observed spatial allele clusters portray the random theoretical pattern as stated in our null hypothesis.

5 Conclusion

We constructed free scale co-expression network of genes encoding for premature stop codons through accession-based co-occurrence matrix. This network reveals several clusters enlightening functional connections between co-occurred genes that are considered to be responsible for local adaptations in various geographical locations having environmental constraints. Moreover, we have inferred that the genes encoding premature stop codons have a key role in regulation of guard cells and closure of stomata either in prevention of water loss in dry conditions or altering flowering time in cold regions. Furthermore, we have proposed a set of genes including PAH1, AT12CYS-1 HSP-81.2, GCR2, NOG1-2, AP1G1, NRPD4, DM2H, DRP4C, ABCA1, CRK41, AIG1, GSL-OH, WRKY16 TRM33, NUDT10, LON4, ICA, UMAMIT28 and FUT8 depicting their key involvement in local adaptations in several cellular processes. In previous studies, these genes have reportedly shown involvement in regulatory pathways related to drought tolerance and flowering time alterations however, to our knowledge there has been no previous data that they are somehow functionally linked, but our data provides candidates for hypothesis testing. Finally, the approach of building co-expression network based on accession data is general and the developed pipeline can be applied to other organisms for example rice.

References

1001 Genomes. (n.d.). 1001 Genomes. https://www.1001genomes.org

Agliassa, C., Narayana, R., Bertea, C. M., Rodgers, C. T., & Maffei, M. E. (2018). Reduction of the geomagnetic field delays Arabidopsis thaliana flowering time through downregulation of flowering-related genes. *Bioelectromagnetics*, *39*(5), 361–374. https://doi.org/10.1002/bem.22123

Ågren, J., & Schemske, D. W. (2012). Reciprocal transplants demonstrate strong adaptive differentiation of the model organism Arabidopsis thaliana in its native range. *New Phytologist*, *194*(4), 1112–1122.

Akutsu, T., Miyano, S., & Kuhara, S. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computational Biology*, *7*(3–4), 331–343.

Alberts B, Johnson A, & Lewis J. (2002). *Molecular Biology of the Cell* (4th Edition). Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26818/

Alonso-Blanco, C., Aarts, M. G., Bentsink, L., Keurentjes, J. J., Reymond, M., Vreugdenhil, D., & Koornneef, M. (2009). What has natural variation taught us about plant development, physiology, and adaptation? *The Plant Cell*, *21*(7), 1877–1896.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., & Ding, W. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, *166*(2), 481–491.

Alonso-Blanco, C., Blankestijn-de Vries, H., Hanhart, C. J., & Koornneef, M. (1999). Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *96*(8), 4710–4717.

Alonso-Blanco, C., & Koornneef, M. (2000). Naturally occurring variation in Arabidopsis: An underexploited resource for plant genetics. *Trends in Plant Science*, *5*(1), 22–29. https://doi.org/10.1016/S1360-1385(99)01510-1

Anastasio, A. E., Platt, A., Horton, M., Grotewold, E., Scholl, R., Borevitz, J. O., Nordborg, M., & Bergelson, J. (2011). Source verification of mis-identified Arabidopsis thaliana accessions. *The Plant Journal*, *67*(3), 554–566.

Andrade, C. (2019). The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. *Indian Journal of Psychological Medicine*, *41*(3), 210–215. PubMed. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., & Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *Journal of Computational Biology : A Journal of*

Computational Molecular Cell Biology, *23*(4), 239–247. PubMed. https://doi.org/10.1089/cmb.2015.0205

Arif, A., Zafar, Y., Arif, M., & Blumwald, E. (2013). Improved growth, drought tolerance, and ultrastructural evidence of increased turgidity in tobacco plants overexpressing Arabidopsis vacuolar pyrophosphatase (AVP1). *Molecular Biotechnology*, *54*(2), 379–392.

Aryal, U. K., Xiong, Y., McBride, Z., Kihara, D., Xie, J., Hall, M. C., & Szymanski, D. B. (2014). A Proteomic Strategy for Global Analysis of Plant Protein Complexes. *The Plant Cell*, *26*(10), 3867–3882. https://doi.org/10.1105/tpc.114.127563

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. PubMed. https://doi.org/10.1038/75556

Atias, O., Chor, B., & Chamovitz, D. A. (2009). Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. *BMC Systems Biology*, *3*(1), 1–22.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., & Hu, T. T. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627–631.

Bach-Pages, M., Homma, F., Kourelis, J., Kaschani, F., Mohammed, S., Kaiser, M., van der Hoorn, R. A. L., Castello, A., & Preston, G. M. (2020). Discovering the RNA-Binding Proteome of Plant Leaves with an Improved RNA Interactome Capture Method. *Biomolecules*, *10*(4). https://doi.org/10.3390/biom10040661

Banta, J. A., Dole, J., Cruzan, M. B., & Pigliucci, M. (2007). Evidence of local adaptation to coarse-grained environmental variation in Arabidopsis thaliana. *Evolution: International Journal of Organic Evolution*, *61*(10), 2419–2432.

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113.

Barboza, L., Effgen, S., Alonso-Blanco, C., Kooke, R., Keurentjes, J. J., Koornneef, M., & Alcázar, R. (2013). Arabidopsis semidwarfs evolved from independent mutations in GA20ox1, ortholog to green revolution dwarf alleles in rice and barley. *Proceedings of the National Academy of Sciences*, *110*(39), 15818–15823.

Barbujani, G. (2000). Geographic patterns: How to identify them and why. *Human Biology*, 133–153.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *ArXiv Preprint ArXiv:1402.1894*.

Beck, J., Schmuths, H., & Schaal, B. (2008). Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics. *Blackwell Publishing Ltd*, *17*(3), 902–915.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The New S Language. Wadsworth & Brooks. *Cole.[Google Scholar]*.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., & Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, *12*(1). https://doi.org/10.1186/gb-2011-12-1-r10

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how? *Journal of Clinical Epidemiology*, *54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Bengtsson, H. (2014). *MatrixStats: Methods that apply to rows and columns of a matrix. R package* (0.10.1) [R].

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. (2003). Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Annals of Botany*, *91*(5), 547–557. https://doi.org/10.1093/aob/mcg057

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., & Rhee, S. Y. (2004). Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. *Plant Physiology*, *135*(2), 745–755. https://doi.org/10.1104/pp.104.040071

Bergelson, J., & Roux, F. (2010). Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics*, *11*(12), 867–879.

Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M., & Caudy, A. A. (2009). Measuring differential gene expression by short read sequencing: Quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, *10*(1), 1–10.

Botella, M. A., Parker, J. E., Frost, L. N., Bittner-Eddy, P. D., Beynon, J. L., Daniels, M. J., Holub, E. B., & Jones, J. D. (1998). Three genes of the Arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants. *The Plant Cell*, *10*(11), 1847–1860.

Bouchabke, O., Chang, F., Simon, M., Voisin, R., Pelletier, G., & Durand-Tardif, M. (2008). Natural variation in Arabidopsis thaliana as a tool for highlighting differential drought responses. *PloS One*, *3*(2), e1705.

Bourdais, G., Burdiak, P., Gauthier, A., Nitsch, L., Salojärvi, J., Rayapuram, C., Idänheimo, N., Hunter, K., Kimura, S., & Merilo, E. (2015). Large-scale phenomics identifies primary and fine-tuning roles for CRKs in responses related to oxidative stress. *PLoS Genetics*, *11*(7), e1005373.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Brennan, A. C., Méndez-Vigo, B., Haddioui, A., Martínez-Zapater, J. M., Picó, F. X., & Alonso-Blanco, C. (2014). The genetic structure of Arabidopsis thaliana in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. *BMC Plant Biology*, *14*(1), 1–14.

Bresinsky, A., & Strasburger, E. (2013). *Lehrbuch der Botanik*.

Buisine, N., Quesneville, H., & Colot, V. (2008). Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, *91*(5), 467–475.

Butte, A. J., & Kohane, I. S. (1999). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000* (pp. 418–429). World Scientific.

Byers, P. H. (2002). Killing the messenger: New insights into nonsense-mediated mRNA decay. *The Journal of Clinical Investigation*, *109*(1), 3–6.

Callahan, H. S., & Pigliucci, M. (2002). Shade-induced plasticity and its ecological significance in wild populations of Arabidopsis thaliana. *Ecology*, *83*(7), 1965–1980.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., & Apweiler, R. (2003). The Gene Ontology Annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, *13*(4), 662–672. https://doi.org/10.1101/gr.461403

Candela, H., Martınez-Laborda, A., & Micol, J. L. (1999). Venation pattern formation inArabidopsis thalianavegetative leaves. *Developmental Biology*, *205*(1), 205–216.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., & Lippert, C. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, *43*(10), 956–963.

Chang, I., Curran, A., Woolsey, R., Quilici, D., Cushman, J. C., Mittler, R., Harmon, A., & Harper, J. F. (2009). Proteomic profiling of tandem affinity purified 14-3-3 protein complexes in Arabidopsis thaliana. *Proteomics*, *9*(11), 2967–2985.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2015). Package 'shiny.' *See Http://Citeseerx. Ist. Psu. Edu/Viewdoc/Download*.

Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry*, *76*(1), 51–74.

Charitou, T., Bryan, K., & Lynn, D. J. (2016). Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, *48*(1), 1–12.

Chen, K., Du, L., & Chen, Z. (2003). Sensitization of defense responses and activation of programmed cell death by a pathogen-induced receptor-like protein kinase in Arabidopsis. *Plant Molecular Biology*, *53*(1), 61–74.

Chen, K., Fan, B., Du, L., & Chen, Z. (2004). Activation of hypersensitive cell death by pathogen-induced receptor-like protein kinases from Arabidopsis. *Plant Molecular Biology*, *56*(2), 271–283.

Chen, Y., Dubois, M., Vermeersch, M., Inzé, D., & Vanhaeren, H. (2021). Distinct cellular strategies determine sensitivity to mild drought of Arabidopsis natural accessions. *Plant Physiology*, *186*(2), 1171–1185. PubMed. https://doi.org/10.1093/plphys/kiab115

Cheng, C., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: A complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal*, *89*(4), 789–804.

Clément, M., Leonhardt, N., Droillard, M.-J., Reiter, I., Montillet, J.-L., Genty, B., Lauriere, C., Nussaume, L., & Noël, L. D. (2011). The cytosolic/nuclear HSC70 and HSP90 molecular chaperones are important for stomatal closure and modulate abscisic acid-dependent physiological responses in Arabidopsis. *Plant Physiology*, *156*(3), 1481–1492.

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, *8*(12), 1229–1231.

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology*, *2010*.

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change*, *3*(1), 52–58.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Depuydt, T., & Vandepoele, K. (2021). Multi-omics network-based functional annotation of unknown Arabidopsis genes. *The Plant Journal*, *108*(4), 1193–1212.

Diaz-Garcia, L., Covarrubias-Pazaran, G., Schlautman, B., & Zalapa, J. (2017). SOFIA: An R Package for Enhancing Genetic Visualization With Circos. *Journal of Heredity*, *108*(4), 443–448. https://doi.org/10.1093/jhered/esx023

Dowle, M., Srinivasan, A., Short, T., & Lianoglou, S. (2019). Data. Table: Extension of 'data. Frame.' *R Package Version*, *1*(8).

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., & Cherry, J. M. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research*, *30*(1), 69–72. https://doi.org/10.1093/nar/30.1.69

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H. A., & Weigel, D. (2018). Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nature Ecology & Evolution*, *2*(2), 352–358. https://doi.org/10.1038/s41559-017-0423-0

Fan Liu-Min, Zhang Wei, Chen Jin-Gui, Taylor J. Philip, Jones Alan M., & Assmann Sarah M. (2008). Abscisic acid regulation of guard-cell K+ and anion channels in Gβ- and RGS-deficient Arabidopsis lines. *Proceedings of the National Academy of Sciences*, *105*(24), 8476–8481. https://doi.org/10.1073/pnas.0800980105

Fernandez-Calvino, L., Faulkner, C., Walshaw, J., Saalbach, G., Bayer, E., Benitez-Alfonso, Y., & Maule, A. (2011). Arabidopsis plasmodesmal proteome. *PloS One*, *6*(4), e18880.

Ficklin, S. P., & Feltus, F. A. (2011). Gene coexpression network alignment and conservation of gene modules between two grass species: Maize and rice. *Plant Physiology*, *156*(3), 1244–1256.

Ficklin, S. P., Luo, F., & Feltus, F. A. (2010). The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks. *Plant Physiology*, *154*(1), 13–24.

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W.-K., & Mockler, T. C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, *20*(1), 45–58.

Fisher, R. (1918). *The correlation between relatives on the supposition of mendelian inheritance. Trans. Roy. Soc.*

Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., & Wilczek, A. M. (2011a). A map of local adaptation in Arabidopsis thaliana. *Science*, *334*(6052), 86–89.

Fournier-Level, A., Korte, A., Cooper, M., Nordborg, M., Schmitt, J., & Wilczek, A. (2011b). Data from: A map of local adaptation in Arabidopsis thaliana. *Dryad Digital Repository*.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*(3–4), 601–620.

Gaudet, P., Livstone, M., & Thomas, P. (2010). *Annotation inferences using phylogenetic trees*.

Gaxiola, R. A., Li, J., Undurraga, S., Dang, L. M., Allen, G. J., Alper, S. L., & Fink, G. R. (2001). Drought-and salt-tolerant plants result from overexpression of the AVP1 H+-pump. *Proceedings of the National Academy of Sciences*, *98*(20), 11444–11449.

Gillis, J., & Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, *8*(3), e1002444–e1002444. PubMed. https://doi.org/10.1371/journal.pcbi.1002444

Glass, K., & Girvan, M. (2014). Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets. *Scientific Reports*, *4*(1), 4191. https://doi.org/10.1038/srep04191

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu. Rev. Phytopathol.*, *43*, 205–227.

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Interpretation of Quantitative Research*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Gookin, T. E., Kim, J., & Assmann, S. M. (2008). Whole proteome identification of plant candidate G-protein coupled receptors in Arabidopsis, rice, and poplar: Computational prediction and in-vivo protein coupling. *Genome Biology*, *9*(7), R120. https://doi.org/10.1186/gb-2008-9-7-r120

Goujon, T., Ferret, V., Mila, I., Pollet, B., Ruel, K., Burlat, V., Joseleau, J.-P., Barrière, Y., Lapierre, C., & Jouanin, L. (2003). Down-regulation of the AtCCR1 gene in Arabidopsis thaliana: Effects on phenotype, lignins and cell wall degradability. *Planta*, *217*(2), 218–228. https://doi.org/10.1007/s00425-003-0987-6

Gregersen, N., Bross, P., Jørgensen, M. M., Corydon, T. J., & Andresen, B. S. (2000). Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders. *Journal of Inherited Metabolic Disease*, *23*(5), 441–447.

Grzybowska, E. A. (2018). Calcium-binding proteins with disordered structure and their role in secretion, storage, and cellular signaling. *Biomolecules*, *8*(2), 42.

Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith, R. K., Maiti, R., Chan, A. P., Yu, C., Farzad, M., & Wu, D. (2005). Complete reannotation of the Arabidopsis genome: Methods, tools, protocols and the final release. *BMC Biology*, *3*(1), 1–19.

Hagège, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., & Forné, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nature Protocols*, *2*(7), 1722–1733. https://doi.org/10.1038/nprot.2007.243

Hampe, A., & Petit, R. J. (2005). Conserving biodiversity under climate change: The rear edge matters. *Ecology Letters*, *8*(5), 461–467.

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., & Bergelson, J. (2011a). Adaptation to climate across the Arabidopsis thaliana genome. *Science*, *334*(6052), 83–86.

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., & Bergelson, J. (2011b). Adaptation to climate across the Arabidopsis thaliana genome. *Science*, *334*(6052), 83–86.

Hansen, B. G., Kerwin, R. E., Ober, J. A., Lambrix, V. M., Mitchell-Olds, T., Gershenzon, J., Halkier, B. A., & Kliebenstein, D. J. (2008). A Novel 2-Oxoacid-Dependent Dioxygenase Involved in the Formation of the Goiterogenic 2-Hydroxybut-3-enyl Glucosinolate and Generalist Insect Resistance in Arabidopsis. *Plant Physiology*, *148*(4), 2096–2108. https://doi.org/10.1104/pp.108.129981

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., & Young, R. A. (2000). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Biocomputing 2001* (pp. 422–433). World Scientific.

Hazbun, T. R., Malmström, L., Anderson, S., Graczyk, B. J., Fox, B., Riffle, M., Sundin, B. A., Aranda, J. D., McDonald, W. H., Chiu, C.-H., Snydsman, B. E., Bradley, P., Muller, E. G. D., Fields, S., Baker, D., Yates, J. R. 3rd, & Davis, T. N. (2003). Assigning function to yeast proteins by integration of technologies. *Molecular Cell*, *12*(6), 1353–1365. https://doi.org/10.1016/s1097-2765(03)00476-3

He, X.-J., Hsu, Y.-F., Pontes, O., Zhu, J., Lu, J., Bressan, R. A., Pikaard, C., Wang, C.-S., & Zhu, J.-K. (2009). NRPD4, a protein related to the RPB4 subunit of RNA polymerase II, is a component of RNA polymerases IV and V and is required for RNA-directed DNA methylation. *Genes & Development*, *23*(3), 318–330.

Heo, J. B., Sung, S., & Assmann, S. M. (2012). Ca2+-dependent GTPase, extra-large G protein 2 (XLG2), promotes activation of DNA-binding protein related to vernalization 1 (RTV1), leading to activation of floral integrator genes and early flowering in Arabidopsis. *Journal of Biological Chemistry*, *287*(11), 8242–8253.

Hereford, J. (2009). A quantitative survey of local adaptation and fitness trade-offs. *The American Naturalist*, *173*(5), 579–588.

HOCHBERG, Y. (1988). *A sharper Bonferroni procedure for multiple tests of significance*. *4*(75), 800–802.

Hoffmann, A. A., & Sgrò, C. M. (2011). Climate change and evolutionary adaptation. *Nature*, *470*(7335), 479–485. https://doi.org/10.1038/nature09670

Hoffmann, M. H. (2002). Biogeography of Arabidopsis thaliana (l.) heynh.(Brassicaceae). *Journal of Biogeography*, *29*(1), 125–134.

Hohmann, N., Schmickl, R., Chiang, T.-Y., Lučanová, M., Kolář, F., Marhold, K., & Koch, M. A. (2014). Taming the wild: Resolving the gene pools of non-model Arabidopsis lineages. *BMC Evolutionary Biology*, *14*(1), 1–21.

Holub, E. B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. *Nature Reviews Genetics*, *2*(7), 516–527.

Holub, E. B. (2007). Natural variation in innate immunity of a pioneer species. *Current Opinion in Plant Biology*, *10*(4), 415–424.

Hölzle, A., Jonietz, C., Törjek, O., Altmann, T., Binder, S., & Forner, J. (2011). A RESTORER OF FERTILITY-like PPR gene is required for 5′-end processing of the nad4 mRNA in mitochondria of Arabidopsis thaliana. *The Plant Journal*, *65*(5), 737–744.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., & Vilhjálmsson, B. J. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, *44*(2), 212–216.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., & Lane, H. C. (2007). DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, *35*(suppl_2), W169–W175.

Hubert, D. A., Tornero, P., Belkhadir, Y., Krishna, P., Takahashi, A., Shirasu, K., & Dangl, J. L. (2003). Cytosolic HSP90 associates with and modulates the Arabidopsis RPM1 disease resistance protein. *The EMBO Journal*, *22*(21), 5679–5689.

Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M., & Kimchi-Sarfaty, C. (2009). Silent (Synonymous) SNPs: Should We Care About Them? In A. A. Komar (Ed.), *Single Nucleotide Polymorphisms: Methods and Protocols* (pp. 23–39). Humana Press. https://doi.org/10.1007/978-1-60327-411-1_2

Huynen, M. A., & Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences*, *95*(11), 5849–5856.

Hwang, I. S., Choi, D. S., Kim, N. H., Kim, D. S., & Hwang, B. K. (2014). The pepper cysteine/histidine-rich DC1 domain protein CaDC1 binds both RNA and DNA and is required for plant cell death and defense response. *The New Phytologist*, *201*(2), 518–530. https://doi.org/10.1111/nph.12521

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, *292*(5518), 929–934.

Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. https://doi.org/10.1038/35048692

Jenssen, T.-K., Lægreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, *28*(1), 21–28.

Jezkova, T., & Wiens, J. J. (2016). Rates of change in climatic niches in plant and animal populations are much slower than projected climate change. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1843), 20162104.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., & Dean, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science*, *290*(5490), 344–347.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd edition). Wiley.

Jones Alexander M., Xuan Yuanhu, Xu Meng, Wang Rui-Sheng, Ho Cheng-Hsun, Lalonde Sylvie, You Chang Hun, Sardi Maria I., Parsa Saman A., Smith-Valle Erika, Su Tianying, Frazer Keith A., Pilot Guillaume, Pratelli Réjane, Grossmann Guido, Acharya Biswa R., Hu Heng-Cheng, Engineer Cawas, Villiers Florent, … Frommer Wolf B. (2014). Border Control—A Membrane-Linked Interactome of Arabidopsis. *Science*, *344*(6185), 711–716. https://doi.org/10.1126/science.1251358

Jones, J. D., & Dangl, J. L. (2006). The plant immune system. *Nature*, *444*(7117), 323–329.

Joo, J. H., Wang, S., Chen, J., Jones, A., & Fedoroff, N. V. (2005). Different signaling and cell death roles of heterotrimeric G protein α and β subunits in the Arabidopsis oxidative stress response to ozone. *The Plant Cell*, *17*(3), 957–970.

Kahle, D. J., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *R J.*, *5*(1), 144.

Kanapin, A., Batalov, S., Davis, M. J., Gough, J., Grimmond, S., Kawaji, H., Magrane, M., Matsuda, H., Schönbach, C., Teasdale, R. D., & Yuan, Z. (2003). Mouse proteome analysis. *Genome Research*, *13*(6B), 1335–1344. https://doi.org/10.1101/gr.978703

Kang, H.-G., Oh, C.-S., Sato, M., Katagiri, F., Glazebrook, J., Takahashi, H., Kachroo, P., Martin, G. B., & Klessig, D. F. (2010). Endosome-Associated CRT1 Functions Early in Resistance Gene–Mediated Defense Signaling in Arabidopsis and Tobacco. *The Plant Cell*, *22*(3), 918–936. https://doi.org/10.1105/tpc.109.071662

Kang, H.-G., Woo Choi, H., von Einem, S., Manosalva, P., Ehlers, K., Liu, P.-P., Buxa, S. V., Moreau, M., Mang, H.-G., Kachroo, P., Kogel, K.-H., & Klessig, D. F. (2012). CRT1 is a nuclear-translocated MORC endonuclease that participates in multiple levels of plant immunity. *Nature Communications*, *3*(1), 1297. https://doi.org/10.1038/ncomms2279

Kawakatsu, T., Huang, S.-S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery, J. R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C.-R., Wang, C., Bemm, F., Becker, C., O'Neil, R., O'Malley, R. C., Quarless, D. X., … Ecker, J. R. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell*, *166*(2), 492–505. https://doi.org/10.1016/j.cell.2016.06.044

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, *7*(12), 1225–1241.

KENDALL, M. G. (1938). A NEW MEASURE OF RANK CORRELATION. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81

Kerwin, R., Feusier, J., Corwin, J., Rubin, M., Lin, C., Muok, A., Larson, B., Li, B., Joseph, B., & Francisco, M. (2015). Natural genetic variation in Arabidopsis thaliana defense metabolism genes modulates field fitness. *Elife*, *4*, e05604.

Kesari, R., Lasky, J. R., Villamor, J. G., Des Marais, D. L., Chen, Y.-J. C., Liu, T.-W., Lin, W., Juenger, T. E., & Verslues, P. E. (2012). Intron-mediated alternative splicing of Arabidopsis P5CS1 and its association with natural variation in proline and climate adaptation. *Proceedings of the National Academy of Sciences*, *109*(23), 9197–9202.

Kim, H., Kojima, M., Choi, D., Park, S., Matsui, M., Sakakibara, H., & Hwang, I. (2016). Overexpression of INCREASED CAMBIAL ACTIVITY, a putative methyltransferase, increases cambial activity and plant growth. *Journal of Integrative Plant Biology*, *58*(11), 874–889.

Kim, M. J., Jang, I.-C., & Chua, N.-H. (2016). The mediator complex MED15 subunit mediates activation of downstream lipid-related genes by the WRINKLED1 transcription factor. *Plant Physiology*, *171*(3), 1951–1964.

Kim, P.-J., & Price, N. D. (2011). Genetic Co-Occurrence Network across Sequenced Microbes. *PLOS Computational Biology*, *7*(12), e1002340. https://doi.org/10.1371/journal.pcbi.1002340

King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., & Roth, F. P. (2003). Predicting gene function from patterns of annotation. *Genome Research*, *13*(5), 896–904.

Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., & Penin, A. A. (2016). A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, *88*(6), 1058–1070.

Kobe, B., & Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, *11*(6), 725–732.

Koch, M. A. (2019). The plant model system Arabidopsis set in an evolutionary, systematic, and spatio-temporal context. *Journal of Experimental Botany*, *70*(1), 55–67.

Koch, M., & Matschinger, M. (2007). *2007 Evolution and genetic differentiation among relatives of Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA 104, 6272-6277.*

Koornneef, M., Alonso-Blanco, C., Peeters, A. J., & Soppe, W. (1998). Genetic control of flowering time in Arabidopsis. *Annual Review of Plant Biology*, *49*(1), 345–370.

Koornneef, M., Alonso-Blanco, C., & Vreugdenhil, D. (2004). NATURALLY OCCURRING GENETIC VARIATION IN ARABIDOPSIS THALIANA. *Annual Review of Plant Biology*, *55*(1), 141–172. https://doi.org/10.1146/annurev.arplant.55.031903.141605

Koornneef, M., & Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal*, *61*(6), 909–921.

Kostaki, K.-I., Coupel-Ledru, A., Bonnell, V. C., Gustavsson, M., Sun, P., McLaughlin, F. J., Fraser, D. P., McLachlan, D. H., Hetherington, A. M., Dodd, A. N., & Franklin, K. A. (2020). Guard Cells Integrate Light and Temperature Signals to Control Stomatal Aperture. *Plant Physiology*, *182*(3), 1404–1419. https://doi.org/10.1104/pp.19.01528

Krannitz, P., Aarssen, L., & Lefebvre, D. (1991). Correction for non-linear relationships between root size and short term Pi uptake in genotype comparisons. *Plant and Soil*, *133*(2), 157–167.

Krannitz, P. G., Aarssen, L. W., & Dow, J. M. (1991). The effect of genetically based differences in seed size on seedling survival in Arabidopsis thaliana (Brassicaceae). *American Journal of Botany*, *78*(3), 446–450.

Kronholm, I., Picó, F. X., Alonso-Blanco, C., Goudet, J., & Meaux, J. de. (2012). Genetic basis of adaptation in Arabidopsis thaliana: Local adaptation at the seed dormancy QTL DOG1. *Evolution: International Journal of Organic Evolution*, *66*(7), 2287–2302.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics*, *17*(1), 21–24.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645.

Kuhlmann, M., Meyer, R. C., Jia, Z., Klose, D., Krieg, L.-M., von Wirén, N., & Altmann, T. (2020). Epigenetic Variation at a Genomic Locus Affecting Biomass Accumulation under Low Nitrogen in Arabidopsis thaliana. *Agronomy*, *10*(5). https://doi.org/10.3390/agronomy10050636

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, *99*(2), 248–256. https://doi.org/10.3732/ajb.1100340

Kwok, P.-Y., Deng, Q., Zakeri, H., Taylor, S. L., & Nickerson, D. A. (1996). Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics*, *31*(1), 123–126.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*(D1), D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Langridge, J., & Griffing, B. (1959). A study of high temperature lesions in Arabidopsis thaliana. *Australian Journal of Biological Sciences*, *12*(2), 117–135.

Lasky, J. R., Des Marais, D. L., Lowry, D. B., Povolotskaya, I., McKay, J. K., Richards, J. H., Keitt, T. H., & Juenger, T. E. (2014). Natural variation in abiotic stress responsive gene expression and local adaptation to climate in Arabidopsis thaliana. *Molecular Biology and Evolution*, *31*(9), 2283–2296.

Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of Arabidopsis thaliana: The roles of geography and climate. *Molecular Ecology*, *21*(22), 5512–5529.

Lease, K. A., Wen, J., Li, J., Doke, J. T., Liscum, E., & Walker, J. C. (2001). A Mutant Arabidopsis Heterotrimeric G-Protein β Subunit Affects Leaf, Flower, and Fruit Development. *The Plant Cell*, *13*(12), 2631–2641. https://doi.org/10.1105/tpc.010315

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, *14*(6), 1085–1094.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., & Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology*, *28*(2), 149–156.

Lee, J.-H., Terzaghi, W., Gusmaroli, G., Charron, J.-B. F., Yoon, H.-J., Chen, H., He, Y. J., Xiong, Y., & Deng, X. W. (2008). Characterization of Arabidopsis and Rice DWD Proteins and Their Roles as Substrate Receptors for CUL4-RING E3 Ubiquitin Ligases. *The Plant Cell*, *20*(1), 152–167. https://doi.org/10.1105/tpc.107.055418

Lee, S., Senthil-Kumar, M., Kang, M., Rojas, C. M., Tang, Y., Oh, S., Choudhury, S. R., Lee, H.-K., Ishiga, Y., Allen, R. D., Pandey, S., & Mysore, K. S. (2017). The small GTPase, nucleolar GTP-binding protein 1 (NOG1), has a novel role in plant innate immunity. *Scientific Reports*, *7*(1), 9260. https://doi.org/10.1038/s41598-017-08932-9

Lee, T.-H., Kim, Y.-K., Pham, T. T. M., Song, S. I., Kim, J.-K., Kang, K. Y., An, G., Jung, K.-H., Galbraith, D. W., & Kim, M. (2009). RiceArrayNet: A database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiology*, *151*(1), 16–33.

Lee, W. Y., Lee, D., Chung, W., & Kwon, C. S. (2009). Arabidopsis ING and Alfin1-like protein families localize to the nucleus and bind to H3K4me3/2 via plant homeodomain fingers. *The Plant Journal*, *58*(3), 511–524.

Lee-Yaw, J. A., Kharouba, H. M., Bontrager, M., Mahony, C., Csergő, A. M., Noreen, A. M., Li, Q., Schuster, R., & Angert, A. L. (2016). A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. *Ecology Letters*, *19*(6), 710–722.

Lim, C. J., Yang, K., Hong, J. K., Choi, J. S., Yun, D.-J., Hong, J. C., Chung, W. S., Lee, S. Y., Cho, M. J., & Lim, C. O. (2006). Gene expression profiles during heat acclimation in Arabidopsis thaliana suspension-culture cells. *Journal of Plant Research*, *119*(4), 373–383.

Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M.-I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., & Barnstead, M. (1999). Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature*, *402*(6763), 761–768.

Linster, E., Stephan, I., Bienvenut, W. V., Maple-Grødem, J., Myklebust, L. M., Huber, M., Reichelt, M., Sticht, C., Geir Møller, S., Meinnel, T., Arnesen, T., Giglione, C., Hell, R., & Wirtz, M. (2015). Downregulation of N-terminal acetylation triggers ABA-mediated drought responses in Arabidopsis. *Nature Communications*, *6*(1), 7640. https://doi.org/10.1038/ncomms8640

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536.

Liu, Z.-W., Shao, C.-R., Zhang, C.-J., Zhou, J.-X., Zhang, S.-W., Li, L., Chen, S., Huang, H.-W., Cai, T., & He, X.-J. (2014). The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. *PLoS Genetics*, *10*(1), e1003948.

Llorente, F., Alonso-Blanco, C., Sánchez-Rodriguez, C., Jorda, L., & Molina, A. (2005). ERECTA receptor-like kinase and heterotrimeric G protein from Arabidopsis are required for resistance to the necrotrophic fungus Plectosphaerella cucumerina. *The Plant Journal*, *43*(2), 165–180.

Lodish H, Berk A, & Zipursky SL. (2000). *Molecular Cell Biology* (4th Edition). W. H. Freeman. https://www.ncbi.nlm.nih.gov/books/NBK21578/

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., & Nizhynska, V. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, *45*(8), 884–890.

Ludlow, M. (1989). *Strategies of response to water stress*.

Lueck, J. D., Yoon, J. S., Perales-Puchalt, A., Mackey, A. L., Infield, D. T., Behlke, M. A., Pope, M. R., Weiner, D. B., Skach, W. R., McCray, P. B., Jr, & Ahern, C. A. (2019). Engineered transfer RNAs for suppression of premature termination codons. *Nature Communications*, *10*(1), 822–822. PubMed. https://doi.org/10.1038/s41467-019-08329-4

Lutz, U., Pose, D., Pfeifer, M., Gundlach, H., Hagmann, J., Wang, C., Weigel, D., Mayer, K. F., Schmid, M., & Schwechheimer, C. (2015). Modulation of ambient temperature-dependent flowering in Arabidopsis thaliana by natural variation of FLOWERING LOCUS M. *PLoS Genetics*, *11*(10), e1005588.

MacLennan, N. K., Dong, J., Aten, J. E., Horvath, S., Rahib, L., Ornelas, L., Dipple, K. M., & McCabe, E. R. (2009). Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice. *Molecular Genetics and Metabolism*, *98*(1–2), 203–214.

Magrath, R., Bano, F., Morgner, M., Parkin, I., Sharpe, A., Lister, C., Dean, C., Turner, J., Lydiate, D., & Mithen, R. (1994). Genetics of aliphatic glucosinolates. I. Side chain elongation in Brassica napus and Arabidopsis thaliana. *Heredity*, *72*(3), 290–299.

Mähler, N., Wang, J., Terebieniec, B. K., Ingvarsson, P. K., Street, N. R., & Hvidsten, T. R. (2017). Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics*, *13*(4), e1006402.

Mahto, A. (2014). Package 'splitstackshape'—Stack and reshape datasets after splitting concatenated values. *CRAN. Available at URL Https://Cran. r-Project. Org/Web/Packages/Splitstackshape/Splitstackshape. Pdf*, *12*.

Mahto, A. (2018). *Splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values. R package. Version 1.4. 6*.

Mao, L., Van Hemert, J. L., Dash, S., & Dickerson, J. A. (2009). Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, *10*(1), 1–24.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, *285*(5428), 751–753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, *402*(6757), 83–86.

Mason Michael G. & Botella José R. (2000). Completing the heterotrimer: Isolation and characterization of an Arabidopsis thaliana G protein γ-subunit cDNA. *Proceedings of the National Academy of Sciences*, *97*(26), 14784–14788. https://doi.org/10.1073/pnas.97.26.14784

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., & Terryn, N. (1999). Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. *Nature*, *402*(6763), 769–777.

Mckay, J. K., Richards, J. H., & Mitchell-Olds, T. (2003). Genetics of drought adaptation in Arabidopsis thaliana: I. Pleiotropy contributes to genetic correlations among ecological traits. *Molecular Ecology*, *12*(5), 1137–1151.

Méndez-Vigo, B., Picó, F. X., Ramiro, M., Martínez-Zapater, J. M., & Alonso-Blanco, C. (2011). Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in Arabidopsis. *Plant Physiology*, *157*(4), 1942–1955.

Menges, M., Hennig, L., Gruissem, W., & Murray, J. A. H. (2003). Genome-wide gene expression in an Arabidopsis cell suspension. *Plant Molecular Biology*, *53*(4), 423–442. https://doi.org/10.1023/B:PLAN.0000019059.56489.ca

Mentewab, A., & Stewart, C. N. (2005). Overexpression of an Arabidopsis thaliana ABC transporter confers kanamycin resistance to transgenic plants. *Nature Biotechnology*, *23*(9), 1177–1180. https://doi.org/10.1038/nbt1134

Mitchell-Olds, T. (1996). Genetic constraints on life-history evolution: Quantitative-trait loci influencing growth and flowering in Arabidopsis thaliana. *Evolution*, *50*(1), 140–145.

Mitchell-Olds, T., & Pedersen, D. (1998). The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. *Genetics*, *149*(2), 739–747.

Mitchell-Olds, T., & Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, *441*(7096), 947–952. https://doi.org/10.1038/nature04878

Mithen, R., Clarke, J., Lister, C., & Dean, C. (1995). Genetics of aliphatic glucosinolates. III. Side chain structure of aliphatic glucosinolates in Arabidopsis thaliana. *Heredity*, *74*(2), 210–215.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, *14*(10), 719–732.

Moissiard Guillaume, Bischof Sylvain, Husmann Dylan, Pastor William A., Hale Christopher J., Yen Linda, Stroud Hume, Papikian Ashot, Vashisht Ajay A., Wohlschlegel James

A., & Jacobsen Steven E. (2014). Transcriptional gene silencing by Arabidopsis microrchidia homologues involves the formation of heteromers. *Proceedings of the National Academy of Sciences*, *111*(20), 7474–7479. https://doi.org/10.1073/pnas.1406611111

Montesinos-Navarro, A., Picó, F. X., & Tonsor, S. J. (2012). Clinal variation in seed traits influencing life cycle timing in Arabidopsis thaliana. *Evolution: International Journal of Organic Evolution*, *66*(11), 3417–3431.

Mostafavi, S., & Morris, Q. (2010). Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, *26*(14), 1759–1765.

Mudgil, Y., Uhrig, J. F., Zhou, J., Temple, B., Jiang, K., & Jones, A. M. (2009). Arabidopsis N-MYC DOWNREGULATED-LIKE1, a Positive Regulator of Auxin Transport in a G Protein–Mediated Pathway. *The Plant Cell*, *21*(11), 3591–3609. https://doi.org/10.1105/tpc.109.065557

Müller, B., Fastner, A., Karmann, J., Mansch, V., Hoffmann, T., Schwab, W., Suter-Grotemeyer, M., Rentsch, D., Truernit, E., & Ladwig, F. (2015). Amino acid export in developing Arabidopsis seeds depends on UmamiT facilitators. *Current Biology*, *25*(23), 3126–3131.

Müller, H., & Mancuso, F. (2008). Identification and analysis of co-occurrence networks with NetCutter. *PloS One*, *3*(9), e3178–e3178. https://doi.org/10.1371/journal.pone.0003178

Murphy, A., & Taiz, L. (1995). A new vertical mesh transfer technique for metal-tolerance studies in Arabidopsis (ecotypic variation and copper-sensitive mutants). *Plant Physiology*, *108*(1), 29–38.

Nakamura Yuki, Koizumi Ryota, Shui Guanghou, Shimojima Mie, Wenk Markus R., Ito Toshiro, & Ohta Hiroyuki. (2009). Arabidopsis lipins mediate eukaryotic pathway of lipid metabolism and cope critically with phosphate starvation. *Proceedings of the National Academy of Sciences*, *106*(49), 20978–20983. https://doi.org/10.1073/pnas.0907173106

Nienhuis, J., Sills, G. R., Martin, B., & King, G. (1994). Variance for water-use efficiency among ecotypes and recombinant inbred lines of Arabidopsis thaliana (Brassicaceae). *American Journal of Botany*, *81*(8), 943–947.

Nikolovski, N., Rubtsov, D., Segura, M. P., Miles, G. P., Stevens, T. J., Dunkley, T. P., Munro, S., Lilley, K. S., & Dupree, P. (2012). Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiology*, *160*(2), 1037–1051.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., & Goyal, R. (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, *3*(7), e196.

Obulareddy, N., Panchal, S., & Melotto, M. (2013). Guard cell purification and RNA isolation suitable for high-throughput transcriptional analysis of cell-type responses to biotic stresses. *Molecular Plant-Microbe Interactions*, *26*(8), 844–849.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple Testing and Statistical Power With Modified Bonferroni Procedures. *Journal of Educational and Behavioral Statistics*, *22*(4), 389–406. https://doi.org/10.3102/10769986022004389

Ordon, J., Martin, P., Erickson, J. L., Ferik, F., Balcke, G., Bonas, U., & Stuttmann, J. (2021). Disentangling cause and consequence: Genetic dissection of the DANGEROUS MIX2 risk locus, and activation of the DM2h NLR in autoimmunity. *The Plant Journal*, *106*(4), 1008–1023.

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 1–10.

Pandey, S., Monshausen, G. B., Ding, L., & Assmann, S. M. (2008). Regulation of root-wave response by extra large and conventional G proteins in Arabidopsis thaliana. *The Plant Journal*, *55*(2), 311–322.

Pant, B. D., Lee, S., Lee, H.-K., Krom, N., Pant, P., Jang, Y., & Mysore, K. S. (2022). Overexpression of Arabidopsis nucleolar GTP-binding 1 (NOG1) proteins confers drought tolerance in rice. *Plant Physiology*, *189*(2), 988–1004. https://doi.org/10.1093/plphys/kiac078

Par˘enicová, L., de Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., Cook, H. E., Ingram, R. M., Kater, M. M., Davies, B., Angenent, G. C., & Colombo, L. (2003). Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World[W]. *The Plant Cell*, *15*(7), 1538–1551. https://doi.org/10.1105/tpc.011544

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, *421*(6918), 37–42.

Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, *202*(3), 615–627.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, *96*(8), 4285–4288.

Pierrelée, M., Reynders, A., Lopez, F., Moqrich, A., Tichit, L., & Habermann, B. H. (2021). Introducing the novel Cytoscape app TimeNexus to analyze time-series data using temporal MultiLayer Networks (tMLNs). *Scientific Reports*, *11*(1), 13691. https://doi.org/10.1038/s41598-021-93128-5

Pigliucci, M. (1998). Ecological and evolutionary genetics of Arabidopsis. *Trends in Plant Science*, *3*(12), 485–489.

Pigliucci, M., Pollard, H., & Cruzan, M. B. (2003). Comparative studies of evolutionary responses to light environments in Arabidopsis. *The American Naturalist*, *161*(1), 68–82.

Pignatta, D., Erdmann, R. M., Scheer, E., Picard, C. L., Bell, G. W., & Gehring, M. (2014). Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife*, *3*, e03198.

Pizzio, G. A., Paez-Valencia, J., Khadilkar, A. S., Regmi, K., Patron-Soberano, A., Zhang, S., Sanchez-Lares, J., Furstenau, T., Li, J., & Sanchez-Gomez, C. (2015). Arabidopsis type I proton-pumping pyrophosphatase expresses strongly in phloem, where it is required for pyrophosphate metabolism and photosynthate partitioning. *Plant Physiology*, *167*(4), 1541–1553.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., & White, J. (2001). The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, *29*(1), 159–164.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramonell, K., Berrocal-Lobo, M., Koh, S., Wan, J., Edwards, H., Stacey, G., & Somerville, S. (2005). Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen Erysiphe cichoracearum. *Plant Physiology*, *138*(2), 1027–1036.

Rao, M. V., & Davis, K. R. (1999). Ozone-induced cell death occurs via two distinct mechanisms in Arabidopsis: The role of salicylic acid. *The Plant Journal*, *17*(6), 603–614.

Rashotte, A. M., Jenks, M. A., Nguyen, T. D., & Feldmann, K. A. (1997). Epicuticular wax variation in ecotypes of Arabidopsis thaliana. *Phytochemistry*, *45*(2), 251–255.

Ratcliffe, D. (1976). Germination characteristics and their inter-and intra-population variability in Arabidopsis. *Arabidopsis Information Service*, *13*, 34–45.

Ream, T. S., Haag, J. R., Wierzbicki, A. T., Nicora, C. D., Norbeck, A. D., Zhu, J.-K., Hagen, G., Guilfoyle, T. J., Paša-Tolić, L., & Pikaard, C. S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Molecular Cell*, *33*(2), 192–203.

Reichel, M., Liao, Y., Rettel, M., Ragan, C., Evers, M., Alleaume, A.-M., Horos, R., Hentze, M. W., Preiss, T., & Millar, A. A. (2016). In Planta Determination of the mRNA-Binding Proteome of Arabidopsis Etiolated Seedlings. *The Plant Cell*, *28*(10), 2435–2452. https://doi.org/10.1105/tpc.16.00562

Reuber, T. L., & Ausubel, F. M. (1996). Isolation of Arabidopsis genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. *The Plant Cell*, *8*(2), 241–249. https://doi.org/10.1105/tpc.8.2.241

Riechmann J. L., Heard J., Martin G., Reuber L., Jiang C.-Z., Keddie J., Adam L., Pineda O., Ratcliffe O. J., Samaha R. R., Creelman R., Pilgrim M., Broun P., Zhang J. Z., Ghandehari D., Sherman B. K., & -L. Yu G. (2000). Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science*, *290*(5499), 2105–2110. https://doi.org/10.1126/science.290.5499.2105

Rockman, M. V. (2008). Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature*, *456*(7223), 738–744.

Roux, F., Touzet, P., Cuguen, J., & Le Corre, V. (2006). How to be early flowering: An evolutionary perspective. *Trends in Plant Science*, *11*(8), 375–381.

RStudio Team. (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc. http://www.rstudio.com/

Rudella, A., Friso, G., Alonso, J. M., Ecker, J. R., & van Wijk, K. J. (2006). Downregulation of ClpR2 Leads to Reduced Accumulation of the ClpPRS Protease Complex and Defects in Chloroplast Biogenesis in Arabidopsis. *The Plant Cell*, *18*(7), 1704–1721. https://doi.org/10.1105/tpc.106.042861

Saez-Aguayo, S., Rondeau-Mouro, C., Macquet, A., Kronholm, I., Ralet, M.-C., Berger, A., Sallé, C., Poulain, D., Granier, F., Botran, L., Loudet, O., de Meaux, J., Marion-Poll, A., & North, H. M. (2014). Local Evolution of Seed Flotation in Arabidopsis. *PLOS Genetics*, *10*(3), e1004221. https://doi.org/10.1371/journal.pgen.1004221

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., & Fartmann, B. (2000). European Union Chromosome 3 Arabidopsis sequencing consortium; institute for genomic research; Kazusa DNA Research Institute. Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. *Nature*, *408*, 820–822.

Savas, S., Tuzmen, S., & Ozcelik, H. (2006). Human SNPs resulting in premature stop codons and protein truncation. *Human Genomics*, *2*(5), 1–13.

Schapire, A. L., Valpuesta, V., & Botella, M. A. (2006). TPR proteins in plant hormone signaling. *Plant Signaling & Behavior*, *1*(5), 229–230.

Schell, T., Kulozik, A. E., & Hentze, M. W. (2002). Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. *Genome Biology*, *3*(3), 1–6.

Schleicher, S., & Binder, S. (2021). In Arabidopsis thaliana mitochondria 5′ end polymorphisms of nad4L-atp4 and nad3-rps12 transcripts are linked to RNA PROCESSING FACTORs 1 and 8. *Plant Molecular Biology*, *106*(4), 335–348.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., & Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, *37*(5), 501–506.

Schmitz, R. J., & Ecker, J. R. (2012). Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends in Plant Science*, *17*(3), 149–154.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., & Schork, N. J. (2013). Patterns of population epigenomic diversity. *Nature*, *495*(7440), 193–198.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., & Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, *20*(16), 2597–2604.

Schwarte, S., Brust, H., Steup, M., & Tiedemann, R. (2013). Intraspecific sequence variation and differential expression in starch synthase genes of Arabidopsis thaliana. *BMC Research Notes*, *6*(1), 84. https://doi.org/10.1186/1756-0500-6-84

Schwartz, C. J., Doyle, M., Manzaneda, A., Rey, P., Mitchell-Olds, T., & Amasino, R. (2010). Natural Variation of Flowering Time and Vernalization Responsiveness in Brachypodium distachyon. *BioEnergy Research*, *3*, 38–46. https://doi.org/10.1007/s12155-009-9069-3

Scott, P. (2000). Resurrection plants and the secrets of eternal leaf. *Annals of Botany*, *85*(2), 159–166.

Segami, S., Tomoyama, T., Sakamoto, S., Gunji, S., Fukuda, M., Kinoshita, S., Mitsuda, N., Ferjani, A., & Maeshima, M. (2018). Vacuolar H+-Pyrophosphatase and Cytosolic Soluble Pyrophosphatases Cooperatively Regulate Pyrophosphate Levels in Arabidopsis thaliana. *The Plant Cell*, *30*(5), 1040–1061. https://doi.org/10.1105/tpc.17.00911

Seo, P. J., Kim, M. J., Park, J., Kim, S., Jeon, J., Lee, Y., Kim, J., & Park, C. (2010). Cold activation of a plasma membrane-tethered NAC transcription factor induces a pathogen resistance response in Arabidopsis. *The Plant Journal*, *61*(4), 661–671.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Sharma, R., Griffing, B., & Scholl, R. (1979). Variations among races of Arabidopsis thaliana (L.) Heynh for survival in limited carbon dioxide. *Theoretical and Applied Genetics*, *54*(1), 11–15.

Shen, Y., Wang, J., Ding, Y. u., Lo, S. W., Gouzerh, G., Neuhaus, J.-M., & Jiang, L. (2011). The Rice RMR1 Associates with a Distinct Prevacuolar Compartment for the Protein Storage Vacuole Pathway. *Molecular Plant*, *4*(5), 854–868. https://doi.org/10.1093/mp/ssr025

Sherrard, M. E., & Maherali, H. (2006). The adaptive significance of drought escape in Avena barbata, an annual grass. *Evolution*, *60*(12), 2478–2489.

Shindo, C., Bernasconi, G., & Hardtke, C. S. (2007). Natural genetic variation in Arabidopsis: Tools, traits and prospects for evolutionary ecology. *Annals of Botany*, *99*(6), 1043–1054.

Siepielski, A. M., Morrissey, M. B., Buoro, M., Carlson, S. M., Caruso, C. M., Clegg, S. M., Coulson, T., DiBattista, J., Gotanda, K. M., & Francis, C. D. (2017). Precipitation drives global variation in natural selection. *Science*, *355*(6328), 959–962.

Single-nucleotide polymorphism. (2015, October 11). Scitable by Nature. https://www.nature.com/scitable/

Somerville, C., & Koornneef, M. (2002). A fortunate choice: The history of Arabidopsis as a model plant. *Nature Reviews Genetics*, *3*(11), 883–889.

Song, H., Zhao, R., Fan, P., Wang, X., Chen, X., & Li, Y. (2009). Overexpression of AtHsp90. 2, AtHsp90. 5 and AtHsp90. 7 in Arabidopsis thaliana enhances plant sensitivity to salt and drought stresses. *Planta*, *229*(4), 955–964.

Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinformatics*, *13*(1), 1–21.

Stapley, B. J., & Benoit, G. (1999). Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Biocomputing 2000* (pp. 529–540). World Scientific.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, *18*(suppl_2), S231–S240.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, *302*(5643), 249–255.

Stuttmann, J., Peine, N., Garcia, A. V., Wagner, C., Choudhury, S. R., Wang, Y., James, G. V., Griebel, T., Alcázar, R., & Tsuda, K. (2016). Arabidopsis thaliana DM2h (R8) within the Landsberg RPP1-like resistance locus underlies three different cases of EDS1-conditioned autoimmunity. *PLoS Genetics*, *12*(4), e1005990.

Suzuki, D. T., Griffiths, A. J., & Miller, JH. (2000). *An introduction to genetic analysis.* (7th Edition). New York: WH Freeman and Company.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., & Ploetz, L. (2007). The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research*, *36*(suppl_1), D1009–D1014.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., Kawashima, K., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Naruo, K., Okumura, S., … Fransz, P. (2000). Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. *Nature*, *408*(6814), 823–826. https://doi.org/10.1038/35048507

Takahashi, T., Naito, S., & Komeda, Y. (1992). Isolation and analysis of the expression of two genes for the 81-kilodalton heat-shock proteins from Arabidopsis. *Plant Physiology*, *99*(2), 383–390.

Tanaka, K., Swanson, S. J., Gilroy, S., & Stacey, G. (2010). Extracellular nucleotides elicit cytosolic free calcium oscillations in Arabidopsis. *Plant Physiology*, *154*(2), 705–719.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., & Koonin, E. V. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, *29*(1), 22–28.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. https://doi.org/10.1038/35048692

The Arabidopsis Information Resource. (n.d.). http://arabidopsis.org

The Institute of Genomic Research. (n.d.). https://www.jcvi.org

Theologis, A., Ecker, J. R., Palm, C. J., Federspiel, N. A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., & Bowman, C. L. (2000). Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. *Nature*, *408*(6814), 816–820.

Thieme, C. J., Rojas-Triana, M., Stecyk, E., Schudoma, C., Zhang, W., Yang, L., Miñambres, M., Walther, D., Schulze, W. X., & Paz-Ares, J. (2015). Endogenous Arabidopsis messenger RNAs transported to distant tissues. *Nature Plants*, *1*(4), 1–9.

Thieme, C. J., Rojas-Triana, M., Stecyk, E., Schudoma, C., Zhang, W., Yang, L., Miñambres, M., Walther, D., Schulze, W. X., Paz-Ares, J., Scheible, W.-R., & Kragler, F. (2015). Endogenous Arabidopsis messenger RNAs transported to distant tissues. *Nature Plants*, *1*(4), 15025. https://doi.org/10.1038/nplants.2015.25

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences*, *102*(23), 8245–8250.

Tian, D., Traw, M., Chen, J., Kreitman, M., & Bergelson, J. (2003). Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature*, *423*(6935), 74–77.

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, *29*(12), 673–680. https://doi.org/10.1016/j.tree.2014.10.004

Trentmann, S. M., & Kende, H. (1995). Analysis of Arabidopsis cDNA that shows homology to the tomato E8 cDNA. *Plant Molecular Biology*, *29*(1), 161–166. https://doi.org/10.1007/BF00019127

Trusov, Y., Rookes, J. E., Chakravorty, D., Armour, D., Schenk, P. M., & Botella, J. R. (2006). Heterotrimeric G proteins facilitate Arabidopsis resistance to necrotrophic pathogens and are involved in jasmonate signaling. *Plant Physiology*, *140*(1), 210–220.

Trusov, Y., Zhang, W., Assmann, S. M., & Botella, J. R. (2008). G γ 1+ G γ 2=/G β: Heterotrimeric G Protein G γ-Deficient Mutants Do Not Recapitulate All Phenotypes of G β-Deficient Mutants. *Plant Physiology*, *147*(2), 636–649.

Tsugama, D., Liu, S., & Takano, T. (2013). Arabidopsis heterotrimeric G protein β subunit, AGB1, regulates brassinosteroid signalling independently of BZR1. *Journal of Experimental Botany*, *64*(11), 3213–3223.

Tuteja, N. (2007). Abscisic Acid and Abiotic Stress Signaling. *Plant Signaling & Behavior*, *2*(3), 135–138. https://doi.org/10.4161/psb.2.3.4156

Urano, D., Phan, N., Jones, J. C., Yang, J., Huang, J., Grigston, J., Philip Taylor, J., & Jones, A. M. (2012). Endocytosis of the seven-transmembrane RGS1 protein activates G-protein-coupled signalling in Arabidopsis. *Nature Cell Biology*, *14*(10), 1079–1088. https://doi.org/10.1038/ncb2568

Vaughn, M. W., Tanurdžić, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P. D., Dedhia, N., McCombie, W. R., Agier, N., & Bulski, A. (2007). Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biology*, *5*(7), e174.

Vercruyssen, L., Gonzalez, N., Werner, T., Schmülling, T., & Inzé, D. (2011). Combining enhanced root and shoot growth reveals cross talk between pathways that control plant organ size in Arabidopsis. *Plant Physiology*, *155*(3), 1339–1352.

Von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., & Bork, P. (2007). STRING 7—Recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, *35*(suppl_1), D358–D362.

Wahde, M., & Hertz, J. (2000). Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, *55*(1–3), 129–136.

Wang, G., Ellendorff, U., Kemp, B., Mansfield, J. W., Forsyth, A., Mitchell, K., Bastas, K., Liu, C.-M., Woods-Tör, A., Zipfel, C., de Wit, P. J. G. M., Jones, J. D. G., Tör, M., & Thomma, B. P. H. J. (2008). A Genome-Wide Functional Investigation into the Roles of Receptor-Like Proteins in Arabidopsis. *Plant Physiology*, *147*(2), 503–517. https://doi.org/10.1104/pp.108.119487

Wang, L., Kazachkov, M., Shen, W., Bai, M., Wu, H., & Zou, J. (2014). Deciphering the roles of Arabidopsis LPCAT and PAH in phosphatidylcholine homeostasis and pathway coordination for chloroplast lipid synthesis. *The Plant Journal*, *80*(6), 965–976.

Wang, S., Assmann, S. M., & Fedoroff, N. V. (2008). Characterization of the Arabidopsis heterotrimeric G protein. *Journal of Biological Chemistry*, *283*(20), 13913–13922.

Wang, S., Narendra, S., & Fedoroff, N. (2007). Heterotrimeric G protein signaling in the Arabidopsis unfolded protein response. *Proceedings of the National Academy of Sciences*, *104*(10), 3817–3822.

Wang, Y., Lyu, W., Berkowitz, O., Radomiljac, J. D., Law, S. R., Murcha, M. W., Carrie, C., Teixeira, P. F., Kmiec, B., Duncan, O., Van Aken, O., Narsai, R., Glaser, E., Huang, S., Roessner, U., Millar, A. H., & Whelan, J. (2016). Inactivation of Mitochondrial Complex I Induces the Expression of a Twin Cysteine Protein that Targets and Affects Cytosolic, Chloroplastidic and Mitochondrial Function. *Molecular Plant*, *9*(5), 696–710. https://doi.org/10.1016/j.molp.2016.01.009

Weigel, D. (2012). Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. *Plant Physiology*, *158*(1), 2–22. https://doi.org/10.1104/pp.111.189845

Weigel, D., & Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, *10*(5), 1–5.

Weigel, D., & Nordborg, M. (2015). Population genomics for understanding adaptation in wild plant species. *Annual Review of Genetics*, *49*(1), 315–338.

Wickham, H. (2011a). Ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185.

Wickham, H. (2011b). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. *R Package Version 0.4*, *3*, p156.

Wickham, H., & Wickham, M. H. (2015). *Package 'reshape.'*

Wickham, H., & Wickham, M. H. (2017). Package 'tidyr.' *Easily Tidy Data with'spread'and'gather ()'Functions*.

Wickham, H., & Wickham, M. H. (2019). *Package 'stringr.'*

Wong, T.-Y., Fernandes, S., Sankhon, N., Leong, P. P., Kuo, J., & Liu, J.-K. (2008). Role of premature stop codons in bacterial evolution. *Journal of Bacteriology*, *190*(20), 6718–6725.

Wuest, S. E., Vijverberg, K., Schmidt, A., Weiss, M., Gheyselinck, J., Lohr, M., Wellmer, F., Rahnenführer, J., von Mering, C., & Grossniklaus, U. (2010). Arabidopsis Female Gametophyte Gene Expression Map Reveals Similarities between Plant and Animal Gametes. *Current Biology*, *20*(6), 506–512. https://doi.org/10.1016/j.cub.2010.01.051

Xiao, Y. (2017). A fast algorithm for two-dimensional Kolmogorov–Smirnov two sample tests. *Computational Statistics & Data Analysis*, *105*, 53–58.

Yáñez, M., Gil-Longo, J., & Campos-Toimil, M. (2012). Calcium binding proteins. *Advances in Experimental Medicine and Biology*, *740*, 461–482. https://doi.org/10.1007/978-94-007-2888-2_19

Yeh, Y.-H., Chang, Y.-H., Huang, P.-Y., Huang, J.-B., & Zimmerli, L. (2015). Enhanced Arabidopsis pattern-triggered immunity by overexpression of cysteine-rich receptor-like kinases. *Frontiers in Plant Science*, *6*, 322.

Yeung, N., Cline, M. S., Kuchinsky, A., Smoot, M. E., & Bader, G. D. (2008). Exploring biological networks with Cytoscape software. *Current Protocols in Bioinformatics*, *23*(1), 8–13.

Yoo, A. B., Jette, M. A., & Grondona, M. (2003). *Slurm: Simple linux utility for resource management*. 44–60.

Yu, T.-Y., Shi, D.-Q., Jia, P.-F., Tang, J., Li, H.-J., Liu, J., & Yang, W.-C. (2016). The Arabidopsis receptor kinase ZAR1 is required for zygote asymmetric division and its daughter cell fate. *PLoS Genetics*, *12*(3), e1005933.

Zhan, X., Dixon, A., Batbayar, N., Bragin, E., Ayas, Z., Deutschova, L., Chavko, J., Domashevsky, S., Dorosencu, A., Bagyura, J., Gombobaatar, S., Grlica, I. D., Levin, A., Milobog, Y., Ming, M., Prommer, M., Purev-Ochir, G., Ragyov, D., Tsurkanu, V., … Bruford, M. W. (2015). Exonic versus intronic SNPs: Contrasting roles in revealing the population genetic differentiation of a widespread bird species. *Heredity*, *114*(1), 1–9. https://doi.org/10.1038/hdy.2014.59

Zhang, H. (2016). *Using the RCircos Package*.

Zhang, H., Meltzer, P., & Davis, S. (2013). RCircos: An R package for Circos 2D track plots. *BMC Bioinformatics*, *14*(1), 1–5.

Zhang, J., & Lechowicz, M. J. (1995). Responses to CO2 enrichment by two genotypes of Arabidopsis thaliana differing in their sensitivity to nutrient availability. *Annals of Botany*, *75*(5), 491–499.

Zhang, T., Xu, P., Wang, W., Wang, S., Caruana, J. C., Yang, H.-Q., & Lian, H. (2018). Arabidopsis G-protein β subunit AGB1 interacts with BES1 to regulate brassinosteroid signaling and cell elongation. *Frontiers in Plant Science*, *8*, 2225.

Zhou, J., Wang, Z., Wang, X., Li, X., Zhang, Z., Fan, B., Zhu, C., & Chen, Z. (2018). Dicot-specific ATG8-interacting ATI3 proteins interact with conserved UBAC2 proteins and play critical roles in plant stress responses. *Autophagy*, *14*(3), 487–504.

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., & van Wijk, K. J. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PloS One*, *3*(4), e1994.

SUPPLEMENTARY DATA

TABLES

Table 15: Gene centric table showing no. of pre-mat stop codons incurred in each gene.

No of PSCs	No of genes	List of genes
1	4741	
2	1876	
3	1124	
4	646	
5	433	
6	292	
7	203	
8	178	
9	99	
10	100	
11	57	
12	56	
13	31	AT1G49250,AT1G50750,AT1G51820,AT1G63880,AT1G65110, AT2G06845,AT2G14288,AT3G09790,AT3G19040,AT3G27600, AT3G29796,AT3G43260,AT3G43890,AT3G44250,AT3G44630, AT3G44670,AT3G44690,AT3G44770,AT3G45510,AT3G45720, AT3G46160,AT3G48770,AT4G15070,AT4G20920,AT5G01050, AT5G07570,AT5G27606,AT5G35715,AT5G41310,AT5G43610, AT5G46490
14	29	AT1G35820,AT1G35860,AT1G41820,AT1G44740,AT1G47660, AT1G49920,AT1G58520,AT2G05360,AT2G07310,AT2G07750, AT2G13450,AT2G27120,AT3G25460,AT3G25510,AT3G27040, AT3G27680,AT3G32960,AT3G44400,AT3G46470,AT3G47130, AT4G09740,AT4G10560,AT4G12330,AT4G16890,AT5G01150, AT5G03360,AT5G28090,AT5G28780,AT5G35604
15	24	AT1G31540,AT1G35150,AT1G35850,AT1G40390,AT1G43950, AT1G60130,AT1G61190,AT1G63870,AT1G65850,AT2G02490, AT2G07190,AT2G24340,AT3G17400,AT3G42550,AT4G05360, AT4G09360,AT4G13880,AT4G14390,AT4G14670,AT4G16960, AT4G23370,AT5G28823,AT5G33393,AT5G42260
16	17	AT1G32140,AT1G60930,AT1G65200,AT1G65990,AT1G66235, AT1G69550,AT2G07760,AT3G16820,AT3G28870,AT3G29830, AT3G45940,AT4G03935,AT4G08593,AT4G09775,AT5G28730, AT5G41740,AT5G43240
17	8	AT1G58602,AT2G01050,AT2G06541,AT2G13510,AT2G15042, AT3G29255,AT4G13760,AT5G35230
18	19	AT1G37113,AT1G41920,AT1G51520,AT1G52940,AT1G59780, AT2G05350,AT2G16040,AT2G18130,AT3G29450,AT3G42770, AT3G43153,AT3G44780,AT4G05080,AT4G10200,AT4G16250, AT4G16920,AT5G26580,AT5G33406,AT5G43740
19	8	AT1G36970,AT1G37020,AT1G59453,AT2G05642,AT2G12900, AT4G03740,AT4G11540,AT5G28190
20	2	AT2G06500,AT4G29090
21	9	AT1G71320,AT2G04810,AT2G14000,AT2G22440,AT3G29638, AT3G30200,AT3G31950,AT3G43470,AT5G32613
22	5	AT1G58390,AT2G15710,AT4G08097,AT5G38190,AT5G42905

23	3	AT1G59620,AT3G29080,AT5G34870
24	1	AT1G22000
25	2	AT1G43730,AT3G30520
26	4	AT2G10260,AT3G30820,AT4G03580,AT4G06526
27	4	AT1G28180,AT1G43722,AT1G46696,AT3G29750
28	7	AT1G34170,AT1G47300,AT1G51172,AT2G13500,AT3G45800, AT3G46120,AT4G13610
29	1	AT2G15110
30	3	AT2G07240,AT3G42870,AT3G46800
31	1	AT2G11010
32	3	AT2G15420,AT4G06688,AT4G18150
33	4	AT1G20400,AT3G30230,AT3G43160,AT4G03830
34	1	AT3G32904
36	1	AT1G43760
40	1	AT5G32590
45	2	AT3G30770,AT5G39770
46	1	AT3G42060
49	1	AT3G43148
54	1	AT3G42723
59	1	AT2G10440