COMPUTATIONAL ANALYSIS OF PEANUT POD FILLING BASED ON COMPARATIVE ANALYSIS OF GENOTYPES

by

CHANDLER MITCHELL SPRUEILL

(Under the Direction of Peggy Ozias-Akins)

ABSTRACT

Cultivated peanut (*Arachis hypogaea* L.) is grown throughout the world for its nutritious subterranean seeds. To advance our knowledge of this crop an RNA-seq experiment was carried out which compares three pod tissues over four reproductive stages in two genotypes. These two genotypes, Tifrunner and NC 3033, represent divergent phenotypes for pod filling of peanut. This phenotype presents as the space occupied by mature peanut seeds within the pod and is an important part of peanut yield. Computational analysis of differential gene expression unearthed a list of 294 genes differentially expressed across one or more factor levels. Further analysis of these expression differences using machine learning, quantitative trait loci comparisons, and ortholog searches of *Arabidopsis thaliana* resulted in a potential interaction network holding keys to increased control of the pod filling phenotype for use by peanut breeders. However, experimental validation of the candidate genes is recommended.

INDEX WORDS: Computational Biology, RNA-seq, Cultivated Peanut, Machine Learning, Gene Expression

COMPUTATIONAL ANALYSIS OF PEANUT POD FILLING BASED ON COMPARATIVE ANALYSIS OF GENOTYPES

by

CHANDLER MITCHELL SPRUEILL

B.S., Oregon State University, 2019

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2021

COMPUTATIONAL ANALYSIS OF PEANUT POD FILLING BASED ON COMPARATIVE ANALYSIS OF GENOTYPES

by

CHANDLER MITCHELL SPRUEILL

Major Professor:	Peggy Ozias-Akins
Committee:	Scott Jackson
	Jason Wallace

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2021

DEDICATION

To my supportive parents Wade and Michelle, to my curious and intelligent brother Henry, to my lovely partner Amber, and to all the friends and family who have supported me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

One of the major challenges facing the larger scientific and agricultural institutions is the paradoxical problem of increasing food production but in a sustainable manner to provide adequate food for the greatly expanding number of humans on the planet. Simultaneously, on a global scale, we need to significantly increase the amount of food produced annually while not significantly increasing the amount of land farmed. The simple solution to this problem is to create crops that yield more consumable materials per unit farmed (FAO, 2017). While simple to state, this solution has an arduous path. Since the beginning of the 20th century, crop production has exploded due in no small part to advances in many different scientific and mathematical fields. The best example of this is the Green Revolution of the 1950s and 1960s. Due to the implementation of new technologies and agricultural development, many world regions saw up to double the crop yield of the previous decades (Hazell, 2009). The Green Revolution combined many different aspects of agriculture, from better pest and soil management to creating better, more uniform plant cultivars (Glaeser, 2011). No single scientific branch could have created such a successful campaign. Therefore, it appears that the best way to find a solution like that of the Green Revolution is to develop a detailed and intricate understanding of crops to then to integrate this knowledge into practical applications.

While it is impossible to miss the massive success of the Green Revolution, critics have justifiably noted some critical cultural and environmental problems both created by the Green Revolution and unsolved by it. Significant difficulty has been seen in implementing Green

Revolution strategies in countries without a robust governmental system. Farmers need government investments through subsidies or loan programs to create yield gains seen through the expensive management processes and seed programs. Widespread adoption of agro-chemicals and mono-cropping endeavors had unforeseen environmental impacts. These issues primarily come from misuse of chemicals; however, this does not change that they present a severe problem (Glaeser, 2011).

Nevertheless, since the 1960s our understanding of biology has grown immensely with the proliferation of computational power, decentralization of information through the internet, and accessibility of large-scale lab equipment. Therefore, the baseline and contributing factors for the new solution are much broader. Genetics plays a huge role in this, primarily through exploring a crop's inner workings. While cliché, there is much truth in calling DNA the blueprint of an organism. Blueprints for a car, a cellphone, or any other machine are the most fundamental way to examine them, diagnose problems, and create valuable modifications. Apply this same idea to crops; through their genetic blueprint, we can analyze why crops behave the way they do and then modify them based on this research to create outcomes superior to anything known before. Again, while simply stated, this solution is far from being trivial. The handful of nucleotides that control a trait of interest are impossible to find without practical guides or evidence. Using tools such as trait mapping to find QTL or regions that may harbor specific traits works well but does not give the resolution necessary to find the genes controlling a trait (Frary et al., 2000). Therefore, the integration of more data is necessary. Processes that determine gene expression are one such data source. Genes associated with the trait of interest are located under the previously identified loci and can be found using computational tools (Carter et al., 2001). However, location does not necessitate function in a pathway. Therefore, functional studies must

be conducted to confirm candidate gene identity. The modern method of functionally analyzing a gene uses gene expression analyses, primarily RNA-seq, which provides datasets containing all transcribed genes from a tissue sample (Chettoor et al., 2014). While the presence of mRNA does not automatically mean the presence of translated proteins, it does give a good indication of which genes are being activated—giving overall a better idea of the basic mechanisms behind a pathway.

This basic idea was used to guide this evaluation of cultivated peanuts. The pod filling phenotype present in some market types of peanuts was analyzed using a computational examination of an RNA-seq experiment in conjunction with previously published knowledge on pod traits (Chavarro et al., 2020). Advanced analysis of expression networks and pathways were evaluated by machine learning on the normalized expression levels of genes found to be differentially expressed between and within analytical factors. This combination analysis provides a list of candidate genes from which further analysis can be undertaken to validate their interaction with the studied phenotype experimentally. If they are influential in the pathway, this phenotype can be exploited to create new cultivars with more desirable characteristics (Bhullar et al., 2010).

Introduction to Peanut

Cultivated peanut (*Arachis hypogaea* L.) is a leguminous grain and oil crop grown worldwide for its highly nutritious seed. Peanut seeds contain many different oils, vitamins, proteins, and organic compounds that, in combination, make peanuts a healthy food. Peanut consumption has been negatively correlated with significant human diseases such as obesity and some forms of cancer (Arya et al., 2016; Toomer, 2018). In comparisons between peanuts and other culinary nuts and legumes, peanuts have competitive levels of protein and oils, placing

them above most nuts in terms of protein and above soybeans in oil content (Venkatachalam and Sathe, 2006; Bonku and Yu, 2020). The components of peanut protein are all 20 amino acids with some variability in methionine and cysteine content. Oil composition varies based on cultivar and breeding heritage; however, large amounts of monounsaturated and polyunsaturated fats such as those in oleic and linoleic acids contribute to a healthier fat content than some comparable foods (Barbour et al., 2015; Arya et al., 2016). Variability in oil contents stems from high-oleic peanut lines, which have been bred to have a larger oleic to linoleic acid ratio. These varieties intend to increase monounsaturated fats providing a slightly healthier oil, but more importantly, a peanut with far greater shelf stability (Jung et al., 2000; Bera et al., 2019). The micronutrient composition of peanuts has high amounts of both water- and fat-soluble vitamins. These are described as either having greater or comparable levels to those found in similar foods. Particular attention is paid to Vitamins B and E, which are essential for animal diets. Peanuts contain various phytochemicals with some medicinal benefits, such as polyphenols, isoflavones, phytosterols, and resveratrol. To summarize, peanuts offer a complex and nutritious food source to animals, leading to human cultivation (Arya et al., 2016).

Peanuts were first domesticated somewhere in the plateaus of central South America, approximately 9,400 years ago (Kalberer and Belamkar, 2014; Bertioli et al., 2019). Cultivated tetraploid peanut was derived from a cross between two diploid species and a spontaneous chromosome doubling associated temporally with the beginnings of domestication. While making it potentially more enticing to humans, this genetic outcome also isolated tetraploid peanut from producing viable crosses with other species of wild peanuts in the area (Bertioli et al., 2019). The role of peanuts in traditional American societies has been debated. The traditional belief was that peanuts were just used as a staple crop for the people living in modern-day Peru

and Bolivia. However, some recent studies speculated that peanuts have held prestige in those societies. Anthropological assessments of food residues from both high- and low-class food preparation areas have shown differential presence favoring peanuts in high-class settings. Paired with archaeological evidence of peanuts incorporated with ceremonial jewelry and iconography, this leads to some evidence that these people's peanuts held ritual power. The reasons for this are pure speculation but may be due to supposed medicinal and other qualities associated with peanut fruit (Masur et al., 2018).

Currently, peanuts are grown to an amount numbering 49 million metric tons worldwide, with more than a third taking place in China followed by India, then Nigeria, Sudan, and the United States (FAO, 2019). In the United States, peanuts are separated into four general market classes based on morphology, seed qualities, and growing region. Runner and Virginia-type peanuts are primarily grown in a region running along the Southern Atlantic coast of the United States from Florida to Virginia. Valencia and Spanish-type peanuts are grown further west from Texas and Oklahoma to New Mexico. Runner-type peanuts are mainly used for processed peanut goods such as peanut butter. Virginia types are mainly grown for whole peanut consumption, either in shell or not. Spanish-type peanuts are noted to have a sweeter taste and are therefore used in candies. Valencia-type peanuts are grown for boiling (Klevorn et al., 2019).

Cultivated peanut comes from a cross between two wild progenitor species, *Arachis duranensis* and *Arachis ipaënsis*, followed by a polyploidization event creating a sexually viable allotetraploid. This process reproductively isolated *A. hypogaea* from most other *Arachis* species in South America, which are diploid (Kalberer and Belamkar, 2014; Bertioli et al., 2016, 2019). *Arachis monticola* is an exception to this, being another tetraploid *Arachis*. However, it appears to be either a traditional wild form of *A. hypogaea* or escapes from *A. hypogaea* cultivation

(Krapovickas et al., 2007). The *A. hypogaea* genome is 2,552 Mb divided into 20 chromosomes, with the genome separable into AA and BB sub-genomes represented respectively by *A. duranensis* and *A. ipaënsis*. Within this genome, there are 67,124 genes encoding 84,714 transcripts. Due to similarities between the sub-genomes, chromosomal rearrangements have been possible and are detectable computationally. In seven cases in the *Arachis hypogaea* genome, chromosome ends have undergone homoeologous recombination resulting in genome segments with four alleles from one sub-genome instead of two from each sub-genome. This process is seen both on a large scale with chromosome ends and with individual alleles. Homoeologous recombination is theorized to be the underlying mechanism. Although early *A. hypogaea* were reproductively isolated from other *Arachis* species, there was enough genetic variation within the species that genetic differences could be phenotypically detected over time (Bertioli et al., 2019).

Interestingly, many other crops, like peanut, are also polyploid. It is significant that a larger percentage of domesticated plants are polyploid compared with undomesticated species. This occurrence may imply that there are features in polyploid plants that led them to be selected by early humans. One such feature may be the slight heterosis observed in polyploid peanut species (Hilu, 1993; Renny-Byfield and Wendel, 2014). However, there is evidence of cultivation in many *Arachis* species (Krapovickas et al., 2007). However, it was *Arachis hypogaea* that was chosen to be the species cultivated on a large scale.

Arachis hypogaea has two named subspecies subsp. *hypogaea* and subsp. *fastigiata*. Most strikingly, the difference between these two is the presence of flowers on the mainstem of the plant. Both have been cultivated by humans, although in different regions of South America, with subsp. *hypogaea* being centralized near Bolivia and subsp. *fastigiata* focused near Peru.

Within subsp. *hypogaea* two varieties have been named var. *hypogaea* and var. *hirsuta*, differentiated by the distribution of hairs on the abaxial side of the leaf. Both runner-type and Virginia-type peanuts are within var. *hypogaea*, while var. *hirsuta* is classified as a Peruvian-runner-type, not typically grown in the United States. The other subspecies branch, subsp. *fastigiata,* contains var. *fastigiata*, var. *peruviana*, var. *aequatoriana*, and var. *vulgaris*. The varieties *peruviana* and *aequatoriana* are rarely, if ever, grown in the United States. The varieties *fastigiata* and *vulgaris* correspond to the Valencia and Spanish market type peanuts (Krapovickas et al., 2007).

Arachis hypogaea superficially looks typical of a legume; however, it harbors distinctions to differentiate it physiologically from other leguminous crops. Above ground, peanuts grow with a main stem and a differential number of horizontally developing plant shoots known as runners. Leaves are alternately produced from nodes, with four leaflets pinnately oriented on each pedicel. Inflorescences and flowers are produced from leaf axils, depending on the peanut type. In some instances, this may include the main upright stem. These flowers are yellow to orange with a conspicuous banner petal above two wings enclosing the keel petal and reproductive organs. These flowers are frequently self-pollinating but do cross-pollinate naturally in approximately 1-6% of total flowers. A key feature of peanut flowers is that after fertilization, the production of a peg, an extension from the bottom of the ovary, begins to grow towards the ground. Growth of the peg continues underground, where pod development begins. Mature fruits of peanut are easily separable into three main sections. These are the outer shell, the papery inner seed coat, and the embryo itself. Peanut, like many other legumes, produces root nodules and makes symbiotic relations with nitrogen-fixing bacteria. These are borne on the true roots of the plant, not on the peg or the fruit. The truly distinguishing feature of peanuts is their

subterranean fruits. A common way of describing this phenomenon is by "aerial flower and subterranean fruit," perfectly encapsulating the process of peanut geocarpy (Smith, 1950).

Growing peanuts in the southeastern United States is a fight against the pathogens that thrive in the hot, humid climate. While aerial pathogens are of concern for peanut management, the subterranean growth habit of the fruits introduces some other possible methods of disease attack on the fruit. Significant work has been done to create high-yielding cultivars while also having solid resistance to the pathogens that will inevitably infect fields (Chappell et al., 2020). One such pathogen is Tomato Spotted Wilt Virus (TSWV) which causes spotted wilt disease in peanuts. This disease infects plants through viral transmission from thrips. The severity of the disease is variable, from chlorosis and ringspots to severe stunting of aerial plant features, including leaves and pegs. Infections can cause significant yield loss, with the most severe seen in Georgia as high as 12% of total yield lost to TSWV (Culbreath and Srinivasan, 2011). Management and cultivar selection can reduce the severity and incidence of TSWV infections, reducing the potential yield costs for growers (Chappell et al., 2020). Plant pathogenic nematodes are challenging to deal with from a management perspective. Therefore, genetic resistance can be a better approach to control yield losses due to field infestations. Although *Arachis hypogaea* does not have known resistances to *Meloidogyne arenaria* (Neal), root-knot nematode, wild relatives in the *Arachis* genus such as *A. cardenasii* and *A. stenosperma*, have been identified with disease resistance loci (Mitkowski and Abawi, 2003; Ballén-Taborda et al., 2019). Backcrossing schemes have been successfully implemented to incorporate disease resistance from wild backgrounds into elite lines (Mitkowski and Abawi, 2003; Chu et al., 2011). Fungal diseases of peanut are described as an "unavoidable union." Many fungal diseases include early and late leaf spots, southern stem rot, and *Cylindrocladium* black rot (Hammons et al.,

1981; Jackson, 1983). While frequent fungicide applications effectively reduce the severity of fungal pathogens (whose yield effects can be as high as 50%), the weather ultimately plays the most critical role in disease severity. Intrinsic resistance is in cultivars is preferable to management for it reduces time and chemicals costs to growers (Brown et al., 2005; Chappell et al., 2020).

This study will analyze cultivated peanut with the use of an RNA-seq experiment derived from two peanut cultivars NC 3033 and Tifrunner. The goal of this analysis is to provide insight into the genes and pathways controlling the pod filling phenotype present in Tifrunner but absent in NC 3033. The RNA-seq data will be processed in accordance with established norms in the field of comparative and functional genomics to find dissimilarity between expression levels of genes. These genes will reflect pod filling since the RNA-seq samples were derived from pod tissues during the reproductive stages where pod filling occurs. Differentially expressed gene sets will be further analyzed using machine learning techniques along with homology searches to effectively estimate function of genes in these pathways. Some external validation of these pathways and genes will be done by comparison to QTL known to affect pod filling (Chavarro et al., 2020). A review of biological and computational components of the project is presented in Chapter 2.

Peanut Pod Filling

Ultimately the goal of any plant breeding program is to create a better yielding cultivar than either parent. While not a simple process, it can be guided towards specific goals by first investigating how traits of interest are achieved. In many other crops, analysis of the seed filling developmental stages has led to better understandings of how energy is accumulated in fruits (Shiraiwa et al., 2004; Hajduch et al., 2010; Yin et al., 2020). Across developmental stages, gene

expression changes, creating new profiles of protein present, changing the size, shape, and nutrition of the seeds. This same idea can be studied in peanuts by analyzing peanut seed size and the pod volume they occupy. Pod-filling is a vital step in establishing yield for a peanut crop; during this stage the plant diverts nutrients and activates metabolic pathways to develop its seeds (Halvey et al., 1987; Gupta et al., 2014). Thus, this developmental stage is essential to produce high-quality peanuts. Analysis of these pathways in *Arabidopsis thaliana* has demonstrated that changing gene expression leads to new pathways to synthesize these compounds (Hajduch et al., 2010). Therefore, similar principles can be applied to the study of this phenomenon in peanuts. However, in peanuts, there are distinct differences amongst some cultivars that do not develop their seeds entirely, ultimately leading to peanut pods containing voids around the seeds (Chavarro et al., 2020).

This phenomenon is termed incomplete pod filling, it is an undesirable trait from the point of view of both growers and plant breeders. While a cultivar may have desirable characteristics, the incomplete pod-filling trait limits the potential success that the cultivar may have in a market (Figure 1.1). An example is NC 3033, a Virginia-type peanut, which, although visually the seeds are good size and the cultivar has high resistance to *Cylindrocladium* black-rot, it exhibits incomplete pod filling (Beute et al., 1976; Chavarro et al., 2020). Theoretically, if this cultivar did not possess this phenotype, it would be more useful in peanut development as a parental line. However, currently, the reduced pod-filling phenotype reduces the cultivar's practical yield. Therefore, understanding the mechanisms leading to complete pod-filling and, therefore, more significant yield potential is vital for the continued development of peanuts.

Previously, a recombinant inbred line (RIL) population for discovering QTL associated with pod filling was developed from a cross between NC 3033 and Tifrunner (Holbrook et al.,

2013; Chavarro et al., 2020). Tifrunner is a runner-type peanut with market-ready pod phenotypes such as complete pod filling (Holbrook and Culbreath, 2007). From this QTL analysis, forty-nine total QTL were identified to be associated with pod filling measured phenotypes. These measurements were for seed size index, kernel percentage, seed weight, pod weight, single-kernel, double-kernel, pod area, and pod density. Phenotypic data from this study was collected across three years in 2013 from 134 $F_{6:8}$, 2014 from 152 $F_{6:9}$, and in 2015 from 160 $F_{6:10}$ lines. The phenotyping method varied by trait but was performed for all lines. Pod weight, seed weight, pod area, and pod density were measured with a sample of 250g of pods. Genetic maps were generated from the phenotyped portion of the RIL population plus an additional 165 $F_{6:7}$ RILs that were genotyped using an Affymetrix Axiom_Arachis SNP array and 111 SSR markers. From these, 1,998 SNPs and 100 SSRs were used for QTL analysis. Many QTL discovered in this analysis were clustered either with direct overlap or close in proximity. Assessing the locations of these QTL with other QTL published for seed and pod traits showed that eleven QTL regions from six studies were closely located with QTL from this population (Chavarro et al., 2020).

Genotypes Under Evaluation

Tifrunner is a runner-type peanut bred for cultivation in the southeastern United States, released in 2007 by the USDA Agricultural Research Service and the University of Georgia to introduce Tomato Spotted Wilt Virus (TSWV) into a new peanut line. Originally collected from a market in Porto Alegre, Brazil, in the early 1950s, PI 203396, one parent of Tifrunner, has been used to bring resistance to late leaf spot, southern stem rot, and TSWV into numerous released cultivars. The other parent is F439-16-10-3, one of the component lines comprising the 'Florunner' cultivar. When Tifrunner was bred, it was competitive with other cultivars in seed

size with 62g/100 seeds and exceeded other commonly grown peanuts in TSWV and leaf spot disease resistances. Tifrunner is important as a cultivar because it is the genotype selected as the first published tetraploid peanut genome sequence. It grows with an erect central stem with the spreading growth habit typical of other runner-type cultivars (Holbrook and Culbreath, 2007).

NC 3033 is a Virginia-type peanut developed in the 1950s and released in 1976 from the North Carolina Agricultural Experiment Station. NC 3033 was created from a cross between 'Ga 207-7' and 'A48' for southern stem rot resistance. Some southern stem rot resistance was observed in addition to high levels of resistance to *Cylindrocladium* black rot (CBR) (Hammons et al., 1981). NC 3033 was selected for this study because it has many characteristics that differentiate Virginia-type peanuts from runner-types. Its growth habit is dissimilar to that of Tifrunner growing in semi-dwarf bunches, low homogeneity between pods, however larger individual seeds. (Beute et al., 1976). Principally incomplete pod filling is characteristic of this line, where the bulk of the peanut seeds do not fill the shell cavity, leaving space within (Chavarro et al., 2020). This void reduces the yield potential of cultivars possessing this characteristic. The selection of these cultivars for analysis of peanut pod filling is based on the differences present in their phenotypes and previous QTL analysis using a RIL population derived from an NC 3033 x Tifrunner cross. In the most apparent sense, Tifrunner exhibits the pod filling phenotype. The seeds of peanut fruits almost wholly occupy the space within the pod. By comparison, NC 3033 seeds occupy a smaller proportion of the entire pod (Chavarro et al., 2020).

Peanut Reproductive Development and Pod Tissues

One of the most distinguishable traits of peanut plants is their geocarpy. Although flowers are borne from shoots above ground, fruit development occurs entirely below ground.

The entire reproductive pathway of peanut, from flowers to fruit, occurs in 9 stages, labeled from R1-R9. Each stage represents unique observable changes in the reproductive structures of peanuts from flower to fruit and is essential in understanding the output of a reproductive cycle in peanuts. Stages are defined as when the average state of a population of peanut plants meets a particular goal. Stage R1 is the beginning of bloom, marked with the opening of at least a single flower as the average in the population. The next stage is marked by the development of a peg in half of the population. The remaining stages R3-R9 are concerned with the development of pod tissues and fruit. R3 through R7 encompass all stages from the beginning of pod development to just before a harvest-ready crop of stage R8. R3 is the beginning of pod creation, marked with the penetration of the peg below ground and enlargement of the ovary tip. Full pod size is achieved in stage R4, where fruit are expanded to their final size. Stage R5 is marked by the discernable presence of cotyledons within the cut pod. R6 is the filling of the seed within the pod tissue, marking a time when the seed becomes more differentiable from the pericarp tissues. This effect is more pronounced in R7, beginning maturity, with color being developed on the interior parts of the pericarp. Here the seed continues to grow, with the seed approaching its final size. Full maturity is at stage R8, marked with the browning of the exterior of pods. This stage represents the completion of the development of fertile seeds. The remaining stage, R9, occurs with the desiccation of the peg and pods retaining their position in the soil, no longer attached to the mother plant. In evaluating the yield of peanut plants, the stages leading up to R8, full maturity, are essential. These stages, notably R3-R7, represent the accumulation of nutrients and the changing of genetics in the fruit leading to maturity. Therefore, studying the effects that contribute to fruit yield should be taken from this perspective (Boote, 1982; Shiraiwa et al., 2004; Gupta et al., 2014).

Dissecting a mature peanut pod naturally lends itself to classifying three main pieces of tissue. The first is the embryo, within the buff seed contained in the center of the pod varying in number from 1-4, each however fitting to a common form. The seed contains two large cotyledons completely encapsulating the plumule while leaving part of the radicle exposed from a small gap between halves. The second tissue is the papery seed coat surrounding the peanut seed, typically a brown or red color; it is easily removed from mature seeds with light force (Young and Schadel, 2004). The final tissue, the hull, formally known as the pericarp is composed of three sub-tissue layers. From the outside of the pericarp in, they are the exocarp, mesocarp, and endocarp. Each layer contains slightly different cell compositions lending each to have unique functions within the tissue. For instance the exocarp lacks the vascular bundles of the mesocarp and endocarp, while the endocarp in part, lacks the sclerified cells of the other sub-tissues on its inner-most surface (Gilman and Smith, 1977; Halliburton et al., 2015).

Gene Expression

Three main molecular stages exist in the pathway from the genotype to the phenotype of an organism. In order, they are from DNA, to RNA, to protein. This pathway is true for most known living organisms, and thus any understanding of the phenotype of an organism can be traced back to the genotype. Gene expression relies on the interactions of many molecules and pathways at each step influencing the effect of a gene. In this pathway, the first step is the regulation of gene expression. This process controls the transcription of a DNA gene sequence into RNA by preventing or allowing the successful binding and operation of RNA polymerase. One primary way this occurs is by binding various proteins and molecules, called transcription factors, to DNA that either encourage or block transcription. This process is critical in regulating genes whose products are situational, as opposed to those essential for homeostasis and therefore

constitutively expressed. Regulation of transcription can be a complicated array of possible outcomes based on the numbers and types of transcription factors, acting positively and negatively on transcription. Once the required criteria for transcription have been met, RNA polymerase transcribes the DNA 5' to 3', creating a single strand of RNA. This RNA is complementary to the DNA it is transcribed from with the added difference of not possessing the nucleotide Thymine, instead using Uracil (Elson, 1965). The new RNA strand is termed pre-mRNA, and it consists of four main parts, a 5' and a 3' untranslated region (UTR) and a combination of introns and exons. Introns are removed by a process known as splicing and the exons and UTR are rejoined in the same order. The UTR are processed by adding a cap to the 5' end and a string of poly-adenine to the 3' end (Buccitelli and Selbach, 2020).

Regulation of this RNA strand can occur in a few places. The first of these is by changing the set of introns spliced from a transcript. This process termed alternate splicing creates unique proteins from the same gene. Select motifs contained within UTR can act as protein targets, either contributing to a change in localization of the mRNA or tagging for degradation. Post-modification, the mRNA is exported from the nucleus and begins the translation process, changing from the code of nucleic acids to that of amino acids. Ribosomes, organelles comprised of both rRNA and protein, capture the 5' end of the mRNA sequence and, after the 5' UTR, begin reading the RNA code in triplicate nucleotide segments called codons. Each codon corresponds to a matching complementary codon present on a tRNA, which carries each new amino acid to the ribosome. As each codon is read through, more amino acids are added to the newly forming polypeptide chain. Chemical properties of the cell environment, other proteins, and the amino acid composition of this chain influence the way it bends and twists as it forms. After translation is complete, the protein self assembles into a conformation that generally results in the lowest

strain on the molecule. This assembly occurs due to the chemical properties of each amino acid restricting the orientation parameters of others (Buccitelli and Selbach, 2020).

The interface between the genotype and the phenotype of an organism is in its mRNA content. The gene expression process culminates in the aggregation of proteins in a cell. Coordinated gene expression within groups of cells gives rise to tissues. These tissues then react to biotic and abiotic stresses by altering gene expression systems, giving rise to different proteins. These dictate what the cell needs for survival. They are also dynamically operating, and any attempt to isolate them is a snapshot into the process they are trying to influence (Benfey and Mitchell-Olds, 2008). For sessile organisms, like plants, gene expression is one of the principal ways to adapt to a changing environment. For instance, a long summer drought may open pathways for the plant to begin transcription and translation of genes whose proteins allow a cell to survive under such conditions. However, in a more general sense, changes in gene expression are one of the reasons there is diversity even within species. Activation of certain genes at specific times can profoundly influence the organization and response of an individual (Chinnusamy et al., 2007).

Comparative Genomics

Evolution as a concept allows for the comparison of species through the lens of changes since their divergence. Before the widespread use of genetic sequences, this was performed to visually identify structures and traits that are similar and dissimilar in each species. While flawed in some ways, the fundamental idea is that homology can be found in the underlying structure or the ontology of a structure. Applying this principle to genetic sequences can produce more detailed predictions of evolutionary relationships than other comparative studies. The genetic lens can be fine-tuned from large-scale analysis of whole chromosomes to small-scale

interpretation of changes in individual nucleotides. Each different genetic lens, tell slightly different stories about the evolution of species, however they all fit under the umbrella of comparative genomics (Koonin et al., 2000).

Broadly, the types of studies done in comparative genomics fit into two categories. These would be studies finding commonalities between species and those finding differences. The intention behind finding commonalities operates on the assumption that conservation indicates function. While experimental validation is necessary to see function directly, usually in expression for genes and subsequently phenotypes, it is still valuable to computationally search for these genes to narrow the search space for the expensive laboratory assays. Where differences lie also have distinct implications for the evolution of organisms. For example, take two closely related species where one has a trait for drought resistance that is not present in the other. Comparative genome analysis can indicate the dissimilar genomic regions between the two species and narrow the search space for alleles or novel genes and pathways that contribute to the new trait. There is also a third type of study which blends the similarity-based approach with the difference-based approach. Here differences would be searched in the regions that are conserved between species. This protocol can lead to discovering new alleles, differences in expression patterns, or novel uses of the same pathway (Paterson et al., 2000).

One of the most prominent uses of comparative genomics is creating phylogenetic trees, which illustrate the potential uses of the field. Before gene sequencing, relationships between organisms were estimated based on observed phenotypes. The most accessible were an organism's physical characters, both internal and external; however, fossils were also used. While for some comparisons, this could give a helpful understanding of the evolutionary paths of these organisms, it is easy to be misled by similarities without shared feature ancestry. This

process was significantly improved with DNA sequencing technologies and algorithms for estimating nucleotide substitution rates. Now phylogenetic relationships could be based on divergence in shared sequence. Take a gene held in common between two species, examining the differences present in the nucleotide sequence paired with knowledge of how fast genetic sequences change allows for the extrapolation of time since species divergence. Scaled up, this concept can be applied to large numbers of sequences or organisms, which, when displayed in a dendrogram, tell the story of evolution for those lineages (Pearson, 2013).

Comparative genomics can also be applied to sequences for which there is no knowledge. Take a novel sequence derived from an organism never studied. While there are intelligent algorithms for discovering genetic features, such as genes, determining the function of undescribed sequences is difficult without expensive physical assays. However, matching these new sequences to ones whose functions have already been determined gives a shortcut to determining function. This functional annotation method relies on the idea that although genetic sequences change, the underlying functionality of the genes tends to change at a slower rate. Meaning changes in function are due to the conglomeration of changes in DNA, so individual changes will not disrupt function. The workflow of analysis such as this is outlined by (Pearson, 2013).

Simply put, there are two steps, finding homologs and then extracting functional information from one gene. Homologous searches can be performed in two manners; one uses whole-genome segment alignment to find regions that share a direct common ancestry, so-called orthologous regions. Genes in each region will be directly descended from a common ancestor, but most importantly, it operates on a genome-scale. Even in cases then where genes or chromosome segments are translocated computationally, they can be associated. The other

operates on individual genes or small genome segments. Here, an alignment algorithm such as BLAST will perform database searches to find genes matching the residues of interest (Pearson, 2013). The advantage to a method like this is that it is computationally lightweight. A database can contain sequence from tens of thousands of different organisms, each of which can be sorted through in at most a few minutes, although if many genes are being assayed, this method can be cumbersome (Wall et al., 2010). The second half of the process, assigning function, relies on the match from the first step and a functional language such as Gene Ontology (GO) to give an accurate understanding of the nature of the function. GO uses three broad categories to divide terminology for gene functions: Biological Products, Molecular Functions, and Cellular Components (Ashburner et al., 2000).

Each contains its own set of terms hosted at graph nodes that become more specific as one moves through GO networks. GO networks are based on directed acyclic graphs (DAGs). These graphs intend to move from node to node without ever forming loops. GO descriptions need to become more specific as the graph is navigated. Logically, each node for a GO annotation limits the number of nodes a path can move to next because of this structure. This structure is necessary because GO describes actual functionality. Two nodes cannot be joined if the functional term of the second node is not nested within the first. Each GO annotation is based on hierarchies of data precision. At the top is experimentally validated functionality, considered to be the most reliable source. At the bottom of the hierarchy are computationally predicted functions, which may not be as precise as others, although they are not intrinsically poor in quality. They are derived from computational comparisons of gene sequences to other genes of known function. They are tagged as "putative" or "predicted" functions, which still give a reasonable indication of gene function (Ashburner et al., 2000).

Functional Genomics and Transcriptomics

Genes control the phenotype of an organism, concerted changes in expression alter individual cells. Organized gene expression contributes to developmental responses, leading to tissues, regions of similar gene expression patterns. The collection of tissues in an organism gives rise to the entire individual. Each tissue serves a unique function and allows adaptation to the environment. The fundamental force driving this process is gene expression; therefore, the study of gene expression explains how an organism is formed and adapts to changes in its surroundings, an essential function of plants with their sessile nature (Benfey and Mitchell-Olds, 2008). Functional genomics and transcriptomics are both methods of interpreting this information. Although they overlap significantly, they also differ in the precise scope of their realms. Functional genomics is a field derived from the study of genes in their role in contributing to the function of an organism or pathway. As the genomics part implies, it is concerned mainly with large-scale analysis of tissues or clusters of expressed genes. It differs from early work with genes that primarily focused on smaller scale analysis, expressing a single gene or small gene family. Expression analysis is performed mostly from the RNA level, meaning that mRNA transcripts are used to determine expression and then the basis of further functional understanding (Cano-Gamez and Trynka, 2020). Transcriptomics is the technical, biological, and mathematical field describing how RNA transcripts are analyzed and interpreted. While the function of genes is undoubtedly in this realm, it also encompasses techniques for determining the probable set of expressed genes (Nayak and Hasija, 2021). Their differences lie in the subset of genes they examine. Functional genomics focuses on the functional aspect of expressed genes while transcriptomics on the entire set of transcripts.

Their similarities cause them to overlap in the technologies used in the evaluation of genes. Both fields came about in the late 20th century and fully developed in the early 21st with the increasing complexity of lab-based quantification of genetic transcripts. The first technology discussed is relevant mainly to transcriptomics. Expressed sequence tags (ESTs) are sequenced versions of mRNA transcripts that had been reverse transcribed into cDNA. mRNA is the liaison between DNA and protein and therefore is an essential part of gene expression. EST libraries contain sets of most genetic transcripts in a tissue at the time of sampling. They are small fragments of genes that rely on an overlapping set of an entire transcript to be obtained to elucidate the mRNA code. They are error-prone from two angles. The first is that there tend to be errors in sequencing the fragments typically around its ends which leaves only about half of the fragment as useful sequence. The second is that it is hard to make quantitative measures from a library. Genes with relatively low expression can be under-represented in the final library and, therefore, sequenced data. Extrapolation cannot be made to compare expression levels of genes for functional genomics and one cannot quantitatively compare gene expression differences across times or tissues. This shortcoming makes assigning functionality to genes much more difficult. However, putative function can be predicted using comparative database searches if related genes are known. What ESTs provide is a good starting set of expressed genes in an organism (Parkinson and Blaxter, 2009).

Serial analysis of gene expression (SAGE) improves upon the goals of ESTs and provides some indication of relative transcript abundance. SAGE's advantages occur through the in-vitro concatenation of individual sequence tags into much longer molecules sequenced using Sanger sequencing. Software is then used to isolate each sub-tag of the longer sequence and produces more reliable tag counts than ESTs. Therefore, some comparisons as to the relative

abundance of gene transcripts can be estimated and compared (Yamamoto et al., 2001). Microarrays can also serve to estimate relative abundance based on nucleotide probes arranged on glass slides or chips. These probes are based on a known genetic sequence and require prior knowledge of the genome of the organism of interest or the sequences of interest. There are a few different platforms for microarrays, but they all tend to operate on the principle that transcripts with higher abundance will recruit more of the fluorescent molecules from the solution and therefore have quantifiably higher light intensity at their probe. Comparisons can be made between light intensity at probes by converting intensity to a mathematical scale to which normalization metrics can be applied (Hoheisel, 2006; Mantione et al., 2014).

The final method of transcript quantification and functional elucidation is RNA-sequencing. Although it is the newest technology, it has quickly become quite prolific, partly due to the diffusion of high-throughput sequencing technologies and the computational capabilities to handle the large-scale outputs of the analysis. Like the other methods, mRNA transcripts are first isolated from samples. However, the samples here can be as small as a single cell. The mRNA is fragmented then transformed into cDNA; this cDNA is PCR amplified then sequenced using high throughput sequencing technologies. Typically, an Illumina-based approach is taken due to low error rates compounded with very high total throughput. Outputs from sequencing are the cDNA fragments transformed into digital storage (Wang et al., 2009; Illumina, 2011; Mantione et al., 2014; Conesa et al., 2016).

RNA-seq data analysis begins by mapping cleaned reads to regions in the genome. The fragment size from sequencers is typically less than the length of a gene. Therefore, special consideration must be taken to ensure each read fragment is being assigned to the correct gene. This process is handled by software known as read mappers, algorithms that seek to effectively

assemble fragments into gene constructs then find the closest sequence match. When working with a small number of sequences, like the long sequences given by Sanger sequencing, little consideration is needed to control a program's memory usage when aligning the read to the genome. However, with the NGS techniques used in RNA-seq, it would be impossible to load the gigabytes of sequences typical of an experiment into memory for alignment (Conesa et al., 2016). Adaptation of the Burrows-Wheeler Transform (BWT) to sequence alignment helped remediate this problem. The BWT is a compression algorithm that, simply put, converts text strings into alternative forms which can be more compactly compressed. BWT usually involves swapping characters out for multiples of identical characters that take up fewer bits than the original character when compressed (Li and Durbin, 2009, 2010). Two programs developed with the BWT as their central focus are still in use today. These are Bowtie and BWA (Langmead et al., 2009; Li and Durbin, 2009).

While useful for some protocols, they are inadequate for others as they cannot handle the added challenge of gaps in sequence alignments brought about by splice variants. This problem gave rise to many new algorithms and programs that more effectively detangle RNA-seq reads while also identifying gene splice variants (Conesa et al., 2016). One such program, STAR, is capable of mapping even with gaps produced by intron splicing. STAR's algorithm is performed in two main steps, seed searching and clustering/stitching/scoring gene alignments. The seed searching step takes each sequenced read and finds the location to which it most closely maps. It should be noted that this region may be in a location where the entirety of the read does not map, leaving a tail of an un-mapped sequence. The second half of the seed step remedies this by taking this tail and aligning it to its maximum scored match. The clustering, stitching, and scoring step then takes the mapped reads and stitches them back together if mapped separately, ensuring that

both sides of the read are mapped concordantly. A score is then determined for each read based on sequence mismatches and introduced gaps. This score is used for evaluating the quality of mapped reads. The dataset of mapped reads is useful for downstream analysis of RNA-seq data for transcriptomics or functional genomics analysis. It indicates which genes and splice variants have been detected while also allowing for comparisons between mapped reads to genes both within and between RNA-seq samples (Dobin et al., 2013).

Computational Biology

As a field, computational biology only exists to allow for the effective processing of large-scale or complicated experimental designs as used in biology. To simply define computational biology, it is the intersection of computer science, statistics, mathematics, and biology that gives rise to algorithms for elucidating valuable data within large experimental datasets. The development of this field then depends on discoveries in each of those fields with special consideration needed for biology and computer science. New protocols for wet lab-based assays of biological features, such as the family of sequencing techniques, create the need for new computational algorithms and programs to deal with the unique quirks of these protocols. This process on a large scale would not be possible without continuously growing computational power and reducing costs (Noble, 2002). Moore's law is frequently cited in this regard, in which it is estimated that approximately every two years, there is a doubling in the computational power purchasable with a single United States dollar. While this has been true in the past, recently computational limits have been reached, with a plateau predicted soon (Moore, 1965). Mathematics and statistics are incorporated into computational biology in applied fashions, maneuvering the parameters of established algorithms or ideas to conform to the problem sets of biology. These problem sets can be quite extensive and draw from highly variable data, for

instance, computer vision and sequence interpretation both fit under the umbrella of computational biology, although they are very distinct problems.

It may be counter-intuitive, but computational biology's many difficulties are in determining frameworks for describing biological phenomena. While creating new algorithms and pipelines is quite difficult, the exact pathways of coming to meaningful conclusions from biological data are not necessarily dependent on that process. Many computational pipelines then are built to accept many different data sources and include paths for the elucidation of various end states (Nussinov, 2015). They also often include simple ways to share this data with other researchers easily and consistently. For instance, say a research group has created a genome sequence for a novel organism. Depending on the organism and the specific sequencer used, an optimal pipeline can be found and used. They need this sequence for their experiments but are willing to share it with other groups working on the same species. A lab doing protein work in this species wants the genome sequence to help identify genes involved in a specific pathway. Sequences can be provided in the consistent and widely accepted FASTA format, which programs have already been built around (Pearson and Lipman, 1988). This helps reduce confusion and save time from having to adapt from a proprietary format.

Recently, there has been a strong push in computational biology to apply machine learning and AI concepts to reduce some of the difficulties typically found when adapting more traditionally structured algorithms to biological data. This is because many machine learning techniques can have a more intricate understanding of the signals present in a dataset because of how they are made (Tarca et al., 2007). Usually, the development of an ML process or algorithm is directly tied to the data of interest. These algorithms adapt to the channels of data with which they are

trained. This allows them to fit the data in a way a human researcher could not find in a typical timescale (Jones, 2019).

The data analytics aspects of transcriptomics, comparative genomics, and functional genomics all fit into problems solvable by computational biology. While the computational power of personal computers has dramatically improved in recent decades, so has the computational complexity of problems. Many of these programs and pipelines are created to be run on a computational cluster computer that has exponentially more resources to draw from than a consumer-grade computer. Many developers then have begun to integrate their technologies into conglomerated web sources. Typically, they differentiate themselves based on a specific area of focus or other factors. To describe this, Galaxy, CoGe, and Ensembl will be compared to show differences and similarities typical of computational biology software. Galaxy is an open-source integration of genomics tools that simplifies their uses by moving their tools from a command line to a webpage-style system of boxes and forms. This removes a lot of the initial confusion of using a new bioinformatics tool that typically comes from having too technical manuals for casual users. To assist further with this, they host various file parsing tools to convert between commonly used data formats. While this might seem trivial, from experience, processes like this can be tedious to deploy effectively. The intention of the platform is for biomedical research though all domains of life are supported. Users can upload their data sources and store them in their repository or have an integrated connection to common internet-based data stores. Galaxy provides an integrated web source for running genomics-based analyses with many features that cater to inexperienced users and improve the quality of life of advanced ones. It is reaching a broad audience, pulling tools from many different domains (Afgan et al., 2018). CoGe, short for Comparative Genomics, shares many similarities with Galaxy while also being

much more refined in the overall scope of tools hosted. As its name suggests, the site focuses on tools to perform comparative analysis of genomes. They claim to hold over 55,000 genomes on the site while additionally allowing user-submitted ones to be uploaded. CoGe's primary tool is the genome browser, a way to visualize the pieces contained within a genome with relative simplicity due to its visual format. Tracks, data sources can be overlaid in the same region, directly showing overlap between two or more features. The other tools present are integrated similarly to Galaxy's, allowing for a step-by-step framework of pipelines from upload to output (Lyons and Freeling, 2008a; b). Ensembl operates similarly to CoGe in that it relies on a genome browser style of interface. How it differs is by focusing almost exclusively on tools and datasets for the evaluation of animals. Principally, comparison can be made between sequences to find SNPs or differential expression of genes within or between tissue types. Like the other tools, it allows user sequences to be uploaded and integrates data from their database (Howe et al., 2021). Each of these tools represents the self-imposed limitations but also the advantages for each platform. Galaxy provides a simple way to use many popular tools for someone without much computational experience. However, it does not possess some of the features present in the other two like a genome browser. Ensembl has chosen to operate on a narrow set of organisms and functions, which makes it advantageous to use in contexts relevant to that data set.

Differential Expression Analysis

While some information can be gained from mapped reads, the intention of an RNA-seq experiment is typically to find the differences in expression between groups of factors (McCarthy et al., 2012). To illustrate this, take two gene expression datasets from the same species under different environmental conditions. If one is grown under standard temperature and pressure while the other is grown near the freezing point of water, it would be expected that

genes controlling the response to the cold stress would only be expressed in the organism under cold stress. Therefore, it follows that differences in gene expression between these two organisms would highlight genes involved with cold stress. While simply stated, determining these genes requires careful experimental planning and statistical rigor (Lee et al., 2005).

RNA-seq experiments are tissue, organism, and environmentally dependent. To avoid unwanted interactions, great care must be taken to isolate the factors of interest from all other factors. For instance, sampling a leaf and root from the same plant allows for comparing gene expression between them. However, sampling a root from one plant and a leaf from a plant under drought stress does not allow for an accurate comparison of gene expression. Therefore, it is simpler to work backwards from a trait to determine the experimental design in many cases. Detecting genes of interest controlling how leaves retain their verdancy under drought will involve sampling from leaves of the same maturity from individuals under different watering conditions (Wolf, 2013).

Differential expression analysis is only possible due to the massive proliferation in computational resources in the last 20 years. This growth, coupled with the development of memory and resource-efficient algorithms, enables complex analyses to be run locally on laptops or through renting computational cloud time (Wall et al., 2010). There are a handful of differential expression analysis programs in use today. Historically, most of these programs operated on the Linux/Unix command line, however, the recent trend is to code them for the statistical language R, allowing more straightforward integration of statistical analysis. Several software packages in everyday use capable of differential expression analysis are Limma, edgeR, DEseq2, Cufflinks, and ALDEx2 (Fernandes et al., 2014; Love et al., 2014; Ghosh and Chan, 2016; Conesa et al., 2016; Costa-Silva et al., 2017; Law et al., 2018). From these programs, only

Cufflinks is not directly being developed as a package for R (Trapnell et al., 2012). Generally, each of these packages attempts to calculate similar results, what subset of genes are more or less expressed in samples. However, they differ in the statistical calculation of these outcomes and how they handle raw read data. Raw reads are heavily subject to bias from the sequencing machines and processes used to make RNA-seq libraries. The most typical way of handling this bias is by normalizing the read counts within and between samples. Traditionally, this has been done with Reads per Kilobase Million (RPKM, sometimes called Fragments per Kilobase Million, FPKM for paired-end data from Illumina experiments). While useful in some circumstances, RPKM has largely been replaced in algorithms because it tends to bias towards longer transcripts (Conesa et al., 2016).

Cufflinks is a piece of software developed for vertical use with TopHat, a read mapper, and CummeRbund to analyze differential expression statistics. While Cufflinks is still used, both TopHat and CummRbund have been, for the most part, superseded by more effective software. Cufflinks operates on raw mapped reads by initially packaging together similar transcripts from each sample. These packages help differentiate the different splice forms of a gene while also allowing for estimations of the representation of that gene between samples. This produces normalized counts whereby it is possible to plot differences in expression between samples. This process is facilitated by adapting a maximum likelihood calculation from a non-negative linear model, which is then used to inform placement in the posterior distribution giving the maximum *a posteriori* estimates for comparisons (Trapnell et al., 2010, 2012). The most recently derived program, ALDEx2, works somewhat similarly to Cufflinks in reporting values based on a posterior distribution. However, it differs by deriving the samples for these distributions based on centered log-ratio transformed gene or transcript counts. These are used in conjunction with

Monte-Carlo-based estimates of technical variation from the multivariate beta distribution. This process produces results that are termed "ANOVA-like" and can be interpreted similarly to an ANOVA test (Fernandes et al., 2014).

Limma, while initially designed to work on microarray data, has been adapted to RNA-seq as well. Limma contains two different methods of differential gene expression calculations, Limma-trend and VOOM, which are selected on the ratio of library size across samples. In the case that there is greater than a threefold difference in size, VOOM is recommended. The first step in both is normalization based on trimmed mean of M-values (TMM). This method generates M-values, the Log_2 ratio of treated sample counts to control sample counts per gene. A-values are also generated by dividing the sum of the Log_2 of treated counts and control counts by 2. Trimming occurs by cutting out the top and bottom 30% of M-values and 5% of A-values. The average M-value is then used to normalize counts. Limma-trend can take this data and using the functions "lmFit" and "eBayes," top differentially expressed genes across the experiment can be extracted. The process is similar when using Voom. However, in this case, the function "Voom" is applied, which calculates more precise model weights based on individual genes, to help control for between sample variances, the same Limma pipeline is then applied (Ritchie et al., 2015; Law et al., 2018).

EdgeR and DESeq2 both operate on a generalized linear model (GLM)-based weighting of factors. EdgeR fits a negative binomial distribution then estimates the biological coefficient of variation (BCV), how much variation in gene abundance is due to biological differences. Gene expression is variable and potentially non-independent between genes; therefore, BCV estimation is based on consistent dispersion within samples for individual genes. Dispersions are squeezed towards the global dispersion trend (Robinson et al., 2010; Law et al., 2018). DESeq2

also uses a GLM with parameters estimated from the negative binomial distribution. The dispersion parameter for this distribution is the difference between the mean of a sample gene count and its variance. The GLM is then fit, and a Wald test is run for parameter significance. Although edgeR and DESeq2 have similar computational steps, they differ in calculating statistical significance, this being done with a likelihood ratio test in the former. DESeq2 differs significantly from the other approaches mentioned thus far in that it relies on raw, unnormalized read counts. The rationale behind this is that if normalization of gene counts occurs before parameters estimation, some model precision is removed. Therefore, only library size needs to be corrected, which is handled automatically. Normalized outputs can be extracted using two protocols. One stabilizes based on dispersion, operating on the Log_2 scale, the other is a simple log_2 approach, giving the scaled fitted values (Love et al., 2014). Evaluations made to the efficacy of each of these programs have generally concluded that each has its operating assumptions and should be used in cases where they are advantageous to the experiment (Wang et al., 2019; Stark et al., 2019).

Machine Learning and Neural Networks

Machine learning applications have greatly expanded due to the increase in available computational power and complex datasets. Since the creation of the field in the mid-20th century, the focus and algorithms in use expanded greatly. Work in the 1960s provided the integration of the ideas of Thomas Bayes into algorithms. Bayes' work emphasized the importance of prior information on determining the probability of outcomes. This integration allowed machine learning scientists to estimate the probability of an outcome if there was an increase in the amount of data provided (Solomonoff, 1964). Later, work in the 1980s and 1990s changed the fundamental working units of machine learning to determine patterns based on

datasets instead of prior knowledge. Data-based approaches are the foundation of the current use of machine learning. Approaches and models are categorized based on the type of learning that is taking place. Two of these methods are supervised and unsupervised learning. They are named such based on whether the input data are provided with guidance on how they map to outputs or without any such instruction. Choice in either is dependent upon both the input data and desired outputs. Supervised learning operates based on known inputs and their corresponding outputs. Here a pathway is created between the inputs and outputs so that new outputs can be predicted from the presence of new inputs (Alpaydin, 2014a). Unsupervised learning operates without this intention. Outputs are hidden patterns that on their own would be too complicated to find. However, the structures found by the model can be powerful indications of data similarities or features (Alpaydin, 2014b). The final category of machine learning is reinforcement learning, whereby a model is influenced based on whether it completes a goal, rewarding it (Alpaydin, 2014c). This approach is not typically applied to machine learning in the context of DNA data. Instead, supervised learning is used with solid knowledge of DNA sequence and the desired output to be predicted, and unsupervised learning where this is not the case.

Self-organizing maps (SOM) or Kohonen maps are a form of unsupervised machine learning first developed by Teuvo Kohonen to arrange a map space to represent data biases based on similarity across a great number of dimensions. Fundamentally, SOM are artificial neural networks, a form of unsupervised learning that creates node sets that become linked based on the criteria to sort input data. SOM allows these nodes to compete for input data, refining the selection by which they operate. Over successive iterations, data are gathered into nodes. These nodes are mapped then as squares or hexagons over a two-dimensional plane. Typically, this plane represents the surface of a sphere or a torus, meaning that opposite edges of the plane are

connected. Each node is assigned an initial coordinate within the plane termed the weight vector. During each successive iteration of the algorithm, the input data adjust the weight vector of each node (Wehrens and Buydens, 2007; Haykin, 2009a; Kohonen, 2012; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019). While the map generated by the algorithm can be directly interpreted because it is in two-dimensional space, other algorithms can be applied to refine the map further. Modeled clustering methods such as Gaussian mixture model can further group nodes into larger but easier to work with clusters (Bishop, 2006; Fraley et al., 2020).

Neural networks and biology have been tied together since their inception. Initially, a neural network intended to emulate the functionality of neurons in sending signals in response to stimuli (McCulloch and Pitts, 1943). While this still exists as a biological neural network, more recent research has been on implementation of artificial neural networks. These networks take the underlying idea of the neural network but remove the objective of actual neuron mimicry. Here input nodes representing different categories of data are passed through nodes that alter this data and eventually give an output. As more data is fed to the network, the weights of these nodes change, leading to differing outputs for differing inputs. The most important part of a neural network is the changing of node weights. These can provide very detailed interpretations of complicated inputs (Haykin, 2009b). Perhaps the best way to interpret a neural network is to tie it back to animal learning. As stimuli are exposed to an animal, it may not know how to interpret it at first, but repeated exposure leads to categorization and increased detail in understanding.

In the exploration of the pod filling phenotype of cultivated tetraploid peanut RNA-seq data was processed to determine the genes and pathways controlling this important component of seed yield. Peanuts are an evolving and important component of the diets of humankind

worldwide, thus an increase in understanding of peanuts is to the benefit of humanity. RNA-seq data was prepared computationally, making use of modern read alignment software such as STAR which operated as the input for DESeq2, a program focused on the exploration of differential gene expression. However, making the leap from differentially expressed genes to pathways and gene clusters of interest will relied on the use of machine learning particularly the application of SOM. Contextually, differentially expressed genes were compared between tissues and stages of pod development in the background of Tifrunner a peanut genotype exhibiting pod-filling, and one NC 3033 that does not. Further analysis of gene functionality occurred computationally through orthologs to *Arabidopsis thaliana* and using Gene Ontology.

References

Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46(W1): W537–W544. doi: 10.1093/nar/gky379.

Alpaydin, E. 2014a. Supervised Learning. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 21–48

Alpaydin, E. 2014b. Introduction. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 1–20

Alpaydin, E. 2014c. Reinforcement Learning. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 517–546

Arya, S.S., A.R. Salve, and S. Chauhan. 2016. Peanuts as functional food: a review. J Food Sci Technol 53(1): 31–41. doi: 10.1007/s13197-015-2007-9.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25(1): 25–29. doi: 10.1038/75556.

Ballén-Taborda, C., Y. Chu, P. Ozias-Akins, P. Timper, C.C. Holbrook, et al. 2019. A new source of root-knot nematode resistance from Arachis stenosperma incorporated into allotetraploid peanut (Arachis hypogaea). Sci Rep 9(1): 17702. doi: 10.1038/s41598-019-54183-1.

Barbour, J.A., P.R.C. Howe, J.D. Buckley, J. Bryan, and A.M. Coates. 2015. Effect of 12 Weeks High Oleic Peanut Consumption on Cardio-Metabolic Risk Factors and Body Composition. Nutrients 7(9): 7381–7398. doi: 10.3390/nu7095343.

Benfey, P.N., and T. Mitchell-Olds. 2008. From Genotype to Phenotype: Systems Biology Meets Natural Variation. Science 320(5875): 495–497. doi: 10.1126/science.1153716.

Bera, S.K., J.H. Kamdar, S.V. Kasundra, S.V. Patel, M.D. Jasani, et al. 2019. Steady expression of high oleic acid in peanut bred by marker-assisted backcrossing for fatty acid desaturase mutant alleles and its effect on seed germination along with other seedling traits. PLOS ONE 14(12): e0226252. doi: 10.1371/journal.pone.0226252.

Bertioli, D.J., S.B. Cannon, L. Froenicke, G. Huang, A.D. Farmer, et al. 2016. The genome sequences of Arachis duranensis and Arachis ipaensis , the diploid ancestors of cultivated peanut. Nat Genet 48(4): 438–446. doi: 10.1038/ng.3517.

Bertioli, D.J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, et al. 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nature Genetics 51(5): 877–884. doi: 10.1038/s41588-019-0405-z.

Beute, M.K., J.C. Wynne, and D.A. Emery. 1976. Registration of NC 3033 Peanut Germplasm1 (Reg. No. GP 9). Crop Science 16(6): cropsci1976.0011183X001600060046x. doi: 10.2135/cropsci1976.0011183X001600060046x.

Bhullar, N.K., Z. Zhang, T. Wicker, and B. Keller. 2010. Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene Pm3: a large scale allele mining project. BMC Plant Biol 10(1): 88. doi: 10.1186/1471-2229-10-88.

Bishop, C.M. 2006. Chapter 9: Mixture Models and EM. Pattern recognition and machine learning. Springer, New York. p. 423–455

Bonku, R., and J. Yu. 2020. Health aspects of peanuts as an outcome of its chemical composition. Food Science and Human Wellness 9(1): 21–30. doi: 10.1016/j.fshw.2019.12.005.

Boote, K.J. 1982. Growth Stages of Peanut (Arachis hypogaea L.)1. Peanut Science 9(1): 35–40. doi: 10.3146/i0095-3679-9-1-11.

Brown, S.L., A.K. Culbreath, J.W. Todd, D.W. Gorbet, J.A. Baldwin, et al. 2005. Development of a Method of Risk Assessment to Facilitate Integrated Management of Spotted Wilt of Peanut. Plant Disease 89(4): 348–356. doi: 10.1094/PD-89-0348.

Buccitelli, C., and M. Selbach. 2020. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21(10): 630–644. doi: 10.1038/s41576-020-0258-4.

Cano-Gamez, E., and G. Trynka. 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Frontiers in Genetics 11: 424. doi: 10.3389/fgene.2020.00424.

Carter, R.J., I. Dubchak, and S.R. Holbrook. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 29(19): 3928–3938.

Chappell, T.M., C.B. Codod, B.W. Williams, R.C. Kemerait, A.K. Culbreath, et al. 2020. Adding Epidemiologically Important Meteorological Data to Peanut Rx, the Risk Assessment Framework for Spotted Wilt of Peanut. Phytopathology® 110(6): 1199–1207. doi: 10.1094/PHYTO-11-19-0438-R.

Chavarro, C., Y. Chu, C. Holbrook, T. Isleib, D. Bertioli, et al. 2020. Pod and Seed Trait QTL Identification To Assist Breeding for Peanut Market Preferences. G3: Genes, Genomes, Genetics 10(7): 2297–2315. doi: 10.1534/g3.120.401147.

Chettoor, A.M., S.A. Givan, R.A. Cole, C.T. Coker, E. Unger-Wallace, et al. 2014. Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biology 15(7): 414. doi: 10.1186/s13059-014-0414-2.

Chinnusamy, V., J. Zhu, and J.-K. Zhu. 2007. Cold stress regulation of gene expression in plants. Trends in Plant Science 12(10): 444–451. doi: 10.1016/j.tplants.2007.07.002.

Chu, Y., C.L. Wu, C.C. Holbrook, B.L. Tillman, G. Person, et al. 2011. Marker-Assisted Selection to Pyramid Nematode Resistance and the High Oleic Trait in Peanut. The Plant Genome 4(2). doi: 10.3835/plantgenome2011.01.0001.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17(1): 13. doi: 10.1186/s13059-016-0881-8.

Costa-Silva, J., D. Domingues, and F.M. Lopes. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE 12(12): e0190152. doi: 10.1371/journal.pone.0190152.

Culbreath, A.K., and R. Srinivasan. 2011. Epidemiology of spotted wilt disease of peanut caused by Tomato spotted wilt virus in the southeastern U.S. Virus Research 159(2): 101–109. doi: 10.1016/j.virusres.2011.04.014.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1): 15–21. doi: 10.1093/bioinformatics/bts635.

Elson, D. 1965. Metabolism of Nucleic Acids (Macromolecular DNA and RNA). Annual Review of Biochemistry 34(1): 449–486. doi: 10.1146/annurev.bi.34.070165.002313.

FAO. 2017. The future of food and agriculture: trends and challenges. Food and Agriculture Organization of the United Nations, Rome.

FAO. 2019. Seeds Food and Agriculture Organization of the United Nations. http://www.fao.org/seeds/en/ (accessed 23 June 2021).

Fernandes, A.D., J.N. Reid, J.M. Macklaim, T.A. McMurrough, D.R. Edgell, et al. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2: 15. doi: 10.1186/2049-2618-2-15.

Fraley, C., A.E. Raftery, L. Scrucca, T.B. Murphy, and M. Fop. 2020. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation.

Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, et al. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289(5476): 85–88. doi: 10.1126/science.289.5476.85.

Ghosh, S., and C.-K.K. Chan. 2016. Analysis of RNA-Seq Data Using TopHat and Cufflinks. In: Edwards, D., editor, Plant Bioinformatics: Methods and Protocols. Springer, New York, NY. p. 339–361

Gilman, D.F., and O.D. Smith. 1977. Internal Pericarp Color as a Subjective Maturity Index for Peanut Breeding1. Peanut Science 4(2): 67–70. doi: 10.3146/i0095-3679-4-2-6.

Glaeser, B. 2011. The Green Revolution Revisited: Critique and Alternatives. Taylor & Francis Group, London, UNITED KINGDOM.

Gupta, K., O. Buchshtab, and R. Hovav. 2014. The Effects of Irrigation Level and Genotype on Pod-Filling Related Traits in Peanut (Arachis hypogaea). Journal of Agricultural Science 7(1): p169. doi: 10.5539/jas.v7n1p169.

Hajduch, M., L.B. Hearne, J.A. Miernyk, J.E. Casteel, T. Joshi, et al. 2010. Systems Analysis of Seed Filling in Arabidopsis: Using General Linear Modeling to Assess Concordance of Transcript and Protein Expression. Plant Physiology 152(4): 2078–2087. doi: 10.1104/pp.109.152413.

Halliburton, B.W., W.G. Glasser, and J.M. Byrne. 2015. An Anatomical Study of the Pericarp of Arachis Hypogaea, with Special Emphasis on the Sclereid Component. Botanical Gazette. doi: 10.1086/336807.

Halvey, J., A. Hartzook, and T. Markovitz. 1987. Foliar fertilization of high-yielding peanuts during the pod-filling period. Fertilizer Research 14: 153–10.

Hammons, R.O., D.K. Bell, and E.K. Sobers. 1981. Evaluating Peanuts for Resistance to Cylindrocladium Black Rot1. Peanut Science 8(2): 117–120. doi: 10.3146/i0095-3679-8-2-10.

Haykin, S.S. 2009a. Self-Organizing Maps. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 425–470

Haykin, S.S. 2009b. Introduction. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 1–46

Hazell, P.B.R. 2009. The Asian Green Revolution. Intl Food Policy Res Inst.

Hilu, K.W. 1993. Polyploidy and the evolution of domesticated plants. American Journal of Botany 80(12): 1494–1499. doi: 10.1002/j.1537-2197.1993.tb15395.x.

Hoheisel, J.D. 2006. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet 7(3): 200–210. doi: 10.1038/nrg1809.

Holbrook, C.C., and A.K. Culbreath. 2007. Registration of ‘Tifrunner’ Peanut. Journal of Plant Registrations 1(2): 124–124. doi: 10.3198/jpr2006.09.0575crc.

Holbrook, C.C., T.G. Isleib, P. Ozias-Akins, Y. Chu, S.J. Knapp, et al. 2013. Development and Phenotyping of Recombinant Inbred Line (RIL) Populations for Peanut (Arachis hypogaea). Peanut Science 40(2): 89–94. doi: 10.3146/PS13-5.1.

Howe, K.L., P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, et al. 2021. Ensembl 2021. Nucleic Acids Research 49(D1): D884–D891. doi: 10.1093/nar/gkaa942.

Illumina. 2011. RNA-Seq Data Comparison with Gene Expression Microarrays. https://www.europeanpharmaceuticalreview.com/wp-content/uploads/Illumina_whitepaper.pdf (accessed 2 July 2021).

Jackson, L.F. 1983. Relative Susceptibilities of Component Lines of Peanut Cultivars Early Bunch and Florunner to Early and Late Leafspots1. Peanut Science 10(1): 3–5. doi: 10.3146/i0095-3679-10-1-2.

Jones, D.T. 2019. Setting the standards for machine learning in biology. Nat Rev Mol Cell Biol 20(11): 659–660. doi: 10.1038/s41580-019-0176-5.

Jung, S., G. Powell, K. Moore, and A. Abbott. 2000. The high oleate trait in the cultivated peanut [Arachis hypogaea L.]. II. Molecular basis and genetics of the trait. Mol Gen Genet 263(5): 806–811. doi: 10.1007/s004380000243.

Kalberer, S., and V. Belamkar. 2014. Arachis duranensis, Arachis ipaensis, and the Origins of Cultivated Peanut (Arachis hypogaea). https://www.peanutbase.org/files/misc/arachis_duranensis_ipaensis_info_v02.pdf (accessed 2 July 2021).

Klevorn, C.M., L.L. Dean, and S.D. Johanningsmeier. 2019. Metabolite Profiles of Raw Peanut Seeds Reveal Differences between Market-Types. Journal of Food Science 84(3): 397–405. doi: 10.1111/1750-3841.14450.

Kohonen, T. 2012. Self-Organizing Maps. Springer Science & Business Media.

Koonin, E.V., L. Aravind, and A.S. Kondrashov. 2000. The Impact of Comparative Genomics on Our Understanding of Evolution. Cell 101(6): 573–576. doi: 10.1016/S0092-8674(00)80867-3.

Krapovickas, A., W.C. Gregory, D.E. Williams, and C.E. Simpson. 2007. TAXONOMY OF THE GENUS ARACHIS (LEGUMINOSAE). Bonplandia 16: 7–205.

Kruisselbrink, R.W. and J. 2019. kohonen: Supervised and Unsupervised Self-Organising Maps.

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3): R25. doi: 10.1186/gb-2009-10-3-r25.

Law, C.W., M. Alhamdoosh, S. Su, X. Dong, L. Tian, et al. 2018. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res 5: ISCB Comm J-1408. doi: 10.12688/f1000research.9005.3.

Lee, B., D.A. Henderson, and J.-K. Zhu. 2005. The Arabidopsis Cold-Responsive Transcriptome and Its Regulation by ICE1. The Plant Cell 17(11): 3155–3175. doi: 10.1105/tpc.105.035568.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26(5): 589–595. doi: 10.1093/bioinformatics/btp698.

Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12): 550. doi: 10.1186/s13059-014-0550-8.

Lyons, E., and M. Freeling. 2008a. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53(4): 661–673. doi: 10.1111/j.1365-313X.2007.03326.x.

Lyons, E., and M. Freeling. 2008b. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53(4): 661–673. doi: 10.1111/j.1365-313X.2007.03326.x.

Mantione, K.J., R.M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, et al. 2014. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res 20: 138–141. doi: 10.12659/MSMBR.892101.

Masur, L.J., J.-F. Millaire, and M. Blake. 2018. Peanuts and Power in the Andes: The Social Archaeology of Plant Remains from the Virú Valley, Peru. etbi 38(4): 589–609. doi: 10.2993/0278-0771-38.4.589.

McCarthy, D.J., Y. Chen, and G.K. Smyth. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40(10): 4288–4297. doi: 10.1093/nar/gks042.

McCulloch, W.S., and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5(4): 115–133. doi: 10.1007/BF02478259.

Mitkowski, N.A., and G.S. Abawi. 2003. Root-knot nematode. Root-knot nematode. https://www.apsnet.org/edcenter/disandpath/nematode/pdlessons/Pages/RootknotNematode.aspx (accessed 23 July 2021).

Moore, G.E. 1965. Cramming more components onto integrated circuits. 38(8): 4.

Nayak, R., and Y. Hasija. 2021. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. Genomics 113(2): 606–619. doi: 10.1016/j.ygeno.2021.01.007.

Noble, D. 2002. The rise of computational biology. Nat Rev Mol Cell Biol 3(6): 459–463. doi: 10.1038/nrm810.

Nussinov, R. 2015. Advancements and Challenges in Computational Biology. PLOS Computational Biology 11(1): e1004053. doi: 10.1371/journal.pcbi.1004053.

Parkinson, J., and M. Blaxter. 2009. Expressed sequence tags: an overview. Methods Mol Biol 533: 1–12. doi: 10.1007/978-1-60327-136-3_1.

Paterson, A.H., J.E. Bowers, M.D. Burow, X. Draye, C.G. Elsik, et al. 2000. Comparative Genomics of Plant Chromosomes. The Plant Cell 12(9): 1523–1539. doi: 10.1105/tpc.12.9.1523.

Pearson, W.R. 2013. An Introduction to Sequence Similarity ("Homology") Searching. Curr Protoc Bioinformatics 0 3: 10.1002/0471250953.bi0301s42. doi: 10.1002/0471250953.bi0301s42.

Pearson, W.R., and D.J. Lipman. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8): 2444–2448.

Renny-Byfield, S., and J.F. Wendel. 2014. Doubling down on genomes: Polyploidy and crop plants. American Journal of Botany 101(10): 1711–1725. doi: 10.3732/ajb.1400119.

Ritchie, M.E., B. Phipson, D. Wu, Y. Hu, C.W. Law, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7): e47. doi: 10.1093/nar/gkv007.

Robinson, M.D., D.J. McCarthy, and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139–140. doi: 10.1093/bioinformatics/btp616.

Shiraiwa, T., N. Ueno, S. Shimada, and T. Horie. 2004. Correlation between Yielding Ability and Dry Matter Productivity during Initial Seed Filling Stage in Various Soybean Genotypes. Plant Production Science 7(2): 138–142. doi: 10.1626/pps.7.138.

Smith, B.W. 1950. Arachis Hypogaea. Aerial Flower and Subterranean Fruit. American Journal of Botany 37(10): 802–815. doi: 10.1002/j.1537-2197.1950.tb11073.x.

Solomonoff, R.J. 1964. A formal theory of inductive inference. Information and Control 7(2): 224–254. doi: 10.1016/S0019-9958(64)90131-7.

Stark, R., M. Grzelak, and J. Hadfield. 2019. RNA sequencing: the teenage years. Nat Rev Genet 20(11): 631–656. doi: 10.1038/s41576-019-0150-2.

Tarca, A.L., V.J. Carey, X. Chen, R. Romero, and S. Drăghici. 2007. Machine Learning and Its Applications to Biology. PLOS Computational Biology 3(6): e116. doi: 10.1371/journal.pcbi.0030116.

Toomer, O.T. 2018. Nutritional chemistry of the peanut (Arachis hypogaea). Crit Rev Food Sci Nutr 58(17): 3042–3053. doi: 10.1080/10408398.2017.1339015.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3): 562–578. doi: 10.1038/nprot.2012.016.

Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5): 511–515. doi: 10.1038/nbt.1621.

Venkatachalam, M., and S.K. Sathe. 2006. Chemical Composition of Selected Edible Nut Seeds. J. Agric. Food Chem. 54(13): 4705–4714. doi: 10.1021/jf0606959.

Wall, D.P., P. Kudtarkar, V.A. Fusaro, R. Pivovarov, P. Patil, et al. 2010. Cloud computing for comparative genomics. BMC Bioinformatics 11(1): 259. doi: 10.1186/1471-2105-11-259.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1): 57–63. doi: 10.1038/nrg2484.

Wang, T., B. Li, C.E. Nelson, and S. Nabavi. 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 20(1): 40. doi: 10.1186/s12859-019-2599-6.

Wehrens, R., and L.M.C. Buydens. 2007. Self- and Super-Organizing Maps in R: The kohonen Package. Journal of Statistical Software 21(5): 1–19. doi: 10.18637/jss.v021.i05.

Wehrens, R., and J. Kruisselbrink. 2018. Flexible Self-Organizing Maps in kohonen 3.0. Journal of Statistical Software 87(7): 1–18. doi: 10.18637/jss.v087.i07.

Wolf, J.B.W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Molecular Ecology Resources 13(4): 559–572. doi: 10.1111/1755-0998.12109.

Yamamoto, M., T. Wakatsuki, A. Hada, and A. Ryo. 2001. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods 250(1–2): 45–66. doi: 10.1016/s0022-1759(01)00305-2.

Yin, S., P. Li, Y. Xu, J. Liu, T. Yang, et al. 2020. Genetic and genomic analysis of the seed-filling process in maize based on a logistic model. Heredity 124(1): 122–134. doi: 10.1038/s41437-019-0251-x.

Young, C.T., and W.E. Schadel. 2004. Microstructure of Peanut Seed: A Review. : 13.

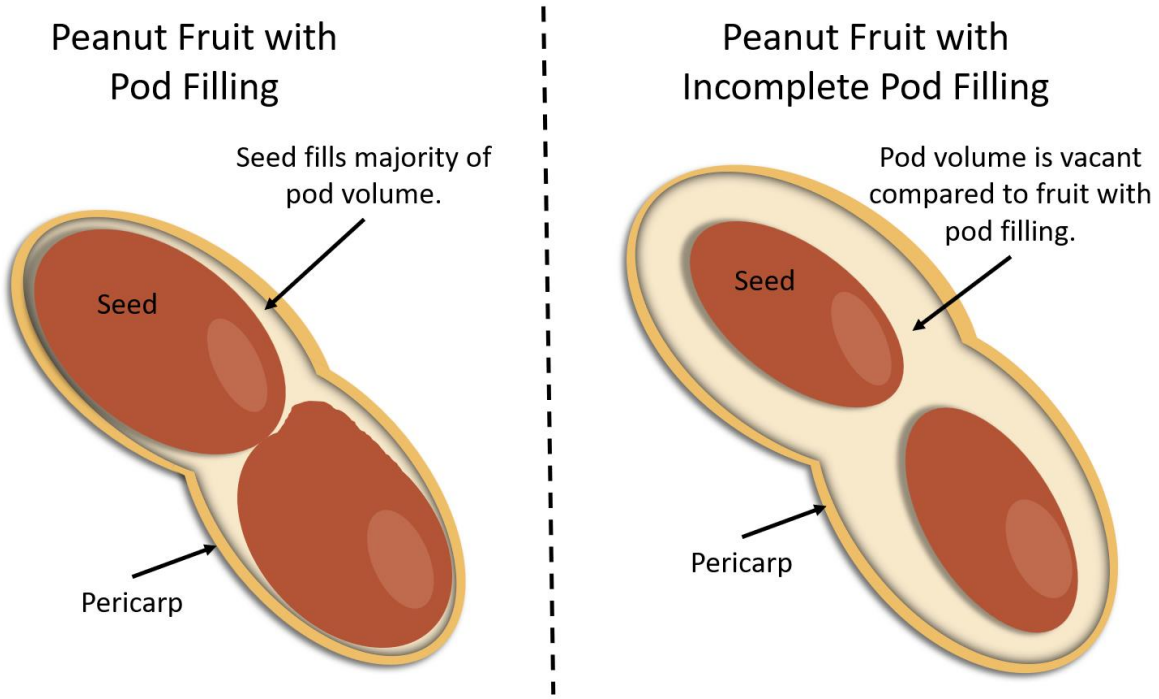

Figure 1.1: Diagram of Pod Filling Phenotypes. These diagrams demonstrate the differences in phenotypes present in genotypes with and without complete pod filling. The left shows complete pod filling. Here most of the volume inside the shell of the peanut fruit is occupied by its seeds. This leads to higher yields per peanut pod. The right image demonstrates peanuts with the incomplete pod filling phenotype. The two seeds pictured occupy a smaller percentage of the total pod volume than the seeds from the complete pod filling example. This implies that if the total volume of both pods were the same the genotype with incomplete pod filling will have lower seed yields on a pod-for-pod basis.

CHAPTER 2

IDENTIFICATION AND ANALYSIS OF GENES ASSOCIATED WITH POD FILLING IN CULTIVATED PEANUT

Introduction

Cultivated peanut (*Arachis hypogaea*) is a crop grown throughout the world for its highly nutritious seed. The annual production of peanuts amounts to 49 million metric tons worldwide, with the primary growing regions being in Asia and the Americas (FAO, 2019). Peanut seeds contain a healthy combination of phytonutrients, vitamins, and minerals, which in conjunction with their high levels of proteins, serve as an efficient food source, especially for developing nations (Toomer, 2018). While consumed whole after cooking, peanut also has roles in processed foods and goods (Arya et al., 2016). Currently, there are four main market types of peanuts grown in the United States. Each type has some unique characteristics that make them more suited for its market niche. The two most widely grown market types are runner-types and Virginia-types. Runner-types, primarily grown in the Southeastern U.S., are known for their high-yielding lines mainly sold to processing plants creating secondary goods from peanuts. These goods are peanut oils, peanut butter, and peanut meals, both for human and animal consumption. Virginia-type peanuts produce large seeds that make good candidates for roasted whole seed secondary products such as cocktail peanuts and roasted in-shell peanuts. Differences in peanut end uses are due to differences in phenotype generally held between market types (Arya et al., 2016; Klevorn et al., 2019).

Cultivated peanut *Arachis hypogaea* is an allotetraploid species with an AABB structured genome. The *Arachis* species that constitute the cross are still found in the wild today, partly because this hybridization event corresponded with the beginning of cultivation only 9,400 years ago. Modern molecular work has shown that the A sub-genome corresponds to the species *A. duranensis* and the B sub-genome to *A. ipaënsis.* Each species is diploid; therefore, there was a spontaneous genome doubling event after hybridization to produce a viable tetraploid (Bertioli et al., 2016, 2019). The majority of naturally occurring *Arachis* species are diploid (Moretzsohn et al., 2004). Hybridization and polyploidization reproductively isolated cultivated peanut from random viable crosses with other diploid *Arachis* species, introducing a significant genetic bottleneck into the species. However, recombination between the sub-genomes of *A. hypogaea* that has occurred over time can be detected. Some homoeologous chromosome ends have transitioned from the AABB structure to an AAAA or BBBB structure, likely due to the relative similarities between the sub-genome sequences. The sequence of *A. hypogaea* contains 2.7 Gb of nucleotides arranged into 20 total chromosomes. These chromosomes contain 67,124 total genes, many of which contain homoeologous copies in both sub-genomes (Bertioli et al., 2019). These genes can be explored using online databases such as PeanutBase, which curates genes, sequences, and known functions (Dash et al., 2016). However, knowledge is limited by previous research regarding the functions of peanut genes. Prior work has been done to search for quantitative trait loci (QTL) associated with various traits of interest in agricultural contexts for peanut.

In grain-like crops such as peanuts, the production and quality of seeds is paramount to having a reliable, marketable, and sustainable crop. Therefore, research into the inner workings of nutrient accumulation and seed growth, a stage known as seed filling, allows grain crops to be

improved to reach their full potential (Shiraiwa et al., 2004; Hajduch et al., 2010; Arya et al., 2016; Yin et al., 2020; Chavarro et al., 2020). Peanuts produce and fill their seeds throughout four reproductive stages, each delimited by differences in pod appearance and overall fruit maturity. Fruit development begins with stage R3, which occurs after flowering, when the peg has grown underground, and the ovary begins to swell. This development continues in stages R4, R5, and R6, where the pod tissues become separable and expand. Pods approach their maximum size in stage R7, whereby they continue to mature, becoming harvest-ready in stage R8 (Boote, 1982). Water and nutrients are absorbed into the fruits throughout this process, leading to their final nutrient levels at harvest. Therefore, the evaluation of seed and pod qualities is best examined during these reproductive stages (Halvey et al., 1987; Hajduch et al., 2010; Gupta et al., 2014).

Some peanut genotypes have a pod and seed phenotype known as incomplete pod-filling. This phenotype results in peanut seeds not wholly reaching the outer shell of a peanut pod, leaving a small void around the seed. These seeds are less producer-friendly in that they are sacrificing seed size for no known benefits. This situation is the case for the Virginia-type cultivar NC 3033. Although the cultivar does contain desirable characteristics for disease resistance to *Cylindrocladium* black rot (CBR), it also possesses incomplete pod filling, making it less desirable as a parental line for plant breeding (Beute et al., 1976; Chavarro et al., 2020). In contrast, the runner-type cultivar Tifrunner exhibits complete pod filling, making it a natural contrast to NC 3033 (Holbrook and Culbreath, 2007). Isolated pod tissues from reproductive stages R4-R7 would intuitively contain the genetic differences that lead to the phenotypic differences in NC 3033 and Tifrunner.

While DNA contains the actual genetic code of a gene, it is not an accurate indication that those genes are expressed. Proteins carry out the functionality of genes but examining them is complex and costly both monetarily and computationally. mRNA, the intermediate between DNA and protein, offers a compelling alternative. Direct associations between sequence and genes allow for mRNA to map against DNA directly, while mRNA is much more cost-effective to study while also indicating which proteins may be present in a cell. However, there are some considerations of working with RNA that factor into experimental design. Care must be taken to ensure that the phenotype of interest is controlled by the tissues from which samples are derived. RNA is more like protein than DNA in this regard. While DNA sequence is conserved throughout an organism, the composition of individual cells and tissues varies regarding which proteins and RNA are contained within. Differentiable cells then are a result of these differences (Wang et al., 2009; Love et al., 2014; Costa-Silva et al., 2017; Buccitelli and Selbach, 2020).

Comparing individuals, this association between RNA and sampled tissues proves true. Similar tissues in closely related genotypes of the same species, for the most part, will have conserved expression of gene sets. However, there will be some variation in expression of others to allow for differences in phenotypes (Buccitelli and Selbach, 2020). This process is the fundamental principle behind this analysis of peanut pod-filling. While the cultivar Tifrunner exhibits the complete pod-filling phenotype, the cultivar NC 3033 does not. These differences, therefore, should be detectable in the mRNA transcripts seen between the two. The location of the trait will be the tissues of the pod, being that the pod tissues are the regions exhibiting the phenotype. Using computational tools, RNA-seq reads from samples across pod tissue-developmental stages will be compared, and differentially expressed genes will be extracted.

These genes will be related to experimentally validated pathways and functional groups to elucidate their role in reproductive development and pod-filling.

Bioinformatics tools are the key to unlocking this information from RNA-based data. Software for developing understandings of differentially expressed genes fits into this category; however, in some cases, the differences found between samples are challenging to assign to functional units such as pathways for analysis. This problem is because the high dimensionality of biological data results in challenges associated with filtering data of interest from background data or noise that may be present (Nussinov, 2015). Remediation can be made by combining the exploratory forces of many types of analysis with the information gained from differential expression programs. Combined, this leads to a directed analysis of function, co-expression, or orthologs. This line of thinking can be expanded even further, leading to combinations of these into more complex but perhaps more informative centers of information.

Machine learning applications to biology are numerous due to the vast amount of highly complex or deeply dimensional data types. Machine learning excels at reclassifying complex data into types more conducive to interpretation (Alpaydin, 2014). Self-organizing maps (SOM) fit this category of computational algorithms. Here, an artificial neural network is formed that learns from a matrix of input data crafting new pathways of interpretation, classifying entire input levels into nodes that accumulate more inputs forming a map of nodes containing inputs with distances calculated between nodes based on the relatedness of the data contained within them. This process takes a surface of n-dimensions and converts it to one of 2-dimensions. Dimensional reduction allows for the use of traditional clustering algorithms, which group sub-spaces of a plane into distinct groups based on the distances between them. Functional

annotation of these clusters yields insight into the potential reasons for similar expression patterns across samples (Haykin, 2009a; b; Kruisselbrink, 2019).

Differences in the pod-filling phenotypes of the *Arachis hypogaea* were evaluated using RNA-seq from samples of pod tissues across the reproductive stages associated with this phenotype. Computational analysis began with differential expression analysis to highlight genes that have detectable changes in expression levels between genotype, stages, tissues, and within them. Differentially expressed genes were annotated for functionality using Gene Ontology, and expression clusters were generated using self-organizing maps (SOM) with further Gaussian mixture model (GMM) clustering. These clusters were then compared to known QTL associated with traits for pod filling. Those clusters with increased overlap with QTL were then annotated using functional pathway analysis of orthologs in *Arabidopsis thaliana*.

Materials and Methods

RNA-seq Experiment

Tifrunner and NC 3033 individuals were grown under greenhouse conditions in Athens, GA. Samples for RNA extraction were taken from each genotype at each of the nine tissue and developmental stages pairing, with four biological replicates. The tissue factors used were the pericarp of the pod, the seedcoat within the pericarp, and the embryo within the seedcoat. The stage factors used were a pooled R4-5 reproductive stage, R6, and R7. Samples of all three tissues were taken for sequencing at each reproductive stage. This design resulted in 72 total RNA samples. Liquid nitrogen was used during maceration to prep sample for RNA extraction using Direct-zol RNA MiniPrep kit (Zymo Research, Irvine, CA, USA). Samples were assessed for quality using an Aligent 2100 Bioanalyzer (Agilent RNA 6000 Nano Kit) and treated with

RNeasy MiniElute Cleanup kit® for low concentration RNA or samples with an RNA integrity number less than eight (Chavarro, 2021).

cDNA libraries were generated based on the True-seq Stranded mRNA library preparation kit (Illumina, San Diego, CA, USA) for 0.1-4µg of RNA. The rationale behind using stranded RNA-seq was to optimize for high percent alignments which accurately identify antisense transcripts (Levin et al., 2010). Library size and quality were assessed with an Agilent DNA 7500 kit on an Agilent 2100 Bioanalyzer (Agilent Genomics, Santa Clara, CA, USA). RNA concentration was obtained using a Qubit 2.0 fluorometer (Thermo Fisher Scientific Inc, Waltham, MA, USA). Sequencing was performed on HiSeq 2 x125 bp in 1 complete flow cell at the University of Colorado Denver, resulting in 20-90 Mb of reads per tissue to stage pairing. In an effort to avoid confounding effects, samples were pooled and divided across sequencing lanes. All work in the laboratory and field contributing to this analysis was planned and performed by Dr. Carolina Chavarro (Chavarro, 2021).

Read Mapping and Data Filtering

Raw RNA-seq reads were processed according to guidelines provided with the STAR read mapper. Version 1.0 of the *Arachis hypogaea* genome was downloaded from PeanutBase and used to create read mapped outputs for each RNA-seq sample (Dash et al., 2016; Bertioli et al., 2019). Raw counts per gene were then enumerated using baseline pipelines for HTSeq with *Arachis hypogaea* gene annotations. This step provided a collection of individual files for each sample with the total read counts per gene. A custom python script was created to compile each file into a single composite matrix with all samples and genes present. Pre-processing of the data was completed using R version 4.1.0 (R Core Team, 2021). Genes were removed based on having a sum read count value of less than ten across all samples. Pre-process quality control

was completed using box plots of each sample both from raw counts and normalized counts. No noticeable outliers were detected from plots of raw counts; however, filtering removed one sample with normalized values of less than 100 total reads across all genes. A visual overview of computational pathways is given in Figure 2.1.

Differential Expression Analysis

DESeq2 was used to find differentially expressed genes from the pre-filtered read data. This process uses a formula to find differences within and between factors provided and identified by the user (Love et al., 2014, 2021). Based on the factors considered during experimental design, evaluations were made based on all three-factor categories, genotype, stage, and tissue, including interaction terms between each factor and the three factors combined. This is represented as a formula similar in form to that of a generalized linear regression model. This formula is given by:

$$\sim 1 + genotype + stage + tissue + genotype{:}stage{:}tissue.$$

Establishing a baseline of 1 using the y-intercept term allows for equal comparisons to all factors. Without this, one factor level will be held as the intercept and be computationally impossible to analyze separately from the model's overall baseline. While this does not need to be considered while evaluating gene expression based on the full factorial model, subsequent analysis of gene expression by factor relies on this distinction. The genotype factor contains two levels, NC 3033 and Tifrunner, the two *Arachis hypogaea* genotypes from which samples were taken. The genotype NC 3033 is held as the baseline value of this term from which the expression patterns in Tifrunner are compared. While this does not affect which genes are considered to be differentially expressed, it does alter the way gene expression levels are reported. When using the "contrasts" function of DESeq2, the only factors considered to be

involved in differential expression are those indicated by the user. The program reports positive expression if a differentially expressed gene is expressed at a higher level in the non-baseline factor level. Negative expression indicates that the differentially expressed gene is expressed more in the baseline factor level. For factor groupings such as stage and tissue which contain more than two factor levels, non-baseline factor levels are compared first to a common baseline level. This provides differential expression in the context of a single pairwise comparison. However, this comparison only accounts for measurements from the baseline of the model. Comparisons of expression levels between non-baseline factor levels are computed based on the differences present in each level's pairwise contrast to the baseline level. For the stage factor, R4-5 was chosen as the baseline as it is the earliest reproductive stage of the three. For the tissue factor embryo was chosen as the baseline level as it was the first of the three levels alphabetically. The methods of factor integration in DESeq2 are further explored in Figures 2.2-2.5.

Statistical tests of estimators were computed using a Wald test with null hypothesis $\beta=0$. Before differential analysis, PCA was performed across the samples with highlighting based on factor levels. Differential expression analysis performed with DESeq2 on the full factorial model compiled a set of 294 differentially expressed genes based on thresholds of Benjamini-Hochberg corrected p-value ≤ 0.05 and LFC ± 1 . Prior to filtering, DESeq2's quality was assessed by examining the shape of both the histogram of gene p-values and the dispersion estimates calculated by DESeq2 to ensure they were within acceptable bounds (Love et al., 2014, 2021). This same process was repeated for each factor level contrast, resulting in different sets of differentially expressed genes based on the same filtering criteria. Variance stabilized gene count data were used to generate a heatmap of genes by RNA-seq samples. Each axis was then

independently clustered using k-means clustering, reordering the genes and samples based in the results.

A custom Python3 script was generated to access the Intermine interface for PeanutMine and extract GO terms for each of the 294 differentially expressed genes (Ashburner et al., 2000; Smith et al., 2012; Kalderimis et al., 2014; Dash et al., 2016; Carbon et al., 2021). Using the TopGO package for R, GO term enrichment was calculated based on a K.S. test of significance p-values computed in the analysis of the GO terminology series contained within the differentially expressed gene set. Scaled enrichment of GO terms was plotted using ggplot2, only the top 10 enriched GO terms for each major GO category were retained (Wickham, 2016, p. 2; Alexa and Rahnenfuhrer, 2021).

Self-Organizing Maps

Differentially expressed genes were further evaluated using the R package Kohonen. This package allows for the training and generation of self-organizing maps (SOM) based on differing numbers of input factors provided by a matrix (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019). Differentially expressed genes from the complete model were extracted and normalized based on a variance stabilized transformation. This was then passed to the Kohonen software pipeline using parameters for mapping area based on the following formula:

$$\sqrt{Number\ of\ Nodes} = \lceil Number\ of\ \ Samples^{2/5} \rceil$$

This formula dictates the length and width parameters of the space. We are seeking to maximize the number of nodes present in the map which cannot exceed, due to limitations of SOMs, the number of input parameters. This formula guarantees to generate the largest possible square SOM. A large number of nodes is desirable for it provides the highest resolution for

dissecting the distribution of genes across the SOM. A square SOM allows for consistency in the dimensionality of the mapping space even across different groups of samples. Therefore, to find the maximum number of nodes on one side of the square, the total number of samples is taken the power of 2/5 where then the number is rounded up to the nearest whole number. The highest whole number is then found using the ceiling function.

Additional computational parameters delineate a toroidal geometry overlaid with hexagonal node shapes (Kruisselbrink, 2019). Initial maps generated were further refined using the R package mclust, which performs Gaussian Mixture Modeled (GMM) clustering (Scrucca et al., 2016; Fraley et al., 2020). The Kohonen process was re-run using functions for isolating experimental factors and providing knowledge of them to the SOM algorithm. The two-factor level function "xyf" was run to compute a SOM using genotype levels. This SOM was further clustered using the same method as above. The "SuperSOM" function was used to evaluate the impact of all factors and levels on SOM construction (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019). Again, mclust was used to further cluster nodes. No significant improvements were noticed with the more complex models. Therefore, the basic SOM was used for the remaining analysis. For each contrast set calculated using DESeq2, SOM analysis was re-run using the same parameters and modeling procedures.

Among contrast models, charts were generated to display the opposing levels of gene expression present between factors within clusters. This was to demonstrate that expression of genes within certain clusters was biased towards one of the factor levels. This was performed using scaled log fold change (LFC) data for genes within clusters. Plotting was done in R using ggplot2 (Wickham, 2016, p. 2; R Core Team, 2021).

Cluster to QTL Analysis

Clusters determined by SOM and GMM were compared to QTL found in a previous study of pod phenotype traits in a RIL population from Tifrunner x NC 3033 (Chavarro et al., 2020). Using the R package GSAQ, gene sets as dictated by SOM and GMM clusters were compared to the entire collection of pod filling associated QTL (Das, 2016). Due to the imprecision in the position of two QTL, estimations were forced to be made regarding their distal genome range. A p-value of ≤ 0.05 reported by GSAQ determined whether there was significant overlap between a cluster and the QTL collection.

Functional Pathway Analysis

Orthologs to *Arachis hypogaea* in *Arabidopsis thaliana* were determined using Orthofinder 2.5.2 on a CentOS v7.9 command line provided by the University of Georgia: Georgia Advanced Computing Resource Center (Emms and Kelly, 2015, 2019). *Arachis hypogaea* genome version 1.0 was used in conjunction with the *Arabidopsis thaliana* Araport11 genome (Cheng et al., 2017; Bertioli et al., 2019). SOM and GMM clusters with significant overlap with pod filling associated QTL were then converted to orthologs in *A. thaliana.* Protein interaction networks were then computed using R based searches of StringDB (Snel et al., 2000; von Mering et al., 2003, 2005; Franceschini et al., 2016; Szklarczyk et al., 2019, 2021). A threshold p-value ≤ 0.05 of interactions estimated by StringDB was used to filter significant interaction networks.

Results

Differential Expression Analysis

The principal intention of differential expression analysis is to understand the genes or gene sets whose expression levels differ across a factor level gradient. This analysis of peanut

pods attempts to determine these values across three discrete sample factor groups. These are the plant's genotype, the sample's tissue, and its reproductive stage (Figures 2.2-2.5). Due to the methodology of DESeq2, the reported number of differentially expressed genes based on cut-off values of Benjamini-Hochberg corrected p-values ≤ 0.05 and LFC ± 1 was able to be explored through multiple factorial comparisons. These results reframe the conception of differentially expressed genes to evaluate changes based on differences in factor levels across samples (Table 2.1). What the "complete model" is describing is the difference in gene expression between the final levels of each factor and the other levels. Therefore, the model shows where gene expression has either increased or decreased compared to the factor levels of: "Tifrunner", "R7' and "seedcoat". This is against the baseline factor levels of: "NC 3033", "R4-5", and "embryo". Therefore, a differentially expressed gene is determined to have had a detectable change in expression from the baseline factor levels to the first set of factor levels. This change is enumerated in the reported fold change of differentially expressed genes. Contrast models then refine the differential expression space to pairwise sets of factor levels. Here, differential expression is based on higher expression of a gene within one of the factor levels. This explains why we observe differences in differentially expressed genes between the complete model and the contrast models. Analysis of p-values reported by DESeq2 lead to the justification for using Benjamini-Hochberg correction for adjusted p-values, the histogram showed a clear bias towards p-values near zero indicating the need for multiple test correction (Figure 2.6). Volcano plots were used to visualize the p-value and LFC cut-offs, the set of Benjamini-Hochberg corrected p-values ≤ 0.05 and LFC ± 1 was plotted and visualized to check for any noticeable patterns or unusual results (Figure 2.7).

Principle component analysis of samples demonstrated that between each of the three factor groups, the tissue factor had the most influence on the variance present in a sample based on PC1 and PC2. These two P.C.s contributed >80% of the total variance present (Figure 2.8).

Simple k-means clustering of the samples and differentially expressed genes of the complete model were plotted in a scaled heatmap against each other (Figure 2.9). From this, it became clear that the influence of tissue factor levels observed in the PCA was preserved. Generally, samples clustered together based more on tissue level identity than by either genotype or reproductive stage. The clustering of differentially expressed genes was much more cryptic. Some gene branch clusters were observed to have slightly variable expression levels corresponding with branch clusters in samples. Only in the case of two branches were there appreciable patterns in the expression of genes across RNA-seq samples. These seemed to correspond with tissue factor level.

All 294 differentially expressed genes from the complete differential expression model were matched with their Gene Ontology (GO) annotations from PeanutMine. Enrichment of GO terms across all main categories (Biological Processes (BP), Cellular Components (CC), and Molecular Functions (M.F)) was measured, and the top 10 enriched terms from each main category were selected and plotted (Ashburner et al., 2000). This showed a clear bias in these genes towards terms relating to BP and MF It should be noted that the most enriched term was "gene expression," relating to the regulation of genes through transcription factors. The subsequent seven terms relate to the creation or degradation of molecular compounds (Figure 2.10).

Self-Organizing Map Clustering

Self-organizing maps were evaluated based on graphical outputs demonstrating distributions of genes, the distance between nodes, and network convergence over algorithmic iterations. Examples of expected results are presented from the SOM evaluation of the full factorial model. This evaluation begins with tracking the graph convergence over the 100 iterations of the SOM algorithm. What is expected is a final plateau resulting from a downward trend seen through the graph. This means that the algorithm is finding ways to create tighter, more closely connected nodes over successive iterations, indicating that the algorithm is learning from the data (Figure 2.11). The mapping plot indicates where genes are distributed throughout the space. It is not easy to directly pull information from this plot except for apparent differences in genes in nodes (Figure 2.12). A more precise measure of the differences in numbers of genes within nodes is seen in the counts plot, which enumerates the members of each node (Figure 2.13). Relationships between nodes are represented in the neighborhood distance plot. This plot denotes the sum of the distances present between each node and its immediate neighbor nodes. A lower value for a node indicates that it is more related to the nodes surrounding it (Figure 2.14). The codes plot shows the codebook vectors initialized by the SOM algorithm. These are the values to which genes are mapped and organized around. The red graphs indicate the influence of a specific input column, each sample from the RNA-seq experiment, on a codebook vector, unfortunately due to the number of samples the charts were abstracted to a line chart instead of pie-charts showing increased influence. This restricts the interpretability of the chart plot however it is still possible to draw lines of comparison between similarly shaped codes lines which indicates that these nodes were generated on similar backgrounds (Figure 2.15). The quality plot demonstrates the intra-connectedness of the genes within a node compared to the

codebook vector. A lower value indicates that a codebook vector better represents the genes within the node (Figure 2.16). The final mapping plot shows the distribution of GMM clusters on the backdrop of the SOM. Colors denote clustered nodes; they may not be immediately adjacent. The mapping space's edges wrap in all directions, which means that a cluster may wrap from top to bottom or left to right (Figure 2.17).

To ensure consistency between the SOM results and the GMM clustering of nodes, visual comparisons were made between the mapping distance and quality plots to the colored clusters of the GMM SOM. As an example, the plots of the full differential expression model's SOM are presented. Neighbor distance plots demonstrate the inter-relatedness of adjacent nodes, while the quality plot shows the intra-connectedness of a node. Therefore, when analyzing GMM clusters, there should be similarities between regions of low neighbor distance and larger clusters. This is observed in the bright red cluster present through the left side of the map (Figure 2.14; Figure 2.17). Therefore, there is agreement in the SOM results and the GMM results. Looking at much greater neighbor distances and where there is tight intra-connectedness, the opposite should be observed (Figure 2.14; Figure 2.16). This trend can be seen in the purple, gray, and light orange clusters of the GMM map (Figure 2.17).

Working with tissue contrasts proved difficult as there were many differentially expressed genes between pod tissues. However, SOM and GMM plots for all other contrasts are included. The genotype contrast created a well-connected map containing seven clusters (Figure 2.18). Each of the stage contrast groups produced noticeably different clustering. They varied in cluster number from six in the R4-5 to R6 contrast, to nine in the R4-5 to R7 contrast, and eight in the R6 to R7 contrast (Figures 2.19-2.21). The map from R4-5 to R6 appears to be the most contiguous, only two nodes are not directly adjacent to the other members of their clusters. This

indicates that the SOM and GMM had the easiest time working with the data from that contrast almost certainly due to the far fewer differentially expressed genes, only 148 compared to 7418 and 3568 in the other two.

Plots of biased expression between factor levels within contrast models allowed for separation of the SOM clusters based on whether there was biased gene expression and to which factor level it was biased. This data is only included for the genotype and stage contrasts as the tissue factors had too many differentially expressed genes to work effectively with these analyses. From the seven clusters of the genotype contrast four show bias towards one factor level. Three are biased towards Tifrunner those would be clusters two, three, and six. The only cluster biased towards NC 3033 is cluster 4 (Figure 2.22). The expression bias of the contrast between R4-5 and R6 is much more pronounced with only a single cluster, cluster two, sharing much expression between the two (Figure 2.23). All other clusters have the majority of differentially expressed genes being in stage R4-5. This is not replicated in the contrast between R4-5 and R7, here of the nine total clusters three have biased expression all favoring R7 (Figure 2.24). A similar trend is observed in the final contrast between R6 and R7 where three of the eight clusters show bias towards R6, none towards R7 (Figure 2.25).

Quantitative Trait Loci Analysis

Each Self-organizing map cluster was compared to 52 QTL generated from a recombinant inbred line (RIL) population created from an initial cross between Tifrunner and NC 3033 (Chavarro et al., 2020; Table 2.2). Clusters that did not clear the significance p-value $\leq$ 0.05 were rejected from further testing. The entire search returned 12 total clusters with significant overlaps with QTL. Analysis of individual clusters from the complete model was reduced to just the 2 clusters (clusters 3 and 4) with significant overlap to QTL from the full

factorial model (Table 2.3). Cluster 3 contained 42 genes 19 of which were found not to have any useful gene or gene family descriptions (either in the complete lack of description, unknown protein, or uncharacterized protein family). One of which, "MW6VFJ" overlaps with QTL. Three other genes overlap with QTL, these were "6UGQ34", "TJI57B", and "EB1RPN". Cluster 4 contained 61 genes, here only 9 genes were without useful description. 9 genes were also associated with QTL. 2 genes were associated with ribosomes or ribonucleoproteins "V6RPLX" and "MZ9EG4" which also overlap the same QTL set. This set contains QTL associated with seed weight, double pods, and pod area (Table 2.4).

Functional Analysis

Orthofinder was used to compute the orthologs of *Arachis hypogaea* in *Arabidopsis thaliana*. Genes within clusters significantly overlapped with QTL were then converted to their orthologs in *A. thaliana*. StringDB networks were plotted, and p-values were calculated based on the number of edges between nodes exceeding expected random amounts. Across the complete model and all contrasts only 12 SOM clusters were found to have significant associations with QTL. Using these 12 clusters, networks of protein interactions were generated. This set was reduced based on a p-value ≤ 0.05 cutoff based on reported significance of edge numbers from StringDB resulting in only 7 networks. One significant network was from the complete model, one was from the genotype contrast, one from the contrast model between the pericarp and seedcoat factor levels, and four from stage contrasts. Of the four, one was from the R4-5 to R6 contrast, one from the R4-5 to R7 contrast and two from the R6 to R7 contrast. From the complete model only cluster 4 created a significant network. This network contains two clusters, one with eleven protein interactions, and one with a single interaction (Figure 2.26). The network from the genotypes contrast contained a central grouping of five genes with interactions. The

tissue contrast network provided 77 total interactions with a strong central interaction cluster. The smallest network is attributed to R4-5 to R6 contrast, it only consists of two interactions between the six genes within the network. The R4-5 to R7 network contained 277 interactions, clustering was difficult to discern taking on a more tree-like pattern than other networks. Many of the genes in the network do not have interactions. The two networks of the R6 R7 contrast are similar in size, one containing 219 interactions, the other with 296 interactions. The smaller network contains one strong central cluster, the larger contains two networks with some interactions between them.

Discussion

Complications of Differential Expression Analysis

Differential expression analysis provides insight into the potential candidate genes that influence the phenotype of an organism. Using a multi-factorial sampling design may accurately describe the biology of a phenotype however, it makes computational analyses difficult. Examining the results of this differential expression analysis shows that some factors have a much stronger influence on a gene's expression levels than others. The most obvious is the differences in gene expression observed in contrast models between tissue types. These yielded tens of thousands of differentially expressed genes far greater than the hundreds or thousands observed in other contrast models. The contribution of these factors can also be observed in the sample PCA. Here, three distinct clusters are observed with clear identities belonging to the three tissue factor levels (Figure 2.8). This indicates that the impact of tissue levels is more significant on the variance between samples than any other factor. This is consistent with the observations made in differentially expressed gene numbers between contrast groups. However, the complete model provided a much more succinct 294 differentially expressed genes. The reason for this is

almost certainly due to the distinct differences between the physical characters of tissue types. The pericarp, seedcoat, and embryo each have distinct biological roles which are not replicated in the other tissues. Logically then, variation in between the genes expressed in these tissues should be great. However, this variation may not play as great of a role in pod filling as the other factor groups do. Biologically, what almost certainly is true is that concerted efforts by all aspects of peanut pods contribute to differential pod phenotypes. While the roles of individual factors are difficult to parse from the full differential expression model, the smaller contrast model could be used to inform the search of the complete model differentially expressed genes list. Therefore, a continuation of this analysis could come from determining which genes identified by the more complex complete model are included in the contrast sets. What this may demonstrate is a common set of genetic factors that differentiate pod phenotypes.

Further Evaluation of Differentially Expressed Genes

A common problem throughout this analysis was the lack of distinct annotations to base recommendations of potential genes to validate. This was seen primarily in two areas; these were Gene Ontology annotations and in available gene descriptions. In many cases, even when there was data present, the data on hand was too vague to distinctly point out any individual genes as being probable candidates for influencing pod filling. Perhaps with more, or better descriptions this would not be the case, but this is not a realistic scenario. What should be undertaken is an evaluation of currently known orthologs contained within better understood related organisms. For example, take the known seed filling pathways in *Glycine max* or *Arabidopsis thaliana*, compute the orthologous genes in *Arachis hypogaea* and then compare these orthologs to sets of differentially expressed genes. One possible problem with this approach is that it may miss genes who do not possess detectable orthologs in those species or genes whose role within the peanut

genome has altered since species divergence. What may help with these problems are the SOM clusters generated here. The genes contained within SOM clusters have been grouped based on similar expression patterns throughout the RNA-seq samples. Similar patterns indicate that genes may be behaving in a coordinated fashion, thus although they may not align with seed filling genes in related species, they may still have a role in peanut.

Expansion of Analysis

Clusters were used for functional pathway analysis. Previous studies established QTL related to the pod filling phenotype; these were used to isolate the search space from all clusters to only ones with significant overlap with QTL. Due to limitations in the organisms with compatibility to current functional pathway tools, the genes of *Arachis hypogaea* needed to be converted to their orthologs in *Arabidopsis thaliana*. This introduces a new level of abstraction in the pathway mapping; however, it gives a far greater understanding of the potential interactions of these genes strictly because of the large amount of work previously done in *A. thaliana.* Significant pathway maps will need to be converted back into *A. hypogaea* genes for further analysis. Further analysis is necessary for this process only highlights potential genes involved with pod filling. However, the role they play needs to be experimentally validated. This is only possible in some cases. Some of the significant gene interaction networks generated by StringDB are far too large to reasonably assess with current technologies. The smaller network generated by the complete model then is a good candidate for beginning these analyses.

This undertaking aimed to narrow the scope of study for determining the genes that control the pod filling phenotype of peanut. The eventual goal of studying this phenotype is to determine a validated set of genes which could act as breeding targets for plant breeders working with cultivated peanut. The pod filling phenotype is a crucial component of market preference

for *Arachis hypogaea* and is a limiting factor in integrating certain genotypes in plant breeding pipelines. Genotypes such as NC 3033 do possess desirable characteristics, such as disease resistances. However, the risk of crossing a line like it with another line such as Tifrunner is that undesirable traits make the offspring of the cross less viable as a marketable product. While this study did succeed in evaluating potential gene clusters that may have a part in controlling the pod filling trait. More study is needed with the eventual aim of having a discrete pathway or gene set to act as a guide for other work.

References

Alexa, A., and J. Rahnenfuhrer. 2021. topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.13).

Alpaydin, E. 2014. Introduction. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 1–20

Arya, S.S., A.R. Salve, and S. Chauhan. 2016. Peanuts as functional food: a review. J Food Sci Technol 53(1): 31–41. doi: 10.1007/s13197-015-2007-9.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25(1): 25–29. doi: 10.1038/75556.

Bertioli, D.J., S.B. Cannon, L. Froenicke, G. Huang, A.D. Farmer, et al. 2016. The genome sequences of Arachis duranensis and Arachis ipaensis , the diploid ancestors of cultivated peanut. Nat Genet 48(4): 438–446. doi: 10.1038/ng.3517.

Bertioli, D.J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, et al. 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nature Genetics 51(5): 877–884. doi: 10.1038/s41588-019-0405-z.

Beute, M.K., J.C. Wynne, and D.A. Emery. 1976. Registration of NC 3033 Peanut Germplasm1 (Reg. No. GP 9). Crop Science 16(6): cropsci1976.0011183X001600060046x. doi: 10.2135/cropsci1976.0011183X001600060046x.

Blighe, K., S. Rana, and M. Lewis. 2021. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Boote, K.J. 1982. Growth Stages of Peanut (Arachis hypogaea L.)1. Peanut Science 9(1): 35–40. doi: 10.3146/i0095-3679-9-1-11.

Buccitelli, C., and M. Selbach. 2020. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21(10): 630–644. doi: 10.1038/s41576-020-0258-4.

Carbon, S., E. Douglass, B.M. Good, D.R. Unni, N.L. Harris, et al. 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research 49(D1): D325–D334. doi: 10.1093/nar/gkaa1113.

Chavarro, C. 2021. Personal Correspondence.

Chavarro, C., Y. Chu, C. Holbrook, T. Isleib, D. Bertioli, et al. 2020. Pod and Seed Trait QTL Identification To Assist Breeding for Peanut Market Preferences. G3: Genes, Genomes, Genetics 10(7): 2297–2315. doi: 10.1534/g3.120.401147.

Cheng, C.-Y., V. Krishnakumar, A.P. Chan, F. Thibaud-Nissen, S. Schobel, et al. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal 89(4): 789–804. doi: 10.1111/tpj.13415.

Costa-Silva, J., D. Domingues, and F.M. Lopes. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE 12(12): e0190152. doi: 10.1371/journal.pone.0190152.

Das, S. 2016. GSAQ: Gene Set Analysis with QTL.

Dash, S., E.K.S. Cannon, S.R. Kalberer, A.D. Farmer, and S.B. Cannon. 2016. Chapter 8 - PeanutBase and Other Bioinformatic Resources for Peanut. In: Stalker, H.T. and F. Wilson, R., editors, Peanuts. AOCS Press. p. 241–252

Emms, D.M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16(1): 157. doi: 10.1186/s13059-015-0721-2.

Emms, D.M., and S. Kelly. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20(1): 238. doi: 10.1186/s13059-019-1832-y.

FAO. 2019. Seeds Food and Agriculture Organization of the United Nations. http://www.fao.org/seeds/en/ (accessed 23 June 2021).

Fraley, C., A.E. Raftery, L. Scrucca, T.B. Murphy, and M. Fop. 2020. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation.

Franceschini, A., J. Lin, C. von Mering, and L.J. Jensen. 2016. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. Bioinformatics 32(7): 1085–1087. doi: 10.1093/bioinformatics/btv696.

Gupta, K., O. Buchshtab, and R. Hovav. 2014. The Effects of Irrigation Level and Genotype on Pod-Filling Related Traits in Peanut (Arachis hypogaea). Journal of Agricultural Science 7(1): p169. doi: 10.5539/jas.v7n1p169.

Hajduch, M., L.B. Hearne, J.A. Miernyk, J.E. Casteel, T. Joshi, et al. 2010. Systems Analysis of Seed Filling in Arabidopsis: Using General Linear Modeling to Assess Concordance of Transcript and Protein Expression. Plant Physiology 152(4): 2078–2087. doi: 10.1104/pp.109.152413.

Halvey, J., A. Hartzook, and T. Markovitz. 1987. Foliar fertilization of high-yielding peanuts during the pod-filling period. Fertilizer Research 14: 153–10.

Haykin, S.S. 2009a. Self-Organizing Maps. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 425–470

Haykin, S.S. 2009b. Neural networks and learning machines. 3rd ed. Prentice Hall, New York.

Haykin, S.S. 2009c. Principal-Component Analysis. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 367–418

Holbrook, C.C., and A.K. Culbreath. 2007. Registration of ‘Tifrunner’ Peanut. Journal of Plant Registrations 1(2): 124–124. doi: 10.3198/jpr2006.09.0575crc.

Kalderimis, A., R. Lyne, D. Butano, S. Contrino, M. Lyne, et al. 2014. InterMine: extensive web services for modern biology. Nucleic Acids Res 42(Web Server issue): W468-472. doi: 10.1093/nar/gku301.

Kassambara, A. 2020. ggpubr: “ggplot2” Based Publication Ready Plots.

Klevorn, C.M., L.L. Dean, and S.D. Johanningsmeier. 2019. Metabolite Profiles of Raw Peanut Seeds Reveal Differences between Market-Types. Journal of Food Science 84(3): 397–405. doi: 10.1111/1750-3841.14450.

Kruisselbrink, R.W. and J. 2019. kohonen: Supervised and Unsupervised Self-Organising Maps.

Levin, J.Z., M. Yassour, X. Adiconis, C. Nusbaum, D.A. Thompson, et al. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7(9): 709–715. doi: 10.1038/nmeth.1491.

Love, M., C. Ahlmann-Eltze, K. Forbes, S. Anders, W. Huber, et al. 2021. DESeq2: Differential gene expression analysis based on the negative binomial distribution. Bioconductor version: Release (3.13).

Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12): 550. doi: 10.1186/s13059-014-0550-8.

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, et al. 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1): 258–261. doi: 10.1093/nar/gkg034.

von Mering, C., L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, et al. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33(Database issue): D433-437. doi: 10.1093/nar/gki005.

Moretzsohn, M. de C., M.S. Hopkins, S.E. Mitchell, S. Kresovich, J.F.M. Valls, et al. 2004. Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biol 4: 11. doi: 10.1186/1471-2229-4-11.

Nussinov, R. 2015. Advancements and Challenges in Computational Biology. PLOS Computational Biology 11(1): e1004053. doi: 10.1371/journal.pcbi.1004053.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Scrucca, L., M. Fop, T.B. Murphy, and A.E. Raftery. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8(1): 289–317.

Shiraiwa, T., N. Ueno, S. Shimada, and T. Horie. 2004. Correlation between Yielding Ability and Dry Matter Productivity during Initial Seed Filling Stage in Various Soybean Genotypes. Plant Production Science 7(2): 138–142. doi: 10.1626/pps.7.138.

Smith, R.N., J. Aleksic, D. Butano, A. Carr, S. Contrino, et al. 2012. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28(23): 3163–3165. doi: 10.1093/bioinformatics/bts577.

Snel, B., G. Lehmann, P. Bork, and M.A. Huynen. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28(18): 3442–3444. doi: 10.1093/nar/28.18.3442.

Szklarczyk, D., A.L. Gable, D. Lyon, A. Junge, S. Wyder, et al. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1): D607–D613. doi: 10.1093/nar/gky1131.

Szklarczyk, D., A.L. Gable, K.C. Nastou, D. Lyon, R. Kirsch, et al. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49(D1): D605–D612. doi: 10.1093/nar/gkaa1074.

Toomer, O.T. 2018. Nutritional chemistry of the peanut (Arachis hypogaea). Crit Rev Food Sci Nutr 58(17): 3042–3053. doi: 10.1080/10408398.2017.1339015.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1): 57–63. doi: 10.1038/nrg2484.

Wehrens, R., and L.M.C. Buydens. 2007. Self- and Super-Organizing Maps in R: The kohonen Package. Journal of Statistical Software 21(5): 1–19. doi: 10.18637/jss.v021.i05.

Wehrens, R., and J. Kruisselbrink. 2018. Flexible Self-Organizing Maps in kohonen 3.0. Journal of Statistical Software 87(7): 1–18. doi: 10.18637/jss.v087.i07.

Wickham, H. 2016. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.

Yin, S., P. Li, Y. Xu, J. Liu, T. Yang, et al. 2020. Genetic and genomic analysis of the seed-filling process in maize based on a logistic model. Heredity 124(1): 122–134. doi: 10.1038/s41437-019-0251-x.

Table 2.1: Differentially Expressed Genes by Model Type.

Differential Expression Model	Number of Differentially Expressed Genes
Complete Model	294
Genotype Contrast NC 3033 to Tifrunner	825
Stage Contrast R4-5 to R6	148
Stage Contrast R4-5 to R7	7418
Stage Contrast R6 to R7	3568
Tissue Contrast Pericarp to Embryo	18287
Tissue Contrast Pericarp to Seedcoat	15933
Tissue Contrast Seedcoat to Embryo	17488

Table 2.2: Quantitative Trait Loci Used in Self Organizing Map Analysis. Trait abbreviations are as follows: [a] 16/64 percentage; [b] kernel percentage; [c] pod weight; [d] seed weight; [e] single-kernel pods; [f] double-kernel pods; [g] pod area; [h] pod density. Table and data adapted from Chavarro et al., 2020.

Trait	QTL	Chromosome	Physical Position Proximal (Mbp)	Physical Position Distal (Mbp)
16/64P[a]	q16/64PA02.1	Aarhy.02	96.45	97.6
	q16/64PA06.1	Arahy.06	4.41	4.92
	q16/64PA10_B04.1	Arahy.14	80.28	101.14
	q16/64PA10_B04.1	Arahy.10	80.28	101.14
KP[b]	qKPA01.1a	Arahy.01	12.33	24.65
	qKPA01.1b	Arahy.01	11.43	48.53
	qKPA07_B07.1	Arahy.17	0.83	1.2
	qKPA07_B07.1	Arahy.07	0.83	1.2
	qKPA03_2.2	Arahy.03	0.06	0.56
	qKPA06.2	Arahy.06	1.32	2.3
	qKPA07_B07.2a	Arahy.17	0.83	1.2
	qKPA07_B07.2b	Arahy.17	1.1	1.2
	qKPA07_B07.3a	Arahy.17	0.83	1.03
	qKPA07_B07.3b	Arahy.17	1.1	1.2
	qKPA07_B07.2a	Arahy.07	0.83	1.2
	qKPA07_B07.2b	Arahy.07	1.1	1.2
	qKPA07_B07.3a	Arahy.07	0.83	1.03
	qKPA07_B07.3b	Arahy.07	1.1	1.2
PW[c]	qPWA04.2a	Arahy.04	126.38	128.54
	qPWA04.2b	Arahy.04	125.1	128.54
	qPWA07_B07.2	Arahy.17	0.83	1.2
	qPWA07_B07.2	Arahy.07	0.83	1.2
	qPWB06_1.2	Arahy.06	146.38	150.86
	qPWA04.3	Arahy.04	126.38	128.54
	qPWB06_1.3	Arahy.06	146.38	150.86
SdW[d]	qSdWA04.2a	Arahy.04	126.38	128.54
	qSdWA04.2b	Arahy.04	125.1	128.54
	qSdWA07_B07.2	Arahy.17	0.63	1.03
	qSdWA07_B07.2	Arahy.07	0.63	1.03
	qSdWB06_1.2	Arahy.16	146.38	150.86
	qSdWA04.3	Arahy.04	126.38	128.54
	qSdWA07_B07.3	Arahy.17	0.63	1.03

	qSdWA07_B07.3	Arahy.07	0.63	1.03
	qSdWB06_1.3	Arahy.16	146.38	150.86
SP[e]	qSPA04.2	Arahy.04	2.66	5.05
	qSPA05_B05.2	Arahy.05	96.58	97.34
	qSPA05_B05.2	Arahy.15	96.58	97.34
	qSPB06_1.2	Arahy.16	140.36	146.39
	qSPB09.2	Arahy.19	150.29	158.02
	qSPA04.3a	Arahy.04	2.66	4.96
	qSPA04.3b	Arahy.04	4.93	5.37
	qSPB01.3	Arahy.11	48.77	99.07
	qSPB09.3	Arahy.19	152.08	158.02
DP[f]	qDPA07_B07.2	Arahy.17	0.83	1.2
	qDPA07_B07.2	Arahy.07	0.83	1.2
	qDPB06_1.2	Arahy.16	146.38	150.86
	qDPA04.3	Arahy.04	126.38	128.54
	qDPB06_1.3	Arahy.16	146.38	150.86
PA[g]	qPAA04.2	Arahy.04	126.21	126.76
	qPAA07_B07.2	Arahy.07	0.63	1.03
	qPAA07_B07.2	Arahy.17	0.63	1.03
	qPAB06_1.2	Arahy.16	146.38	150.86
	qPAA04.3a	Arahy.04	126.22	126.7
	qPAA04.3b	Arahy.04	126.38	128.54
	qPAA07_B07.3	Arahy.07	0.63	1.03
	qPAA07_B07.3	Arahy.17	0.63	1.03
	qPAB03.3	Arahy.13	1.8	2.66
	qPAB06_1.3	Arahy.16	146.38	150.86
PD[f]	qPDA03_B03.2	Arahy.03	141.72	150
	qPDA03_B03.2	Arahy.13	141.72	150
	qPDA09.2	Arahy.09	108.04	111.64
	qPDA04.3	Arahy.04	120.76	125.14
	qPDB06_1.3	Arahy.16	130.49	136.39

Table 2.3: Complete model Self Organizing Map to QTL Comparison. From the complete model only two clusters were identified to be significantly associated with QTL. These were clusters 3 and 4. The "p-value" column of this table denotes the test for overlap using GSAQ, a statistical package for R. More information on QTL is present in Table 2.2. More information on genes within clusters is in Table 2.4.

Cluster	P-value	Overlapping QTL
1	1	qKPA01.1a , qKPA01.1b , qPDA04.3 , qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
2	1	qPDA03_B03.2 , qPDA03_B03.2_2
3	1.05E-16	qPDA03_B03.2 , q16/64PA10_B04.1 , qPDA03_B03.2_2, qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
4	1.26E-126	qSPA04.2 , qSPA04.3a , qPDA04.3 , qPDA04.3 , q16/64PA10_B04.1 , q16/64PA10_B04.1 , qSPB01.3 , qPDA03_B03.2_2, qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3 , qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
5	0.1007798	qKPA01.1b , qSPB09.2 , qSPB09.3
6	1	NA
7	0.9711734	NA
8	0.9711734	NA
9	1	qKPA03_2.2 , qPDA03_B03.2 , qPDA09.2

Table 2.4: Genes Within Clusters Found to be Significantly Associated with QTL. This table was created from a database search for gene family descriptions from PeanutMine. Values of "N.A." were included where there was no inclusion of gene family description. Overlapping QTL are further described in Table 2.2. SOM clusters refer to the cluster generated by SOM and GMM application to the 294 complete model differentially expressed genes. Only the clusters 3 and 4 are present for they were the only clusters significantly associated with pod filling QTL.

Gene	SOM Cluster	Gene Family	Overlapping QTL
8L03H9	3	Uncharacterised protein family (UPF0497)	
162RMF	3	unknown protein	
DMWP0X	3	proteasome assembly chaperone-like protein	
0YUQ4V	3	uncharacterized protein LOC100783517 [Glycine max]	
TJI57B	3	acyl-activating enzyme 17, peroxisomal protein, putative	qPDA03_B03.2
SG7PGU	3	SET domain-containing protein	
091D1Z	3	CASP-like protein 6 [Glycine max]	
FZWW2A	3	Unknown protein	
Z1S22E	3	plasma membrane H+-ATPase	
UB40IU	3	Small nuclear ribonucleoprotein family protein	
73J6R1	3	NA	
2C1UIX	3	NA	
IWZ446	3	Zinc-binding ribosomal protein family protein	
Y0FL4G	3	NA	
EB1RPN	3	2-oxoisovalerate dehydrogenase subunit alpha	q16/64PA10_B04.1
0FR3XP	3	NA	
Z5A5F8	3	NA	
XH2DY7	3	NA	
CGD4UQ	3	transcription factor Pcc1	

L1HKPT	3	protein NLP8-like isoform X3 [Glycine max]	
P9B1J4	3	uncharacterized protein LOC100811474 [Glycine max]	
99W4V5	3	Unknown protein	
8H36PU	3	NA	
TDQV0A	3	hypothetical protein	
HEPM80	3	Sas10/Utp3/C1D family protein	
MW6VFJ	3	uncharacterized protein LOC100783517 [Glycine max]	qPDA03_B03.2_2
0TS7M9	3	NHL domain-containing protein	
YHM9KM	3	Encodes protein with unknown function whose expression is repressed by inoculation with Agrobacterium tumerifaciens.	
XJ8ERT	3	methionine aminopeptidase 1B	
3Z4R3A	3	arabinogalactan peptide 16-like [Glycine max]	
Q2AYB8	3	uncharacterized protein LOC100779434 [Glycine max]	
6UGQ34	3	Flavin-binding monooxygenase family protein	qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
255VJ8	3	ATP binding microtubule motor family protein	
IZ2YN8	3	small nuclear ribonucleoprotein associated protein B	
9ST4S6	3	uncharacterized protein LOC100789572 isoform X1 [Glycine max]	
LSZ6GE	3	telomere repeat-binding protein 3-like isoform X1 [Glycine max]	
PJ36E7	3	NA	
BRJ78C	3	NA	
UPL4LC	3	Defensin MtDef4.7	
ED2J4B	3	lipid-A-disaccharide synthase	
GH8915	3	NA	
59G793	4	syntaxin of plants 51	
XWP5FV	4	Protein kinase superfamily protein	
DA8MAP	4	small nuclear ribonucleoprotein F	
2YK8S2	4	Ribosomal protein L39 family protein	
TT7XH0	4	cullin-associated NEDD8-dissociated protein	

50HBEA	4	unknown protein	
594W3X	4	poly(A) polymerase 1	
40255N	4	Protein of unknown function (DUF3317)	
DN4CQF	4	protein SCARECROW-like [Glycine max]	
1BZV33	4	nudix hydrolase homolog 3	
NJG8XE	4	phospholipase D P2	qSPA04.2 , qSPA04.3a
SSHM9A	4	translocon-associated protein beta (TRAPB) family protein	
SLD8LW	4	Kef-type K+ transport system, membrane component n=1 Tax=Methylophaga aminisulfidivorans MP RepID=F5SYA9_9GAMM	
MZXR1C	4	succinate dehydrogenase [ubiquinone] iron-sulfur subunit	qPDA04.3
Z4V8RQ	4	protein ARABIDILLO 1-like isoform X2 [Glycine max]	qPDA04.3
FI4HQ3	4	GDA1/CD39 nucleoside phosphatase family protein	
S449H3	4	uncharacterized protein LOC102666492 isoform X3 [Glycine max]	
9639NF	4	Protein of unknown function, DUF538	
7T3N35	4	J domain-containing protein DDB_G0295729-like isoform X2 [Glycine max]	
QX4V5J	4	NA	
KAQB8D	4	protein HASTY 1-like isoform X1 [Glycine max]	
E9IKCM	4	40S ribosomal protein S28-1	
GS8BAY	4	transducin family protein / WD-40 repeat family protein	
3ZC2CN	4	Small nuclear ribonucleoprotein family protein	
P4PJLA	4	60S ribosomal protein L38-like [Glycine max]	
G3L9EH	4	ribosomal protein S27	
C36CE0	4	zinc finger MYM-type protein 1-like [Glycine max]	
NU92WK	4	pollen protein Ole E I-like protein	
7NSI0A	4	60S ribosomal protein L38-like [Glycine max]	
F1HK63	4	60S acidic ribosomal protein family	
2F0WHB	4	Potassium transporter family protein	q16/64PA10_B04.1
G3H90Z	4	actin-11	q16/64PA10_B04.1
9M8SYH	4	exosome complex exonuclease RRP6	

FP7IK1	4	Protein kinase superfamily protein	qSPB01.3
F0HWYS	4	cullin-associated NEDD8-dissociated protein	
6AX92Y	4	DEAD-box ATP-dependent RNA helicase	
N7IJYH	4	Ribosomal protein L39 family protein	
W452X0	4	ribosomal RNA processing protein 1 homolog B-like isoform X1 [Glycine max]	
1UBI84	4	Likely PAP/25A associated domain containing protein/Poly(A) RNA polymerase cid11 n=1 Tax=Blumeria graminis f. sp. hordei (strain DH14) RepID=N1JI17_BLUG1	
YZL4IN	4	autophagy 3 (APG3) protein	
EZF9NP	4	ATP-dependent Clp protease ATP-binding subunit	
HKPY8E	4	poly(A) polymerase 1	
4217G9	4	Protein of unknown function (DUF3317)	qPDA03_B03.2_2
DU9C9C	4	ethylene-responsive transcription factor 1B	
M8JYJC	4	bZIP transcription factor bZIP109 isoform X1 [Glycine max]	
RD3H93	4	protein HASTY 1-like isoform X1 [Glycine max]	
2P4BCH	4	1-deoxy-D-xylulose 5-phosphate synthase 1	
MZ9EG4	4	Small nuclear ribonucleoprotein family protein	qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
V6RPLX	4	40S ribosomal protein S28-1	qSdWB06_1.2 , qSdWB06_1.3 , qDPB06_1.2 , qDPB06_1.3 , qPAB06_1.2 , qPAB06_1.3
V4B8PK	4	Gibberellin-regulated family protein	
S7XF8N	4	Ethylene insensitive 3 family protein	
K5AEM0	4	NA	
C1V79H	4	60S ribosomal protein L38-like [Glycine max]	
H6IXD1	4	Preprotein translocase Sec, Sec61-beta subunit protein	
J1UBKH	4	NA	
LVH8ZT	4	NA	

NNA8KD	4	Light-sensor Protein kinase n=2 Tax=Ceratodon purpureus RepID=PHY1_CERPU	
GHTC3Q	4	Low temperature and salt responsive protein family	
IT1C9S	4	60S ribosomal protein L38-like [Glycine max]	
973ZBX	4	transcription factor GTE12-like [Glycine max]	
5V04YF	4	60S acidic ribosomal protein family	
S92YTY	4	exosome complex exonuclease RRP6	

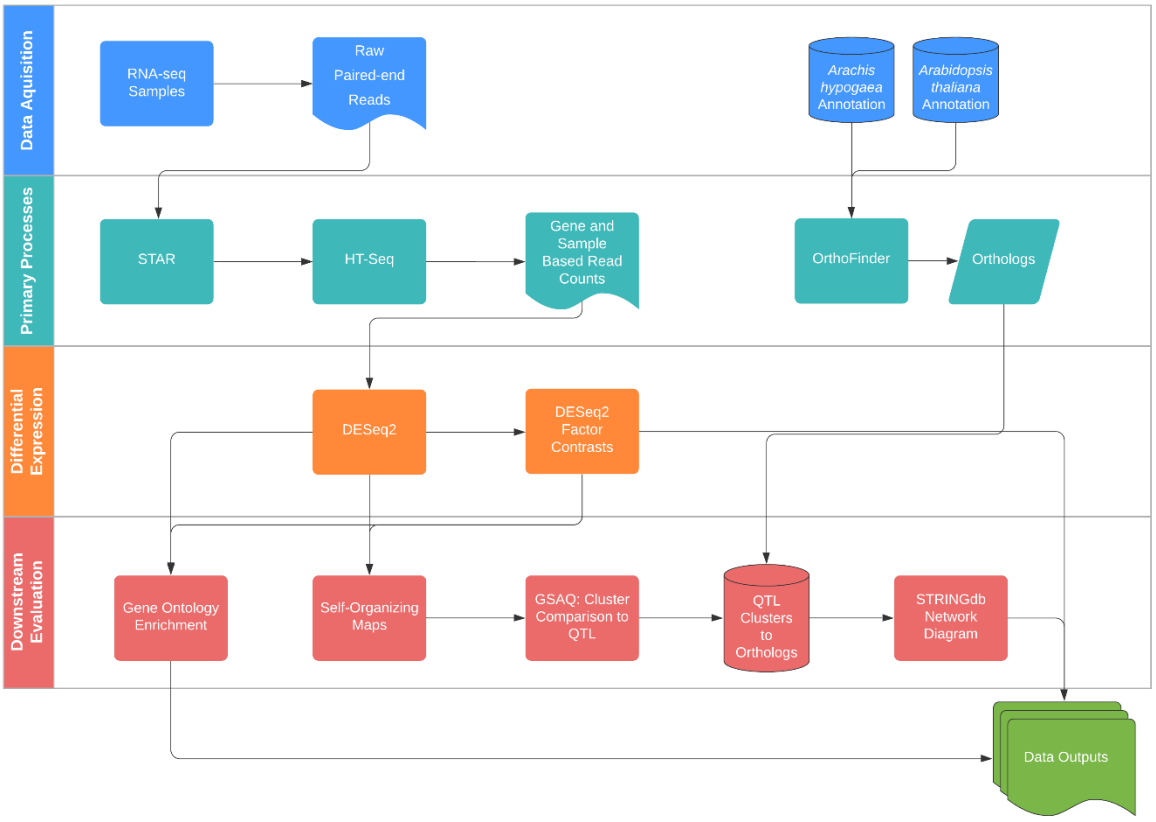

Figure 2.1: Swimming Lanes Chart of Computational Pipeline. Flow chart showing the flows of data from raw RNA-seq reads and annotation to output. Created in Lucidchart (www.lucidchart.com).

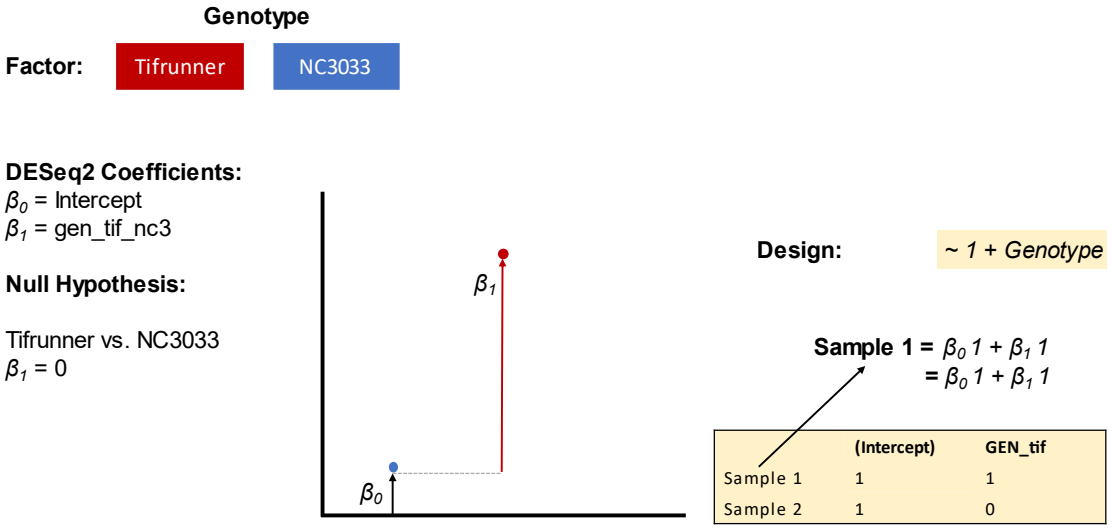

Figure 2.2: Genotype Factor Diagram. Demonstration of how the genotype factor is applied to the differential expression model. DESeq2 computes the β values based on the contribution of gene expression throughout samples with a switching term based on the presence of the Tifrunner genotype in the sample. If the β value is significantly more than zero then Tifrunner is an applicable factor in estimating the expression level of a gene (Love et al., 2014). To measure the contribution of Tifrunner, the genotype factor level NC 3033 is being held as the baseline.

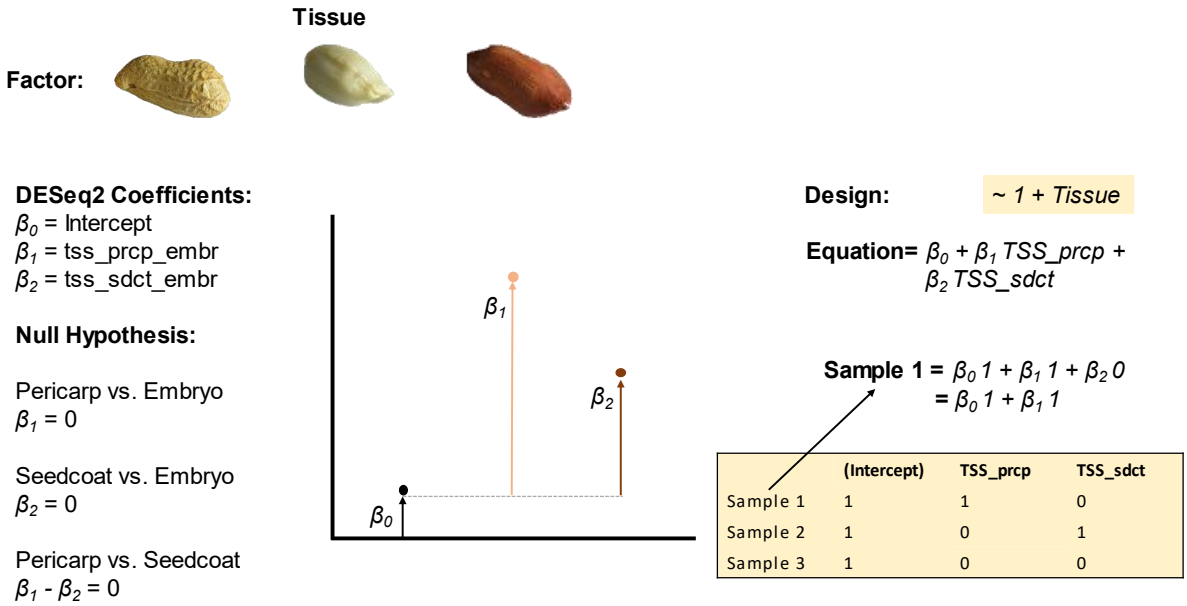

Figure 2.3: Tissue Factor Diagram. Three factor levels operate on a similar principle to the two factor levels of the genotype factor. However, here two of the factors must be measured from a common baseline tissue factor. Being used as the baseline factor level does not influence the outcome of an experiment. Here initially embryo is used as the baseline factor level to measure the effect of pericarp and seedcoat. The contrast of the non-baseline factor levels, in this case seedcoat and pericarp is measured by the differences they hold to the baseline level. B values indicate a non-zero effect of that factor on the expression level of a gene (Love et al., 2014).

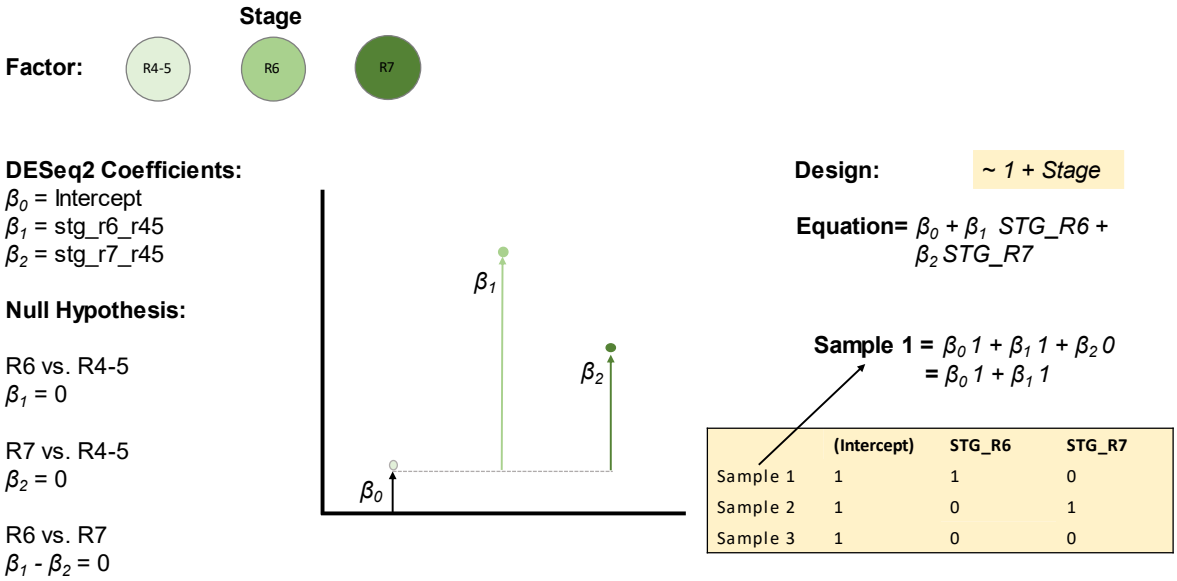

Figure 2.4: Stage Factor Diagram. The stage factor operates in the same manner as the tissue factor. Where there is an effect of the stage on the expression of a gene the β value is found not to be zero. Here R4-5 is being used as the baseline level (Love et al., 2014).

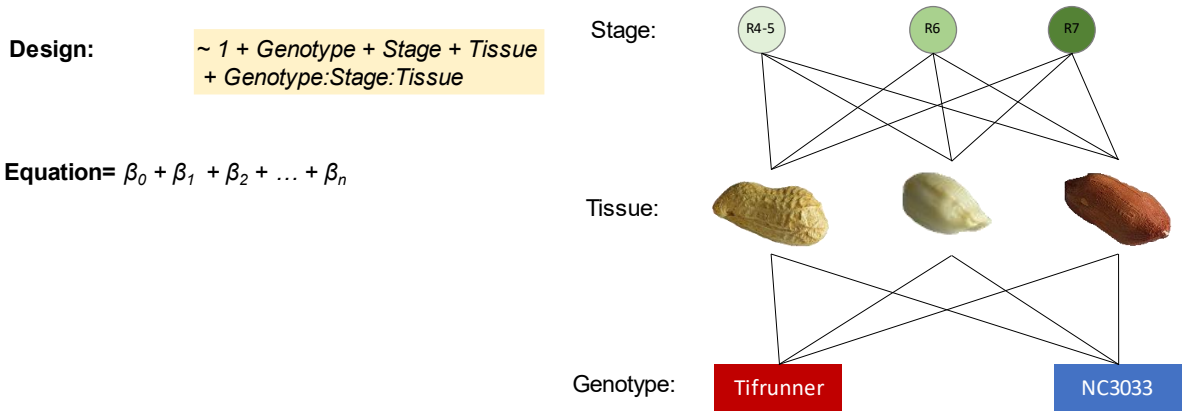

Figure 2.5: Full Differential Expression Model Diagram. A visualization of the complete model given to DESeq2. Each RNA-seq sample contains a genotype factor, a tissue factor, and a stage factor. Due to the effect of these samples being estimated from the same sample, terms for the interactions between factors and the three way genotype:tissue:stage interaction have β values in the complete model equation (Love et al., 2014, 2021, p. 2).

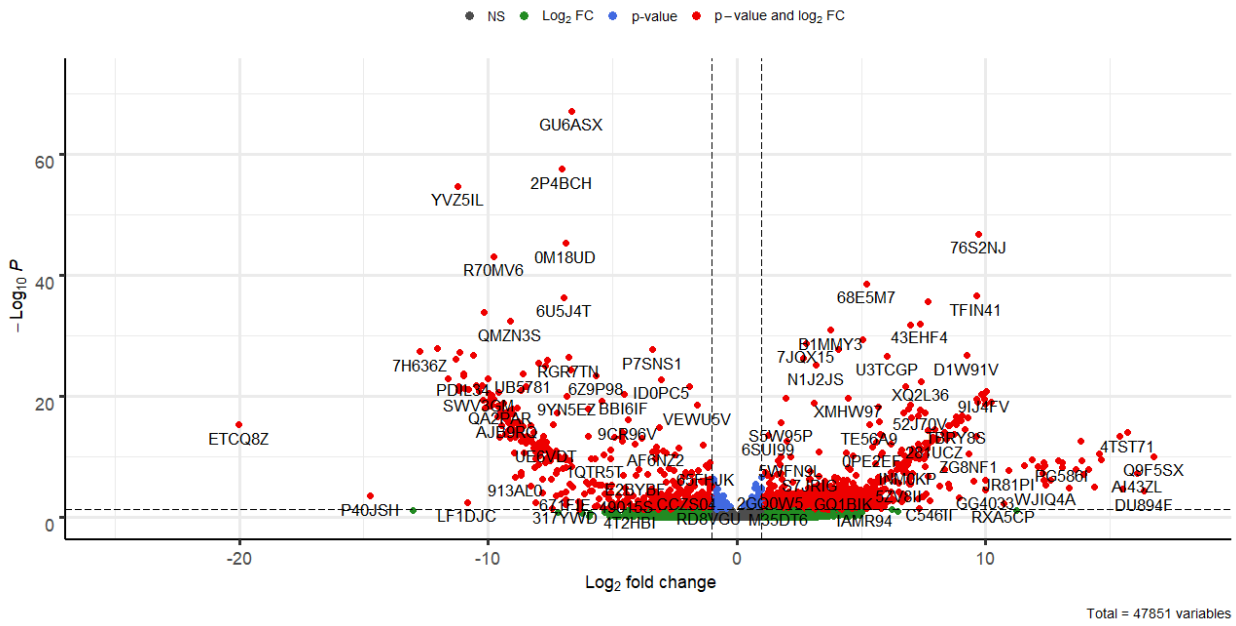

Figure 2.6: Volcano Plot of Differentially Expressed Genes. Volcano plot generated using the R package EnhancedVolcano. This plot shows genes relative to the selection criteria of corrected p-value ≤ 0.05 and LFC ± 1. From the 47851 original genes only 294 remained after criterion filtering. This dual criterion has filtered all but the red points from further evaluation. The blue region are genes that meet the p-value criteria, but do not exceed the LFC criteria. The reciprocal is true for the green region. This is an effect of the dual p-value and LFC filter, refining the gene set beyond what an independent p-value or LFC filter could (Blighe et al., 2021).

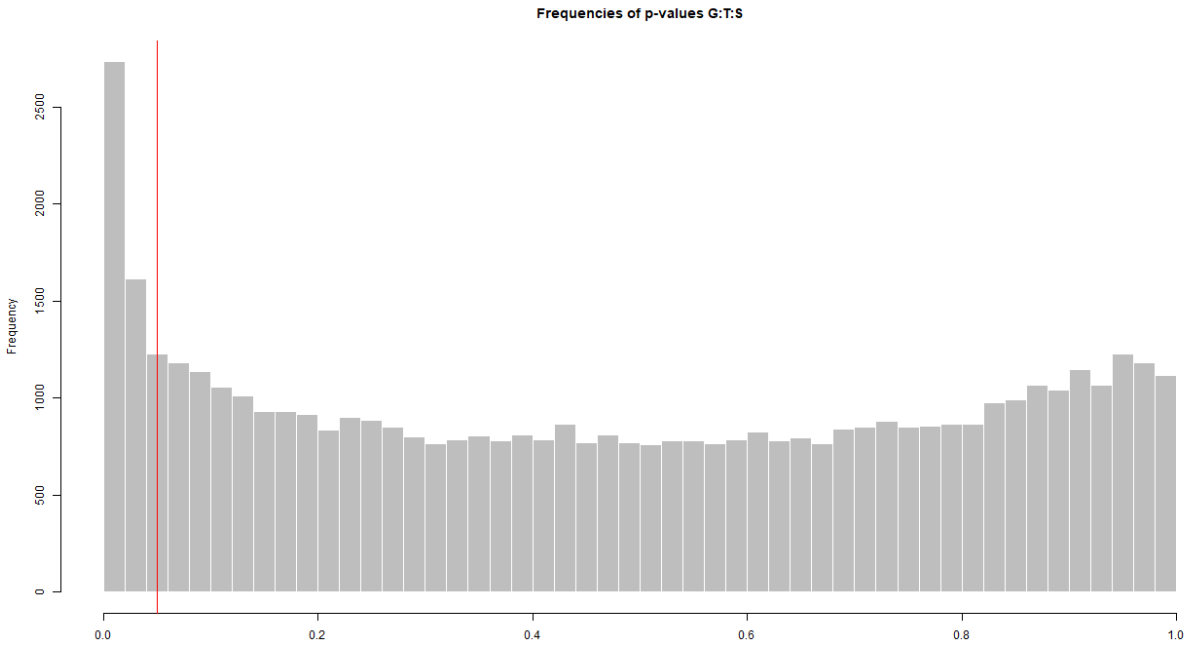

Figure 2.7: Histogram of DESeq2 P-values. Plotting of non-adjusted p-values for genes reported by DESeq2. What is observed is an overall skew and peak towards 0. This justifies the need for multiple test correction, the spike at 0 may come from false positive p-values. Without correction, approximately 5% of genes who qualify with p-value ≤ 0.05 would be false positives (Love et al., 2014).

Figure 2.8: Principal Component Analysis of RNA-seq Samples. A-C: Charts displaying the same PCA computation with PC1 on the horizontal axis and PC2 on the vertical axis. Each datapoint refers to individual RNA-seq samples. A highlights the distribution of tissues and genotypes. There appears to be random distribution of genotype within clusters which are defined by tissue. This phenomenon is echoed in plot C, with stage and tissue identified. PC1 separates the embryo tissue from pericarp and seedcoat, while PC2 pulls seedcoat from pericarp and embryo. B shows genotype and stage highlights; no discernible patterns are present. D shows the percent variance explained by the first six principal components. Over 80% of the variance in samples is explained by PC1 and PC2 (Haykin, 2009c; Wickham, 2016, p. 2; Kassambara, 2020, p. 2).

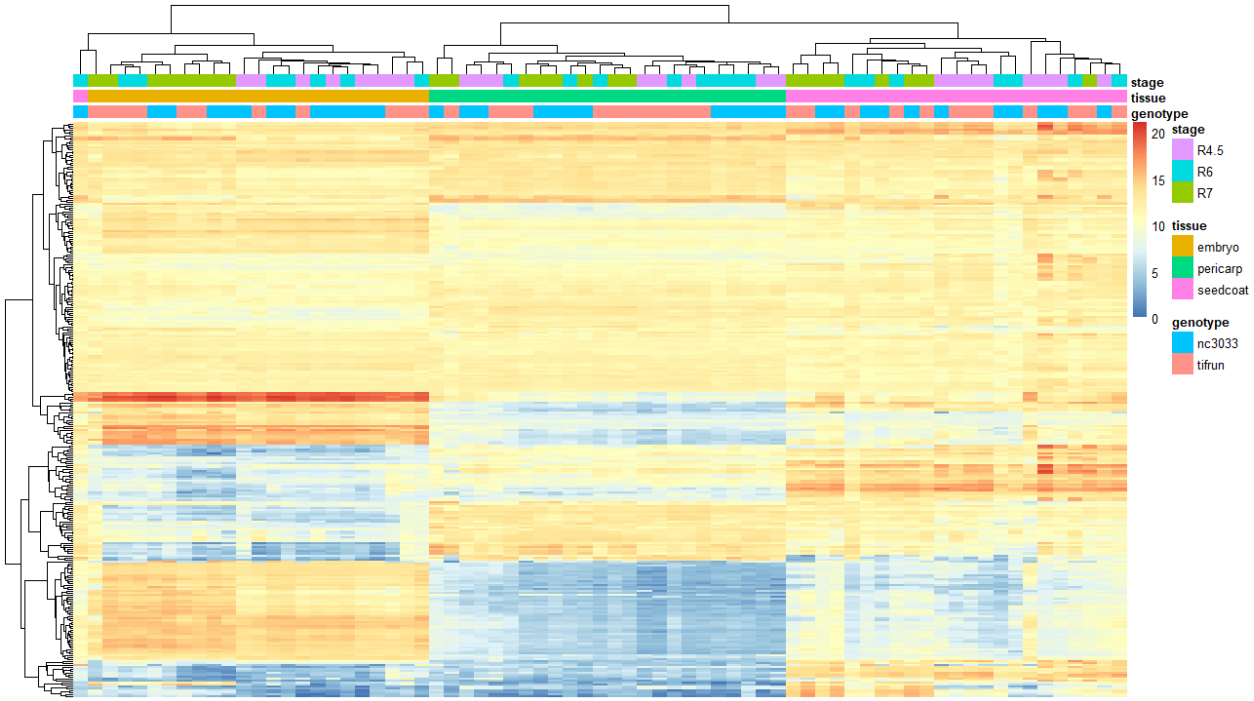

Figure 2.9: K-means Clustering of Samples and Differentially Expressed Genes. Heatmap color denotes the scaled expression value of a gene within a sample or gene. Along the horizontal axis are k-means clustering of RNA-seq samples identified by color-coded factor level. Across the vertical axis are the 294 differentially expressed genes identified by DESeq2. There appears to be a few clusters of genes which exhibit distinguishable changes in expression across samples, particularly those in the bottom third of the chart. Here a block can be observed with more expression that appears to correlate with the embryo factor level. This trend is repeated for the seedcoat factor level in the bottom-most block of genes. Samples appear to be clustered based on tissue factor level, with some possible sub-clustering based on stage and genotype throughout (Love et al., 2014, 2021, p. 2).

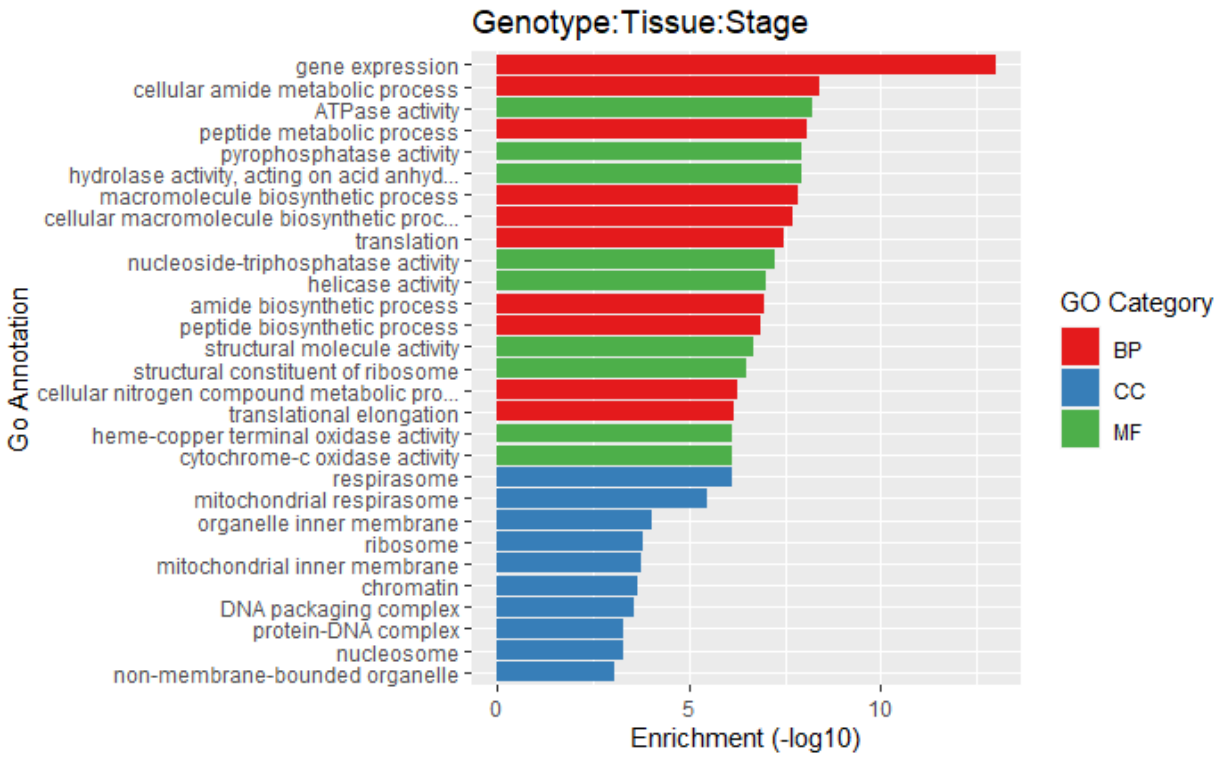

Figure 2.10: Gene Ontology Enrichment in Differentially Expressed Genes. Enrichment is based on K.S. test of adjusted p-values from differentially expressed genes. Terms were derived from Intermine (Smith et al., 2012; Kalderimis et al., 2014). There is an obvious split between the Cellular Components terms and the other two broad categories. Strong enrichment is noticed for terms involved in gene regulation and metabolite processing. This captures the most common functionality of GO annotated differentially expressed genes (Ashburner et al., 2000; Alexa and Rahnenfuhrer, 2021; Carbon et al., 2021).

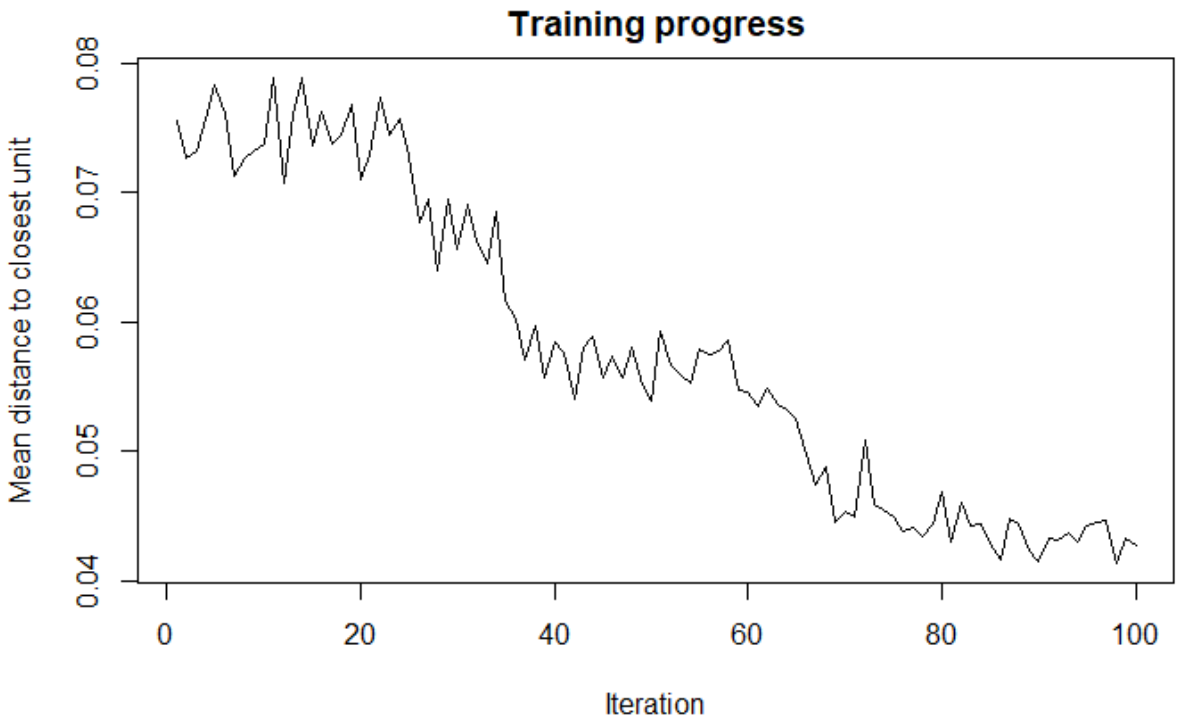

Figure 2.11: Self Organizing Map Training Chart. SOM algorithm training over the 100 iterations the program was run. What is desired is a downward trend to the line meaning that on average a gene is becoming more associated to a node over successive iterations of the SOM, indicating that the neural network is becoming better at classifying the multi-dimensional gene data (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

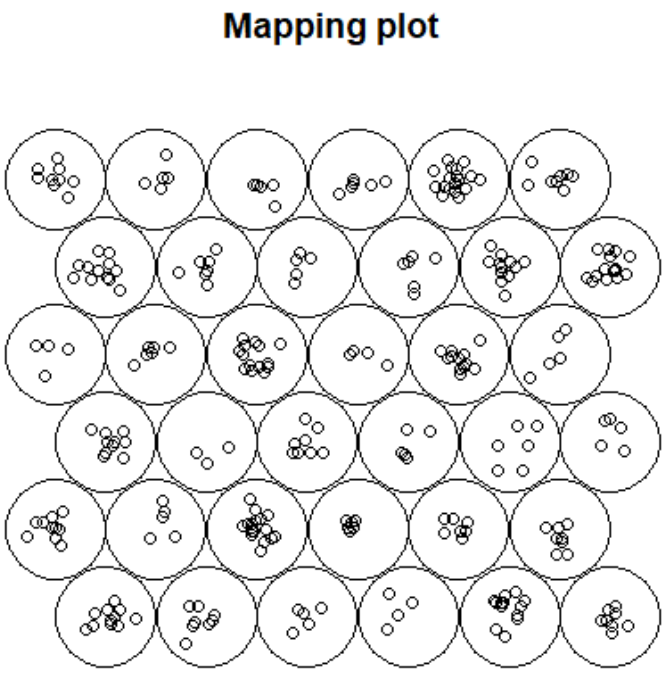

Figure 2.12: Self Organizing Map of Differentially Expressed Genes. Distribution of all 294 complete model differentially expressed genes across the 2-dimensional surface of the SOM. Each dot is a single gene, sides are continuous from left to right and top to bottom (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

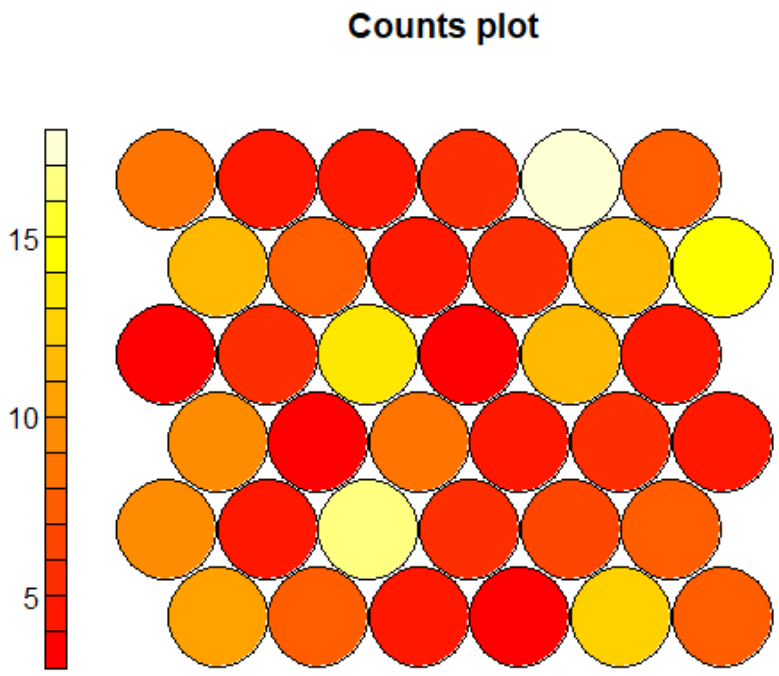

Figure 2.13: Gene Counts within Self Organizing Map. Heatmap overlaid upon the SOM enumerating the numbers of genes within each node. While many nodes contain four to 10 genes, two nodes have more than 15 (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

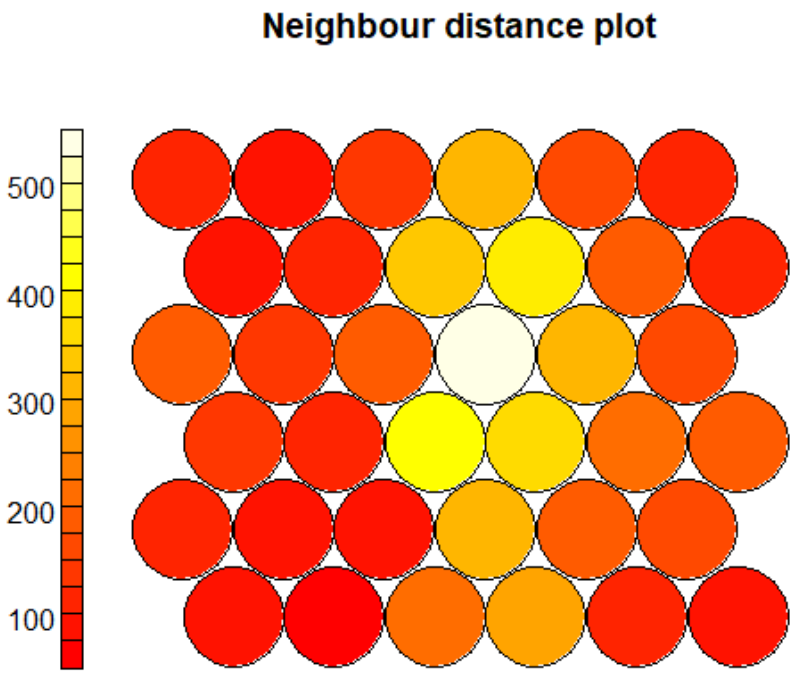

Figure 2.14: Distance to Neighbor Nodes within Self Organizing Map. This plot illustrates the similarities of each node to the six nodes immediately surrounding it. Higher values show more dissimilarity between nodes. A single node shows a distance of >500 indicating that it is very dissimilar to its neighbors (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

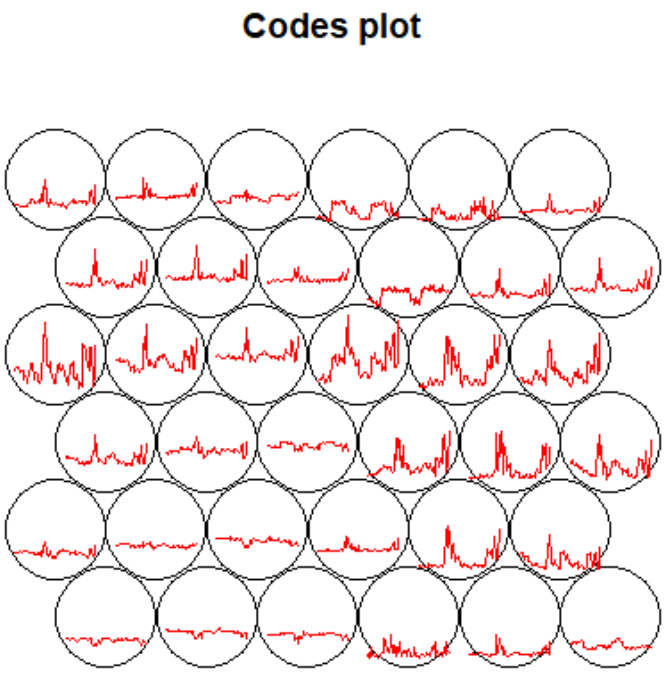

Figure 2.15: Self Organizing Map Codebook Vectors. Graphs showing the contributions of samples on the codebook vectors which define each node. Graphs are meant to be read relationally, meaning that codebook vectors are derived from more similar data the closer in shape they are. There appears to be three main types of graphs, one with a flat central line, one with numerous peaks, and one with two plateaus with low sections (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

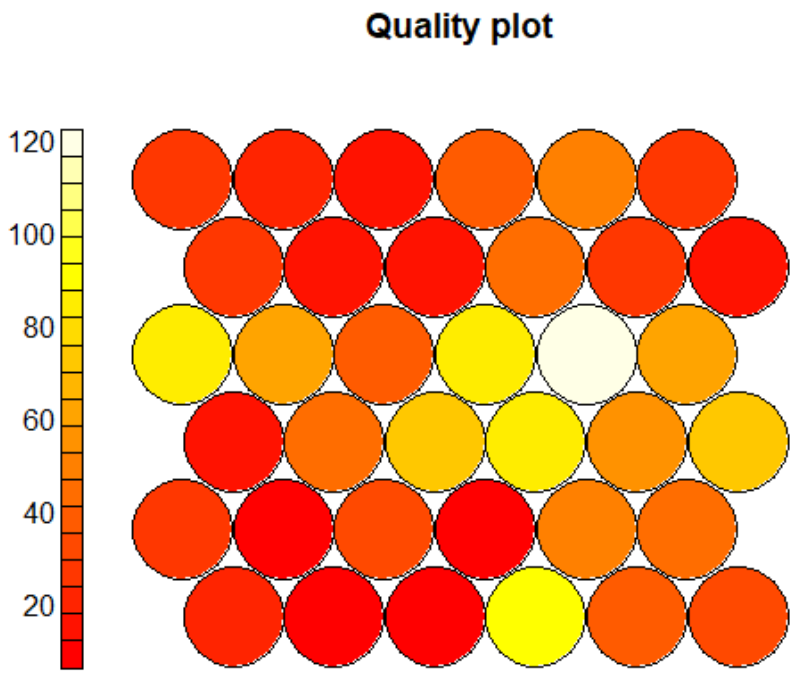

Figure 2.16: Quality of Codebook Vectors. Mean distance of genes within node to the codebook vector of the node. Lower values therefore show where the codebook vector better describes the data within the node. A single node shows very high quality, meaning poor description of the genes within the node (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

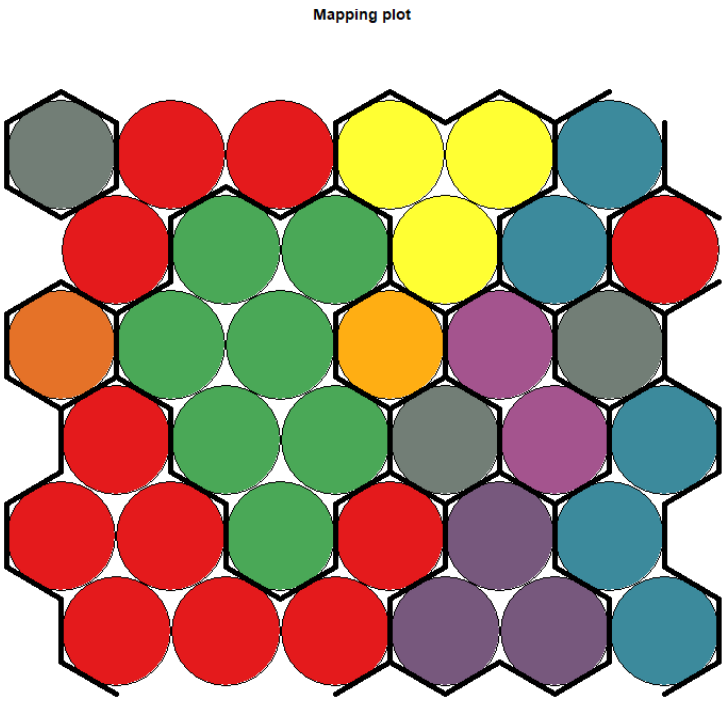

Figure 2.17: Gaussian Mixture Model Clustering of Self Organizing Map. Colors show clusters found by GMM clustering over the 2-d surface of the SOM. Edges connect across, moving across the left edge of the plane places one on the right edge, the same holds for moving from the top to the bottom. GMM found 9 clusters, they appear to agree with the findings of the previous SOM diagnostic plots. The only cluster which does not share borders with itself is the grey one. It may appear that the nodes of this cluster are farther in space than they truly are. Instead, there is only a single node separating the member nodes (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

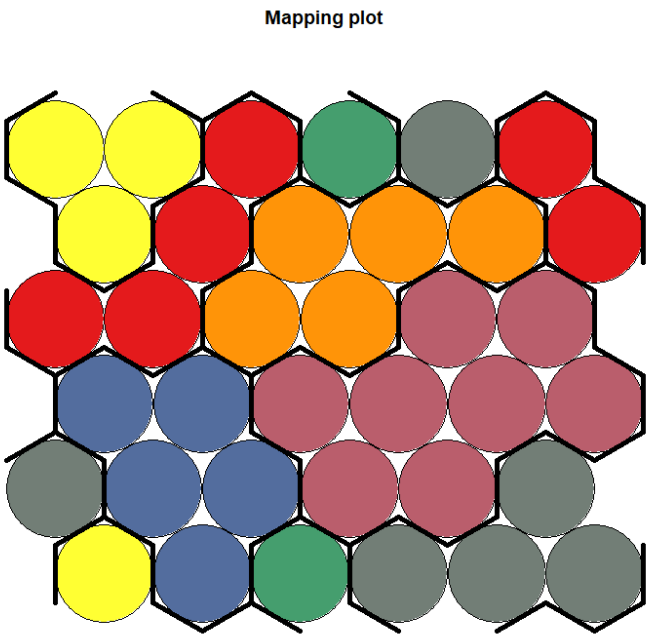

Figure 2.18: Self Organizing Map and Clustering of Genotype Contrast Tifrunner to NC 3033. Seven clusters of nodes were found. All clusters are continuous, meaning all clusters within the map share one adjacent node with other members of its assigned GMM cluster. The largest cluster contains eight nodes while the smallest has only two (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

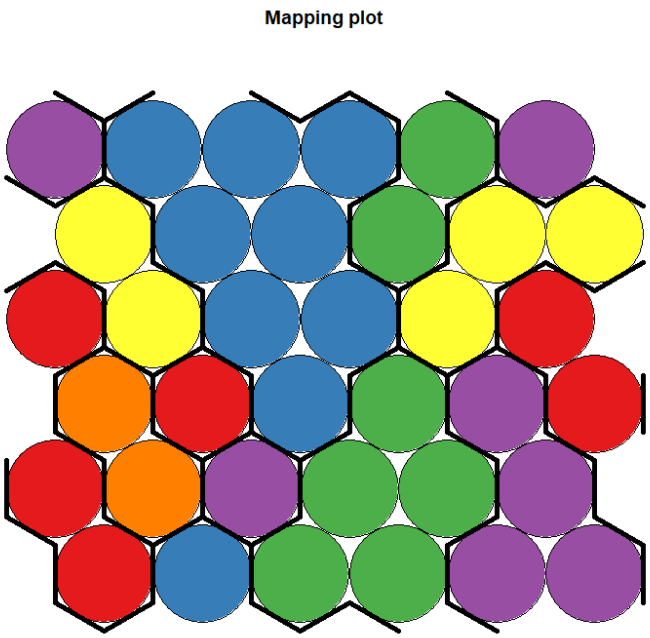

Figure 2.19: Self Organizing Map and Clustering of Reproductive Stage Contrast R4-5 to R6. Six clusters of nodes were found. Two clusters were discontinuous, purple and red, which each have a single node not adjacent to the other members of their clusters. The largest cluster contains nine nodes while the smallest has two (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

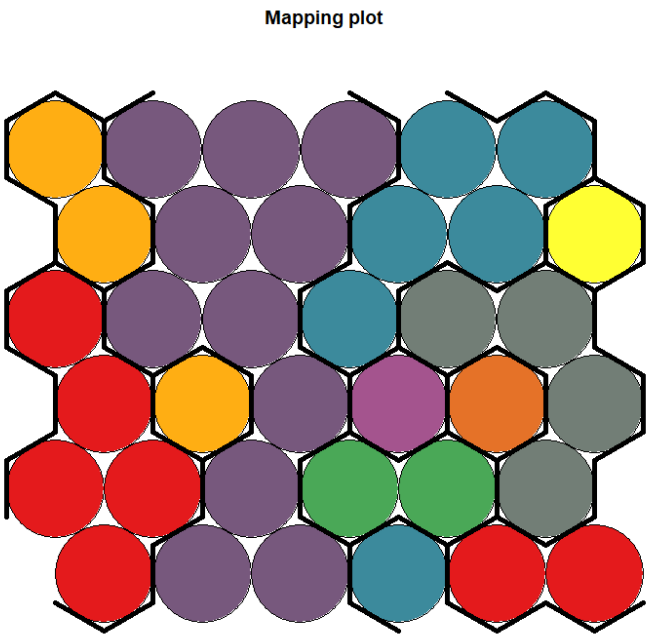

Figure 2.20: Self Organizing Map and Clustering of Reproductive Stage Contrast R4-5 to R7. Nine total node clusters were found. Three clusters consist of only a single node, yellow, dark orange, and light purple. The light orange cluster had a single node non-adjacent to the other two. The largest cluster, purple, hold eleven nodes (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

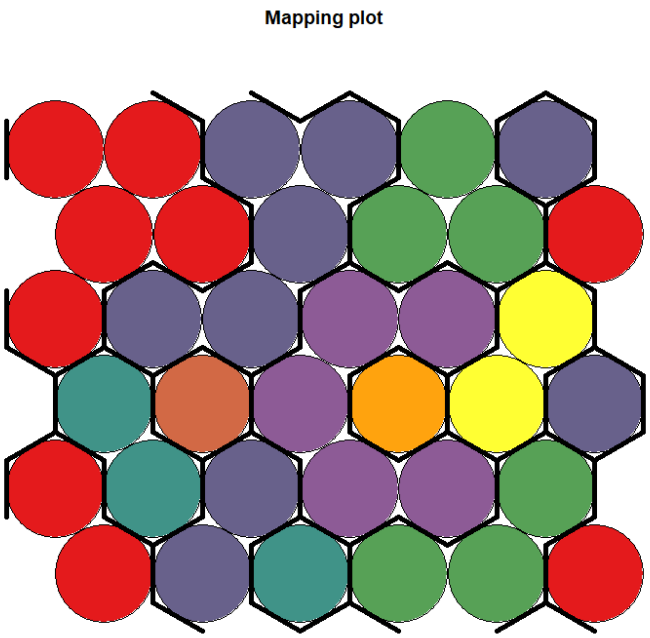

Figure 2.21: Self Organizing Map and Clustering of Reproductive Stage Contrast R6 to R7. Nine total node clusters were found. Two clusters consisted of a single node, while the largest contained nine nodes. This cluster, purple, also had two nodes non-adjacent to the remaining nine nodes. A single node from the cyan cluster was non-adjacent to the others (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018; Kruisselbrink, 2019).

Figure 2.22: Scaled Expression of Genes within Self Organizing Map Clusters of Genotype Contrast Tifrunner to NC 3033. Vertical axis shows scaled log fold change, while horizontal axis denotes the SOM cluster that the plotted points (genes) are derived. Positive scaled LFC is higher expression in the Tifrunner genotype, while negative is higher expression NC 3033. There is clear bias towards Tifrunner as a part of clusters 2, 3, and 6. Only a single cluster, cluster 4, shows bias towards NC 3033. While it is evident that in many clusters the scaled LFC of genes varies strongly amongst its cluster, between clusters there does not appear to be observable differences in the mean scaled LFC values.

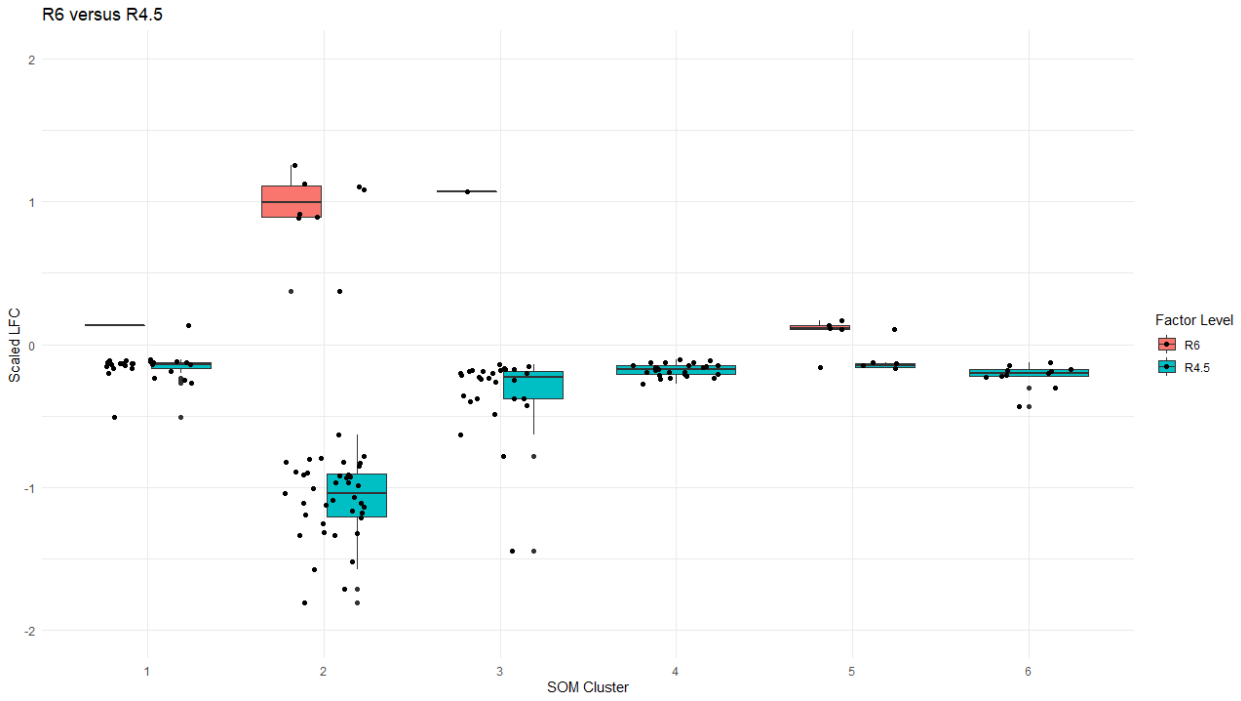

Figure 2.23: Scaled Expression of Genes within Self Organizing Map Clusters of Reproductive Stage Contrast R4-5 to R6. Vertical axis shows scaled log fold change, while horizontal axis denotes the SOM cluster that the plotted points (genes) are derived. Positive scaled LFC is higher expression in the R6 reproductive stage, while negative is higher expression in R4-5. All but cluster 2 and 5 show clear bias towards expression in R4-5. Cluster 2 interestingly has much larger scaled LFC values than the other clusters, having a higher mean scaled LFC than is seen across all clusters for R4-5 and across all but one cluster for R6.

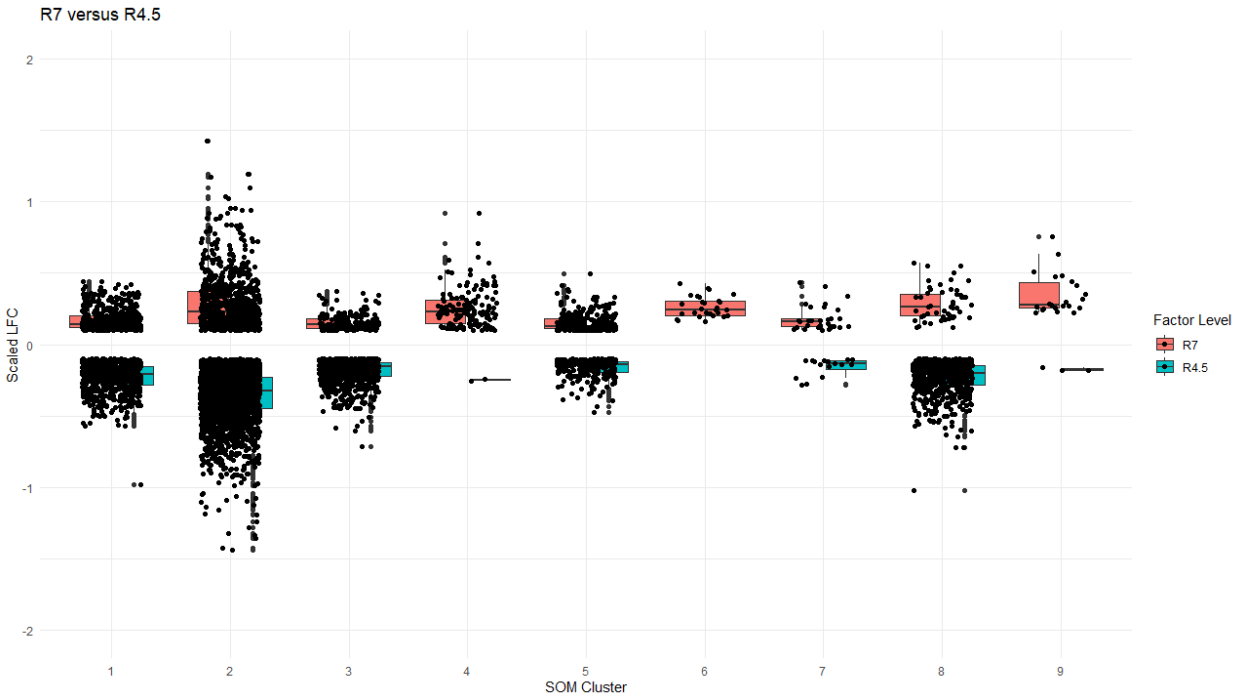

Figure 2.24: Scaled Expression of Genes within Self Organizing Map Clusters of Reproductive Stage Contrast R4-5 to R7. Vertical axis shows scaled log fold change, while horizontal axis denotes the SOM cluster that the plotted points (genes) are derived. Positive scaled LFC is higher expression in the R7 reproductive stage, while negative is higher expression in R4-5. Scaled LFC bias seems to be anchored in reproductive stage R7 with all three biased clusters, pointing towards R7.

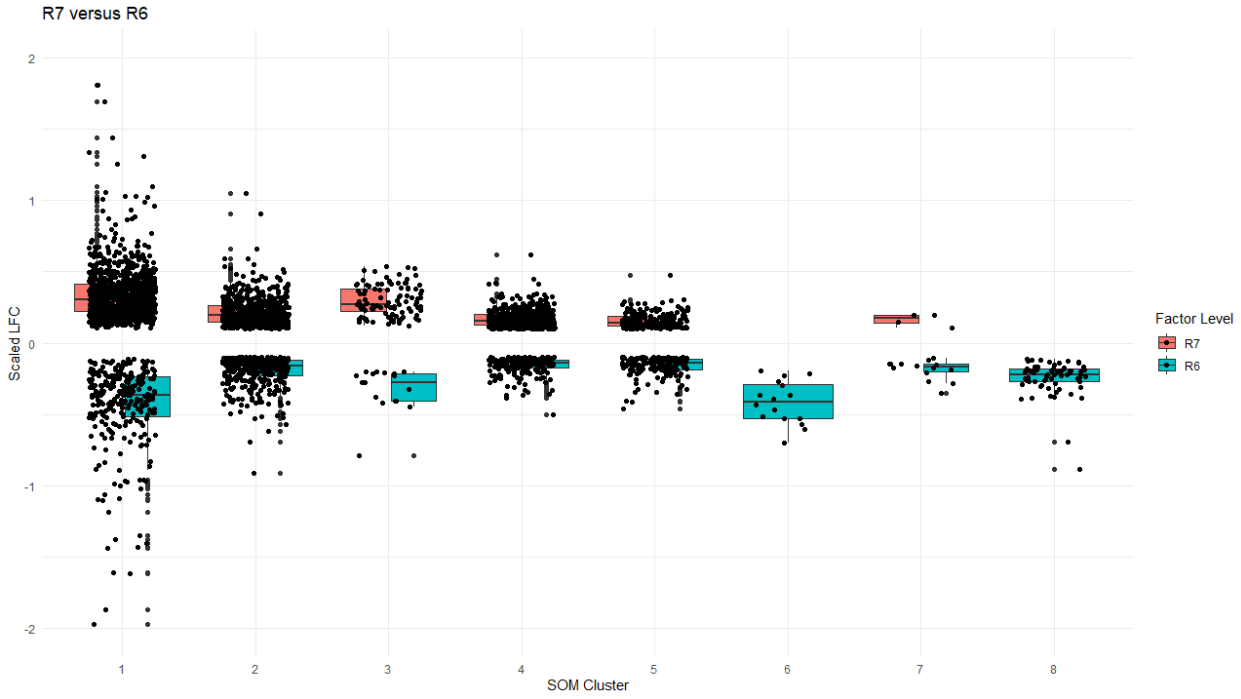

Figure 2.25: Scaled Expression of Genes within Self Organizing Map Clusters of Reproductive Stage Contrast R6 to R7. Vertical axis shows scaled log fold change, while horizontal axis denotes the SOM cluster that the plotted points (genes) are derived. Positive scaled LFC is higher expression in the R7 reproductive stage, while negative is higher expression in R6. There are two biased clusters, 6 and 8, which are pointed towards R6.

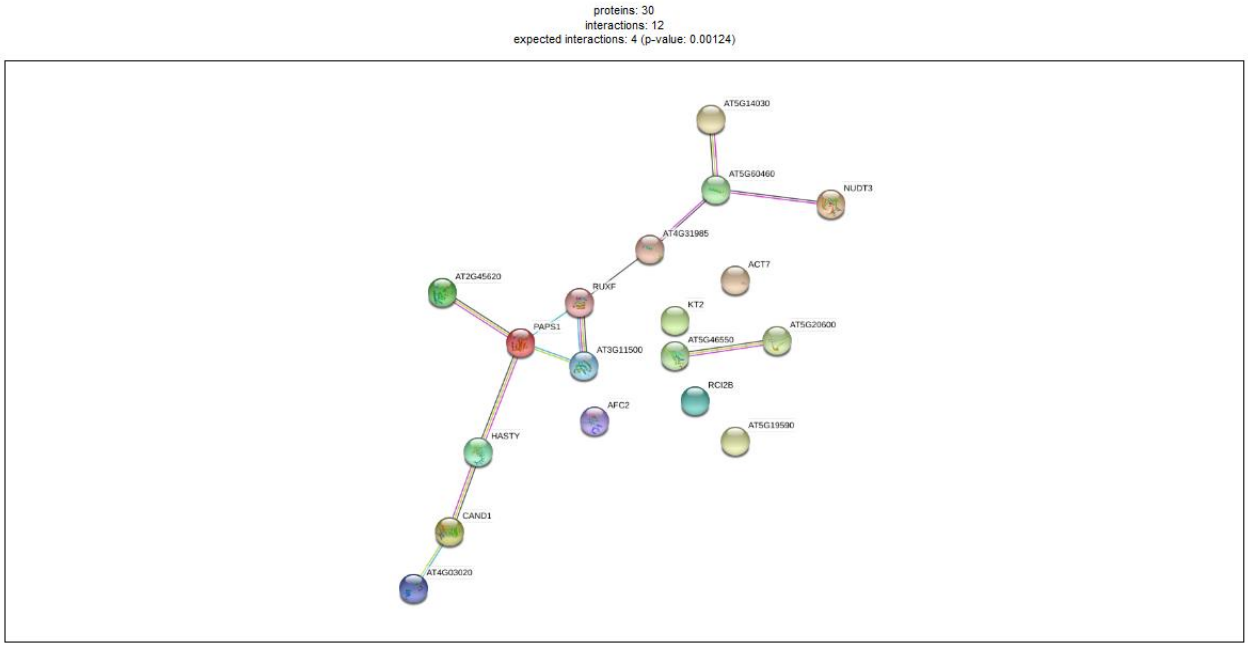

Figure 2.26: StringDB Network of Significant Gene Cluster. Protein network of *Arabidopsis thaliana* orthologs from complete model SOM cluster 4. There are two disconnected clusters one containing 11 genes the other 2 genes. 5 genes are disconnected from both clusters. A p-value of 0.00124 reinforces this map for the interactions between nodes are greater than what would be expected by random chance (Snel et al., 2000; von Mering et al., 2003, 2005; Franceschini et al., 2016; Szklarczyk et al., 2019, 2021). Orthologs were derived based on a SOM cluster of *Arachis hypogaea* genes found to have significant overlap with QTL known to be associated with pod filling (Das, 2016; Chavarro et al., 2020).

CHAPTER 3

SUMMARY

Plant breeding is a science that incorporates numerous different aspects of plant, biological, and mathematical sciences to produce new genotypes of plants with improvements over their parents. One of the backbones of this process is developing a greater understanding in the controls of desirable phenotypes to benefit established plant breeding pipelines. Thus, this analysis used RNA-seq and machine learning to pull apart the potential genes that control the pod filling phenotype present in two distinct *Arachis hypogaea* genotypes Tifrunner and NC 3033. Tifrunner represents many of the qualities typical of a commercial peanut cultivar in that it is high yielding and possesses a complete pod filling phenotype (Holbrook and Culbreath, 2007). NC 3033 while having some desirable characteristics such as disease resistance is less desirable due to incomplete pod filling, where the seed fails to occupy the entire peanut pod interior, making it less profitable for a producer (Beute et al., 1976; Chavarro et al., 2020).

A multifactorial RNA-seq design, sampling from both Tifrunner and NC 3033 and three pod tissues across three grouped reproductive stages allowed for the reduction of gene analysis to 294 genes. Coupled with machine learning algorithms in the form of Self Organizing Maps these genes were grouped into clusters based on expression levels across all samples. Each cluster was compared to quantitative trait loci known to be associated with the pod filling phenotype which reduced the cluster search space down to only those with significant overlap. Due to limitations in the functional annotation of *A. hypogaea* genes functional network analysis of the cluster was

performed using orthologs from *Arabidopsis thaliana*. A single significant network was found. Functional analysis of the genes within the significant network will need to be performed to validate the gene set for use in peanut improvement.

References

Beute, M.K., J.C. Wynne, and D.A. Emery. 1976. Registration of NC 3033 Peanut Germplasm1 (Reg. No. GP 9). Crop Science 16(6): cropsci1976.0011183X001600060046x. doi: 10.2135/cropsci1976.0011183X001600060046x.

Chavarro, C., Y. Chu, C. Holbrook, T. Isleib, D. Bertioli, et al. 2020. Pod and Seed Trait QTL Identification To Assist Breeding for Peanut Market Preferences. G3: Genes, Genomes, Genetics 10(7): 2297–2315. doi: 10.1534/g3.120.401147.

Holbrook, C.C., and A.K. Culbreath. 2007. Registration of 'Tifrunner' Peanut. Journal of Plant Registrations 1(2): 124–124. doi: 10.3198/jpr2006.09.0575crc.

REFERENCES

Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46(W1): W537–W544. doi: 10.1093/nar/gky379.

Alexa, A., and J. Rahnenfuhrer. 2021. topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.13).

Alpaydin, E. 2014a. Supervised Learning. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 21–48

Alpaydin, E. 2014b. Introduction. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 1–20

Alpaydin, E. 2014c. Reinforcement Learning. Introduction to machine learning. Third edition. The MIT Press, Cambridge, Massachusetts. p. 517–546

Arya, S.S., A.R. Salve, and S. Chauhan. 2016. Peanuts as functional food: a review. J Food Sci Technol 53(1): 31–41. doi: 10.1007/s13197-015-2007-9.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25(1): 25–29. doi: 10.1038/75556.

Ballén-Taborda, C., Y. Chu, P. Ozias-Akins, P. Timper, C.C. Holbrook, et al. 2019. A new source of root-knot nematode resistance from Arachis stenosperma incorporated into allotetraploid peanut (Arachis hypogaea). Sci Rep 9(1): 17702. doi: 10.1038/s41598-019-54183-1.

Barbour, J.A., P.R.C. Howe, J.D. Buckley, J. Bryan, and A.M. Coates. 2015. Effect of 12 Weeks High Oleic Peanut Consumption on Cardio-Metabolic Risk Factors and Body Composition. Nutrients 7(9): 7381–7398. doi: 10.3390/nu7095343.

Benfey, P.N., and T. Mitchell-Olds. 2008. From Genotype to Phenotype: Systems Biology Meets Natural Variation. Science 320(5875): 495–497. doi: 10.1126/science.1153716.

Bera, S.K., J.H. Kamdar, S.V. Kasundra, S.V. Patel, M.D. Jasani, et al. 2019. Steady expression of high oleic acid in peanut bred by marker-assisted backcrossing for fatty acid desaturase mutant alleles and its effect on seed germination along with other seedling traits. PLOS ONE 14(12): e0226252. doi: 10.1371/journal.pone.0226252.

Bertioli, D.J., S.B. Cannon, L. Froenicke, G. Huang, A.D. Farmer, et al. 2016. The genome sequences of Arachis duranensis and Arachis ipaensis , the diploid ancestors of cultivated peanut. Nat Genet 48(4): 438–446. doi: 10.1038/ng.3517.

Bertioli, D.J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, et al. 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nature Genetics 51(5): 877–884. doi: 10.1038/s41588-019-0405-z.

Beute, M.K., J.C. Wynne, and D.A. Emery. 1976a. Registration of NC 3033 Peanut Germplasm1 (Reg. No. GP 9). Crop Science 16(6): cropsci1976.0011183X001600060046x. doi: 10.2135/cropsci1976.0011183X001600060046x.

Beute, M.K., J.C. Wynne, and D.A. Emery. 1976b. Registration of NC 3033 Peanut Germplasm1 (Reg. No. GP 9). Crop Science 16(6): cropsci1976.0011183X001600060046x. doi: 10.2135/cropsci1976.0011183X001600060046x.

Bhullar, N.K., Z. Zhang, T. Wicker, and B. Keller. 2010. Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene Pm3: a large scale allele mining project. BMC Plant Biol 10(1): 88. doi: 10.1186/1471-2229-10-88.

Bishop, C.M. 2006. Chapter 9: Mixture Models and EM. Pattern recognition and machine learning. Springer, New York. p. 423–455

Blighe, K., S. Rana, and M. Lewis. 2021. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Bonku, R., and J. Yu. 2020. Health aspects of peanuts as an outcome of its chemical composition. Food Science and Human Wellness 9(1): 21–30. doi: 10.1016/j.fshw.2019.12.005.

Boote, K.J. 1982. Growth Stages of Peanut (Arachis hypogaea L.)1. Peanut Science 9(1): 35–40. doi: 10.3146/i0095-3679-9-1-11.

Brown, S.L., A.K. Culbreath, J.W. Todd, D.W. Gorbet, J.A. Baldwin, et al. 2005. Development of a Method of Risk Assessment to Facilitate Integrated Management of Spotted Wilt of Peanut. Plant Disease 89(4): 348–356. doi: 10.1094/PD-89-0348.

Buccitelli, C., and M. Selbach. 2020. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21(10): 630–644. doi: 10.1038/s41576-020-0258-4.

Cano-Gamez, E., and G. Trynka. 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Frontiers in Genetics 11: 424. doi: 10.3389/fgene.2020.00424.

Carbon, S., E. Douglass, B.M. Good, D.R. Unni, N.L. Harris, et al. 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research 49(D1): D325–D334. doi: 10.1093/nar/gkaa1113.

Carter, R.J., I. Dubchak, and S.R. Holbrook. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 29(19): 3928–3938.

Chappell, T.M., C.B. Codod, B.W. Williams, R.C. Kemerait, A.K. Culbreath, et al. 2020. Adding Epidemiologically Important Meteorological Data to Peanut Rx, the Risk Assessment Framework for Spotted Wilt of Peanut. Phytopathology® 110(6): 1199–1207. doi: 10.1094/PHYTO-11-19-0438-R.

Chavarro, C. 2021. Personal Correspondence.

Chavarro, C., Y. Chu, C. Holbrook, T. Isleib, D. Bertioli, et al. 2020. Pod and Seed Trait QTL Identification To Assist Breeding for Peanut Market Preferences. G3: Genes, Genomes, Genetics 10(7): 2297–2315. doi: 10.1534/g3.120.401147.

Cheng, C.-Y., V. Krishnakumar, A.P. Chan, F. Thibaud-Nissen, S. Schobel, et al. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal 89(4): 789–804. doi: 10.1111/tpj.13415.

Chettoor, A.M., S.A. Givan, R.A. Cole, C.T. Coker, E. Unger-Wallace, et al. 2014. Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biology 15(7): 414. doi: 10.1186/s13059-014-0414-2.

Chinnusamy, V., J. Zhu, and J.-K. Zhu. 2007. Cold stress regulation of gene expression in plants. Trends in Plant Science 12(10): 444–451. doi: 10.1016/j.tplants.2007.07.002.

Chu, Y., C.L. Wu, C.C. Holbrook, B.L. Tillman, G. Person, et al. 2011. Marker-Assisted Selection to Pyramid Nematode Resistance and the High Oleic Trait in Peanut. The Plant Genome 4(2). doi: 10.3835/plantgenome2011.01.0001.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17(1): 13. doi: 10.1186/s13059-016-0881-8.

Costa-Silva, J., D. Domingues, and F.M. Lopes. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE 12(12): e0190152. doi: 10.1371/journal.pone.0190152.

Culbreath, A.K., and R. Srinivasan. 2011. Epidemiology of spotted wilt disease of peanut caused by Tomato spotted wilt virus in the southeastern U.S. Virus Research 159(2): 101–109. doi: 10.1016/j.virusres.2011.04.014.

Das, S. 2016. GSAQ: Gene Set Analysis with QTL.

Dash, S., E.K.S. Cannon, S.R. Kalberer, A.D. Farmer, and S.B. Cannon. 2016. Chapter 8 - PeanutBase and Other Bioinformatic Resources for Peanut. In: Stalker, H.T. and F. Wilson, R., editors, Peanuts. AOCS Press. p. 241–252

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1): 15–21. doi: 10.1093/bioinformatics/bts635.

Elson, D. 1965. Metabolism of Nucleic Acids (Macromolecular DNA and RNA). Annual Review of Biochemistry 34(1): 449–486. doi: 10.1146/annurev.bi.34.070165.002313.

Emms, D.M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16(1): 157. doi: 10.1186/s13059-015-0721-2.

Emms, D.M., and S. Kelly. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20(1): 238. doi: 10.1186/s13059-019-1832-y.

FAO. 2017. The future of food and agriculture: trends and challenges. Food and Agriculture Organization of the United Nations, Rome.

FAO. 2019. Seeds Food and Agriculture Organization of the United Nations. http://www.fao.org/seeds/en/ (accessed 23 June 2021).

Fernandes, A.D., J.N. Reid, J.M. Macklaim, T.A. McMurrough, D.R. Edgell, et al. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2: 15. doi: 10.1186/2049-2618-2-15.

Fraley, C., A.E. Raftery, L. Scrucca, T.B. Murphy, and M. Fop. 2020. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation.

Franceschini, A., J. Lin, C. von Mering, and L.J. Jensen. 2016. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. Bioinformatics 32(7): 1085–1087. doi: 10.1093/bioinformatics/btv696.

Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, et al. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289(5476): 85–88. doi: 10.1126/science.289.5476.85.

Ghosh, S., and C.-K.K. Chan. 2016. Analysis of RNA-Seq Data Using TopHat and Cufflinks. In: Edwards, D., editor, Plant Bioinformatics: Methods and Protocols. Springer, New York, NY. p. 339–361

Gilman, D.F., and O.D. Smith. 1977. Internal Pericarp Color as a Subjective Maturity Index for Peanut Breeding1. Peanut Science 4(2): 67–70. doi: 10.3146/i0095-3679-4-2-6.

Glaeser, B. 2011. The Green Revolution Revisited: Critique and Alternatives. Taylor & Francis Group, London, UNITED KINGDOM.

Gupta, K., O. Buchshtab, and R. Hovav. 2014. The Effects of Irrigation Level and Genotype on Pod-Filling Related Traits in Peanut (Arachis hypogaea). Journal of Agricultural Science 7(1): p169. doi: 10.5539/jas.v7n1p169.

Hajduch, M., L.B. Hearne, J.A. Miernyk, J.E. Casteel, T. Joshi, et al. 2010. Systems Analysis of Seed Filling in Arabidopsis: Using General Linear Modeling to Assess Concordance of Transcript and Protein Expression. Plant Physiology 152(4): 2078–2087. doi: 10.1104/pp.109.152413.

Halliburton, B.W., W.G. Glasser, and J.M. Byrne. 2015. An Anatomical Study of the Pericarp of Arachis Hypogaea, with Special Emphasis on the Sclereid Component. Botanical Gazette. doi: 10.1086/336807.

Halvey, J., A. Hartzook, and T. Markovitz. 1987. Foliar fertilization of high-yielding peanuts during the pod-filling period. Fertilizer Research 14: 153–10.

Hammons, R.O., D.K. Bell, and E.K. Sobers. 1981. Evaluating Peanuts for Resistance to Cylindrocladium Black Rot1. Peanut Science 8(2): 117–120. doi: 10.3146/i0095-3679-8-2-10.

Haykin, S.S. 2009a. Self-Organizing Maps. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 425–470

Haykin, S.S. 2009b. Introduction. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 1–46

Haykin, S.S. 2009c. Neural networks and learning machines. 3rd ed. Prentice Hall, New York.

Haykin, S.S. 2009d. Principal-Component Analysis. Neural networks and learning machines. 3rd ed. Prentice Hall, New York. p. 367–418

Hazell, P.B.R. 2009. The Asian Green Revolution. Intl Food Policy Res Inst.

Hilu, K.W. 1993. Polyploidy and the evolution of domesticated plants. American Journal of Botany 80(12): 1494–1499. doi: 10.1002/j.1537-2197.1993.tb15395.x.

Hoheisel, J.D. 2006. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet 7(3): 200–210. doi: 10.1038/nrg1809.

Holbrook, C.C., and A.K. Culbreath. 2007a. Registration of 'Tifrunner' Peanut. Journal of Plant Registrations 1(2): 124–124. doi: 10.3198/jpr2006.09.0575crc.

Holbrook, C.C., and A.K. Culbreath. 2007b. Registration of 'Tifrunner' Peanut. Journal of Plant Registrations 1(2): 124–124. doi: 10.3198/jpr2006.09.0575crc.

Holbrook, C.C., T.G. Isleib, P. Ozias-Akins, Y. Chu, S.J. Knapp, et al. 2013. Development and Phenotyping of Recombinant Inbred Line (RIL) Populations for Peanut (Arachis hypogaea). Peanut Science 40(2): 89–94. doi: 10.3146/PS13-5.1.

Howe, K.L., P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, et al. 2021. Ensembl 2021. Nucleic Acids Research 49(D1): D884–D891. doi: 10.1093/nar/gkaa942.

Illumina. 2011. RNA-Seq Data Comparison with Gene Expression Microarrays. https://www.europeanpharmaceuticalreview.com/wp-content/uploads/Illumina_whitepaper.pdf (accessed 2 July 2021).

Jackson, L.F. 1983. Relative Susceptibilities of Component Lines of Peanut Cultivars Early Bunch and Florunner to Early and Late Leafspots1. Peanut Science 10(1): 3–5. doi: 10.3146/i0095-3679-10-1-2.

Jones, D.T. 2019. Setting the standards for machine learning in biology. Nat Rev Mol Cell Biol 20(11): 659–660. doi: 10.1038/s41580-019-0176-5.

Jung, S., G. Powell, K. Moore, and A. Abbott. 2000. The high oleate trait in the cultivated peanut [Arachis hypogaea L.]. II. Molecular basis and genetics of the trait. Mol Gen Genet 263(5): 806–811. doi: 10.1007/s004380000243.

Kalberer, S., and V. Belamkar. 2014. Arachis duranensis, Arachis ipaensis, and the Origins of Cultivated Peanut (Arachis hypogaea).

https://www.peanutbase.org/files/misc/arachis_duranensis_ipaensis_info_v02.pdf (accessed 2 July 2021).

Kalderimis, A., R. Lyne, D. Butano, S. Contrino, M. Lyne, et al. 2014. InterMine: extensive web services for modern biology. Nucleic Acids Res 42(Web Server issue): W468-472. doi: 10.1093/nar/gku301.

Kassambara, A. 2020. ggpubr: "ggplot2" Based Publication Ready Plots.

Klevorn, C.M., L.L. Dean, and S.D. Johanningsmeier. 2019. Metabolite Profiles of Raw Peanut Seeds Reveal Differences between Market-Types. Journal of Food Science 84(3): 397–405. doi: 10.1111/1750-3841.14450.

Kohonen, T. 2012. Self-Organizing Maps. Springer Science & Business Media.

Koonin, E.V., L. Aravind, and A.S. Kondrashov. 2000. The Impact of Comparative Genomics on Our Understanding of Evolution. Cell 101(6): 573–576. doi: 10.1016/S0092-8674(00)80867-3.

Krapovickas, A., W.C. Gregory, D.E. Williams, and C.E. Simpson. 2007. TAXONOMY OF THE GENUS ARACHIS (LEGUMINOSAE). Bonplandia 16: 7–205.

Kruisselbrink, R.W. and J. 2019. kohonen: Supervised and Unsupervised Self-Organising Maps.

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3): R25. doi: 10.1186/gb-2009-10-3-r25.

Law, C.W., M. Alhamdoosh, S. Su, X. Dong, L. Tian, et al. 2018. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res 5: ISCB Comm J-1408. doi: 10.12688/f1000research.9005.3.

Lee, B., D.A. Henderson, and J.-K. Zhu. 2005. The Arabidopsis Cold-Responsive Transcriptome and Its Regulation by ICE1. The Plant Cell 17(11): 3155–3175. doi: 10.1105/tpc.105.035568.

Levin, J.Z., M. Yassour, X. Adiconis, C. Nusbaum, D.A. Thompson, et al. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7(9): 709–715. doi: 10.1038/nmeth.1491.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26(5): 589–595. doi: 10.1093/bioinformatics/btp698.

Love, M., C. Ahlmann-Eltze, K. Forbes, S. Anders, W. Huber, et al. 2021. DESeq2: Differential gene expression analysis based on the negative binomial distribution. Bioconductor version: Release (3.13).

Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12): 550. doi: 10.1186/s13059-014-0550-8.

Lyons, E., and M. Freeling. 2008a. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53(4): 661–673. doi: 10.1111/j.1365-313X.2007.03326.x.

Lyons, E., and M. Freeling. 2008b. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53(4): 661–673. doi: 10.1111/j.1365-313X.2007.03326.x.

Mantione, K.J., R.M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, et al. 2014. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res 20: 138–141. doi: 10.12659/MSMBR.892101.

Masur, L.J., J.-F. Millaire, and M. Blake. 2018. Peanuts and Power in the Andes: The Social Archaeology of Plant Remains from the Virú Valley, Peru. etbi 38(4): 589–609. doi: 10.2993/0278-0771-38.4.589.

McCarthy, D.J., Y. Chen, and G.K. Smyth. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40(10): 4288–4297. doi: 10.1093/nar/gks042.

McCulloch, W.S., and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5(4): 115–133. doi: 10.1007/BF02478259.

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, et al. 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1): 258–261. doi: 10.1093/nar/gkg034.

von Mering, C., L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, et al. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33(Database issue): D433-437. doi: 10.1093/nar/gki005.

Mitkowski, N.A., and G.S. Abawi. 2003. Root-knot nematode. Root-knot nematode. https://www.apsnet.org/edcenter/disandpath/nematode/pdlessons/Pages/RootknotNematode.aspx (accessed 23 July 2021).

Moore, G.E. 1965. Cramming more components onto integrated circuits. 38(8): 4.

Moretzsohn, M. de C., M.S. Hopkins, S.E. Mitchell, S. Kresovich, J.F.M. Valls, et al. 2004. Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biol 4: 11. doi: 10.1186/1471-2229-4-11.

Nayak, R., and Y. Hasija. 2021. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. Genomics 113(2): 606–619. doi: 10.1016/j.ygeno.2021.01.007.

Noble, D. 2002. The rise of computational biology. Nat Rev Mol Cell Biol 3(6): 459–463. doi: 10.1038/nrm810.

Nussinov, R. 2015. Advancements and Challenges in Computational Biology. PLOS Computational Biology 11(1): e1004053. doi: 10.1371/journal.pcbi.1004053.

Parkinson, J., and M. Blaxter. 2009. Expressed sequence tags: an overview. Methods Mol Biol 533: 1–12. doi: 10.1007/978-1-60327-136-3_1.

Paterson, A.H., J.E. Bowers, M.D. Burow, X. Draye, C.G. Elsik, et al. 2000. Comparative Genomics of Plant Chromosomes. The Plant Cell 12(9): 1523–1539. doi: 10.1105/tpc.12.9.1523.

Pearson, W.R. 2013. An Introduction to Sequence Similarity ("Homology") Searching. Curr Protoc Bioinformatics 0 3: 10.1002/0471250953.bi0301s42. doi: 10.1002/0471250953.bi0301s42.

Pearson, W.R., and D.J. Lipman. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8): 2444–2448.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Renny-Byfield, S., and J.F. Wendel. 2014. Doubling down on genomes: Polyploidy and crop plants. American Journal of Botany 101(10): 1711–1725. doi: 10.3732/ajb.1400119.

Ritchie, M.E., B. Phipson, D. Wu, Y. Hu, C.W. Law, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7): e47. doi: 10.1093/nar/gkv007.

Robinson, M.D., D.J. McCarthy, and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139–140. doi: 10.1093/bioinformatics/btp616.

Scrucca, L., M. Fop, T.B. Murphy, and A.E. Raftery. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8(1): 289–317.

Shiraiwa, T., N. Ueno, S. Shimada, and T. Horie. 2004a. Correlation between Yielding Ability and Dry Matter Productivity during Initial Seed Filling Stage in Various Soybean Genotypes. Plant Production Science 7(2): 138–142. doi: 10.1626/pps.7.138.

Shiraiwa, T., N. Ueno, S. Shimada, and T. Horie. 2004b. Correlation between Yielding Ability and Dry Matter Productivity during Initial Seed Filling Stage in Various Soybean Genotypes. Plant Production Science 7(2): 138–142. doi: 10.1626/pps.7.138.

Smith, B.W. 1950. Arachis Hypogaea. Aerial Flower and Subterranean Fruit. American Journal of Botany 37(10): 802–815. doi: 10.1002/j.1537-2197.1950.tb11073.x.

Smith, R.N., J. Aleksic, D. Butano, A. Carr, S. Contrino, et al. 2012. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28(23): 3163–3165. doi: 10.1093/bioinformatics/bts577.

Snel, B., G. Lehmann, P. Bork, and M.A. Huynen. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28(18): 3442–3444. doi: 10.1093/nar/28.18.3442.

Solomonoff, R.J. 1964. A formal theory of inductive inference. Information and Control 7(2): 224–254. doi: 10.1016/S0019-9958(64)90131-7.

Stark, R., M. Grzelak, and J. Hadfield. 2019. RNA sequencing: the teenage years. Nat Rev Genet 20(11): 631–656. doi: 10.1038/s41576-019-0150-2.

Szklarczyk, D., A.L. Gable, D. Lyon, A. Junge, S. Wyder, et al. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1): D607–D613. doi: 10.1093/nar/gky1131.

Szklarczyk, D., A.L. Gable, K.C. Nastou, D. Lyon, R. Kirsch, et al. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49(D1): D605–D612. doi: 10.1093/nar/gkaa1074.

Tarca, A.L., V.J. Carey, X. Chen, R. Romero, and S. Drăghici. 2007. Machine Learning and Its Applications to Biology. PLOS Computational Biology 3(6): e116. doi: 10.1371/journal.pcbi.0030116.

Toomer, O.T. 2018. Nutritional chemistry of the peanut (Arachis hypogaea). Crit Rev Food Sci Nutr 58(17): 3042–3053. doi: 10.1080/10408398.2017.1339015.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3): 562–578. doi: 10.1038/nprot.2012.016.

Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5): 511–515. doi: 10.1038/nbt.1621.

Venkatachalam, M., and S.K. Sathe. 2006. Chemical Composition of Selected Edible Nut Seeds. J. Agric. Food Chem. 54(13): 4705–4714. doi: 10.1021/jf0606959.

Wall, D.P., P. Kudtarkar, V.A. Fusaro, R. Pivovarov, P. Patil, et al. 2010. Cloud computing for comparative genomics. BMC Bioinformatics 11(1): 259. doi: 10.1186/1471-2105-11-259.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1): 57–63. doi: 10.1038/nrg2484.

Wang, T., B. Li, C.E. Nelson, and S. Nabavi. 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 20(1): 40. doi: 10.1186/s12859-019-2599-6.

Wehrens, R., and L.M.C. Buydens. 2007. Self- and Super-Organizing Maps in R: The kohonen Package. Journal of Statistical Software 21(5): 1–19. doi: 10.18637/jss.v021.i05.

Wehrens, R., and J. Kruisselbrink. 2018. Flexible Self-Organizing Maps in kohonen 3.0. Journal of Statistical Software 87(7): 1–18. doi: 10.18637/jss.v087.i07.

Wickham, H. 2016. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.

Wolf, J.B.W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Molecular Ecology Resources 13(4): 559–572. doi: 10.1111/1755-0998.12109.

Yamamoto, M., T. Wakatsuki, A. Hada, and A. Ryo. 2001. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods 250(1–2): 45–66. doi: 10.1016/s0022-1759(01)00305-2.

Yin, S., P. Li, Y. Xu, J. Liu, T. Yang, et al. 2020. Genetic and genomic analysis of the seed-filling process in maize based on a logistic model. Heredity 124(1): 122–134. doi: 10.1038/s41437-019-0251-x.

Young, C.T., and W.E. Schadel. 2004. Microstructure of Peanut Seed: A Review. : 13.