Development and Application of Statistical Methods for Biological Network Inference from Gene Expression Profiles

ウォン, プイ, シャン

https://doi.org/10.15017/1932015

出版情報：九州大学, 2017, 博士（農学）, 論文博士
バージョン：
権利関係：

Development and Application of Statistical Methods for Biological Network Inference from Gene Expression Profiles

Wong Pui Shan

2018

Contents

Introduction

Experiments relying on the continuous observation of gene expression used to be very time consuming and expensive to quantify. The techniques depended on hybridization approaches using fluorescent tags and microarrays. They required the preparation of a number of materials that increased the cost of the experiment, particularly if the research required custom-made microarrays for specific genomes, enzymes or metabolites [1]. They were also limited by the number of existing genome data that was being researched, the small detection ranges of signals that were measured and the background noise in the data. These limits became a soft threshold on what type of research could be carried out.

When John Craig Venter first ventured into genome sequencing, it signaled the start of a new era in biology [2] [3]. The first genome sequencing project cost roughly US $3 billion and took about 13 years with additional time taken for accuracy analyses. It used a combination of Sanger sequencing and shotgun sequencing [4] [5] [6], as well as other techniques which were improved upon by James Watson who took the next step and completed the next genome sequencing in four months and cost less than US $1.5 million to complete [7] [8]. The last decade has brought about dramatic changes to the landscape of research involving genomes, resulting in the incredulous efficiency of performing a sequencing run in one to two days and costing US $1,000 to $5,000 [9] [10]. The developments achieved by those first generation sequencing methods have helped to generate the concepts that led to the success of the current next-generation sequencing methods (NGS).

NGS is the successor of the technology pioneered by Sanger sequencing [2] and is now routinely used in research, and clinical and diagnostic applications for more involved studies such as transcriptome profiling (miRNA-Seq) [11], chromatin immunoprecipitation with sequencing (high resolution ChIP-sequencing) [12] [13] [14] and locating genetic variants such as single nucleotide polymorphisms (SNP) [15]. The initial complication of genome sequencing was due a number of factors including the size of genomes, in particular eukaryote genomes, the preparation and use of bacterial artifi-

cial chromosomes (BAC) and related operational difficulties, and the computing power that was available at the time to align and store all the data. While Sanger sequencing is highly accurate, it has a limited throughput which is throttled by the use of gels or polymers. It is still used today for deep sequencing as the long fragments make it possible to resolve repeat regions but NGS has become the primary tool of choice for the genomics field in general.

NGS was generated from an idea conceived from shotgun sequencing; the idea of massive parallel sequencing reactions. By building up on existing knowledge, it has several advantages over its predecessors such as its cell free system, and cyclical and parallel sequencing which gives higher coverage and throughput. NGS requires less reagents overall, as sequencing has become possible by using single strands, and sequencing reactions are able to be run on a single chip. These features work well with the short overlapping repeat approach, which leads to better coverage and respectable accuracy despite shorter reads since each region is sequenced many times. The improvement in computational power and sequencing algorithms has also contributed in the popularity of NGS as better computing power has made it possible to map many short reads to scaffolds quickly, as reflected by Moore's Law which shows a trend with the doubling of computing power every two years.

NGS is available through several commercially available platforms. These include the Roche GS-FLX 454 Genome Sequencer (454 sequencing), the Illumina Genome Analyzer (Solexa), the Applied Biosystems SOLiD analyzer, Polonator G.007 and the Helicos HeliScope platforms. The platforms also come with supporting alignment and assembly tools as well as analysis software such as Sequence Analysis Viewer for assessing sequence quality, and Hiseq Analysis Software for alignments from Illumina platforms. There are platform independent software available as well, with an active community of user support available for them due to their popularity amongst researchers such as MAQ [16], Bowtie 2 [17] [18], Tophat [19] [20] [21] and Cufflinks [22], as well as full pipeline programs like QuickNGS [11]. These companies and software developers have had an important impact on NGS development and strongly influences on where the technology will go in the future [23].

The advantages offered by NGS has resulted in a more democratized sequencing landscape where most laboratories around the world have access to affordable and efficient genome analyses technologies. By allowing more researchers to perform sequencing experiments, NGS has invigorated many fields in biology such as biotechnology, forensics, agrigenomics and clinical medicine, and advanced several new fields of study such as personalized medicine, epigenomics and metabolomics. It has allowed the *de*

novo assembly of many genomes to be possible including the panda [24], the Pacific bluefin tuna [25], the sunflower [26] and the Antarctic midge [27], as well as the re-sequencing of genomes that have already been processed such as *Escherichia coli* [28] and *Sesamum indicum* [29].

Most importantly, the technological development of NGS has precipated major developments in other research that use sequences such as detecting genetic elements of diseases and transcriptome analyses [30]. The NGS methods that rely on expressed transcripts have been readily coopted in transcriptome experiments that typically use RNA-Sequencing (RNA-Seq) to quantify cell transcripts at the time of sampling, providing a real-time snapshot of transcript levels. Since NGS is supremely efficient over former methods, it has recently become more feasible to perform extensive studies that incorporate multiple conditions during an extended period of time. The increased resolution to the data produces measurements containing more detail that is needed when studying processes like gene expression regulation since it is a multi-component complex system.

This influx of data being generated from NGS research has called attention to the need for better data management for biologists from issues such as data security as well as data sharing among international research groups. A significant part of the big data dilemma is the ability to analyze data to keep up with the rate at which it is is being produced and the ability to manipulate much larger data sets than in the past. The size of standard data files can now include several thousand species in microbiome studies or contain the expression of thousands of genes in experiments with three or more variables, all the while, many journals call for multiple replicates for robustness. The number of data points make it difficult to visualize the full experiment in one diagram because of the limit of a physical screen size. Even a simple task of graphing the raw data points may require a wait time for rendering. This culminates in data that reach the barriers of what traditional hypothesis tests are able to analyze. This can be demonstrated by the computing time it takes for some calculations to complete such as linear models, bayesian inferences or covariance matrices. Consequently, new methods should be developed alongside the progression of nucleic acid sequencing.

This analysis task is a major objective in the field of bioinformatics which applies statistical and computational analysis to biological data. The development of more appropriate and convenient data analysis tools is the key to understanding and exploring the large amounts of data being outputted from sequencers by investigating what relationships were recorded in the data and how the results can be interpreted in the context of cellular signaling and metabolic processes. The investigations should yield testable

hypotheses to lead the way for future projects.

NGS data presents some particular restrictions to analyses tools. In terms of the data structure, NGS data tends to hold numerous and complex correlations between the thousands of genes, as well as commonly having a noisy background signal compared to a weak foreground signal if the investigation is focused on a small subset of genes. The general statistical term is the large p, small n problem where the large number of variables to small number of replicates make it difficult to identify true signals with an appropriate level of false discovery rate [31]. It also makes it difficult to create models without over fitting them since adding more variables to a model will make the model fit the data better but it will be too specific and less useful as a model for other data sets.

Current methods generally utilize strategies relying on linear or Bayesian statistics and have been developed from the elements of other types of experiments such as detecting differential gene expression in microarray analyses [32] [33]. Apart from the methodology of the analysis itself, another key step of NGS data analysis is the choosing or developing of the correct data analysis techniques appropriate for the number of samples and variables of the project. This is largely solved by using multiple tools, either in series to supplement the results of upstream tools such as gene ontologies (GO) [34] or in parallel as a multi-faceted approach such as a combination of metabolic pathway tools [35], gene network tools [36] and interactome tools [37].

The results also need to be presented visually so that the essential elements can be quickly comprehended by viewers, such as the standard use of heat maps in microarray data, sunburst compound graphs for showing a mixture of genomic information [38] or metabolic pathways [39] [40] [41]. Many detailed visualization tools were created for popular model organisms such as humans and mouse [42] [43] while more universal tools tend be simple such as Venn diagrams [44], so projects on novel organisms have to work quickly to annotate the genome as much as possible so that they can use notable database accessions or create tools of their own. Finally, with the volume of data to be analyzed, it is important that analysis tools are easily accessible, have a reasonably fast run time and are intuitive to use. This is why many favorable and popular methods are either internet tools like protein localization tool WoLF PSORT [45], complete software packages like Geneious [46] or are part of a repository of tools like Bioconductor [47].

Due to the prevalence of NGS, studies on gene expression and regulation have advanced a great deal so that it is now possible to examine the complex processes of eukaryotes instead of simpler prokaryotes [1]. As more genomes are being sequenced, the landscape of cellular biology has progressed from identifying genes and classifying

functions to transcriptomes, epigenomes and metabolomes. This technique has already been applied to several standard eukaryotes such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, mouse and human cells [48] [49] [50] [51] [52] [53].

Transcriptomics study the type and quantities of the different RNA molecules that are found in a cell or population of cells at a given time. The aim is to create a record of transcripts that exist for all species including sequence variations in populations [51] [53], to determine which sections of genes are transcribed and to quantify how the quantity of transcripts change at different times and conditions. Most commonly, mRNA quantity is analysed but other non-coding RNA and small RNA can be important to the end product as well. For many complex eukaryotes, there is also the difficulty of cataloguing the different splicing variants for each gene and this is where the precision of RNA-Seq can be of help. This often ties transcriptomics into proteomics and there is much active research into how activity in the transcriptome lead to the proteome that is observed as a result [54].

Epigenomics is the study of the processes that produce different phenotypes from a single genotype such as gene expression regulation through chemical changes to DNA that result in the restriction or access to translation binding sites so that genes are transcribed [55]. It plays a large role in the initiation of different stages of development and tissue differentiation where no biological information is transmitted to following generations of cells but there are observable physical differences between them [56]. Both areas of study have the unique characteristic where the changes in gene expression are dynamically influenced by environmental conditions compared to the genome itself which is relatively static throughout the lifetime of an organism.

Transcriptomics has been particularly influenced by the progress of RNA-Seq by allowing researchers to assess a whole transcriptome. Before NGS, it was very daunting to undertake any research on the transcriptome of a novel organism as the methods required that the genes and regulatory regions were already known, such as a fully annotated and sequenced genome. NGS has removed many of those barriers so research in transcriptomics has surged. Due to the complexity of interactions between genes and their participation of multiple systems, time series experiments are the most appropriate RNA-Seq approaches so far. The multiple time points are able to capture the dynamics of whole regulatory networks by providing data to compare to, and enabling the identification of possible dependencies and regulators [57]. These types of experiments give the best results when evaluating drug responses, cell product yields and development of diseases.

There are different approaches to designing time series experiments with varying time

points and conditions, but the one thing in common they all have is that the time points increase the dimension of the data set, thus increasing complexity during data analysis [58]. The most common type of analysis is identifying differentially expressed genes. Due to the difficulty presented by the dimensionality of the data set and the recent trend in time series experiments, there are only a small number of standard analysis methods such as Next maSigPro [59] and EBSeq-HMM [60]. The common techniques uses amongst them rely on Bayesian statistics or Gaussian processes. In contrast, there are not so many downstream analysis methods that interpret the data in terms of function and context within what is already known. So far, current approaches have considered clustering and enrichment analyses, such as hierarchical Gaussian process modeling [61] and functional enrichment analysis in FGNet [62].

There is currently a scarcity of analysis methods that cater to time series RNA-Seq data that focuses on the biology behind the data. This leads to a gap in the analysis pipeline in contrast to experiments such as microarrays that have an established workflow from analysis methods developed through much time and research. The current status of data analysis development presents an opportune time to offer novel approaches in an effort to complete a standard analysis pipeline. I will present three different analysis methods that fills in the gap and show how they work using appropriate time series RNA-Seq data. To that end, I have developed the methods that evaluates data in a different way. Instead of averaging gene expression values over all time points, my methods make use of the information from the change in expression as genes are up-regulated and down-regulated at each time point. The methods rely on using model organisms as a base template to compare to and to set up the initial framework. They also use network inference as a model to explain and process the results of RNA-Seq data. Each procedure can be successfully applied to a model organism to show how it operates from a different direction of approach, although they are constructed to function for experiments with limited samples from novel organisms.

The target organism chosen to demonstrate the featured analysis methods is the oleaginous diatom, *Fistulifera* sp. strain JPCC DA0580 or *F. solaris*. This pennate diatom was discovered in the junction of Sumiyo River and Yakugachi River, Kagoshima, Japan (Matsumoto et al., 2010, 2014) with initial observations indicating an advantageous propensity to produce biofuel compounds for industrial purposes [63]. It was chosen due to the increased interest in diatoms as a source of bioactive compounds and its potential use as a source of renewable biofuel as a recently discovered organism. Typically, algae demonstrate an ability to efficiently recycle carbon emissions and are, as a group, responsible for approximately 20% of the global carbon fixation while diatoms

themselves make up about 40% of marine sources [64] [65] [66] [67].

Interest in biofuel development derived from algae has risen due to several inherent benefits [68] [69] [70]. Algal cultivation plants would require less land and yield more biomass in contrast to terrestrial crops such as soybean, *Jatropha*, oil palm, sunflower and corn [71] [72] [73] [74]. The development of cultivation technology enables algae to be farmed in open tanks or closed columns where they do not deplete soil meant for agricultural use. Depending on the species, they can use water sources such as waste or saline water, further decreasing competition with agricultural crops. Oil accumulation in microalgae is also typically higher in magnitude than that of biofuel crops as well as faster in terms of rate production. They are also easier to work with in terms of genetic engineering as their genomes are not as complex and their cell structure is easier to manipulate[68] [69] [70] [75] [76]. These points are important for large-scale industrial production to minimize competition with the production of consumable food or with the preservation of neighboring habitats. However, in order for biofuels to be a realistic source of alternative renewable energy, lipid production capabilities need to be improved substantially. Production costs are still relatively high at US $300 - $2600 per barrel) compared to petroleum costs (US $40 - $80 per barrel) which leads to intensive research on commercial scaling and decrease costs [77] [78] [79] [80]. Previous investigations have inspected heterokonts, chlorophytes, dinoflagellates, haptophytes and rhodophytes [81] [82] [83] [84] along with recent studies on *Chlamydomonas reinhardtii* and *Nannochloropsis oceanica*. These studies have contributed much insight into metabolic pathways and genes that could be targeted for possible changes to increase optimum lipid production.

Oleaginous algae, as well as other land plants and various bacteria, can metabolize sunlight and carbon dioxide into chemical energy by reducing carbon molecules into long-chain fatty acids. They accumulate these compounds naturally as a means of energy storage similar to the storage in developing seeds, fruits and leaves [85]. The major product for biofuel production however, are neutral lipids such as triacylglycerol (TAG). The lipid production capabilities can be induced and enhanced in various stress conditions such as low nutrition, low nitrogen, salt stress and sulfur deprivation. This normally occurs at the expense of cell growth as the organism diverts glucose towards energy storage [86] [87] [88]. The reduction in growth and increase in lipid accumulation vary substantially depending on the duration and magnitude of stress from the environment as well as the type of algae [89] [90]. *F. solaris* is distinctly different with regards to oil accumulation and has the ability to accumulate lipids while undergoing logarithmic growth [91]. Additionally, it is able to reach a high neutral lipid content

while doing so (40% to 60%, w/w) [63] [91] [92]. The lipids produced by *F. solaris* are mainly composed of methyl palmitate (C16:0) and methyl palmitoleate (C16:1), both of which can be used as biodiesel fuel [93].

Discovering the mechanisms by which *F. solaris* is able to achieve such a unique feature can lead to the genetic modification of other algae to bestow similar production performance. Current methods rely on the over expression of positive regulation genes or the knock-out of those genes involved in negative regulation [89] [94] [95]. While they have met with some success, the effectiveness of such changes nonetheless rely on the understanding of the effect of the target genes in the relevant metabolic processes attributed to lipid accumulation [96].

The application of these methods on the *F. solaris* data set will illustrate that the analyses can be performed on small sample data from a novel organism collected at multiple time points. The results will be combined to show how a combination of methods will contribute to the overarching biological processes that can be extracted from a time series RNA-Seq experiment.

Chapter 1

Materials

The featured methods are applied to several RNA-Seq gene expression data sets from three organisms; *F. solaris*, *Phaeodactylum tricornutum* and *Arabidopsis thaliana*. *F. solaris* is a novel diatom that is closely related to the model diatom *P. tricornutum* while *A. thaliana* is the model plant organism. The *F. solaris* data is featured in all methods while the *P. tricornutum* data is used in section 2.1.1 and *A. thaliana* data is used in section 4.1.2.

The *F. solaris* genome was sequenced by the Roche 454 GS FLX Titanium DNA Pyrosequencer and the library is built by the GS FLX Titanium General Library Preparation Kit as per manufacturer instructions. The RNA-Seq expression data was measured using Illumina Genome Analyzer IIx and sampled while it was cultured in control and treatment substrates. The treatment substrate of artificial sea water is the f/2 medium (Guillard and Ryther, 1962) (75 mg $NaNO_3$, 6 mg $Na_2HPO_4 \cdot H_2O$, 0.5 μg vitamin B_{12}, 0.5 μg biotin, 100 μg thiamine HCl, 10 mg $Na_2SiO_3 \cdot 9H_2O$, 4.4 mg Na_2-EDTA, 3.16 mg $FeCl_3 \cdot 6H_2O$, 12 μg $CoSO_4 \cdot 5H_2O$, 21 μg $ZnSO_4 \cdot 7H_2O$, 0.18 mg $MnCl_2 \cdot 4H_2O$, 70 μg $CuSO_4 \cdot 5H_2O$, and 7 μg $Na_2MoO_4 \cdot 2H_2O$ per litre of artificial seawater). The control substrate is a 10 fold dilution of the above substrate [93]. The treatment substrate induces oil accumulation over a period of three days so the cultures are initially in the control substrate and subsequently introduced to the treatment substrate, and sampled at successive the time points 0, 24, 48 and 60 hours after the introduction. The RNA-Seq data is reported in Reads Per Kilobase per Million (RPKM) [50] and aligned to the genome using Bowtie v0.12.7 [17] [18]. The RPKM values were then calculated by ERANGE software v3.2 [50].

P. tricornutum is one of the main model organisms for pennate diatoms and is studied for its ease of transformation and annotated genome information [97] [98] [99]. The expression data for *P. tricornutum* is measured from samples cultured in control and

treatment substrates taken at 48 hours after being introduced to the substrates [100]. The substrate is a f/2-Si medium from Daya Bay, Huizhou, China (75 mg $NaNO_3$, 5.65 mg $NaH_2PO_4 \cdot 2H_2O$, 4.16 mg Na_2 EDTA, 3.15 mg $FeCl_3 \cdot 6H_2O$, 0.01 mg $CuSO_4 \cdot 5H_2O$, 0.022 mg $ZnSO_4 \cdot 7H_2O$, 0.01 mg $CoCl_2 \cdot 6H_2O$, 0.18 mg $MnCl_2 \cdot 4H_2O$, 0.006 mg $Na_2MoO_4 \cdot 2H_2O$, 0.0005 mg vitamin B_{12}, 0.1 mg vitamin B_1 and 0.0005 mg Biotin per litre of natural seawater). The nitrogen was removed for the treatment substrate and introduced into the control substrate.

The *A. thaliana* expression data is measured during the transition from vegetative development to flowering stages [101]. The full data set spans 10 days with a sample taken each day. The methods are applied to data measured from the M4 sample because this model organism is known to start floral transition around that time while earlier time points are during vegetative growth. The data was measured by Illumina HiSeq2000 and sampled from carefully synchronized plants so that samples were at the same growth stage as much as possible.

The expression data for all organisms were normalized to decrease the effect of very high or very low RPKM values in the following analysis methods. This is necessary as RNA-Seq is a competitive profiling method such that extreme values can distort the analysis. The RPKM values are corrected using the sRAP package in R [102] [103] with the recommended threshold of 0.1 for minimizing the influence of the number of reads. The normalization process also log transforms the RPKM values so that they become logged gene expression. Sequences with RPKM values of 0 for all time points and conditions are excluded from all analyses.

For all methods, the *F. solaris* genome was annotated with KEGG annotations by sequence alignment software SSEARCH that is run with the MIQS substitution matrix [40] [104] [41]. The cut off E-value for match establishment is 0.0001. The KEGG annotations is used in conjunction with the full KEGG database for annotations to other databases such as UNIPROT.

Chapter 2

Homolog Comparison in the *F. solaris* Genome

Initial investigations provide a broad overview of a novel organism so genome comparisons are a strategic way to start that. Through some of these initial investigations, it has been noted that *F. solaris* is related to other known diatoms such as *Thalassiosira pseudonana* and also particularly to the model diatom *P. tricornutum*. By relying on the information gathered from annotated genomes that are close in evolutionary and sequence distance, the annotation of *F. solaris* can be assembled faster.

A comparison of their genome sizes suggests that *F. solaris* underwent duplication events since the *F. solaris* genome has 20,470 genes while *P. tricornutum* has 10,402 genes and *T. pseudonana* has 11,776 genes. This is further emphasized when comparing the *F. solaris* genome to itself where it was noted that many genes are present in duplicates of highly conserved but not identical sequences. It is common to observe genome duplication in plants, such as maize, where the duplicated genes bestow increased genetic variation and viability since duplications lead to more sequence space for evolution to act on [105] [106] [107]. Although many duplicated genes can lose their functions and become lost, there are known duplicate genes that have retained function and continue to contribute to organism fitness [108] [109]. It is also possible that some gene duplicates can acquire new functions that contribute to an organisms survival so examining duplicates can be a worthwhile pursuit [110]. In algae, the duplications appear to increase adaptations to stress response mechanisms such as nutrient deprivation [111] which in turn makes the algae valuable as a biofuel source as the process typically involves environmental stress.

Close inspection of genome similarity also reveal that the majority of homologous *F.*

solaris and *P. tricornutum* genes include genes that are known to affect lipid accumulation such as those in photosystem I and II [112] [113]. These genes and other pathways related to lipid metabolism have overlapping processes among *P. tricornutum* and other diatoms [99] [114], indicating that the underlying mechanisms are closely related. This association in lipid production between *F. solaris* and *P. tricornutum* has helped determine putative pathways for fatty acid desaturation in *F. solaris*. However, although *F. solaris* and *P. tricornutum* share similar lipid metabolic genes and pathways, the precise mechanisms that control the level and duration of expression is unknown. The differences in lipid production and accumulation suggests that the lipid metabolic genes are being expressed under different regulation mechanisms between the two species. These factors and metabolic cellular processes are likely to be involved in the unique capacity of *F. solaris* to simultaneously grow and accumulate lipids.

These genomic features of *F. solaris* establishes a base for comparison analyses. This method is made to target the difference in gene expression between two data sets by relying on the existence of homologous genes and their involvement in identical pathways. The data sets that will be used are from *F. solaris* and *P. tricornutum* while they were cultured in similar nutrient deprived conditions that induces the accumulation of lipids. By using data sets that are as similar as possible, the power of detection can be increased for better results. Attention was specifically focused on the expression of the homologous genes only as they contain the fundamental differences between the two gene regulation mechanisms. The method establishes that differences in regulation exists and then isolates the genes related to the observed differences. They are examined for particular expression patterns with a strong focus on identifying expression patterns present in one but not the other data set. Homologous genes exhibiting different levels of expression with particularly polarizing patterns are sorted into groups where they are characterized by functions and pathways to describe the patterns observed in the analysis. In this application of this method on *F. solaris* data, the identified groups that are of most interest are the ones associated with lipid metabolism.

2.1 Method

2.1.1 Comparative Genomics

The sequences of *F. solaris* were compared to sequences of other species to identify homologs between *F. solaris* and *P. tricornutum*. The initial search was performed against the full KEGG database [41] [40] and the species of the top matching sequence were noted while *F. solaris* sequences without a suitably stringent match was removed from

further analysis. The remaining *F. solaris* sequences with KEGG matches were screened against *P. tricornutum* and *T. pseudonana* sequences where it was confirmed that the best species to use for comparative study was *P. tricornutum*. The resulting sequences from *F. solaris* and *P. tricornutum* were assigned as homologs and used in the rest of the analysis.

2.1.2 Expression Data

The fold change expression of the *F. solaris* homologs with four time points were separated from the full set of expression data. These were checked with the homologs found in subsection 2.1.1 since some sequences were not found in the expression data and vice versa. The comparison data was the fold change expression data of *P. tricornutum* with one time point. The fold change of the *P. tricornutum* homologs were extracted from the full data set and was also checked since the same *P. tricornutum* homolog could be the homolog to multiple *F. solaris* sequences.

The *F. solaris* data contained more than one fold change value per sequence while the *P. tricornutum* data contains one fold change value per sequence. The fold change value chosen to represent *F. solaris* in the rest of the analysis was decided by comparing the fold change at each time point with the fold change of *P. tricornutum*. The fold change profile most similar to *P. tricornutum* was the fold change at 60 hours which is consistent with previous observations [91]. This was then chosen to be used for analysis together with the expression data for *P. tricornutum* taken at 48 hours. These fold change values were checked using the lsmeans package in R [102] [115] to compute linear combinations of more than one mean with species, condition, and homology as factors.

The *F. solaris* genome contains duplicate sequences so there were *F. solaris* homologs that were best matched to the same *P. tricornutum* sequence. The fold change of *F. solaris* homologs were averaged across each matching *P. tricornutum* homolog using Equation 2.1.1 such that each homolog was in the analysis only once.

$$F_x = \frac{\Sigma_{\forall i} s_i}{n} \tag{2.1.1}$$

where F_x is the average fold change for *F. solaris* homologs that were best matched to *P. tricornutum* sequence x, s_i is the ith *F. solaris* homolog that is best matched to *P. tricornutum* sequence x and n is the number of *F. solaris* homolog that were best matched to *P. tricornutum* sequence x.

The prepared fold changes for each *F. solaris* F_x and *P. tricornutum* x homolog were then

used to calculate the difference in fold change between the two diatoms as detailed in Equation 2.1.2. The difference in fold change was used as the criterion to quantify the variation between *F. solaris* and *P. tricornutum* in the rest of the analysis.

$$D_x = P_x - F_x \tag{2.1.2}$$

where D_x is the difference between the average fold change of *F. solaris* sequences homologous to *P. tricornutum* sequence x and the fold change of *P. tricornutum* sequence x, P_x is the fold change of *P. tricornutum* sequence x and F_x is the average fold change for *F. solaris* homologs that were best matched to *P. tricornutum* sequence x.

2.1.3 Significance Test

A threshold test was applied to differentiate fold change differences from that of the background of expected differences. The differences were approximated by a normal distribution through the Central Limit Theorem so they were standardized by transforming the differences in fold change D_x into z-scores by using Equation 2.1.3.

$$z_x = \frac{D_x - E(D)}{SD(D)} \tag{2.1.3}$$

where z_x is the z-score of D_x, D_x is the difference between the average fold change of *F. solaris* sequences homologous to *P. tricornutum* sequence x and the fold change of *P. tricornutum* sequence x, $E(D)$ is the expected value of D_x for all x and $SD(D)$ is the standard deviation of D_x for all x.

The threshold was calculated by setting a significance level of 1%, and using the distribution of z_x to infer a cut off value. Fold change differences between *F. solaris* and *P. tricornutum* homologs outside the threshold were selected as being significantly higher or lower than expected.

2.1.4 Gene Ontology

The gene ontologies of the significant homologs were tested for significance using the hypergeometric test in the GOstats package in R [102] [116]. The GOstats package was used due to the inclusion of an option to perform a conditional hypergeometric test. It avoids an issue created when testing gene ontologies that are in a hierarchical structure of the gene ontology graph. The resulting p-values were corrected for multiple testing

using Bonferroni's correction method. Gene ontologies with a p-value < 0.05 were selected to represent their respective group.

2.2 Results

2.2.1 Comparative Genomics

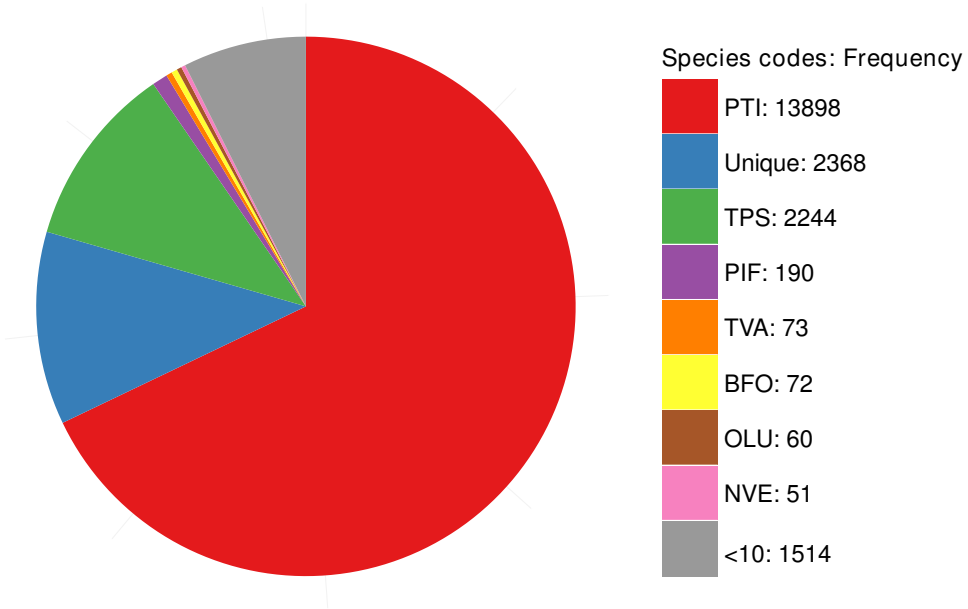

Figure 2.1: The species assignment of 20,470 *F. solaris* genes from searching the KEGG database. The species codes corresponds to the following species, in order from top to bottom: *Phaeodactylum tricornutum* (PTI), *Thalassiosira pseudonana* (TPS), *Phytophthora infestans* (PIF), *Trichomonas vaginalis* (TVA), *Branchiostoma floridae* (BFO), *Ostreococcus lucimarinus* (OLU) and *Nematostella vectensis* (NVE). Assignments to species with less than 10 matches were consolidated into the <10 group for clarity.

A total of 13,898 out of 20,470 *F. solaris* genes were observed to be most similar to *P. tricornutum* genes (68%) against the KEGG database (Figure 2.1). This was the largest majority of matched results. It was followed by unmatched or novel genes, and then by *T. pseudonana* genes. The remaining genes were matched with those from other algae and microorganisms. Since the majority of the matches were *P. tricornutum* genes, it confirmed that they should be kept and used as the homolog genes for comparisons for the remainder of the analysis.

To further validate the use of the *P. tricornutum* genome for comparing gene expression data, another sequence alignment was performed against the *P. tricornutum* and *T. pseudonana* genomes only (Figure 2.2). The majority of the *F. solaris* sequences (73%) are

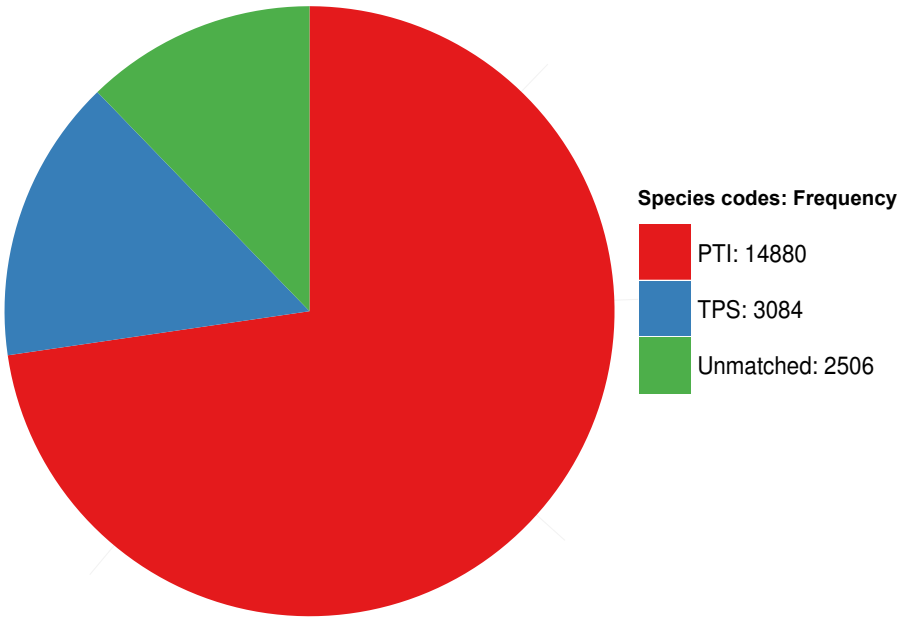

Figure 2.2: The species assignment of 20,470 *F. solaris* genes from searching the *P. tricornutum* and *T. pseudonana* genomes only. When searching in this reduced space of two diatom genomes, there are more *F. solaris* matches to *P. tricornutum* and *T. pseudonana* than in Figure 2.1, and more *T. pseudonana* matches compared to unmatched sequences.

more similar to *P. tricornutum*, affirming that the best organism to perform a comparative analysis on is *P. tricornutum*.

Due to gene duplication in many of the *F. solaris* genes, there were genes that were matched to the same *P. tricornutum* genes multiple times. Overall, there were 13,898 *F. solaris* genes which best matched to 6,589 *P. tricornutum* genes. The matching frequency showed two distinct patterns (Table 2.1). As the number of matched *F. solaris* genes increased, the frequency of the match type decreased. For example, there were 438 *F. solaris* genes which best matched the same *P. tricornutum* gene 3 times. In contrast to that, 26 *F. solaris* genes were best matched the same *P. tricornutum* gene 26 times. Additionally, the frequency of matches of even numbers were often higher than the matches of odd numbers. For example, there were 1044 *F. solaris* genes matching the same *P. tricornutum* gene 4 times but only 150 *F. solaris* genes matched the same *P. tricornutum* gene 5 times.

After completing the genome comparison, the gene expression was also investigated to verify that the data sets were analogous. The difference in gene expression between *F. solaris* and *P. tricornutum* was examined with an exploratory analysis utilizing visual graphs and distribution tests (Figure 2.3). The difference in gene expression was

examined across species, condition and homology using an unweighted ANOVA. The significance test found a significant interaction between species and homology (p-value < 0.05). The differences between the means of species and homology were inspected in more detail and no statistical significance in gene expression was found between the homologous genes of both diatoms (p-value < 0.05), and homologous *F. solaris* genes and non-homologous *P. tricornutum* genes. The comparisons between gene expression showed that only homologous *F. solaris* sequences are comparable with *P. tricornutum*.

No. of *F. solaris* sequences to one *P. tricornutum* sequence	**Frequency of occurence**
1	656
2	10,792
3	438
4	1,044
5	150
6	288
7	98
8	56
9	27
10	100
11	44
12	48
13	52
14	28
15	30
21	21
26	26

Table 2.1: A summary of the number of sequence matches of *F. solaris* sequences on the *P. tricornutum* genome. The left column shows the number of *F. solaris* sequences that best matched one identical *P. tricornutum* sequences. The right column shows the number of *F. solaris* sequences for which such a match occurred. For example, in the first row there were 656 *F. solaris* sequences that best matched one *P. tricornutum* sequence each, in the second row, there were 10,792 *F. solaris* sequences that were best matched to one *P. tricornutum* sequence for each pair of *F. solaris* sequences and in the third row, there were 438 *F. solaris* sequences that were best matched to one *P. tricornutum* sequence for every three *F. solaris* sequences, etc. The sequence matching shows that even numbers of matches were more common than odd numbers of matches and that most *P. tricornutum* sequences were matched by two different *F. solaris* sequences.

After the exploratory analysis, the expression data for both *F. solaris* and *P. tricornutum* was filtered to only include homologous sequences from each species. The rest of the analysis was performed on 13,898 *F. solaris* entries and 6,589 *P. tricornutum* entries identified from the previous analysis. *F. solaris* sequences matched to the same *P. tricornutum* sequence had their RPKM values averaged as shown in Equation 2.1.1, resulting in 6,589 pairs of sequence expression data.

2.2.2 Significance Test

The difference in fold change was examined between homologous *F. solaris* genes and *P. tricornutum* genes to identify gene pairs with a difference in fold change in *F. solaris* compared to *P. tricornutum*. Following the threshold calculation detailed in Equation 2.1.3, the significance percentage was set at 1% in order to select sequences only if their calculated z-scores were lower than -2.5758 or higher than 2.5758. The presence of any interaction was checked for first, between the fold change of the two diatoms by plotting them on each axis. However, the data points were observed to cluster around the center, showing that the majority of the fold change in both organisms are very close to 0 (Figure 2.4).

After applying the threshold on the data set, the effect was visualized by highlighting the points that were over the threshold (Figure 2.4). The points that were identified were often extreme pairings of gene expression where *F. solaris* fold change was high and *P. tricornutum* fold change was low and vice versa. The threshold did not identify genes with very low high fold change or very low fold change in both data sets.

Overall, 194 *F. solaris* genes and 91 *P. tricornutum* genes were chosen from the data set.

These genes were categorized into four groups according to the direction of the fold change of the genes in *F. solaris* and *P. tricornutum*. Those with positive fold change in both diatoms were placed in group 1, those with positive fold change in *F. solaris* and negative fold change in *P. tricornutum* were placed in group 2, those with negative fold change in *F. solaris* and positive fold change in *P. tricornutum* were placed in group 3, and those with negative fold change in both diatoms were placed in group 4. They represent the quadrants in Figure 2.4 if the graph was quartered at the x and y axes. The largest group formed was group 3 followed by group 2, 4 and 1.

The gene expressions of each group were plotted separately to inspect the degree of differences in fold change (Figure 2.5). The expressions of genes in group 1 were distinguished by a larger fold change in *P. tricornutum* compared to *F. solaris*. The *P. tricornutum* genes also had a smaller range of gene expression than *F. solaris* genes, although this could have been attributed to the smaller number of *P. tricornutum* genes. The most relevant group of interest was group 2 as it was composed of up regulated *F. solaris* genes and down regulated *P. tricornutum* genes. Here, the absolute fold change of *P. tricornutum* genes were larger than *F. solaris* genes but in the opposite direction to group 1. Additionally, the mean gene expressions of *F. solaris* genes and the mean control expression of *P. tricornutum* genes were more similar than the mean treatment expression of *P. tricornutum*. In contrast to group 2, group 3 consisted of down regulated *F. solaris* genes and up regulated *P. tricornutum* genes. Although the mean fold change directions are reversed, the degree of differences are similar in value. The gene expressions in group 4 were similar to group 1 where the *P. tricornutum* fold change values were larger than *F. solaris* fold change values. However, the gene expressions for both diatoms had a larger range than those in group 1.

2.2.3 Characterization by Gene Ontology

The final part of the method sorted the genes into groups by gene expression so that they were distinct from each other. To find how the genes within each group were related to each other, an analysis was used that employed GO terms. Each group was characterized by the over-represented ontologies found by testing the terms using the hypergeometric test (Table 2.2).

Group 1 contained 6 *F. solaris* genes and 3 *P. tricornutum* genes that shared 6 gene ontologies between them. The significant over-represented gene ontologies that were identified were 4-aminobutyrate transaminase activity, pyridoxal phosphate binding, transferase activity (transferring nitrogenous groups), amino acid transport, organic acid transport and organic anion transport (Table 2.2).

Group 2 was larger than Group 1 and contained 35 *F. solaris* genes and 16 *P. tricornutum* genes represented by 20 gene ontologies. The hypergeometric test narrowed the list of gene ontologes down to 8 terms (Table 2.2). In general, they are related to cAMP-dependent protein kinases for regulating glycogen, sugar and lipid metabolism, hydrolase activity, isomerase activity, GTP cyclohydrolase activity, protein phosphorylation and transferase complex.

Group 3 was the largest group with 121 *F. solaris* genes and 57 *P. tricornutum* genes and it had the highest number of representative gene ontologies at 94 terms. As a result, some of the significant gene ontologes consisted of broad terms such as localization and membrane. The full list contained seven types of dehydrogenase activities, three reductase activities and three transmembrane transporter activities (Table 2.2). There were four specific singular terms, N2-acetyl-L-ornithine:2-oxoglutarate 5-aminotransferase activity, methylcrotonoyl-CoA carboxylase activity, oxygen-dependent protoporphyrinogen oxidase activity, phytochromobilin:ferredoxin oxidoreductase activity, and three broader terms, steroid binding, lactate transport and tetrapyrrole metabolic process. Also of note are the presence of terms related to mitochondria and chlorophyll in this group.

Group 4 was smaller than groups 1 and 2 and was made up of 32 *F. solaris* genes and 15 *P. tricornutum* gene. Altogether there were 16 gene ontologies between them. The significant gene ontologies were L-ascorbate peroxidase activity, oxidoreductase activity (acting on peroxide as acceptor), antioxidant activity, inorganic anion exchanger activity, response to oxidative stress, photosynthesis (light harvesting) and photosynthesis (Table 2.2). A few terms were related to oxidative stress and photosynthesis while others were related to transporter and peroxidase activity.

Group	**Gene Ontology**	**Corrected P-value**
1	4-Aminobutyrate transaminase activity	3.86×10^{-5}
1	Pyridoxal phosphate binding	4.25×10^{-4}
1	Transferase activity, transferring nitrogenous groups	9.91×10^{-3}
1	Amino acid transport	0.0133
1	Organic acid transport	0.0157
1	Organic anion transport	0.0197
2	GTP cyclohydrolase I activity	6.76×10^{-4}
2	cAMP-dependent protein kinase complex	9.82×10^{-4}
2	cAMP-dependent protein kinase regulator activity	2.59×10^{-3}
2	Phosphomannomutase activity	4.04×10^{-3}
2	Cyclohydrolase activity	4.04×10^{-3}

2	Transferase complex	0.0120
2	Protein phosphorylation	0.0130
2	Kinase regulator activity	0.0346
3	Epoxide dehydrogenase activity	4.21×10^{-5}
3	5-Exo-hydroxycamphor dehydrogenase activity	4.21×10^{-5}
3	2-Hydroxytetrahydrofuran dehydrogenase activity	4.21×10^{-5}
3	Mevaldate reductase activity	4.21×10^{-5}
3	3-Keto sterol reductase activity	4.21×10^{-5}
3	3-Ketoglucose-reductase activity	4.21×10^{-5}
3	Membrane	1.16×10^{-4}
3	Gluconate dehydrogenase activity	1.25×10^{-4}
3	C-3 sterol dehydrogenase (C-4 sterol decarboxylase) activity	1.25×10^{-4}
3	Isocitrate dehydrogenase activity	1.25×10^{-4}
3	Ammonium transmembrane transporter activity	2.90×10^{-4}
3	Steroid dehydrogenase activity	1.04×10^{-3}
3	Transport	1.60×10^{-3}
3	Localization	1.91×10^{-3}
3	Anion transmembrane transporter activity	6.48×10^{-3}
3	Lactate transmembrane transporter activity	8.53×10^{-3}
3	Lactate transport	8.53×10^{-3}
3	N2-acetyl-L-ornithine:2-oxoglutarate 5-aminotransferase activity	8.53×10^{-3}
3	Steroid binding	8.53×10^{-3}
3	Methylcrotonoyl-CoA carboxylase activity	8.53×10^{-3}
3	Oxygen-dependent protoporphyrinogen oxidase activity	8.53×10^{-3}
3	Phytochromobilin:ferredoxin oxidoreductase activity	8.53×10^{-3}
3	Tetrapyrrole metabolic process	0.0121
4	L-ascorbate peroxidase activity	7.53×10^{-10}
4	Oxidoreductase activity, acting on peroxide as acceptor	2.01×10^{-5}
4	Antioxidant activity	5.38×10^{-5}
4	Response to oxidative stress	3.20×10^{-4}
4	Photosynthesis, light harvesting	7.60×10^{-4}
4	Photosynthesis	1.40×10^{-3}
4	Inorganic anion exchanger activity	2.12×10^{-3}

Table 2.2: The list of significant gene ontology terms resulting from hypergeometric testing on the gene ontologies in each gene expression group shown in Figure 2.5. The p-values were corrected for multiple testing using Bonferroni's correction method. These significant gene ontologies were selected to represent the function and metabolism of their respective group.

2.3 Discussion

The genomes of *F. solaris* and *P. tricornutum* were confirmed to be very similar both in sequence and expression, and it has been shown that some of the shared genes still perform the same functions and are part of the same pathways [112] [113]. The comparison analysis showed that *F. solaris* has many duplicated genes and that they are homologs to *P. tricornutum* genes at an individual level so that several *F. solaris* genes can be homologs of one *P. tricornutum* gene. Gene duplication can act as a buffer for genetic mutations and also facilitate the creation of new gene function by assisting in adaptation and increase fitness so it may be an important feature for oil accumulation in *F. solaris* [105] [106] [107]. The large numbers of even numbered *F. solaris* genes to individual *P. tricornutum* genes indicate that most of the duplication events were doubling events and that there were also a small number of higher duplications for which the effect is still unknown. The duplicate genes should be investigated further in the future.

The comparison of gene expression between *F. solaris* and *P. tricornutum* as they were grown in low nutrient and control conditions found that the expression of homologous genes are generally similar across species. However, it was observed that the gene expression of homologous *F. solaris* genes are more similar to the gene expression of *P. tricornutum* genes than its own non-homologous genes. It indicates that some of the non-homologous *F. solaris* genes may have originated from another diatom such as *T. pseudonana*, and that they were most likely influenced by different regulatory mechanisms than *P. tricornutum* genes. Some of those genes would have also included the novel genes as previously identified in Figure 2.1. While those genes may hold a key component responsible for the difference in phenotype, their functions would require further experiments as there are a relatively large number of them.

After applying the threshold method to the data sets, I identified 194 *F. solaris* genes that had varying degrees of fold change differences between *F. solaris* and *P. tricornutum*. Since there were a relatively large number genes to investigate, I classified them based on their fold change values in *F. solaris* and *P. tricornutum*. Each group could then show

what the metabolic differences were and then interpretation can be concentrated on the different responses to the low nitrogen environment that was represented by the classifications.

The first group of genes that were of interest are in Group 2 where the genes were up regulated in *F. solaris* and down regulated in *P. tricornutum*. Some of the gene ontologies of the genes in this group were related to cell regulation such as cAMP-dependent protein kinase complex and protein phosphorylation. The other distinguishing gene ontologies were related to hydrolase, isomerase and transferase activity through parent terms like GTP cyclohydrolase I activity and phosphomannomutase activity. The degree of difference between the control and treatment gene expression was important to note as well. The genes in this group showed a large amount of down regulation in *P. tricornutum* compared to negligible levels of up regulation in *F. solaris*. It indicates that the difference in phenotype could be attributed to the down regulation of these processes in *P. tricornutum* rather than the up regulation in *F. solaris*. There were five *F. solaris* genes with the most pronounced difference in fold change between the two diatoms. These were fso:g2859, fso:g2860, fso:g12378, fso:g9753 and fso:g9752, which were homologous with the *P. tricornutum* gene estExt_Phatr1_ua_kg.C_chr_70081. Although there were no gene ontology annotations, the *P. tricornutum* homolog is noted for being up regulated in response to iron deficiency and is thought to be part of a secretory pathway [99]. The gene, fso:g9752, also contains a domain found in the *C. reinhardtii* gene, FEA1 where it seems to be required for growth in iron deficient conditions and was up regulated in *C. reinhardtii* [117].

The genes in group 3 were in the largest group where the genes were down regulated in *F. solaris* and up regulated in *P. tricornutum*. There were many gene ontologies representing this group and they can be broadly divided into stress response in low nitrogen conditions, energy management, and intracellular transport and localization. While the up regulation of these genes are expected in low nitrogen condition, it is important to note that these genes were down regulated in *F. solaris*. The degree of down regulation is comparatively small but it was the largest for *F. solaris* in all 4 groups. The difference between the two diatoms is distinct due to the treatment means while the control means are quite similar. The large difference in gene expression between *F. solaris* and *P. tricornutum* can be separated into two patterns; the fold change is large in *P. tricornutum* compared to *F. solaris*, or small in *P. tricornutum* compared to *F. solaris*. Among the genes with large fold change differences, there were five *F. solaris* genes with small fold changes compared to the homologous three *P. tricornutum* genes. These were fso:g11687, fso:g15667, fso:g5409, fso:g17137 and fso:g20250 in

F. solaris, and fgenesh1_pg.C_chr_21000194, estExt_fgenesh1_pg.C_chr_10568 and estExt_fgenesh1_pg.C_chr_30465 in *P. tricornutum*. The *P. tricornutum* gene, estExt_fgenesh1_pg.C_chr_30465, has been annotated with the calcium ion binding gene ontology due to the presence of a calcium-binding domain but the homologous *F. solaris* genes also contain a domain associated with 5'-Nucleotidase/apyrase. It is a family of enzymes that catalyzes the hydrolysis of nucleotide molecules that is present in many species and whose metabolic importance depends on its cellular location. For the second pattern where there is a large fold change in *F. solaris* compared to *P. tricornutum*, there were two *F. solaris* genes found that were homologous to one *P. tricornutum* gene. These genes were fso:g9841 and fso:g6241, and the homologous *P. tricornutum* gene was e_gw1.4.342.1. The *P. tricornutum* gene is described by two gene ontologies related to photosynthesis; oxidoreductase activity and phytochromobilin biosynthesis. Additionally, the *P. tricornutum* and *F. solaris* sequences contain domains associated with ferredoxin-dependent bilin reductase that synthesize phytobilin from heme.

The last two groups of genes represent genes whose fold changes were in the same direction but the degree of fold change is very different between *F. solaris* and *P. tricornutum*. In both groups, the degree of fold change is small in *F. solaris* and the degree of fold change is large in *P. tricornutum*. In Group 1 where the genes were up regulated in both diatoms, these genes were associated with metabolizing amino acids to keto acids. However the small size of this group may have affected the selection of gene ontologies. The last group, Group 4, is of a better size and contains genes that were down regulated in both diatoms. The gene ontologies for this group were mainly associated with response to oxidative stress as well as photosynthesis. Although the gene expression was down regulated in both organisms, the fold change in *F. solaris* is much smaller than in *P. tricornutum*. In particular, the four genes with the largest difference in fold change that was annotated was attributed the response to oxidative stress gene ontology. The genes were fso:g10215, fso:g4650, fso:g7681 and fso:g19255, and their homolog was estExt_fgenesh1_pg.C_chr_130213 in *P. tricornutum*. Interestingly, the gene with the greatest difference in fold change in this group was fso:g18615 or estExt_fgenesh1_pg.C_chr_140189 in *P. tricornutum* [99]. They contain a ferritin-related domain with similar descriptions to the domain in FEA1 in *C. reinhardtii* as described for fso:g9752 in Group 2. Upon closer inspection, the *F. solaris* fold change for fso:g18615 is so small, it is quite possible that it was misassigned from Group 2 due to biological variation.

In addition to GO terms, I pursued the investigation further by checking the KEGG pathway membership of the annotated genes in each group in an attempt to find un-

derlying metabolic processes in the results. From the list of significant genes identified in my analysis, 12 *F. solaris* genes had a KEGG annotation belonging in a metabolic pathway. These genes matched five KEGG orthologs that are present in 8 pathways however I excluded the highest level pathway called metabolic pathway. Importantly, the remaining pathways were closely associated with lipid metabolism. There were two *F. solaris* genes from group 3 that matched the PPOX KEGG ortholog in the porphyrin and chlorophyll metabolism pathway. This gene is involved in the biosynthesis of cholorophyll and is present as two isoenzymes in plants [118]. Another KEGG ortholog, L-ascorbate peroxidase, was matched by four *F. solaris* genes in group 4. This ortholog is active in antioxidant activities in glutathione metabolism, and ascorbate and aldarate metabolism pathways. Specifically, it is in a reversible reaction that metabolizes ascorbate into dehydroascorbate during glutathione metabolism and it undergoes the opposite reaction in ascorbate and aldarate metabolism [119]. The next KEGG ortholog was matched by two *F. solaris* genes in group 2. This ortholog is the GCH1 gene which is closely involved in THF tetrahydrofolate synthesis that is needed in the metabolism of amino acids and nucleic acids [120]. Similarly, the riboflavin metabolism pathway was also represented by the ACP KEGG ortholog that was matched by two *F. solaris* but these were in group 3. This gene takes part in a reaction that metabolizes cardiolipin into a fatty acid [121]. The last pathway was glycerophospholipid metabolism which was represented by the CLD1 KEGG ortholog that matched two *F. solaris* genes in group 3. The CLD1 gene is part of a reaction that metabolizes FMN into riboflavin [122]. Although it was unfortunate that there were only 12 annotated genes, they turned out to contain a lot of relevant information.

Overall, *F. solaris* showed little change when it came to genes associated photosynthesis and response to oxidative stress while *P. tricornutum* down-regulated those genes. When observing the genes that were down-regulated in *F. solaris* and up-regulated in *P. tricornutum*, the *F. solaris* genes didn't exhibit as strong of a stress response as *P. tricornutum* genes.

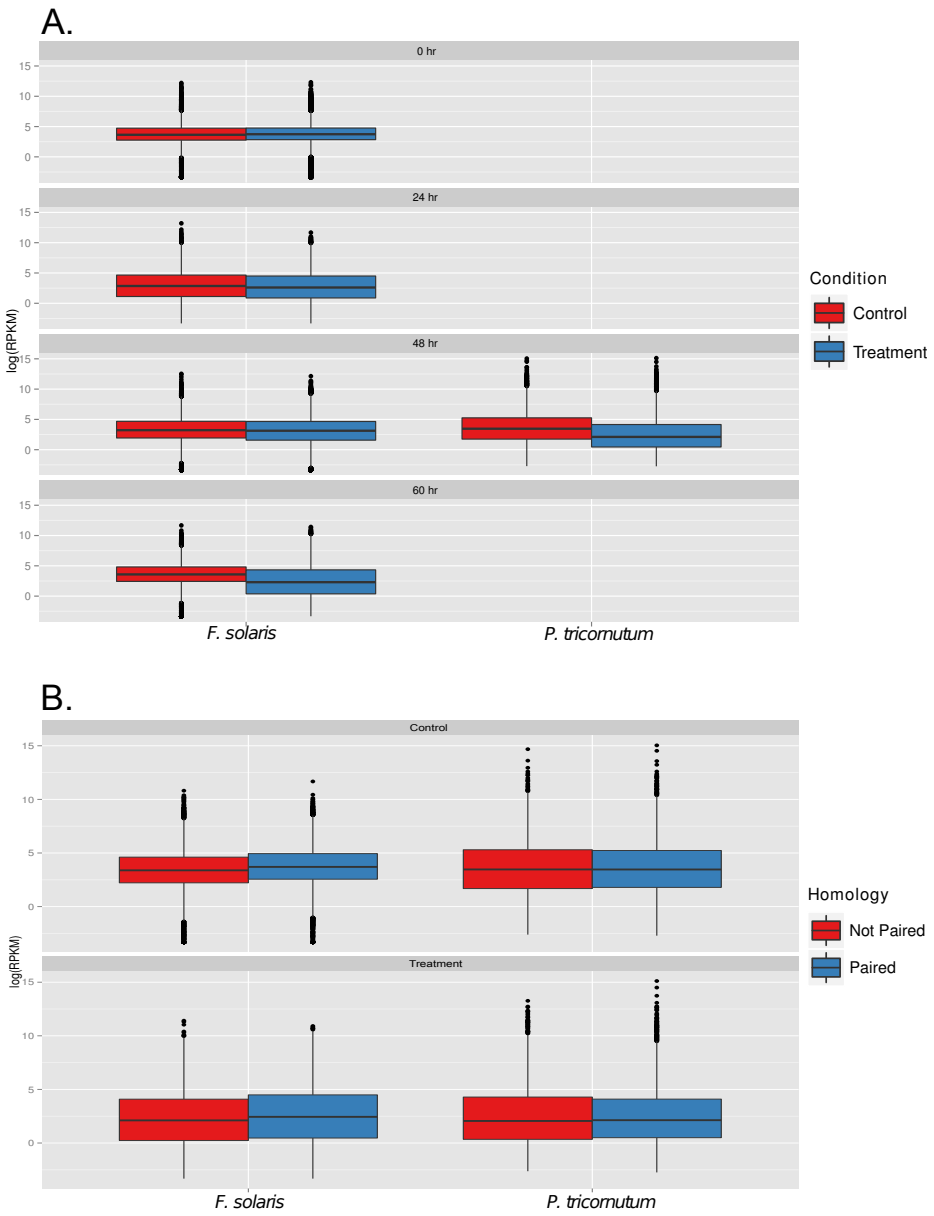

Figure 2.3: A comparison of all expression data from *F. solaris* and *P. tricornutum*. A) The expression of *F. solaris* and *P. tricornutum* separated by time. The expression of *P. tricornutum* at 48 hours is most similar to the expression of *F. solaris* at 60 hours. B) The expression of *F. solaris* at 60 hours next to *P. tricornutum* at 48 hours, separated by homology. *F. solaris* homologs show a more similar expression profile to *P. tricornutum* expression compared to non-homologous *F. solaris* genes. An unweighted ANOVA indicated that there was a significant interaction effect between species and homology (p-value < 0.05). A pairwise comparison showed that there was no significant difference between the homologous genes of both diatoms, and no significant difference between homologous *F. solaris* genes and non-homologous *P. tricornutum* genes (p-value < 0.05)

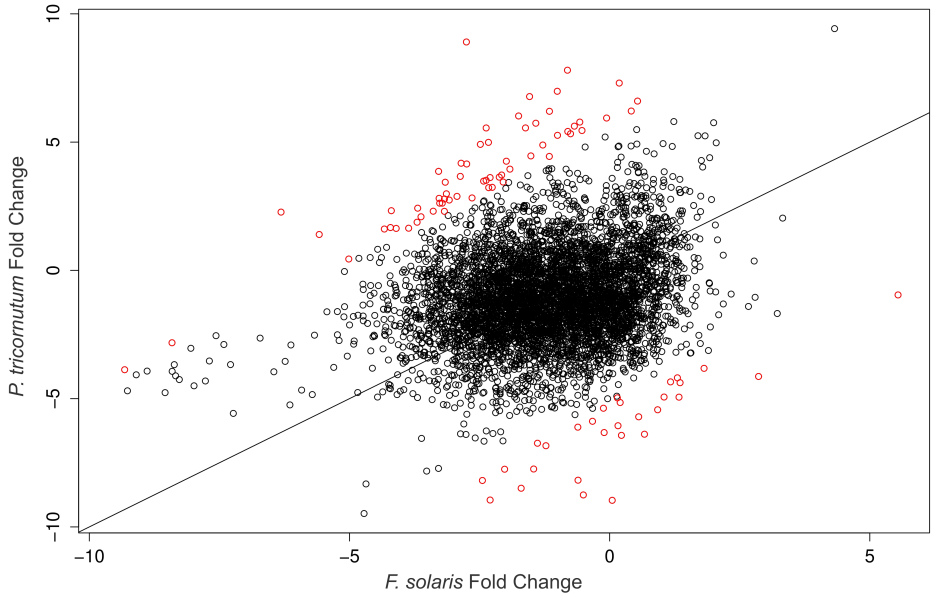

Figure 2.4: The fold change of homologous *F. solaris* genes versus *P. tricornutum* genes. The graph shows that the differences in fold change for most pairs of *F. solaris* and *P. tricornutum* homologs are centered around 0. After applying the threshold, the genes that were identified are highlighted in red. The threshold has identified these genes that exhibit a larger difference in fold change between the two data sets compared to the rest of the data. The threshold method selected 194 *F. solaris* genes and 91 *P. tricornutum* genes in connection to lipid accumulation.

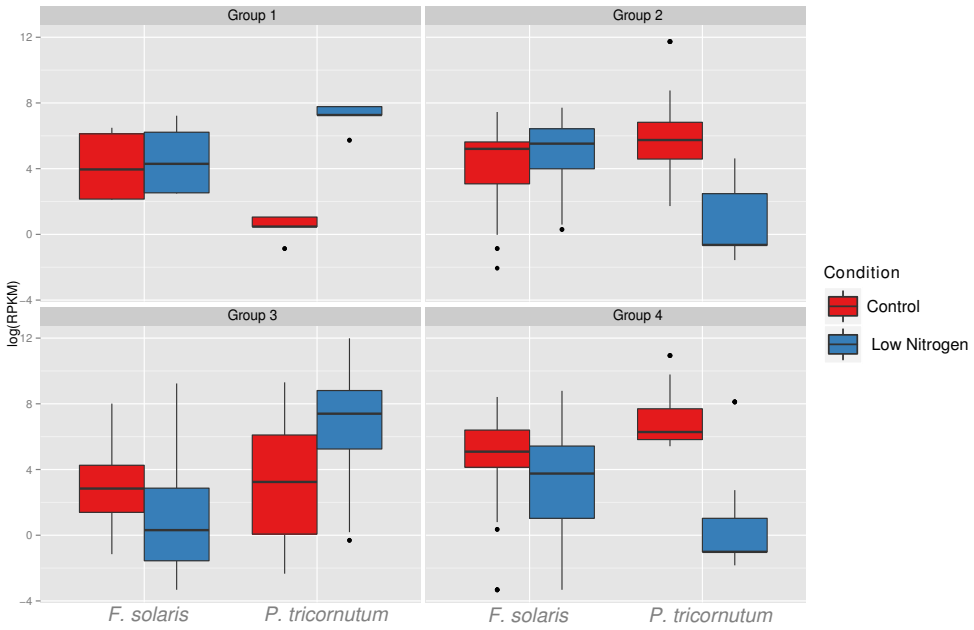

Figure 2.5: The expression data of the 194 significant *F. solaris* genes and 91 *P. tricornutum* genes separated into four groups. Group 1 contains genes that were up regulated in both diatoms and has 6 *F. solaris* genes and 3 *P. tricornutum* genes. Group 2 contains genes that were up regulated in *F. solaris* and down regulated in *P. tricornutum* and has 35 *F. solaris* genes and 16 *P. tricornutum* genes. Group 3 contains genes that were down regulated in *F. solaris* and up regulated in *P. tricornutum* and has 121 *F. solaris* genes and 57 *P. tricornutum* genes. Group 4 contains genes that were down regulated in both diatoms and has 32 *F. solaris* genes and 15 *P. tricornutum* genes.

Chapter 3

Activated Pathway Analysis for Triacylglycerol Biosynthesis

A lot of collaborative work has culminated in the metabolic pathway information that is available to researchers today. Through the work of many scientists, the elemental pathways of common cellular processes have been deduced in a wide range of organisms. As such, these model metabolic pathways contain the sequence of reactions that occur within a cell that starts from one compound to another, e.g. glucose to TAG. They give the framework of possible reactions and compounds available within a cell so investigating the genes that make up the framework will aid in the understanding of the compound synthesis of TAG.

There are several options for analyzing groups of genes and one of the most common tools is gene set enrichment analysis (GSEA). GSEA is a suitable tool when the data needs to be related to previous knowledge and works well for processes that involve a modest number of genes such as metabolism. GSEA approaches the data analysis by looking for associations between predefined groups of genes and a phenotype of interest. It tests for over represented relationships in groups, for example, at the top of a ranked list of genes. This type of method is more sensitive at detecting small but coordinated differential gene expression compared to linear modeling. There are a variety of GSEA tools available for analyzing high-throughput sequencing data [123]. They are available as online services like DAVID [124] [125], statistical packages for R like EBSeq [126] and SPIA [127], or standalone scripts like PAGE [128].

For the purposes of studying oil accumulation, the existing GSEA tools are insufficient for analyzing the time course data from *F. solaris*. For single time point experiments, tools like FuncAssociate and GOEAST are able to handle data from non model organ-

isms [129] [130]. ErmineJ and GAGE are even able to accommodate user specified annotation schemes if there is only one data point per gene [131][132]. These tools use either gene ranks or average gene expression across samples in their calculations. These basic ideas behind those methods are employed in my method with some added adjustments so that the methodology can preserve the differences in time series data and avoids reducing the amount of data further.

This modified approach to GSEA is made for analyzing time series data and thereby allows for the investigation of oil accumulation and growth metabolism in the *F. solaris* data set, a genome with a few annotated genes. It is also suitable for studies that aim to research metabolic pathways where a fuller annotated genome will be more advantageous. To accommodate time series data, the time points are treated as variables and GSEA is performed in high dimensions. To overcome some of the difficulties with working with multivariate distributions, resampling is used to create empirical cumulative distributions from which the p-value is calculated for enrichment. The results are interpreted and visualized using pathway graphs originating from KEGG [40] [41] so the algorithm uses KEGG ortholog annotations in the analysis. This approach allows the results to be clearly displayed with the changes in gene expression. Although there are existing pathway visualization tools [39] [40] [41] [133] [134], they were not suitable for showing the expression of genes from a novel organism on the pathway enrichment results in a way which focuses on the compounds instead of genes. The graphs from this method are created specifically to cater to this approach so that a hypothesized pathway of reactions can be calculated from one compound to another.

By first finding the metabolic pathways that is most regulated during oil accumulation, this method can then effectively show which compounds and genes should be targeted for further research. The application of this method on the *F. solaris* data will discover more about oil accumulation and will focus on the reactions that turn an energy source, glucose, into a biofuel lipid, TAG.

3.1 Methods

3.1.1 Expression Data

The *F. solaris* sequences were searched against the full KEGG database [40] [41] to identify and match KEGG ortholog annotations to the *F. solaris* sequences. Sequences with an accompanying KEGG annotation were used for the analysis. Some of these sequences were assigned to the same identifier so their gene expression values were

averaged using Equation 3.1.1. The *F. solaris* expression data was made from four sampling points so that there were four expression values per sequence. These values were put into a vector in order to retain as much data as possible for each sequence.

$$\mathbf{RPKM}_x = \frac{\Sigma_{\forall i}\mathbf{v}_i}{n} \tag{3.1.1}$$

where $\mathbf{RPKM}_x$ is a vector of logged expression values for ortholog x, $\mathbf{v}_i$ is the ith vector of logged expression values for ortholog x and n is the number of vectors with KEGG ortholog x.

The fold change $\mathbf{FC}_x$ of ortholog x was calculated using the control and treatment RPKM vectors $\mathbf{RPKM}_{x_{control}}$ $\mathbf{RPKM}_{x_{treatment}}$ during the normalization procedure as detailed in Equation 3.1.2.

$$\mathbf{FC}_x = \mathbf{RPKM}_{x_{treatment}} - \mathbf{RPKM}_{x_{control}} \tag{3.1.2}$$

where $\mathbf{FC}_x$ is the log fold change of ortholog x, $\mathbf{RPKM}_{x_{control}}$ is the control logged expression vector of ortholog x and $\mathbf{RPKM}_{x_{treatment}}$ is the treatment logged expression vector of ortholog x.

Consequently, from this point the data was then handled using the orthologs instead of individual genes as vectors of log fold change.

3.1.2 Gene Set Enrichment Analysis

The gene sets to be used in the analysis were established by considering the study objective. Generally, genes that share attributes of interest are grouped together to create gene sets such as functional groups or cellular location groups.

The objective of the *F. solaris* data set was to observe gene expression relating to oil accumulation so the categories of pathways in the KEGG database [40] [41] was used to chose 15 pathways associated with carbohydrates, 8 pathways associated with energy and 17 pathways associated with lipid metabolism. This resulted in a total of 40 gene sets for the analysis. Most crucially, they included the glycolysis and glycerolipid metabolism pathways which contains the two compounds central to oil accumulation, glucose and TAG.

Gene Set Enrichment Analysis Algorithm

Step 1: Subset a portion of fold change data that belong to gene set s. This is an observed fold change matrix of dimension n x 4 where n is the number of fold change vectors in the gene set and 4 is the number of time points in the *F. solaris* data.

Step 2: Calculate the mean fold change u, of gene set s. This observed value is a vector that will be used to calculate the p-value.

Step 3: Resample fold change values from the full data set, n times. This is a sample fold change matrix which has the same dimensions as the one in step 2.

Step 4: Calculate the mean fold change of the resampled data. This is a sample mean from one resampling cycle. It has the same dimension as the observed value in step 3.

Step 5: Repeat steps 4 and 5 6000 times. The sample means from each iteration are stored as rows resulting in a matrix of dimension 6000 x 4.

Step 6: Calculate the p-value by using the empirical cumulative distribution. The empirical cumulative distribution is defined by the following function

$$\hat{F}_s(\mathbf{u}) = \frac{\Sigma_{\forall i}\mathbb{I}(FC_i[1] \leq u_1, FC_i[2] \leq u_2, FC_i[3] \leq u_3, FC_i[4] \leq u_4)}{n} \tag{3.1.3}$$

where $\hat{F}_s$ is the empirical cumulative distribution of gene set s, $\mathbf{u}$ is the observed vector calculated in step 3, $\mathbb{I}$ is the indicator matrix, $\mathbf{FC_i}$ is the fold change vector for ortholog i in gene set s and n is the number of fold change vectors in gene set s.

Step 7: Calculate the p-value. It is the probability observing values as extreme as the observed vector $\mathbf{u}$.

The algorithm was implemented in R [102] and the empirical cumulative distribution was calculated using the mecdf package [135].

3.1.3 Enriched Pathway Graphs

The significantly enriched pathways were identified from the GSEA results and plotted to visualize the level of gene expression associated with the reactions of the compounds within them. The generic pathway and enzyme KGML files were downloaded from the KEGG database [40] [41] and read into R [102]. They were parsed using the KEGGgraph package [136] using the default data structure which depicted nodes as KEGG orthologs and edges as reactions. The default graph was restructured so that the nodes represented compounds and the edges represented KEGG orthologs. The

separate graphs of each pathway were then merged into one and converted into an igraph object for plotting and access to network analyses such as *get.all.shortest.paths* [137]. The merging reduced duplicate reactions and improved visualization by showing all the reactions in one diagram. Unconnected compounds were removed to reduce clutter in the final plot.

3.2 Results

3.2.1 Gene Expression

The *F. solaris* genome was aligned with sequences in the KEGG database. In order to make use of KEGG pathways for this method, the *F. solaris* sequences were annotated using KEGG ortholog matches. The sequence search returned matches to 2,873 orthologs from 20,470 *F. solaris* sequences.

The gene expression data was then filtered to remove non-ortholog sequences while the remaining data was processed into fold change and turned into RPKM vectors of size 4.

3.2.2 Gene Set Enrichment Analysis

Pathway Name	P-value
Photosynthesis	0*
Photosynthesis - antenna proteins	0*
Pentose phosphate pathway	0*
Carbon fixation in photosynthetic organisms	0*
Fatty acid biosynthesis	0*
Amino sugar and nucleotide sugar metabolism	0.0195
Fatty acid metabolism	0.0195
Methane metabolism 00680	0.026
Glycolysis	0.026
Oxidative phosphorylation	0.0325
Biosynthesis of unsaturated fatty acids	0.0455

Table 3.1: The list of enriched pathways resulting from GSEA and their enriched p-values. There were 11 pathways enriched out of 39 pathways tested.
*P-value <0.0001.

The GSEA method was performed on the expression data and it identified 11 significantly enriched pathways whose differential expression was significantly different between oil accumulation and non-accumulating conditions (Table 3.1). Importantly, the most significant pathway included photosynthesis related pathways and fatty acid pathways.

The photosynthesis and photosynthesis antenna protein pathways were two of the most significantly enriched pathways with p-values <0.0001. There was a positive relationship between log fold change and time, indicating that there is increased energy synthesis via photosynthesis during oil accumulation. The median log fold change at 60 hours in the photosynthesis antenna proteins pathway is the highest of all pathways at 8.0. Further investigation reveals that this is due to the log fold change values of light-harvesting complex I chlorophyll a/b binding proteins; LHCA1, LHCA2 and LHCA4.

The other prominent pathways are related to cellular energy metabolism; glycolysis, the pentose phosphate pathway and oxidative phosphorylation were significantly enriched. Both the average differential expression in glycolysis and the pentose phosphate pathway increased over time, particularly from 48 hours. In contrast, the differential expression in the oxidative phosphorylation pathway showed a small decrease in the 24 and 48 hour period before returning to approximately the same expression level it started at.

The other significant pathways are more closely related to synthesizing the materials for TAG and growth; they are fatty acid biosynthesis, biosynthesis of unsaturated fatty acids and amino sugar and nucleotide sugar metabolism. Interestingly, while genes involved in fatty acid biosynthesis were generally up regulated, the gene expression in biosynthesis of unsaturated fatty acids was not so consistent.

The next significantly enriched pathway, carbon fixation in photosynthetic organisms, has several overlapping genes with pyruvate metabolism, glycolysis and the pentose phosphate pathway. Noticeably, the pyruvate metabolism pathway was not significantly enriched individually as a gene set even though they are closely related to pyruvate metabolism.

The methane pathway was unexpectedly significantly enriched. Upon further investigation, it was discovered that the expression data within the methane pathway were also found within the other enriched pathways. For example, both glycolaldehyde dehydrogenase (ALDA) and 6-phosphofructokinase 1 (pfkA) are in the pentose phosphate pathway while (2R)-3-sulfolactate dehydrogenase (comC) is also found in the cystein and methionine metabolism where it takes part in reactions that make pyru-

vate.

3.2.3 Enriched Pathway Plots

To better visualize the results from GSEA, the enriched pathways were plotted as network graphs (Figure 3.1). The graph's nodes were set up as compounds as the focus was on the reactions and compounds instead of the usual approach using genes. Because the focus was shifted to compounds, the glycerolipid pathway needed to be added so that the key compound, TAG, was included.

The graph consisted of 353 compounds and 661 reactions. Most compounds were unique to their pathway but there were 18 compounds that were found in two pathways and 13 compounds that were found in three pathways. These included pyruvate, oxaloacetate and ADP. Expectedly, they are found in glycolysis, pentose phosphate metabolism and other related processes.

Once the graph was constructed, the shortest path of reactions from glucose to TAG was calculated. As the graph was created from pathways that showed a significant relationship with oil accumulation, it can be considered a hypothesized path of metabolic reactions that starts from glucose and can potentially produce TAG. Two shortest paths were found with a length of 11 compounds (Figures 3.2 and 3.3); the same path presented in KEGG contains 15 compounds.

The graph showed that the genes along the hypothesized paths were up regulated by plotting the differential expression direction on the edges of the graph. When viewed next to each other, the differential expression at each time point shows which reactions change differential expression direction (Figure 3.4). The genes along the identified shortest paths were identified and observed to be up regulated during the 60 hours of the experiment.

3.3 Discussion

Predictably, most glycolysis genes were up regulated through the 60 hour duration, although there were notable exceptions; phosphoglucomutase (PGM), phosphoglycerate kinase (PGK) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH). PGM transfers a phosphate group to and from the 1′ position to the 6′ position in α-D-glucose so its down regulation suggests that *F. solaris* is getting its source of α-D-glucose 6-phosphate elsewhere. PGK and GAPDH are used in two reversible reactions to make glycerate 3-phosphate which is a key molecule for TAG production [138]. However,

this reaction can be done in one irreversible step by glyceraldehyde-3-phosphate dehydrogenase (NADP) which was up regulated in the data. The substrate for that reaction, glyceraldehyde 3-phosphate, is used in the pentose phosphate shunt to make nucleic and amino acids like deoxyribose, 2-Deoxy-D-ribose 1-phosphate and D-ribulose 5-phosphate. The genes involved in those reactions were found to be up regulated in the data; they were ribokinase (rbsK), phosphopentomutase (PGM2), 6-phosphogluconate dehydrogenase (PGD) and 3-hexulose-6-phosphate synthase (hxlA). So it seems that *F. solaris* relies on glucose to produce TAG, nucleic and amino acids to achieve accumulation and growth at the same time while using a proton pump to power the reactions under low nitrogen conditions.

Along with with the glycolysis pathway, the oxidative phosphorylation pathway was also significantly enriched. Some of the proteins in that pathway form a membrane protein, the V-type ATPase. It is a proton pump responsible for ATP turnover in mitochondria and was up regulated in the data. There is some evidence of a relationship between increased C16-C18 length fatty acids, which are used in TAG production, and increased hydrolytic activity of V-ATPase [139]. Along with a gradual down regulation of NADH dehydrogenase, it would seem that *F. solaris* focuses on recycling ATP instead of reducing NADP+ for its energy requirements during oil accumulation.

Two of the significantly enriched pathways that were found were the photosynthesis and photosynthesis antenna protein pathways. They showed an increase in fold change together with the increase in time, showing that *F. solaris* increases photosynthesis during oil accumulation. The positive fold change was mainly attributed to the light-harvesting complexes. Specifically, most of the log fold change of proteins in light-harvesting complex II is lower than in light-harvesting complex I. The preference of light-harvesting complex I may be due to the highly efficient nature of photosystem I [140], although *F. solaris* is using both systems simultaneously in this case.

The positive fold change of photosynthesis related pathways related closely to the enriched pathways responsible for synthesizing the materials for TAG and growth. It was observed in the fatty acid biosynthesis pathway, biosynthesis of unsaturated fatty acids pathway and the amino sugar and nucleotide sugar metabolism pathway that the up regulated genes were involved in fatty acid elongation while the down regulated genes were involved in dehydrogenation. This is consistent with synthesis of fatty acyl residues in TAG. Gene expression in amino sugar and nucleotide sugar metabolism also had a positive trend through time. Their up regulation further suggests that sugars are being metabolized for growth during oil accumulation. For example, two of the up regulated genes are glucokinase (glk) and glucose-6-phosphate isomerase (GPI) which

are involved in reversible reactions that convert glucose into fructose which leads to the production of nucleotide sugars. As they are part of reversible reactions, it was difficult to discern whether the forward or backward reaction was dominant without further data but their up regulation means that there was a considerable amount of converting occurring.

The analysis also identified the interesting genes in the carbon fixation in photosynthetic organisms pathway as the genes that are associated with pyruvate. Here, malate dehydrogenase (decarboxylating) was up regulating the reaction that turns malate into pyruvate. In contrast malate dehydrogenase (oxaloacetate-decarboxylating) was down regulated. It could be due to the reactant, NADP, being used in other reactions, such as photosynthesis, that there seems to be a preference for the decarboxylating reaction. Although pyruvate itself was not enriched, it shares seven reactions with the carbon fixation in photosynthetic organisms pathway and is directly linked to 13 other pathways. Therefore, it can be said that oil accumulation affects the reactions in the carbon fixation pathway as a whole, instead of pyruvate specifically.

The visualization step of this method effectively showed the change in fold change through time. The increase in the number of green edges compared to red edges quickly summarized the trend in the data. The shortest path diagrams also assisted in visualization as the important reactions were isolated and made clearer. In addition to the general up regulation trend, the diagrams also showed that the up regulation occurred in sections along the path instead of being concerted. This suggests that gene expression of a phenotype does not change for every gene along the reaction path at a single time point. Instead, the change in gene expression occurs in sections which eventually leads to the up regulation of the full path.

Finally, the method itself was successful in reaching its aim although several issues did arise. The overlap of genes between gene sets can cause problems with detection, especially if the genes in one pathway has a strong signal as seen by the significant enrichment of methane. In this case, the genes in the pentose phosphate pathway have strongly defined differential expression values that is nullifying the expression pattern from the other genes. Although it is fairly reasonable for some genes to be present in multiple pathways, it should be checked if the overlapping genes are making biased contributions. The effect is further amplified in the data as the number of annotated genes are few.

Additionally, the two shortest paths that were found were very similar to each other, mainly differing between the use of glycerol or glycerone. Although a shorter path of reactions is possible, it is unknown where the reactions take place in the cell. If the

proteins are located close to each other, the hypothesized path of reactions could very well be how *F. solaris* produces TAG from glucose. Future experiments on metabolite quantity of the compounds along the path would also provide adequate evidence for the hypothesis.

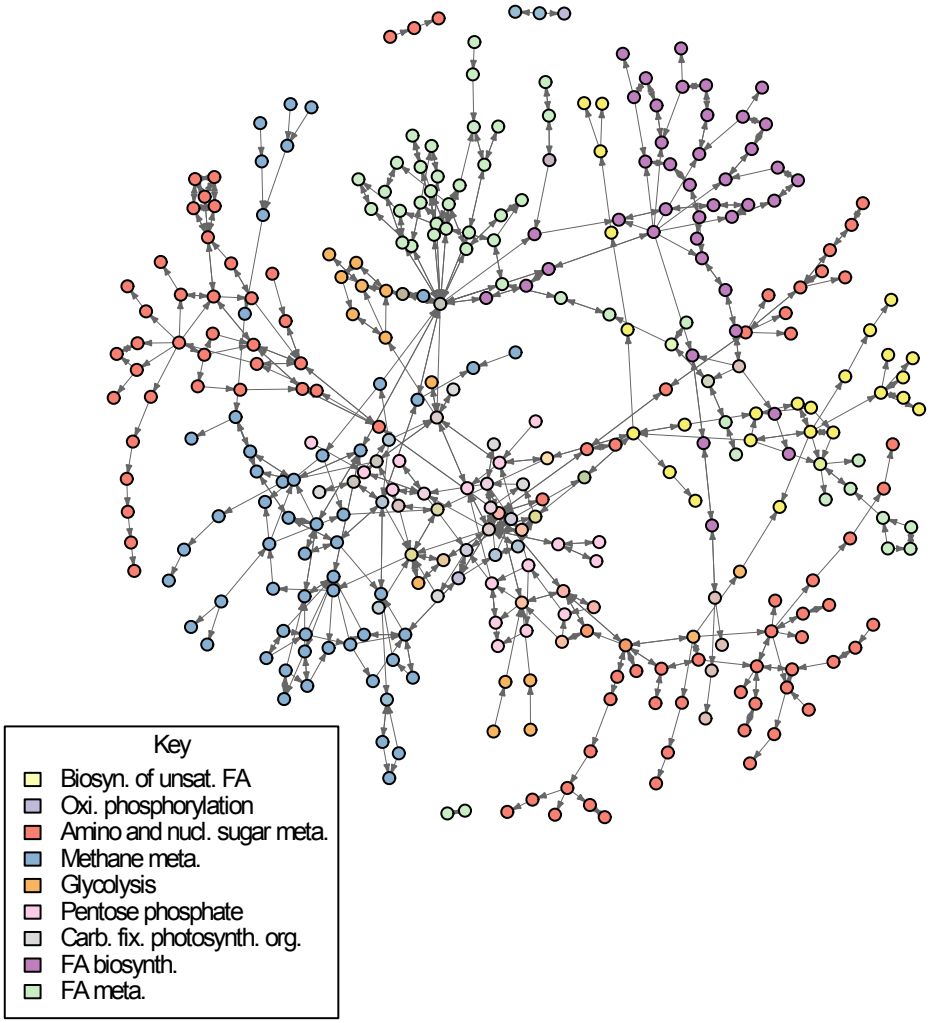

Figure 3.1: The graph of the significantly enriched pathways found using my GSEA method combined with the glycerolipid pathway. The full network contains 353 compounds and 661 reactions but compounds without reaction data were removed from the plot to reduce clutter. The graph is plotted with compounds shown as nodes and reactions shown as edges. The compounds are colored by their pathway membership; compounds belonging to 2 or more pathways are a mixture of the pathway colors. There were 13 compounds belonging to three pathways, 18 compounds belonging to two pathways and 159 compounds that were unique to their pathway. Many of the shared compounds are concentrated in the center of the graph and are closely related to glycolysis and pentose phosphate metabolism.

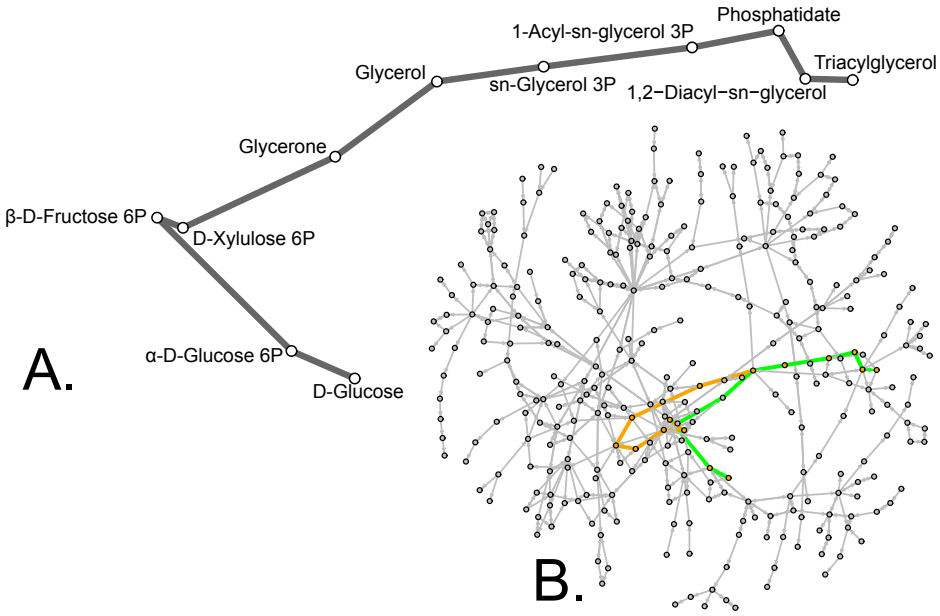

Figure 3.2: The first shortest path found in Figure 3.1 between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. In contrast, the path presented in KEGG is highlighted in orange. The shortest path contains 11 compounds while the KEGG path contains 15 compounds.

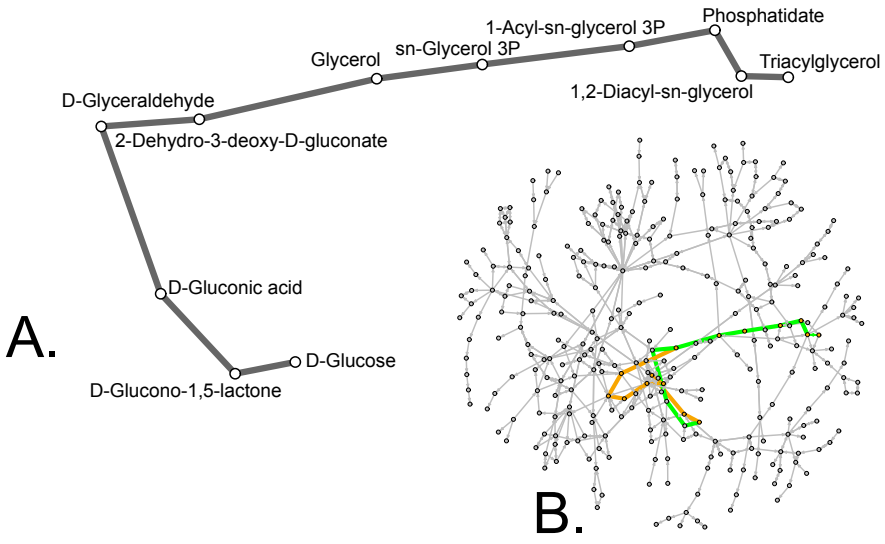

Figure 3.3: The second shortest path found in Figure 3.1 between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. In contrast, the path presented in KEGG is highlighted in orange. The shortest path contains 11 compounds while the KEGG path contains 15 compounds.

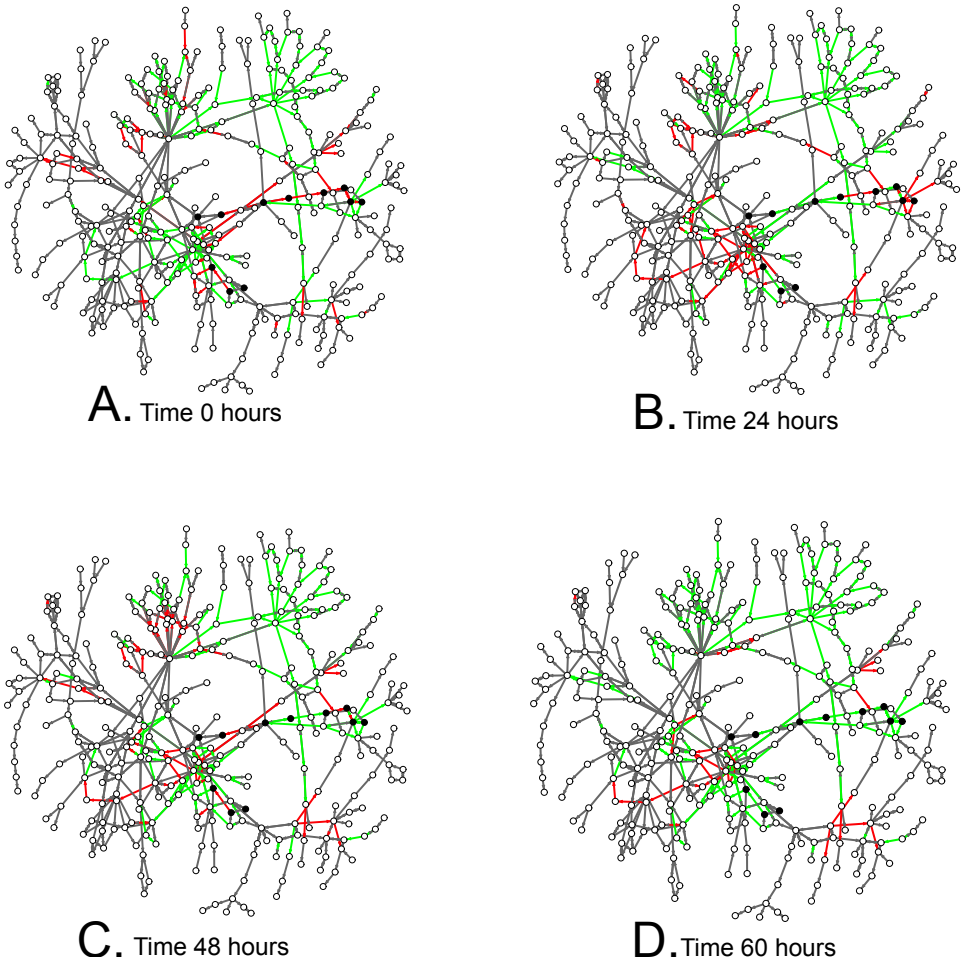

Figure 3.4: These graphs show the differential expression direction at each time point in response to oil accumulating conditions. The differential gene expression direction is shown as edges connecting two compounds. Genes that were up regulated during oil accumulation are highlighted in green while red edges represent down regulation. Genes for which there was no data were left as gray. The nodes highlighted in black are compounds in the second shortest path found between glucose and triacylglycerol (Figure 3.3). The genes along the path switch from red to green at different time points until the majority of genes along the path are green.

Chapter 4

Investigation of Transcriptional Regulation Mechanisms

Transcription factors are proteins that regulate gene transcription and thus affect the quantity, timing and pattern of gene expression in a cell. They are one of the main key components involved in gene regulation because they activate or suppress gene transcription activity, typically by binding to a transcription factor binding site close to a gene. This makes them an important part of understanding how organisms function. When a transcription factor is activated, it can affect the expression of other genes in a many-to-many relationship such as when an activated gene product acts as a catalyst for a reaction that will in turn activate other genes, or if the gene is another transcription factor that may regulate the expression of a separate process. Subsequently, the action of a transcription factor can be coordinated between other genes and transcription factors to affect many other gene expressions further downstream in many other metabolic pathways. This type of relationship happens frequently in a cell and is observed in many situations where there is a response to stimuli, such as metabolites, ligands, or environmental queues, such as light [141] [142].

A network is a type of model that is useful for analyzing many-to-many interactions such as the effects of transcription factors. As networks are analysis as well as visualization tools, they are very useful for examining gene regulation systems where one action affects many components. They are also more applicable for modeling complex relationships than linear models. NGS is able to capture an extraordinary amount of data and has the potential to probe living systems, particularly when done in time series experiments. Analysis of expression data from RNA-Seq take advantage of how the methods measure all of the data as a whole. This is reflected in the varied methods of detecting significant differences in gene expression or identifying genes associated

with a trait for example. Although not directly measured, gene expression regulation is present in the data as well and this method attempts to uncover it by using the full data set as a background while focusing on gene expression regulation via transcription factors.

This analysis method uses gene expression patterns to construct a network that models the interaction of transcription factors within a cellular system. By converting expression values into discrete units, called expression patterns, this method is able to infer a transcription factor network on temporal, small sample RNA-Seq data, extracting information about interactions occurring within the organism. It is an extension of the phylogenetic profile clustering method used for assigning protein functions [143], altered so that it can work on limited samples to produce a clear summary of the relationships between expression patterns. This method is first applied to expression data from *F. solaris* during 60 hours of its lipid accumulation phase in order to understand the process of lipid accumulation as it is an outstanding candidate for biofuel production. The method is then applied to expression data from *A. thaliana* during floral transition to show that the method corroborates the activity of known transcription factors as well as identifies other transcription factors that also provides assistance to expression regulation.

For applications to novel organisms such as *F. solaris*, this method is able to identify transcription factors that can be studied for increasing yield. It signifies decisive genes to research for better understanding that may be useful in biotechnology. The network uses only expression data when lipid accumulation is taking place and identifies gene regulation elements important in lipid accumulation, showing which genes were regulated first and which genes are affected by their regulation.

The application of this method on *A. thaliana* will create a model that shows the expression of transcription factors and its associations with floral transition [101]. The apical meristem growth expression data covers the transition from vegetative development to flowering is controlled by transcription factors responding to genetic and environmental stimuli such as hormones and light [144]. Several major transcription factors were outlined with the *A. thaliana* data [101]. Some of them positively regulate transitioning and flowering like SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1) [145] [146] and FLOWERING LOCUS D (FD) [147] [148]. Some of them negatively regulate flowering like FLOWERING LOCUS C (FC) [149] and SHORT VEGETATIVE PHASE (SVP) [150]. Other transcription factors initiate different stages during the transitioning period such as initiating flowering or controlling floral organ identity. These include LEAFY (LFY) [151] [152], CAULIFLOWER (CAL) [153] [154], PISTILLATA (PI) [155],

AGAMOUS and various SQUAMOSA PROMOTER BINDING PROTEIN-LIKE family (SPL) [156] [157] [158] [159] that go on to regulate APETALA1 (AP1) [160] [161]. The inferred network will show how the floral transition transcription factors are associated with each other and will confirm the identification of additional transcription factors that affect or are affected by known regulators. Additionally, the direction and weight of the network edges quantify the relationships between groups of transcription factors to highlight important regulation events at each time. Lastly, the transition process will be broadly described by the network using an enrichment analysis on the groups of transcription factors.

4.1 Method

4.1.1 Identifying Transcription Factors

For the application of this method, the aim was to model only transcription factor associations so the initial selection of transcription factor sequences was performed first with rigorous criteria.

For the novel organism, *F. solaris*, transcription factor candidates were determined by taking the intersection from the results of two methods to decrease false positive selection as much as possible (Figure 4.1). The first list of transcription factor candidates was generated by searching sequences for PFAM [162] and Superfam [163] domains and then cross-checking the domains in the DBD: Transcription Factor Prediction database [164]. The PFAM and Superfam search results were filtered by using an E-value cut off of 0.05 or if the sequences were known enzymes instead of transcription factors. The resulting sequences that qualified were marked as potential transcription factor candidates. The second list of transcription factor candidates was created by comparison to lists of transcription factors previously identified in the two related diatoms, *P. tricornutum* and *T. pseudonana* [165]. The *F. solaris* sequences annotated with the *P. tricornutum* and *T. pseudonana* transcription factors were marked as potential transcription candidates for this second list. The final list of transcription factor candidates was produced by taking the union of *F. solaris* sequences in both lists that also had expression data.

The process of identifying transcription factors in model organisms can be performed by accessing their annotations in online databases. With regards to the *A. thaliana* data set, the transcription factors were identified using the listings in the Plant Transcription Factor Database [166] which in turn utilized data from the TAIR database, release version 8.0 [167].

4.1.2 Expression Pattern Creation

The purpose of the expression patterns was to convey the change in time in gene expression for each gene so the raw expression data needed to be processed.

The *F. solaris* data was measured in control and treatment conditions at 0, 24, 48 and 60 hours. The processed data conveyed the change in gene expression by computing the log fold change between the control and treatment conditions which was calculated by Equation 4.1.1.

$$F_t(x) = T_t(x) - C_t(x) \tag{4.1.1}$$

where $F_t(x)$ is the log fold change of sequence x at time t, $T_t(x)$ is the treatment log RPKM of sequence x at time t and $C_t(x)$ is the control log RPKM of sequence x at time t.

The *A. thaliana* data set [101] was already normalized across time points so the difference in expression between adjacent time points was calculated directly by taking the difference of gene expression for adjacent time points. The time intervals that were used were M4-M5, M5-M6, M6-M7, M7-M8, M8-M9, M9-M10. Each time point was one day apart starting from 10 days after germination at time point M4 to 16 days after germination at time point M10.

The expression values for each time interval were binned into -1, 0 and 1 values, where -1 indicated decreasing expression, 0 indicated no change in expression and 1 indicated an increase in expression. To compensate for the experimental and biological variation in expression values, a threshold was used to determine whether a non-zero expression was large enough to be considered -1 or 1. As most genes are not related to the processes of interest, they do not generally exhibit as varied gene expression changes so their gene expression was used as a guide for the baseline change in expression. For each gene, the standard deviation in gene expression was calculated across all time intervals and then the median of all the standard deviations was calculated, and the positive and negative value of that was used as the threshold boundary for the 0 bin.

The threshold found for the *F. solaris* data set was 0.9152. The threshold found for the *A. thaliana* data set was 90.29. Succinctly, this put all gene expression between 0.9152 and -0.9152 into the 0 bin in *F. solaris* and -90.29 and 90.29 in *A. thaliana*. Values outside of those intervals were binned into 1 or -1 if they were above or below the interval respectively (Figure 4.2 and Figure 4.3).

Consequently, a vector was assembled for each gene using the binned values, thereby

creating the expression patterns that were entirely comprised of the 3 values, 0, 1 and -1. The elements of each vector were ordered in chronological order so that the first value was the 0 hour value for the *F. solaris* data and the M4-M5 value for the *A. thaliana* data.

4.1.3 Network Construction

Graph Structure

A network was constructed by using the expression patterns as nodes. Edges in the network were added to signify a relationship between the two patterns while the edge weights quantified the relationship.

A pair of expression patterns, nodes u and v, are denoted as $\vec{u}$ and $\vec{v}$ and are made up of values u_i and v_i at each position i within each pattern. All u_i and v_i have possible values of -1, 0 and 1. When $\vec{u}$ contains at least two different values, the first position where u_i is different from u_1 is named a. As $u_1 \neq u_a$, it stands to reason that $a \neq 1$. For example, in (0, 0, 1, 0, 0, 0), $u_{[1]} = 0$ and $u_{[a]} = 1$ so $a = 3$. An edge connecting $\vec{u}$ to $\vec{v}$ is made if there is only one u_i and v_i that is not equal to each other. The position where this difference is located is named b. For example, in $\vec{u} = (1, 0, 0, 1, 0, 0)$ and $\vec{v} = (1, 0, 0, 0, 0, 0)$ the difference is at u_4 and v_4, so therefore $b = 4$.

Edge Properties

The edge direction was decided by considering the expression of transcription factors at early time points affecting the expression of transcription factors at later time points. The following algorithm was developed to decide whether an edge started from u and ended at v.

Step 1: Remove the edge from u to v if either of the vectors are made up of only one value. These are the patterns consisting of only 0s or only 1s or only -1s and so is unrelated to time.

Step 2: Compare $\vec{u}_{[1:a]}$ and $\vec{v}_{[1:a]}$. If they are the same, then the edge direction from u to v is false because it means that the difference in expression pattern is after a. If they are different, continue to step 3.

Step 3: For u and v where $a = b$, check if $v_1 = 0$ and $v_a = 0$. If $v_1 = 0$ and $v_a = 0$, then the edge direction from u to v is true as the pattern in $u_{[1:a]}$ contains 1 or -1 while $v_{[1:a]}$ is 0. If they are both not equal to 0, then the edge direction is false.

Step 4: For patterns where a is different from b, check if $u_b \neq 0$ and $v_b = 0$. If $u_b \neq 0$

and $v_b = 0$, then the edge direction from u to v is true as the expression pattern 1 or -1 happens in u before v. If $u_b = 0$ and $v_b \neq 0$, then the edge direction is false.

The edge direction algorithm has been summarized in Figure 4.4 as a decision tree with examples to show how each step is intended to help classify the orientation of an edge.

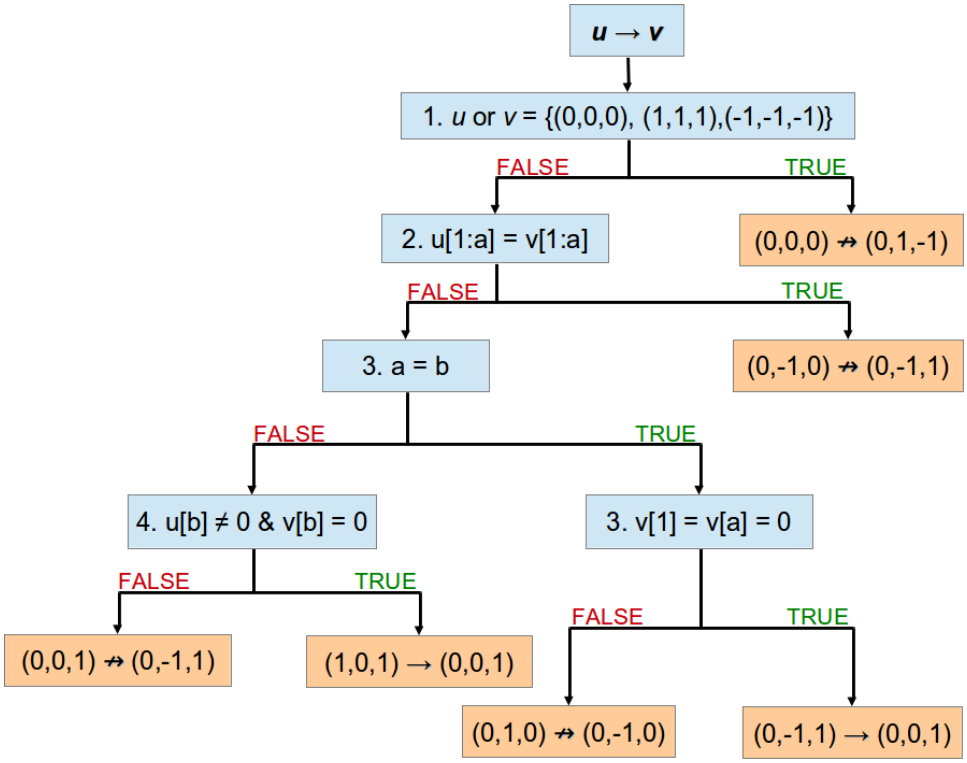

Figure 4.4: A detailed flow chart showing the application of the edge direction algorithm at each step in Section 4.1.3 with examples. The flow chart starts from the top to the bottom. The blue boxes show a step in the algorithm and the orange boxes show an example of a node pair applicable at that stage in the algorithm. The example uses the shortest vector for which this algorithm is applicable but can also be used for longer vectors such as a vector of size 4 for *F. solaris* and a vector of size 6 for *A. thaliana*.

After establishing edge directions, the edge weights were calculated using Equation 4.1.2. The weights are a measure of the expression between genes at nodes u and v, particularly at position u_a and v_a. As there are usually many genes per expression patterns, the median was chosen to represent the average expression value to avoid biases from outliers as much as possible.

$$W(u,v) = \frac{\mathbb{M}_u(F_d(x))}{\mathbb{M}_v(F_d(x))} \frac{\mathbb{M}_v(F_{\forall t}(x))}{\mathbb{M}_u(F_{\forall t}(x))} \tag{4.1.2}$$

where $W(u,v)$ is the weight of the edge from node u to v, d is the time at which the

expression pattern differs between u and v, and $\mathbb{M}_u(F_d(x))$ is the median expression of genes with pattern u at time d. Similarly, $\mathbb{M}_v(F_d(x))$ is the median expression of genes with pattern v at time d. The following $\mathbb{M}_u(F_{\forall t}(x))$ and $\mathbb{M}_v(F_{\forall t}(x))$ is the median of expression at all time points of genes with pattern u and v respectively. In the case that a median is 0, a small number is added to all medians such as $1x10^{-10}$.

Undirected edges and unconnected nodes were then removed from the full network, leaving a subnetwork or several subnetworks of connected nodes and directed edges.

4.1.4 Network Visualization

The networks were created and drawn in R using the igraph package [102] [137].

4.1.5 Enrichment Analysis

This step was performed on the *A. thaliana* results only due to the availability of annotated transcription factors. The gene list of each subnetwork was used in an enrichment analysis at The Plant GSEA [168]. The gene sets used were the three Gene Ontology (GO) components (biological process, cellular component and molecular function) with the species set to *A. thaliana*.

The list of gene ontologies were then summarized using REViGO [169] to collapse redundant terms. This was done using the small setting with the FDR corrected p-values from the enrichment analysis. The *A. thaliana* GO term database was used with the SimRel semantic similarity measure.

4.2 Results

4.2.1 Transcription Factor Selection

The transcription factor candidates for *F. solaris* were selected from 20,470 coding sequences by using the union of two methods (Figure 4.1). The first method relied on known transcription factor domains in PFAM [162], Superfam [163], and DBD: Transcription factor prediction database [164]. There were 6,278 *F. solaris* sequences identified which contained transcription factor domains using the three databases. The second method relied on transcription factors found in the relatives of *F. solaris*; *P. tricornutum* and *T. pseudonana* [165]. There were 360 *F. solaris* sequences identified which had closest sequence matches to the 208 *P. tricornutum* and 250 *T. pseudonana* transcription factors. The number of false positives were minimized by only selecting the *F.*

solaris sequences identified by both methods. By using a combination of searches, the transcription factor selection processes lead to the selection of 187 *F. solaris* sequences chosen as transcription factor candidates for the remainder of the analysis. Among the 187 sequences were 34 transcription factor domains (Table 4.1).

Domain	**Number**
Winged helix DNA-binding domain	120
HSF-type DNA-binding	113
bZIP transcription factor	42
Basic region leucine zipper	11
HLH, helix-loop-helix DNA-binding domain	11
C2H2 and C2HC zinc fingers	6
DNA-binding domain	6
Helix-loop-helix DNA-binding domain	6
Zn2/Cys6 DNA-binding domain	6
Cold-shock DNA-binding domain	5
E2F/DP family winged-helix DNA-binding domain	5
TAZ zinc finger	5
CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B	4
Transcriptional factor tubby, C-terminal domain	4
Tub family	4
A DNA-binding domain in eukaryotic transcription factors	2
AP2 domain	2
Bacterial regulatory proteins, luxR family	2
Conserved core of transcriptional regulatory protein vp16	2
DDT domain	2
Homeodomain-like	2
MIZ zinc finger	2
SART-1 family	2
SGT1 protein	2
SRF-like	2
ssDNA-binding transcriptional regulator domain	2
STAT	2
Sugar fermentation stimulation protein	2
YL1 nuclear protein	2
Zinc finger, C2H2 type	2
CCR4-Not complex component, Not1	1

DNA-binding protein Tfx	1
Helix-turn-helix	1
SAND domain-like	1

Table 4.1: A list of transcription factor domains found in 187 *F. solaris* transcription factor candidates and their frequencies. Some domains may be present in a sequence multiple times and some sequences may have multiple domains so that the frequencies do not have to sum to the number of sequences.

For the *A. thaliana* data set, the database of transcription factors contained 2,193 transcription factor entries, excluding splicing variants, of which 1,402 was present in the data set.

4.2.2 Gene Expression Patterns

The gene expressions of the transcription factor candidates were processed and organized into vectors so that the difference in gene expression over the duration of the experiments could be organized together. The vectors were named expression patterns because the values in the vector came in permutations of 3 values which represented the chronological changes in gene expression of a transcription factor. The permutation values indicated the direction of change while the location within the vector indicated when the change took place.

The application of a threshold on both data sets successfully increased the number of 0 binned values in contrast to 1 and -1 values (Figure 4.2 and Figure 4.3). The number of 1 values increased with time with a decrease in 0 values in the *F. solaris* data as expected, reflecting positive regulation in relation to the increase in TAG related transcripts and the increase in TAG production. There were more 1 and -1 values in the gene expression difference between time points in the initial time intervals in the *A. thaliana* data as gene expression increased to the crucial time period for floral transition, 12 to 13 hours after germination. These differences were decreased in later time points and stayed at a stable level until the end of the experiment.

Since the data from *F. solaris* had four time points, the expression patterns had 4 permutations of 3 values which allowed for a possible 81 combinations of expression patterns however only 28 patterns were observed (Table 4.2). The most common pattern was (0, 0, 0, 0) and was observed in 19% of the transcription factor candidates. The remaining commonly observed patterns contained many zero values such as (0, 0, 0, 1) and (0, 1, 0, 0). In contrast, the uncommon patterns contained a combination of all three values

or had many non-zero values such as (0, 1, -1, 0) or (1, -1, 1, 1). Notably, there were very few observed patterns beginning with a 1 or -1.

Pattern	Membership
(0,0,0,0)	36
(0,0,0,1)	35
(0,1,0,1)	22
(0,0,1,1)	17
(0,0,-1,0)	14
(0,1,0,0)	11
(0,1,1,1)	10
(0,-1,1,1)	6
(0,0,-1,1)	4
(0,-1,0,0)	4
(0,-1,0,1)	4
(0,-1,-1,0)	3
(0,0,1,0)	3
(0,-1,-1,-1)	3
(0,1,-1,1)	2
(0,1,-1,0)	2
(0,1,0,-1)	2
(0,0,0,-1)	1
(0,0,-1,-1)	1
(0,-1,1,0)	1
(1,0,0,0)	1
(1,0,0,1)	1
(1,0,-1,0)	1
(1,1,0,1)	1
(1,-1,1,1)	1
(1,-1,-1,-1)	1
(-1,1,0,0)	1
(-1,-1,-1,0)	1

Table 4.2: Transcription factor expression patterns in *F. solaris* and the number of transcription factors that had those patterns in descending order. The top patterns mostly begin with 0 and contain two or more 0s within the pattern while the bottom patterns mostly begin with 1 or -1.

Similarly, the data from *A. thaliana* had six time points, meaning that the expression patterns had 6 permutations of 3 values. This combination allowed a possible 729 expression patterns, however only 242 patterns were observed. The top three most common patterns were (0, 0, 0, 0, 0, 0) seen in 40% of the transcription factors, (-1, 1, 0, 0, 0, 0) seen in 2.3% of transcription factors and (-1, 0, 0, 0, 0, 0) seen in 2.1% of the transcription factors. In contrast, there were 111 patterns seen in one gene each (46%) and 51 patterns represented by two genes each (21%). The common patterns had more 0s on average compared to uncommon patterns, however they seemed to contain approximately the same frequency of 1s (Table 4.3).

Prevalence of pattern	-1	0	1
10+ genes	2.1	3.6	2
1 or 2 genes	1.8	2.1	1.9

Table 4.3: Summary of binned value representation in transcription factor expression patterns of *A. thaliana*. This shows the average number of each permutation value in two types of expression patterns. The prevalence of pattern represents the number of genes that had the expression pattern. Commonly observed patterns were seen in 10 or more genes while uncommonly observed patterns were seen in only 1 or 2 genes. The number of times 1 and -1 values are observed in expression patterns is similar in both common and uncommon patterns but there is a higher frequency of 0 values in common expression patterns compared to uncommon expression patterns.

Particularly, the floral transition transcription factors from *A. thaliana* were only represented by 15 expression patterns. The prevalence of each binned value in the patterns revealed that the most important time interval for these transcription factors was M6-M7. This time interval was where eight patterns had a 1 value, followed by six patterns with a 0 value and only one pattern with -1. When compared to the frequency counts for the other time periods, it has the highest frequency of the 1 value and the lowest frequency of the -1 value. The other time intervals had a mixture of 7, 6 and 2 or 6, 5 and 4 frequencies for each binned value.

4.2.3 Transcription Factor Network

The transcription factor expression patterns were used to create a network model of the expression relationships between expression patterns. The difference in the size and number of the expression patterns between *F. solaris* and *A. thaliana* resulted in different network structures.

The shorter patterns from the *F. solaris* data set created a smaller, fully connected network with 27 vertices connected by 51 edges (Figure 4.5.A). Since the connecting edges were between two expression patterns that differed by only one value, the connections were more likely with shorter patterns. The most connected vertices were observed in (0, 0, 0, 0), (0, 0, 0, 1) and (0, 1, 0, 1) with degrees of 7. There were four vertices with a degree of 1 and the most common degrees were 3 and 4 which were observed in 6 vertices each. All the vertices with a degree of 1 or 2 began with a 1 or -1 value. The edge direction algorithm was performed and successfully determined the direction of 20 edges connecting 24 vertices (Figure 4.5.B). That is, approximately 88% of the edges had their direction determined.

The expression patterns from the *A. thaliana* data set were longer and more numerous. Consequently, the initial undirected network was larger and not fully connected (Figure 4.6.A.). The inferred network was made up of 242 vertices connected by 680 edges. The most connected vertex was (1, -1, 0, 1, -1, 0) and had a degree of 12, and it was closely followed by (0, 0, 0, 0, 0, 0) which had a degree of 11. There were twelve vertices with a degree of 2, eight vertices with a degree of 1 and one unconnected vertex, (-1, -1, 1, 1, 0, 1). Unlike the *F. solaris* network, the expression patterns for uncommon vertices did not all begin with a 1 or -1. There were seven patterns that started with a 0 with a vertex degree of 2 and three patterns that started with a 0 with a vertex degree of 1. The most common degree was 5 which was observed in 49 vertices. This was followed by 4 degrees which was observed in 39 vertices and 6 which was observed in 38 vertices. Despite the larger number of edges, the edge direction algorithm performed quickly and successfully in determining the direction of 123 edges which connected 136 vertices (Figure 4.6.B.). This only represented 18% of the original undirected network which was lower than the result from the *F. solaris* network.

The networks were then separated into subnetworks by removing undirected edges and any unconnected nodes resulting from that so that the informative sections of the network were isolated. The direction was determined by the temporal effect of gene regulation, while the weights were ratios signifying the impact of gene expression at the time where the expression patterns differed.

The paring down of the *F. solaris* network yielded five subnetworks of varying sizes (Figure 4.7). The number of nodes in each subnetwork, in descending order, were 11 nodes, 5 nodes, 4 nodes, 2 nodes and 2 nodes. Due to the way the edge directions were determined, most subnetworks contained a central sink vertex where all the connecting edges were incoming edges. The sink vertex patterns tended to have many zero values such as (0,0,0,1) and were represented by many transcription factors compared to the

leaf vertex patterns that were not so well represented.

In the largest network, subnetwork 1, there were 3 paths longer than 1 edge leading to the sink vertex where the edge weight decreased as the path was closer to the sink vertex (Figure 4.11). In contrast, there were two paths where the edge weights increased as the path was closer to the sink vertex. However, both paths originated from the same node (1, 1, 0, 1). This node represented one transcription factor with motifs HSF-type DNA-binding and Winged helix DNA-binding domain. The other nodes were (1, -1, 1, 1) which represent one transcription factor with the Basic region leucine zipper motif, (0, 1, -1, 1) which represent two transcription factor with the same motifs as (1, 1, 0, 1) and also Helix-loop-helix DNA-binding domain, and (0, 1, 1, 1) which represented 10 transcription factors that included the same motifs as (1, 1, 0, 1) and also CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B, bZIP transcription factor, Helix-loop-helix DNA-binding domain and E2F/DP family winged-helix DNA-binding domain. The remaining subnetworks were too small for comparisons between long paths as there was only 1 path longer than 1 edge in each network.

The largest edge weight was found in subnetwork 3 from (0, -1, -1, 0) to (0, 0, -1, 0) where the change in expression at 24 hours was exceedingly up regulated compared to the other time points as the weight was 325.0844. This was due to a large negative fold change in g13229 relative to the other two genes it shared with the pattern (0, -1, -1, 0). The transcription factor motifs found in that gene were HSF-type DNA-binding and Winged helix DNA-binding domain.

The network inference method applied to the *A. thaliana* network produced 17 separate subnetworks that contained 136 vertices and 123 edges within them (Figure 4.9). There were 6 subnetworks that were made up of only two vertices and an edge, while the largest subnetwork was made up of 36 vertices and 36 edges. Like the *F. solaris* subnetworks, the sink vertices in the *A. thaliana* subnetworks also contained many 0 values such as (0, 0, 0, 0, 0, 1) and were represented by many transcription factors compared to the leaf vertices. While the smaller subnetworks with five or less vertices were mainly linear, the large subnetworks were mainly star-like with fairy separate branches joined to a central spine. A rectangular structure was visible in three of the large subnetworks consisting of four vertices and four edges.

Most of the edge weights were quite small with a median of 1.76 and a mean of -1.40. There were eight large weights with values over 100 and these were found in the two largest subnetworks. There were two paths of large edge weights with three vertices where the largest weights were between the last two vertex of each path. These were (1, -1, 0, 0, 1, 1) $\rightarrow$ (0, -1, 0, 0, 1, 1) $\rightarrow$ (0, 0, 0, 0, 1, 1) and (-1, 1, 0, 0, 1, 1) $\rightarrow$ (0, 1, 0, 0, 1, 1)

$\rightarrow$ (0, 0, 0, 0, 1, 1). The second path indicates that a regulation step at M5-M6 is related to transcription factor activity at later time points and that the relation is stronger at M4-M5 than at M5-M6.

Specifically, when checking for the known floral transcription factors, it was observed that there were 17 floral transition transcription factors and nine of them were found in three of the subnetworks; 1, 4 and 7. Subnetwork 1 was the largest among them and contained seven transcription factors (Figure 4.10). The pathway begins with SPL4 and LFY in patterns (1, -1, 1, 1, 1, 1) and (-1, 0, 1, 1, 1, 1), and then flows on to AP1, AP3, PI and SPL3, with patterns (0, 0, 0, 1, 1, 1), (0, 0, 0, 0, 1, 1) and (0, 0, 0, 0, 0, 1). Of the two initial transcription factors, the edge from LFY had a larger edge weight than SPL4. The other transcription factors identified in the subnetworks were CAL in subnetwork 4 and FLC in subnetwork 7. The remaining unconnected transcription factors were SVP, CO1, FD, SPL9, SPL5, SPL15, MYB3R1 and MYB3R4.

4.2.4 Enrichment Analysis

The genes of the *A. thaliana* subnetworks were put through a GO enrichment analysis since they were better summarized by function in contrast to transcription factor motifs. The four largest subnetworks had 199, 170, 119 and 74 enriched terms respectively. They shared 54 common terms which made up a 73% majority for the fourth largest subnetwork. The enrichment analysis was successful for all subnetworks except for subnetwork 14 which did not contain enough annotated genes.

The resulting lists of enriched gene ontologies were summarized so that they could be compared. Many of the summaries included general terms such as biological regulation or metabolic processes, however there were several unique and indispensible ontologies that were observed.

Subnetwork 1 was defined by transcription factors annotated with rhythmic process, circadian rhythm, response to light stimulus, protein acetylation, regulation of multicellular organismal process, reproduction and reproductive structure development.

Subnetwork 2 was a lot smaller and was summarized the two terms, long-chain fatty acid metabolic process and defense response to insects.

Subnetwork 3 was the largest network and was distinguished by the terms, immune system process, response to endogenous stimulus, glucuronoxylan metabolic process, bract development, reproductive process, positive regulation of biological process and multiple organism process.

Subnetwork 4 contained moderate number of genes but was only defined with DNA-

templated regulation of transcription and response to gibberellin.

Subnetwork 5 was smaller and was represented by nitrogen compound metabolic process and positive regulation of biological process.

Subnetwork 6 was roughly the same size as subnetwork 5 and was described by response to salicyclic acid, respiratory burst and positive regulation of biological process.

Subnetwork 7 was summarized by the terms that were more specific like, nitrogen compound metabolic process, chloroplast relocation and negative regulation of flower development.

Subnetwork 8 had only two vertices and its summarizing terms were DNA-templated regulation of transcription, response to gibberellin and embryo development ending in seed dormancy.

Similarly subnetwork 9 was the same size and had several unique terms in its summary of response to auxin, multicellular organismal process, developmental process, DNA-templated positive regulation of transcription and gynoecium development.

Subnetwork 10 and 11 were also two vertex networks where subnetwork 10 was described by reproduction, response to abscisic acid, seed germination, peptidyl-histidine modification and reproductive process while subnetwork 11 was described by general transcription factor terms like nucleic acid binding, sequence specific DNA binding transcription factor activity and transcription regulator activity.

Subnetwork 12 was a more moderately sized network distinguished by the terms immune effector process, post embryonic morphogenesis, RNA metabolic process and multicellular organismal process.

Subnetwork 13 is only defined by two terms, regulation of metabolic process and cotyledon morphogenesis.

Subnetwork 15, 16 and 17 are very small structures where subnetwork 15 is involved in demethylation, subnetwork 16 in transcription regulator activity and subnetwork 17 in cellular response to glucose stimulus.

4.3 Discussion

Gene expression is a chronologically dependent process where a change in expression in gene A at one time point is related to the change in expression in gene B at a later time point. This method was developed from that position which led to the creation of expression patterns and the edge building algorithm.

4.3.1 *F. solaris*

A preliminary search for transcription factors in *F. solaris* yielded 160 transcription factors [113] however, my method of using three databases and two evolutionary close genomes produced 187 transcription factors. A PFAM search [162] was also performed alone and yielded 190 transcription factors. Expectedly, different search methods and parameters return different numbers of matches so selecting transcription factors is an indefinite process. Since transcription factors of higher plants has shown a marked divergence from that of microalgae, using databases primarily built up from those organisms and other higher organisms can skew results [170]. While no microalgae specific transcription factors have been discovered, those shared transcription factors may be harder to detect. Ultimately caution must be taken and cut off thresholds decided in order to proceed with the analysis. The inclusion of false positives would most likely alter the results so I decided that only sequences with known transcription factor domains and those matching transcription factors in *P. tricornutum* and *T. pseudonana* would suffice. Although this process may have excluded several transcription factors, I felt that my finished list was most appropriate for the analysis. Indeed, some key transcription factor domains were present in the group of genes that were selected including the AP2 DNA-binding domain which is an important regulator of oil accumulation in *A. thaliana* seeds [171] [172] [173].

One of the most common transcription factor domains found in *F. solaris* was the HSF-type DNA-binding domain (PF00447) from PFAM (Table 4.1). The heat-shock factor (HSF) family of transcription factor domains are very diverse in plants where it is encoded by 21 genes in *A. thaliana* [174] and 52 loci in soybean [175]. Plant HSFs show considerable functional diversification, indicated by the difference in regulation and have shown themselves to be principal regulators in responding to abiotic stresses [176]. Although *F. solaris* is a microalgae, the HSF-type DNA-binding domain was found in more than 50% of the selected transcription factor sequences. This large ratio of HSF domains present in the population of transcription factors indicate that *F. solaris* HSFs are likely to be involved in abiotic stress response also. Additionally, 87% of them are homologs to HSFs found in *P. tricornutum* and *T. pseudonana* where two *F. solaris* transcription factors were homologous to one *P. tricornutum* or *T. pseudonana* transcription factor. For example, the two *F. solaris* genes, g13791 and g14253, are homologs of one *P. tricornutum* gene, PHATRDRAFT_47952. The significance of this means that a majority of *F. solaris* HSFs are present twice in the genome while only the homolog is present once in *P. tricornutum* or *T. pseudonana* genomes. This feature of transcription factor presence is likely to play a large part in the difference in lipid accumulation between

the diatoms.

The expression patterns of *F. solaris* transcription factors had distinct characteristics that summarized a circumstance when a small number of transcription factors were expressed or repressed early in the experiment and a large number of transcription factors were only expressed or repressed later in the experiment. This has been observed in different situations of cell differentiation and embryonic development [177] as well as in the *Nannocloropsis* microalgae [170]. Interestingly, when compared with the patterns of preliminary TAG related genes, the frequency and types of patterns appear to be reversed with TAG genes possessing more patterns beginning with 1 or -1 and very little patterns containing any 0 value. This could be due to the strong fold change exhibited by TAG genes as expected while it was under induction during the experiment. However, similar to the transcription factor patterns, most of the common patterns held many duplicate values such as (-1, 1, 1, 1). It is likely that genes that are steadily regulated are more influential to TAG production as they were observed to be more numerous than genes with fluctuating expression patterns.

The inferred transcription factor subnetworks were relatively small and limited due to the small number of transcription factors included in the analysis (Figure 4.7). However, the largest subnetwork, subnetwork 1, identified a small group of genes which showed early activity and were related to a larger group of transcription factors that were activated later. The edge weights of those genes indicates that the strength of the positive regulation processes is quite strong at the beginning and subsided at later time points. These thirteen genes make suitable candidates for investigations into the initiation of oil accumulation metabolism (g11326, g2962, g13476, g14216, g766, g4004, g5628, g5748, g9763, g6962, g9546, g5631, g14217).

The other candidate transcription factors which show potential for further research were determined by using the edge weights. Unusually large edge weights indicate a strong change in expression between two patterns. The two largest weights shared a sink vertex (0, 0, -1, 0), signifying a strong relationship between the transcription factors with those expression patterns. The genes largely responsible for the size of the edge weights were g13229 and g2870.

A final network was created to illustrate the connectivity between the subnetworks by collapsing the subnetworks into vertices 4.12. The reduced network clearly shows a connection between the three largest subnetworks, 1, 2 and 3. This connection is formed a triangular base in the reduced network. The smaller subnetworks are linked on two sides off the triangular base, with subnetwork 5 being linked to subnetwork 1 and 2, and subnetwork 4 being linked to subnetwork 1 and 3. The sequences in

each subnetwork do not have any KEGG pathway or ortholog annotations however the transciption factor domains show a commonality between them such as the bZIP transcription factor domain, as well as unique domains per subnetwork such as the Cold-shock DNA-binding domain in subnetwork 1, the CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B domain in subnetwork 2, the Zn2/Cys6 DNA-binding domain in subnetwork 3 and the C2H2 and C2HC zinc finger domain in subnetwork 5.

4.3.2 *A. thaliana*

The initial creation of the gene expression patterns for *F. solaris* was done by observing the gene expression for all genes in a chronological manner. The variance of gene expression was found to be increased as time increased and that this variation followed a gamma distribution. The threshold was determined and it worked well, however it did not fit as well when this method was applied to *A. thaliana*. This is mainly due to the type of data that was used. The *F. solaris* data was a fold change value between control and treatment while the *A. thaliana* data was a difference in expression value between two adjacent time points. Thus, they followed different distributions. The method was then adjusted to what is presented here so that it would be applicable for various types of data.

Although the *A. thaliana* genome has more annotations than *F. solaris*, not all the transcription factors in the database were found in the expression data set. Normally, transcription factors are only present at very low copy numbers per cell because, as regulatory elements, they do not need necessarily have to be expressed at high levels [178]. Since RNA-Seq is a competitive sequencing method, they may not have been detected. This could account for some of the absent transcripts as well as exclusion during the sequence quality control step or possible count errors due to alternate splicing [179]. Because RNA-Seq does not sequence a full transcript in a continuous pass, the raw data is made up of many fragments of sequences between 30-400 base pairs in length depending on the sequencing platform. Although many algorithms has the ability to adjust for this [22], some fragments may not be used in an alignment or may be aligned incorrectly, particularly with the possibility of alternate splicing. Most platforms also have a quality control procedure built in that will disregard low quality reads.

When the inferred network was applied to floral transition data, it successfully identified established connections between floral transition transcription factors, such as the positive regulation of AP1 by LFY [152]. The general arrangement of each subnetwork, where outlying leaf patterns were represented by fewer genes than the central patterns, verified the cascading effect initiated by a few genes that affect many down-

stream genes [170] [177].

The expression patterns were able to make examining important expression patterns much clearer and their use was able to help identify additional, related genes. Expression pattern importance can be determined through the experiment design or previous research, such as the function of LFY in the floral transition data. The expression pattern for LFY was (-1, 0, 1, 1, 1, 1) which was shared by two other transcription factors, ATREM1 and BETA HLH PROTEIN 93 (BHLH093). ATREM1 has been observed from the vegetative apical meristem to the inflorescence meristem and binds to AP1, AP3, PI and SVP [180] [181]. This is confirmed in the network where AP1, AP3 and PI are downstream from ATREM1. BHLH093 has been observed in several different development stages such as the final stage of leaf development, expanded cotyledon stage, and flowering [182] [183]. It is possible that BHLH093 is expressed as a regulator of a concurrent process with floral transitioning but is not directly part of the process.

Similarly, connections to identified expression patterns also helped to other genes that are part of related functions. In the network, LFY is connected to AP1 by an additional pattern, (0, 0, 1, 1, 1, 1), which represents the two transcription factors, SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 8 (SPL8) and FOREVER YOUNG FLOWER (FYF). SPL8 has been seen to act as a developmental regulator with gibberellins and flowering time [184] [185] while FYF is a repressor of floral organ senescence [186] [187].

The final network summarized the data into 17 subnetworks, each distinguishable by enriched GO terms. To illustrate the connectivity between the subnetworks, they were collapsed into vertices (Figure 4.12). The reduced network clearly shows a connection between the two largest subnetworks, 1 and 3. This connection is formed by two connecting patterns, indicating a relatively weak link. In contrast, there are 8 and 7 links between subnetwork 1 and 4, and 3 and 4 respectively. This suggests that gene regulation of rhythmic processes, response to light and other processes found in subnetwork 1 is related to immune responses, response to endogenous stimulus and other processes found in subnetwork 3 primarily through transcription factors related to the plant hormone, gibberellin.

Although the inferred network included many floral transition transcription factors, there were several patterns of important transcription factors that remained unconnected such as SOC1 [145]. The edge determination for the network is heavily influenced by the number of time points so that the increase in the number of time points will result in a less connected network due to the higher number of patterns needed to make connections. This effect can be decreased by focusing on a narrower range of time

points, as done by excluding earlier time periods prior to floral transition. Although it will not always be sufficient, it remains that the remaining transcription factors were able to be connected into the network with undirected edges.

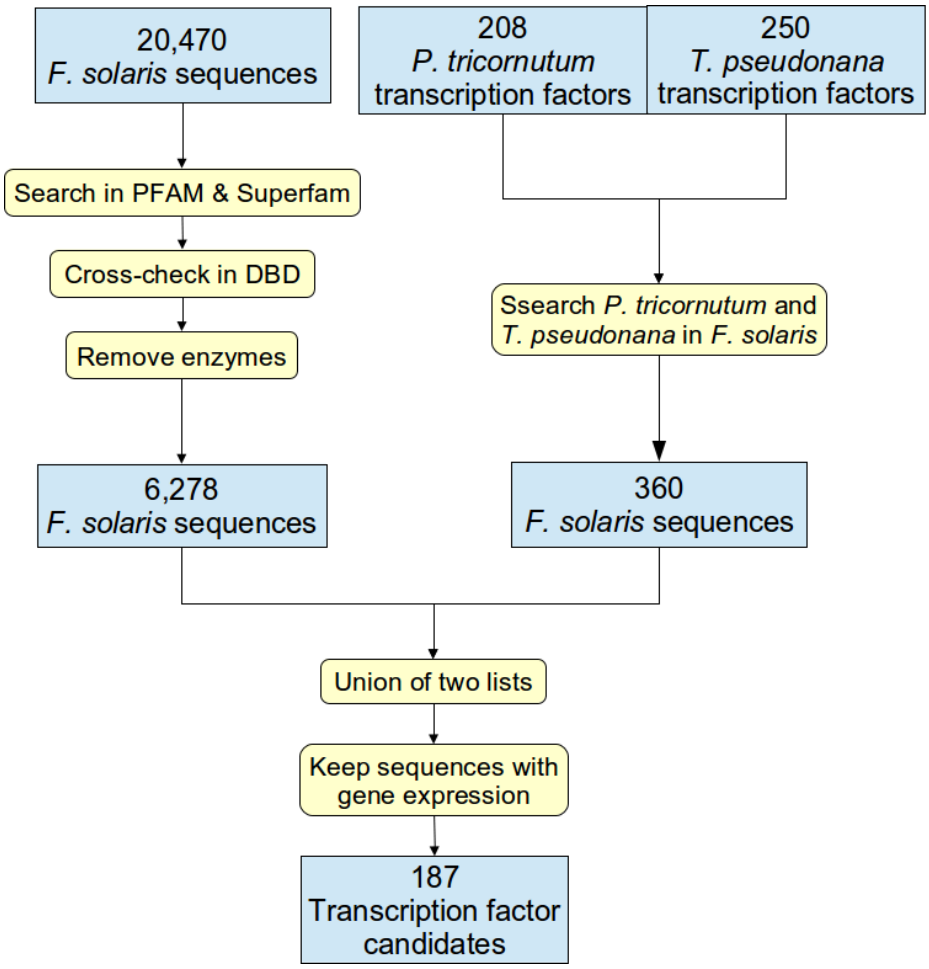

Figure 4.1: An overview of the process to create a list of transcription factor candidates for *F. solaris*. The number of sequences at each step are highlighted in blue boxes and the actions taken to filter the sequences are highlighted in yellow boxes. The two methods start from the top of the diagram and join up in the middle to signify when the transcription factors were chosen by taking the union of two lists in order to reduce the number of false positives.

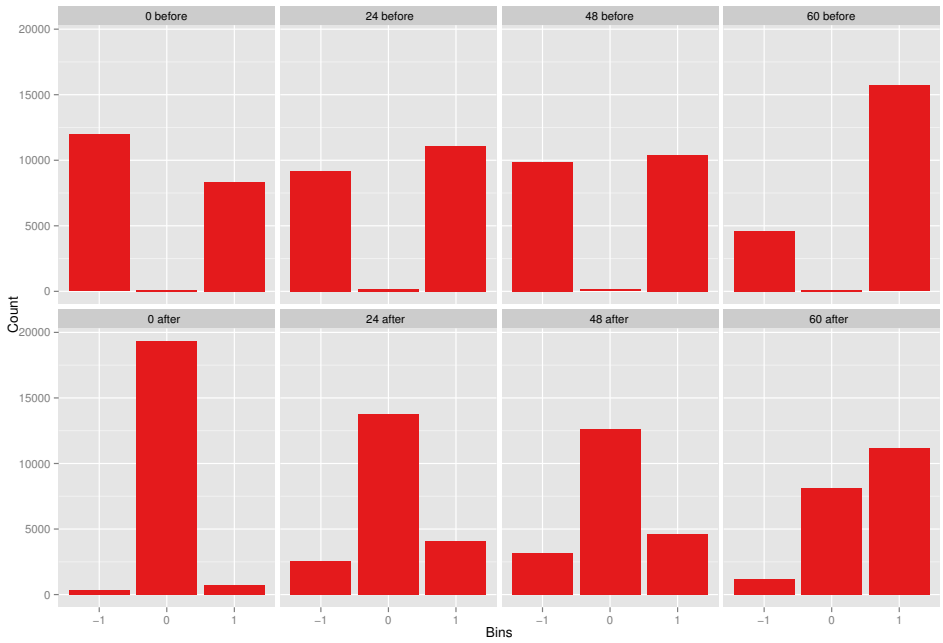

Figure 4.2: The effect of threshold application to the binning processes as part of the method application for *F. solaris* data. The number of genes with a positive fold change increases with time and this pattern is preserved even with the application of the threshold. The number of genes with a neutral fold change decreases with time as more genes undergo differential gene expression during the experiment.

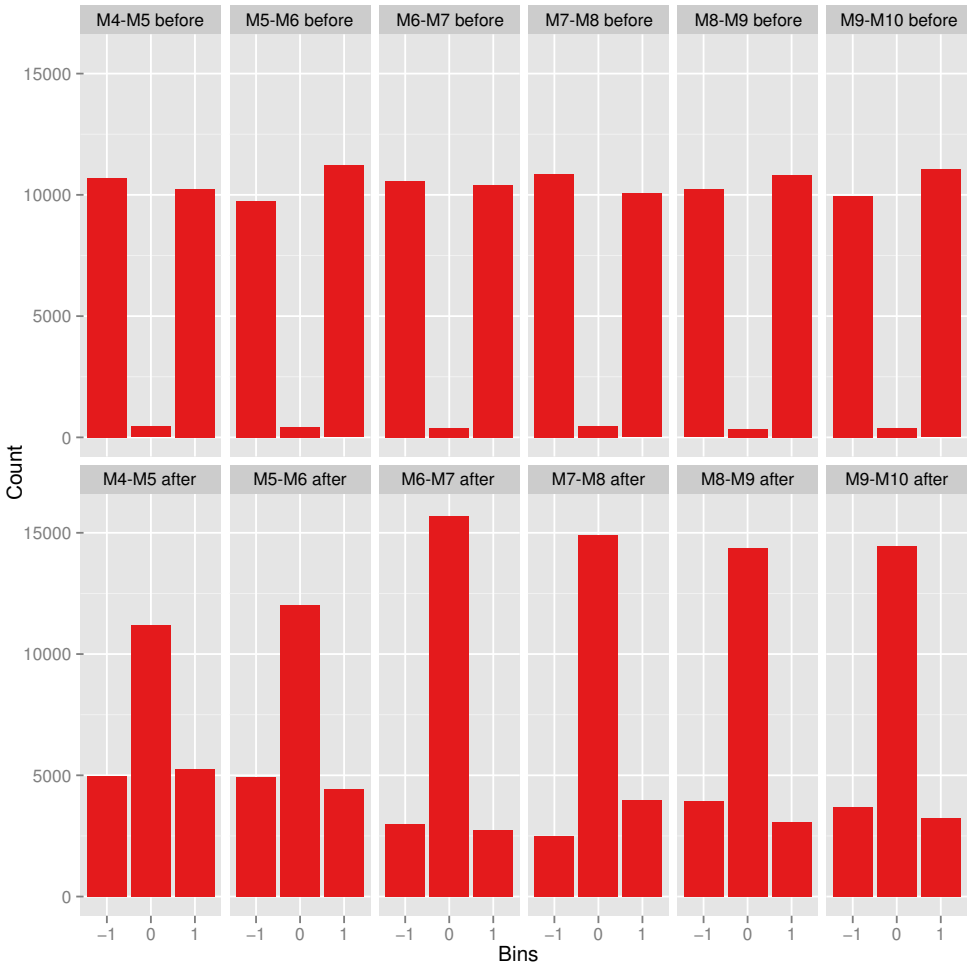

Figure 4.3: The effect of threshold application to the binning processes as part of the method application for *A. thaliana* data. There was no noticeable difference between each time point before threshold application however, a subtle difference is uncovered after threshold application. There is an observable increase in the number of genes with a neutral difference in gene expression at M6-M7. This change is kept for the remaining time points.

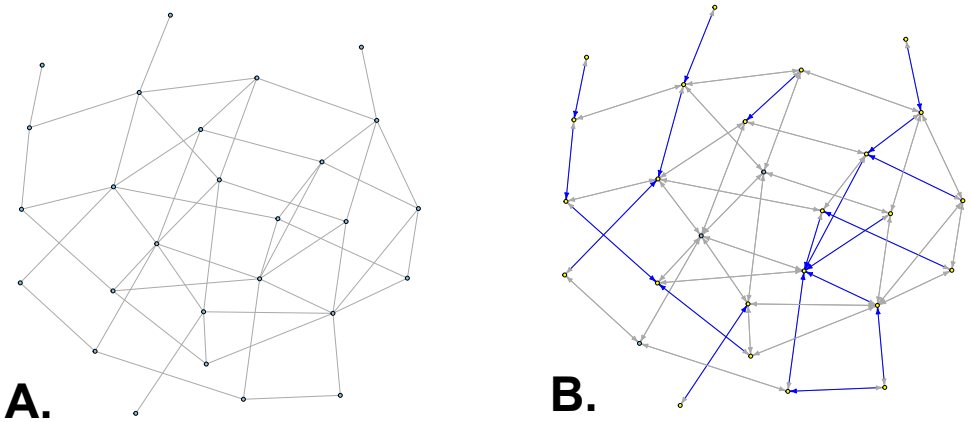

Figure 4.5: The inferred network during construction in the first two stages for the *F. solaris* data set. A. The vertices are joined together to create an undirected network. The vertices are in light blue and are connected by gray edges. B. The edge directions are established and highlighted in dark blue. Vertices with at least one directed edge are in yellow. The final network includes the yellow vertices and dark blue edges only.

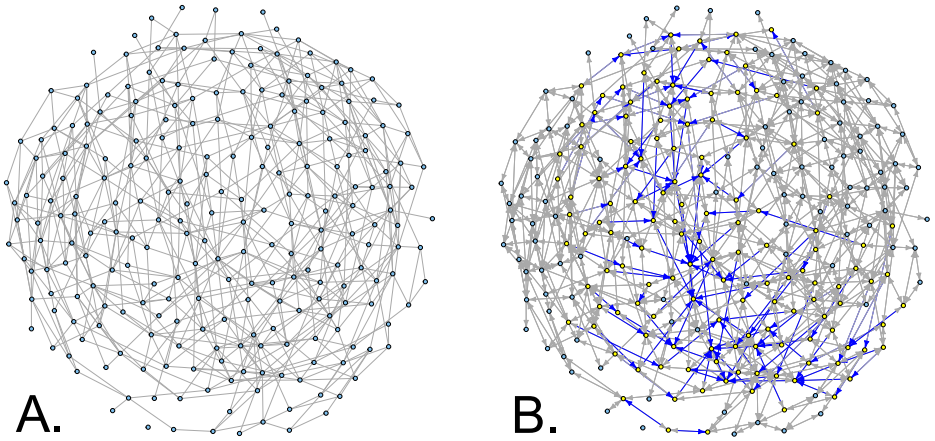

Figure 4.6: The inferred network during construction in the first two stages for the *A. thaliana* data set. A. The vertices are joined together to create an undirected network. The vertices are in light blue and are connected by gray edges. B. The edge directions are established and highlighted in dark blue. Vertices with at least one directed edge are in yellow. The final network includes the yellow vertices and dark blue edges only.

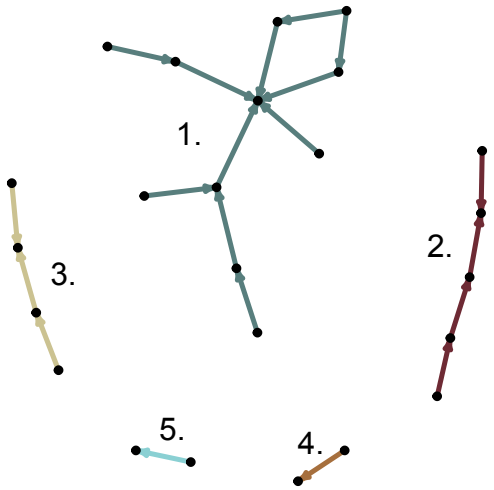

Figure 4.7: The final network with undirected edges removed from Figure 4.5 B. There are 5 unconnected subnetworks labeled 1-5 of varying sizes. In descending order, they are 1 (11 vertices), 2 (5 vertices), 3 (4 vertices), 4 (2 vertices) and 5 (2 vertices).

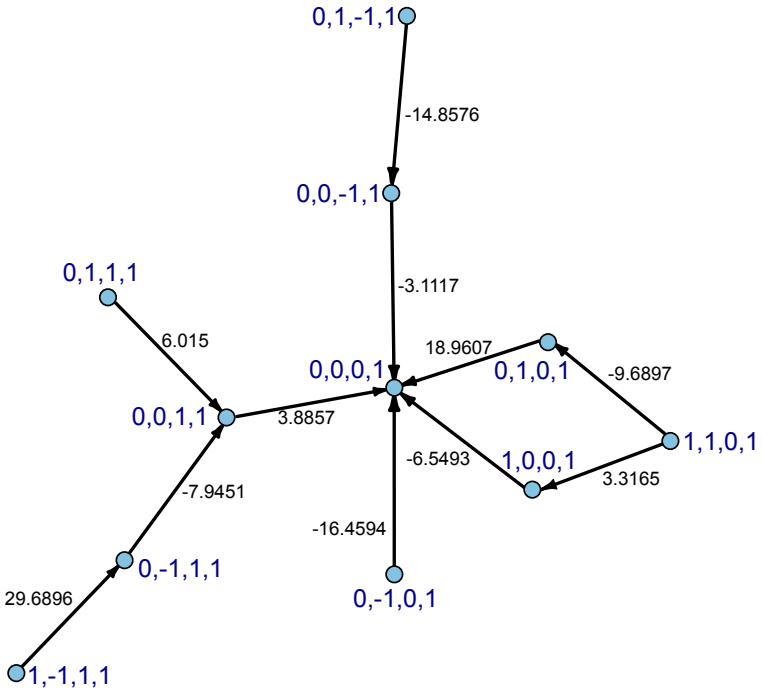

Figure 4.8: The weights and vertex patterns of subnetwork 1. The weights are written in black next to the applicable edge and the patterns are written in dark blue next to the applicable vertex.

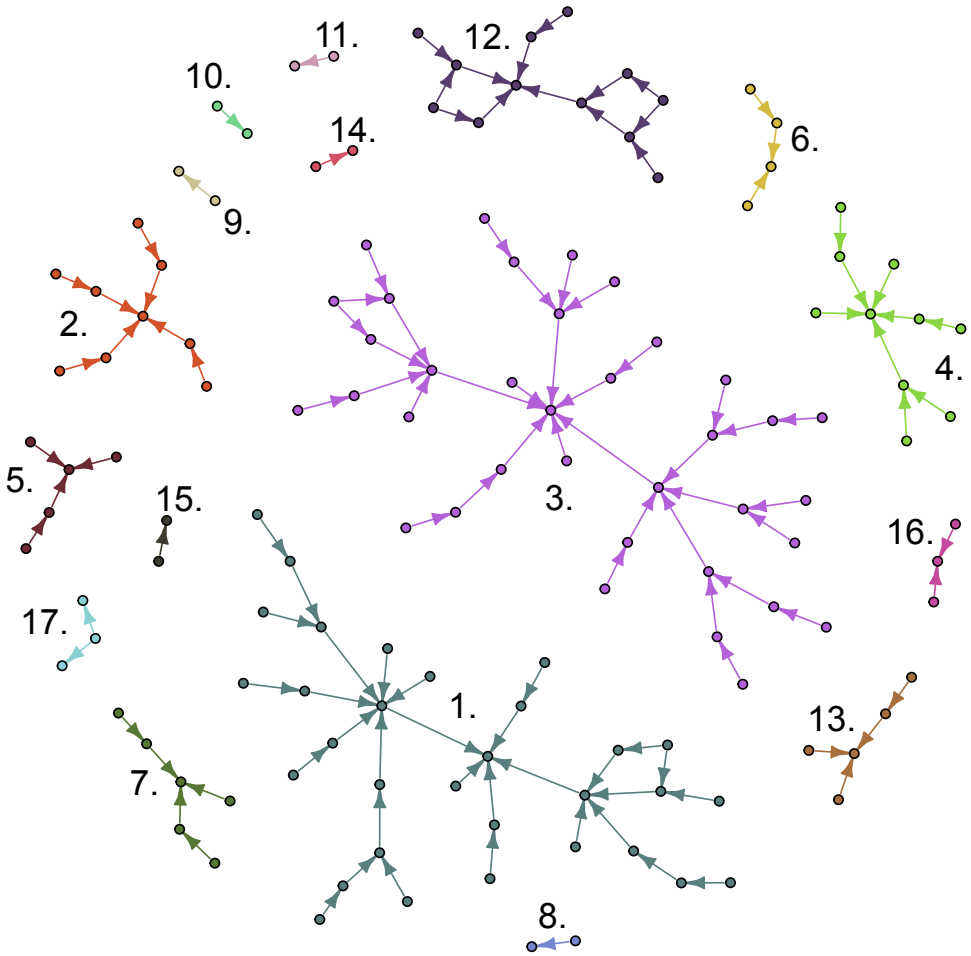

Figure 4.9: The final network with undirected edges removed from Figure 4.6 B. There are 17 unconnected subnetworks labeled 1-17 of varying sizes. In descending order, they are 3 (36 vertices), 1 (31 vertices), 12 (12 vertices), 4 (10 vertices), 2 (9 vertices), 7 (6 vertices), 5 (5 vertices), 13 (5 vertices), 6 (4 vertices), 16 (3 vertices), 17 (3 vertices), 8 (2 vertices), 9 (2 vertices), 10 (2 vertices), 11 (2 vertices), 14 (2 vertices) and 15 (2 vertices). The floral transition transcription factors are in subnetworks 1, 4, and 7.

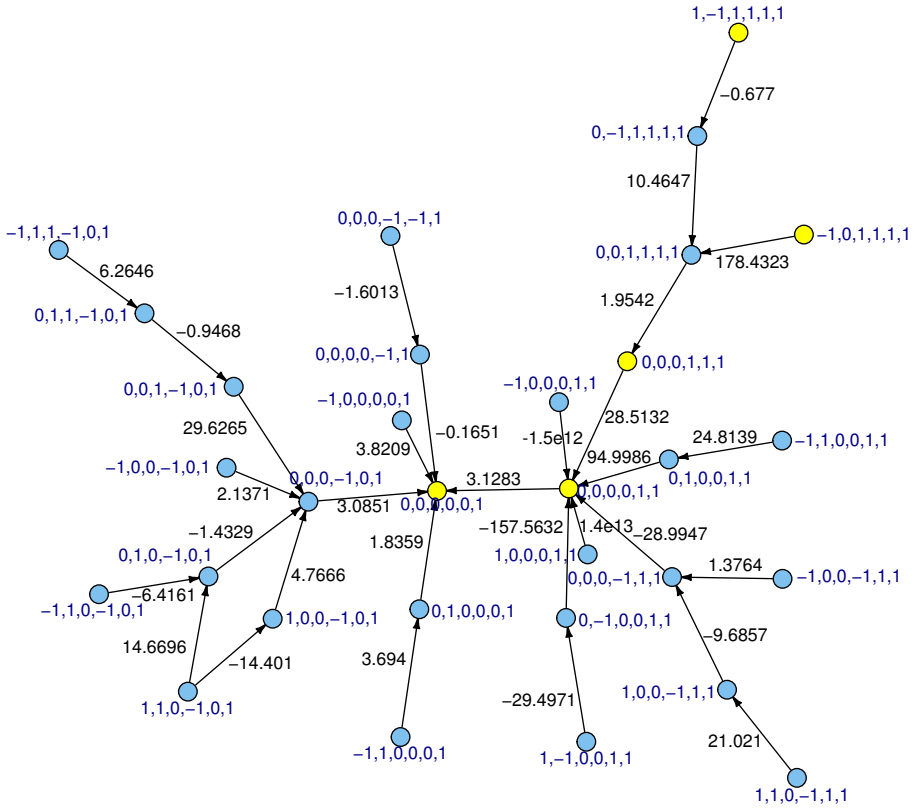

Figure 4.10: The weights and vertex patterns of subnetwork 1. The weights are written in black next to the applicable edge and the patterns are written in dark blue next to the applicable vertex. This major subnetwork contains 9 of the 17 floral transition transcription factors. Their expression patterns are highlighted by yellow vertices.

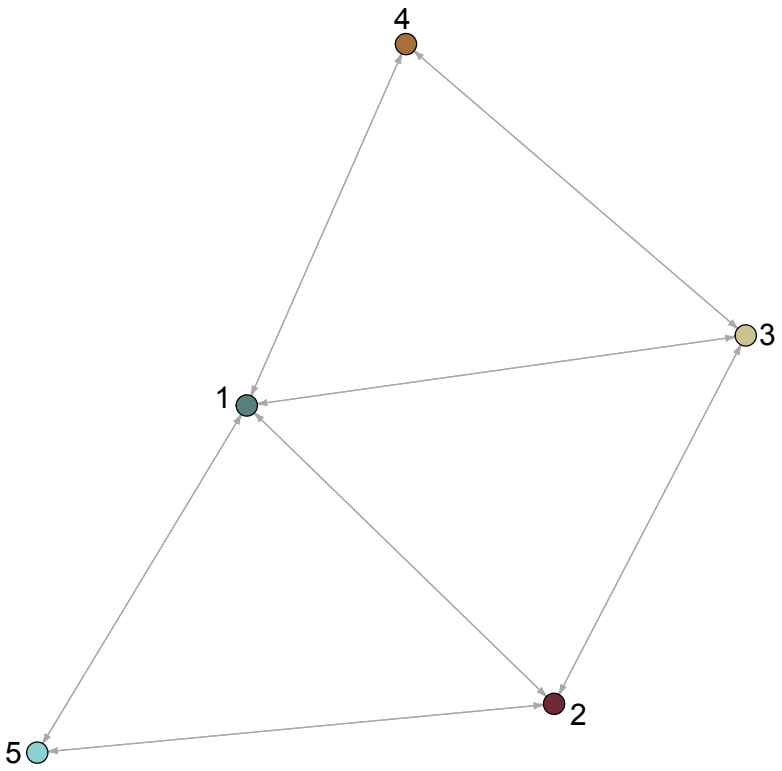

Figure 4.11: A simplified view of the *F. solaris* subnetworks when connected by the minimum number of undirected edges. Each subnetwork vertex and edge connection is merged into one vertex, keeping the edges that lie along the shortest path between subnetworks. The subnetworks are colored as they were in Figure 4.7. Joining vertices that were not part of a subnetwork are marked in light blue and labeled with the expression pattern.

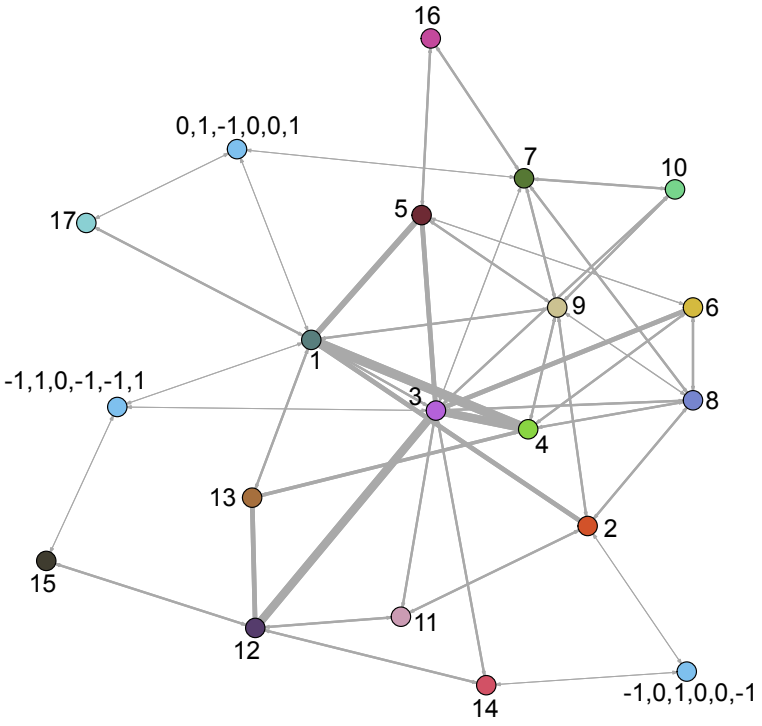

Figure 4.12: A simplified view of the *A. thaliana* subnetworks when connected by the minimum number of undirected edges. Each subnetwork vertex and edge connection is merged into one vertex, keeping the edges that lie along the shortest path between subnetworks. The subnetworks are colored as they were in Figure 4.9. Joining vertices that were not part of a subnetwork are marked in light blue and labeled with the expression pattern.

Chapter 5

Discussion

RNA-Seq is an accessible experimental NGS methodology that has given researchers the opportunity to study changes in gene expression over multiple continuous time points. It has the potential to discover gene expression regulation systems and explain the gene regulation processes but is restrained by the lack of data analysis methods. Although there are a small number of choices available for identifying differentially expressed genes [59] [60], there is a noticeable absence of downstream analyses that enables interpretation of the results in a biological framework [188].

Essentially, biologists would find it helpful to have more analyses available that can contextualize the essence of what the data recorded and also allow them to visualize the results. Methodologies that can assist biologists in this way will aid in discovering more insight to the effects that were present in the experiment since better perception will improve result interpretation and also give better direction for future experiments. I have presented three methods that utilize a different aspect of an organisms biology that each allows for a different biological interpretation of the same data set. They can be used separately if the organism or experiment is not suitable for that particular analysis, for example, if the organism is entirely novel with no close relative for homolog comparisons or the research is exploratory with no particular focus on a particular metabolic pathway. They can also be used concurrently to draw more comprehensive conclusions from all the results, as shown in the case of the *F. solaris* expression data.

Firstly, working with novel organisms can be a challenge because of the absence of past research to rely on and any corresponding annotations that arise from them. Also, their studies are usually limited by time and funding during which a lot of effort is focused on discovering favorable and advantageous traits that can contribute to current knowledge. Consequently, the most suitable type of research are comparative studies because they can quickly provide potential starting points for follow up studies. Since genome

sequencing has become required for genomic studies of any organism, there are more genomes being sequenced and being made publically available. I chose to make use of comparative studies by looking specifically at homology because the current landscape of biology makes it a reasonably advantageous tool to use for studying novel organisms, such as *F. solaris*, where it is more likely that there is a suitable genome to compare to than in the past before NGS [170].

Homologous genes are often used in phylogenetics and structural proteomics as a tool to measure relatedness where the sequence similarity is used to give information about the evolutionary history of the gene. Because they normally share similar structures and often have similar functions, they are one of the first genes to be annotated in a novel organisms genome [189]. Comparisons and inferences can be made using homologous genes between closely related species because they usually do not have many genetic changes so the functions and regulation elements are the same or very similar. Although genetic recombination can complicate the way in which homologous genes are identified, I showed that the genome of *F. solaris* was comparable with the model diatom *P. tricornutum* when I compared the gene expression of the homologous genes and non-homologous genes between the two organisms, and found that the homologous genes were expressed more similarly across species than within each of the respective organisms. This strongly indicated that the homologous genes has kept the same function and regulatory elements which is something that is commonly observed in development genes since they regulate such vital functions [190]. Since oil accumulation is a natural process of environmental stress response for diatoms, the homologous genes activated during biofuel production is probably a strongly conserved function between diatom species that has been preserved into an essential part of their metabolism.

Aside from providing a general examination with data exploration, I also used gene homology to target genes exhibiting changes in expression that could be related to a target metabolic process which was oil accumulation in *F. solaris*. Even though the difference in gene expression of homologous genes in *F. solaris* were smaller than non-homologous genes, the homologous gene expression also showed some distinct differences between the two diatoms. My analysis method targeted those differences when it categorized them by the differences in fold change instead of trying to analyze them individually. I chose to separate the homologous genes by their difference in fold change so that the different types of regulation mechanisms would be more apparent. Genes that were up regulated in one diatom and down regulated in another diatom show a different regulation system compared to genes that were up regulated in both at significantly

different expression levels. Through this type of comparison, I was able to show that *F. solaris* was better able to adjust to low nitrogen conditions through the expression of key genes that were down regulated when compared to the expression regulation that occurred in *P. tricornutum*. Even though the stress response in algae has been shown to be a distinct and conserved trait between some species, this *F. solaris* expression data has shown that it has small but significant differences [111]. This difference is possibly part of what gives *F. solaris* its unique oil accumulation traits.

This method of analysis has the advantage of being fast and simple to perform. Provided that there is a suitable genome to compare to, most of the computer processing was used during the sequence alignment process that found and identified the homologs in each genome. Calculating the differences in fold change and finding the significantly expressed genes run in linear time which is possible for any modern computer. The gene ontology and KEGG pathway information are freely available online and therefore accessible to researchers. Downloading the information for local analyses from KEGG may present a problem, however, as access to the KEGG FTP by non-academic researchers requires a license. Otherwise, this method is easy to implement and does not require any particular processing system. The use of absolute cutoffs for categorizing could be improved upon by excluding or reassigning genes that were very close to the threshold. The interpretation of some of the groups was also difficult with the presence of unannotated genes or differences in annotations between *F. solaris* genes and the homologous *P. tricornutum* genes. However, the characterization of each group can still be used to select genes of interest for future study and annotation as shown here.

Another way of handling large data sets like NGS data is to focus on particular elements of interest, for example, TAG synthesis. RNA-Seq records data for all transcripts at a given time so that it includes information on all processes that were transpiring. Many of the processes involve regular cellular function such as those including housekeeping genes, and are therefore unrelated to TAG synthesis so including them would decrease the power of the analysis. Selecting a smaller subset of data to analyze emphasizes focus so that it creates results that are easier to understand and interpret in the context of the experiment and current knowledge. I chose to investigate metabolic pathways first as it was suitable within the context of compound synthesis for the *F. solaris* data. My method made use of the compounds and genes already identified from previous investigations to select a subset of the data known to represent TAG metabolism such as carbohydrate metabolism pathways and energy metabolism pathways [113] [191]. I chose to use GSEA to perform the selection as it operates well when

used to enrich genes by their annotations, gene ontologies or metabolic membership. Since existing GSEA methods are built for single time experiments, I created my GSEA method to perform an enrichment analysis in situations where the difference in gene expression is strongly related to time. Thus, I was able to identify metabolic pathways that I could use to subset the expression data in order to focus the rest of the analysis on important TAG genes and reactions.

After dividing the data into more manageable pieces, my method focused on the visualization of the TAG expression pathway because it helps for better understanding than individual compound and gene names and yields testable hypotheses for future analyses. Since there were a smaller number of genes to present at once, the visualization was clearer. Metabolic network and pathway reconstruction and visualization is an idea used by a number of tools that are available online [39] [40] [41] [192] because they can help understand and predict metabolic processes and pathways [193]. However my aim was to focus on the reactions between compounds, and account for reactions that were present in multiple pathways. When I decided to rearranging the graph differently than conventional pathway diagrams, it improved visual pattern finding and I noticed that it showed the change in fold change through time clearer because depicting them as lines drew more attention to them compared to circles. By presenting the time component in the data in this way, it brought more impact to the visualization which was demonstrated when my method clearly showed a concerted up regulation in genes that were implicated in TAG metabolism. The inferred network graph was also able to show the reactions which were not part of the up regulation event with contrasting red lines that stood out considerably from the rest of the network.

Once the inferred network was created, it was possible to apply existing network analyses on it which was my other reason for choosing networking. For example finding bottleneck areas which consists of compounds and reactions that are needed to access compounds from different areas of the network. I chose to search for the shortest path between two compounds important to TAG; glucose and TAG. Although there could be other factors that would restrict such reaction chains from taking place, such as the location of each component within the cell, it was useful in generating a hypothesis for the metabolic pathway of TAG synthesis.

While GSEA can be computationally intensive, this was offset in my method by choosing to only perform it on selected groups and consequently reducing the number of enrichment calculations that needed to be performed. Certainly, the *F. solaris* data set only had four time points for which the inferred probability distribution was relatively simple to fit. There was a considerable increase in run time when the time points exceeded

six points and I was not able to complete the analysis on a randomly generated data set of 10 time points. The run time could be improved by using a standard multi dimensional distribution instead of approximating a unique distribution for each data set. The choice in distribution is limited by the irregular shape of the distribution of RNA-Seq data as different types of experiments would have different shapes. For example, a fold change data set comparing a set of conditions would have most values around 0 but a single condition experiment looking at growth or development would contain more differences. Because I used pathway membership as the gene sets, there was an issue with some genes overlapping several gene sets so that their data was analyzed multiple times. It didn't pose a concern for the calculation and most of the overlaps were small. Rather, some of the choices of gene sets or pathways had a high overlap such as glycolysis, pyruvate metabolism and the citrate cycle (TCA cycle) which also limited interpretation of the data as the vast majority of pathway memberships in *F. solaris* genes is unknown, even with the advantage of a smaller portion data to handle.

The visualization of the inferred network was subsequently crucial to the analysis and was successful at presenting the large amount of expression data even after data division. As humans, we are better at understanding information when it is given in a visual manner so the graphs performed well and presented viewers know the key points of information quickly without needing to remember what the expression was like at each time point. Although there are other existing metabolic pathway visualization tools, the main reason I continued to use R for visualization was to keep the analysis pipeline within the same analysis tool to keep the process simple. One of the challenges with using different tools is exporting the data from the previous tool and importing them into the next tool in the correct format. As the enrichment section of the analysis was carried out in R, there were no data conversion issues to continue to the visualization section. Overall, this method has a straightforward idea behind it so it is easily programmed although it has a reliance of certain R packages which would mean that it would not age well if the packages are not updated.

In addition to data division by metabolic pathways, the data can be subset by gene function as well. In the context of gene regulation, I chose to focus on genes that were transcription factors. These genes are a part of a cells gene expression regulatory system, controlling when transcripts are produced and how much to produce, and since transcription factors affect the regulation of other transcription factors in a chronological manner, it followed that the most suitable method of analysis was to model their activity with a network. For example, if the activation of transcription factor A creates a product that initiates the activity of transcription factor B, gene expression of time

series data will show that with a delay in transcript abundance changes. Simply, the relationships my method can model need not be direct effects but can be indirect effects separated by several time points if it was the case. This is beneficial as many gene regulatory systems are not yet known however, my method can still detect them because it looks for the effect of their actions. Because this type of inference relies on a time delay, it can only be performed with time series data and consequently justifies how advantageous these types of experiments can be.

My method created a way to infer a network that could enable the presentation of temporal changes in gene expression in an effective summary. The process of inferring a network established which genes and expression patterns were important in initiating many of the gene expression changes observed in the data. The main methodology that was crucial to this was the use of a combination of binning and patterns to turn quantitative data into qualitative data in order to make network nodes into discrete variables. Additionally, the resulting network was able to represent two features of transcription factors concurrently; up regulated versus down regulated expression responses, and early versus late expression responses to lipid induction. My method also managed to preserve the qualitative values in the forms of edge weight as the information they represented were crucial to showing the size of the effects from the initiating patterns to the downstream patterns. The weights were effective in showing which stage of the regulation had the biggest change and therefore, had the biggest impact on expression levels. This allowed for the identification of important regulation time points as shown by the *A. thaliana* data.

The results from applying the method to two data sets showed the advantages it has on different types of experiments. When applied on *F. solaris*, it successfully created an inferred transcription factor network on genes containing transcription factor motifs and were very likely to be transcription factors. Using the network, a suitable number of transcription factors were identified for further research as the network marked them based on their influential gene expression on downstream genes. The smaller number of genes can help experimenters construct more focused tests in contrast to pursuing a larger group of genes. The method was also applied to floral transition data from *A. thaliana* and identified the relationships of transcription factors, LFY, AP1, AP3, AGAMOUS, PI and SPL. These were confirmed to play pivotal roles in floral transition and my method correctly placed them in the inferred transcription factor network. By investigating shared and connected patterns, I was able to identify other transcription factors that showed a strong association with the previously mentioned transcription factors that regulate floral transition. The identified transcription factors assisted in

the detection of an association between rhythmic process regulation and immune response regulation via regulation of gibberellins through the use of GO enrichment on the inferred network.

From the success of using visualization from the pathway method, I chose to do the same with the results of this method. By representing the relationships between patterns on a graph, it was clearer and quicker to see which patterns initiated changes and which associations had the most effect.

The concept behind the analysis was quite simple and could be easily implemented in any statistics environment. The only additional data it relied on was transcription factor identification. When applying this method to *F. solaris*, I looked for transcription factor motifs which affected the interpretation of the results. Despite how I applied very stringent criteria when choosing transcription factor candidates, the quality of interpretation was hindered by the limited annotation that was available. In contrast, the application of this method on to *A. thaliana* data showed that this method worked well when the transcription factors were readily identifiable. The network inference is sensitive to the initial gene list so including non transcription factor gene expressions would interfere with the network inference and edge weight calculations. This could also be why the inference produced better results for the *A. thaliana* data compared to the *F. solaris* data. Although the use of discrete expression patterns made network inference possible, it also hindered the scalability of this method. The way the patterns were networked together made it difficult for any connections to be made as the patterns got longer with more time points, resulting in sparser, disconnected graphs. It was for this reason that I chose to perform this analysis on a limited time frame for the *A. thaliana* expression data. This effect could have been diminished with some adjustments made to the edge making decision that would take into consideration the variability of patterns in the time point locations on either side of the main connecting time point. The resulting pathway was still successful in identifying transcription factor relationships and determining which transcription factors would be good candidates for future research in gene regulation.

Although TAG biosynthesis is the leading process for oil accumulation in *F. solaris*, my analyses have also identified other processes that can influence the rate of accumulation. These can include abiotic stress response genes, photosynthesis genes, carbon fixation genes and cell cycle regulators. The effect of other processes has also been observed in brown alga *Ectocarpus siliculosus* [194] and also in other microalgae, *Chlorerlla* [195] and *C. reinhardtii* [111] [196]. However, my analyses only showed little change in photosynthesis related metabolism in *F. solaris* compared to a more noticeable decrease

in other diatoms, indicating that the observation of chloroplast breakdown observed in *C. reinhardtii* which halted cell growth, does not occur or is very limited [197] [198]. My analyses also discovered notable increases in oxidative stress activity which concur with a similar decrease in O_2 evolution activity and acetate utilization in *C. reinhardtii* [199], indicating that *F. solaris* is better at coping with the oil induction environment than currently known diatoms. Lastly, by using my method, I was able to show success with identifying transcription factors of microalgae using comparative analysis [170], and identifying potential transcription factors for further study based on their gene expression data and implicated the importance of the HSF transcription factor domain in regulating abiotic stress response [176]. Since there has been some success with manipulating transcription factors to increase oil accumulation, the transcription factors unveiled by my method could be used as targets for similar metabolic engineering approaches in *F. solaris* [200].

The combination of analysis methods used on the *F. solaris* data here shows how data analysis can be performed on a novel organism with limited available data and NGS gene expression data taken in time series. The type of research involving *F. solaris* particularly focuses on the change in expression over time which is what all my methods used to their advantage. This analysis approach can be applicable to other research topics such as gene expression response to the introduction of a drug, the change in gene expression during development or the impact of diseases on gene expression as it develops. As more ambitious NGS projects are undertaken, there will be more of a need for methods that help biologists investigate and interpret genes and expression in the context of a dynamic living system and accordingly, make new hypothesis based on what was observed in these types of preliminary findings.

Acknowledgements

I would like to express my appreciation to Dr. Aburatani Sachiyo at AIST for her valuable expertise and guidance that greatly assisted in the research of this manuscript. Her willingness to give her time and experience so generously has been very much appreciated.

I would also like to express gratitude to Professor Kuhara Satoru and Associate Professor Tashiro Kosuke at Kyushu University for their support and encouragement of this manuscript.

Additionally, I would like to thank Professor Tanaka Tsuyoshi, Associate Professor Yoshino Tomoko and Assistant Professor Maeda Yoshiaki at TAT, Assistant Professor Tanaka Masayoshi at Tokyo Institute of Technology, and other CREST members at TAT for their advice and assistance on the work with *Fistulifera* sp. strain JPCC DA0580.

My grateful thanks are also extended to Professor Fujibuchi Wataru and Assitant Professor Tanaka Michihiro at Kyoto University for their valuable technical support and consultation on the *Fistulifera* sp. strain JPCC DA0580 data.

Also, I would like to give a special thank you to Mr. Taniguchi Takeaki at MRI for his constructive recommendations on this project.

Last but not the least, I would like to thank all members at CBRC for their support and encouragement throughout my study.

References

[1] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1):57–63, Jan 2009. doi: 10.1038/nrg2484.

[2] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, Dec 1977. doi: 10.1073/pnas.74.12.5463.

[3] The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature*, 409:934–941, Feb 2001. doi: 10.1038/35057157.

[4] J. C. Venter, S. Levy, T. Stockwell, K. Remington, and A. Halpern. Massive parallelism, randomness and genomic advances. *Nature Genetics*, 33:219–227, 2003. doi: 10.1038/ng1114.

[5] M. Margulies, M. Egholm, W. E. Altman, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, Sep 2005. doi: 10.1038/nature03959.

[6] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–32, Sep 2005. doi: 10.1126/science.1117389.

[7] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876, Apr 2008. doi: 10.1038/nature06884.

[8] M. Wadman. James Watson's genome sequenced at high speed. *Nature*, 452 (7189):788, Apr 2008. doi: 10.1038/452788b.

[9] A. M. Thayer. Next-Gen Sequencing Is A Numbers Game. *Chemical and Engineering News*, 92(33):11–15, Aug 2014.

[10] E. Hayden. Is the $1,000 genome for real. *Nature News*, 2014.

[11] P. Wagle, M. Nikolić, and P. Frommolt. QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. *BMC Genomics*, 16:487, Jul 2015. doi: 10.1186/s12864-015-1695-x.

[12] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007. doi: 10.1016/j.cell.2007.05.009.

[13] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502, Jun 2007. doi: 10.1126/science.1141319.

[14] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–7, Aug 2007. doi: 10.1038/nmeth1068.

[15] A. K. Daly. Pharmacogenetics and human genetic polymorphisms. *The Biochemical Journal*, 429(3):435–49, Aug 2010. doi: 10.1042/BJ20100522.

[16] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–8, Nov 2008. doi: 10.1101/gr.078212.108.

[17] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, Mar 2012. doi: 10.1038/nmeth.1923.

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

[19] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, Mar 2009. doi: 10.1093/bioinformatics.

[20] D. Kim and S. L. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):R72, Aug 2011.

[21] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, Apr 2013. doi: 10.1186/gb-2013-14-4-r36.

[22] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–5, May 2010. doi: 10.1038/nbt.1621.

[23] J. Zhang, R. Chiodini, A. Badr, and G. Zhang. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3):95–109, Mar 2011. doi: 10.1016/j.jgg.2011.02.003.

[24] R. Li et al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463:311–317, Jan 2010. doi: 10.1038/nature08696.

[25] Y. Nakamura et al. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proceedings of the National Academy of Sciences of the United States of America*, 110(27):11061–11066, Jul 2013. doi: 10.1073/pnas.1302051110.

[26] V. Pegadaraju, R. Nipper, B. Hulke, L. Qi, and Q. Schultz. De novo sequencing of sunflower genome for SNP discovery using RAD (restriction site associated DNA) approach. *BMC Genomics*, 14:556, Aug 2013. doi: 10.1186/1471-2164-14-556.

[27] J. L. Kelley, J. T. Peyton, A. Fiston-Lavier, N. M. Teets, M. Yee, J. S. Johnston, C. D. Bustamante, R. E. Lee, and D. L. Denlinger. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nature Communications*, 5(4611), Aug 2014. doi: 10.1038/ncomms5611.

[28] S. Atsumi, T. Wu, I. M. P. Machado, P. Chen, M. Pellegrini, and J. C. Liao. Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. *Molecular Systems Biology*, 6(1):449, Dec 2010. doi: 10.1038/msb.2010.98.

[29] L. Wang, X. Han, Y. Zhang, D. Li, X. Wei, X. Ding, and X. Zhang. Deep resequencing reveals allelic variation in Sesamum indicum. *BMC Plant Biology*, 14:225, Aug 2014. doi: 10.1186/s12870-014-0225-3.

[30] A. Grada and K. Weinbrecht. Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology*, 133:e11, Aug 2013. doi: 10.1038/jid.2013.248.

[31] M. R. Kosorok and S. Ma. Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *The Annals of Statistics*, 35(4): 1456–1486, Aug 2007. doi: 10.1214/009053606000001433.

[32] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[33] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

[34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000. doi: 10.1038/75556.

[35] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42:D459–71, Jan 2014. doi: 10.1093/nar.

[36] M. Kotera, Y. Yamanishi, Y. Moriya, M. Kanehisa, and S. Goto. GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Research*, 40:W162–7, Jul 2012. doi: 10.1093/nar.

[37] H. Dinkel, S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jödicke, M. A. Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-Aryamontri, N. Haslam, G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella, and T. J. Gibson. ELM–the database of eukaryotic linear motifs. *Nucleic Acids Research*, 40:D242–51, Jan 2012. doi: 10.1093/nar.

[38] Y. Zhu, L. Sun, A. Garbarino, C. Schmidt, J. Fang, and J. Chen. PathRings: a web-based tool for exploration of ortholog and expression data in biological pathways. *BMC Bioinformatics*, 16(165), May 2015. doi: 10.1186/s12859-015-0585-1.

[39] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, Nov 2003. doi: 10.1101/gr.1239303.

[40] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[41] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.*, 40: D109–D114, 2012.

[42] R. Aboukhalil, B. Fendler, and G. Atwal. Kerfuffle: a web tool for multi-species gene colocalization analysis. *BMC Bioinformatics*, 14(22), Jan 2013. doi: 10.1186/1471-2105-14-22.

[43] A. Lachmann and A. Ma'ayan. Lists2Networks: integrated analysis of gene/protein lists. *BMC Bioinformatics*, 11(87), Feb 2010. doi: 10.1186/1471-2105-11-87.

[44] T. Hulsen, J. Vlieg, and W. Alkema. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(488), Oct 2008. doi: 10.1186/1471-2164-9-488.

[45] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35: W585–7, Jul 2007. doi: 10.1093/nar.

[46] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28 (12):1647–9, Jun 2012. doi: 10.1093/bioinformatics.

[47] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, M. Dudoit, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004. doi: 10.1186/gb-2004-5-10-r80.

[48] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–9, Jun 2008. doi: 10.1126/science.1158441.

[49] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J Rogers, and J. Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–43, Jun 2008. doi: 10.1038/nature07002.

[50] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7): 621–8, Jul 2008. doi: 10.1038/nmeth.1226.

[51] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–9, Jul 2008. doi: 10.1038/nmeth.1223.

[52] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, Sep 2008. doi: 10.1101/gr.079558.108.

[53] R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45(1):81–94, Jul 2008. doi: 10.2144/000112900.

[54] D. Gawron, K. Gevaert, and P. Van Damme. The proteome under translational control. *Proteomics*, 14(23-24):2647–62, Dec 2014. doi: 10.1002/pmic.201400165.

[55] R. Bonasio. The expanding epigenetic landscape of non-model organisms. *Epigenetics in Comparative Physiology*, 218:114–122, 2015. doi: 10.1242/jeb.110809.

[56] B. M. Nugent and T. L. Bale. The omniscient placenta: Metabolic and epigenetic regulation of fetal programming. *Frontiers in Neuroendocrinology*, S0091-3022(15): 30004–2, Sep 2015. doi: 10.1016/j.yfrne.2015.09.001.

[57] Z. Bar-Joseph, A. Gitter, and I. Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews. Genetics*, 13 (8), Jul 2012. doi: 10.1038/nrg3244.

[58] S. Oh, S. Song, N. Dasgupta, and G. Grabowski. The analytical landscape of static and temporal dynamics in transcriptome data. *Frontiers in Genetics*, 5(35), Feb 2014. doi: 10.3389/fgene.2014.00035.

[59] M. J. Nueda, S. Tarazona, and A. Conesa. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, 30(18):2598–602, Sep 2014. doi: 10.1093/bioinformatics.

[60] N. Leng, Y. Li, B. E. McIntosh, B. Duffin, S. Tian, J. A. Thomson, C. N. Dewey, R. Stewart, and C. Kendziorski. EBSeq-HMM: a Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics*, 31 (16):2614–22, Aug 2015. doi: 10.1093/bioinformatics.

[61] J. Hensman, N. D. Lawrence, and M. Rattray. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14(252), Aug 2013. doi: 10.1186/1471-2105-14-252.

[62] S. Aibar, C. Fontanillo, C. Droste, and J. De Las Rivas. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, 31(10):1686–8, May 2015. doi: 10.1093/bioinformatics.

[63] M. Matsumoto, H. Sugiyama, Y. Maeda, R. Sato, T. Tanaka, and T. Matsunaga. Marine diatom, Navicula sp. strain JPCC DA0580 and marine green alga, Chlorella sp. strain NKG400014 as potential sources for biodiesel production. *Applied Biochemistry and Biotechnology*, 161(1-8):483–90, May 2010.

[64] P. G. Falkowski, R. T. Barber, and V. V. Smetacek. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, 281(5374):200–7, Jul 1998.

[65] C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281 (5374):237–40, Jul 1998.

[66] J. C. Goldman. Potential role of large oceanic diatoms in new primary production. *Deep Sea Research Part I: Oceanographic Research Papers*, 40(1):159–168, Jan 1993. doi: 10.1016/0967-0637(93)90059-C.

[67] P. G. Falkowski, R. T. Barber, and V. V. Smetacek. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, 281(5374):200–206, Jul 1998. doi: 10.1126/science.281.5374.200.

[68] Y. Chisti. Biodiesel from microalgae. *Biotechnology Advances*, 25(3):294–306, May-Jun 2007. doi: 10.1016/j.biotechadv.2007.02.001.

[69] M. K. Lam and K. T. Lee. Microalgae biofuels: A critical review of issues, problems and the way forward. *Biotechnology Advances*, 30(3):673–90, May-Jun 2012. doi: 10.1016/j.biotechadv.2011.11.008.

[70] T. M. Mata, A. M. António, and N. S. Caetano. Microalgae for biodiesel production and other applications: A review. *Biotechnology Advances*, 14(1):217–232, Jan 2010. doi: 10.1016/j.rser.2009.07.020.

[71] J. W. Moody, C. M. McGinty, and J. C. Quinn. Global evaluation of biofuel potential from microalgae. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8691–6, Jun 2014.

[72] Teresa M. Mata, Antonia A. Martins, and Nidia S. Caetano. Microalgae for biodiesel production and other applications: A review. *Renewable and Sustainable Energy Reviews*, 14(1):217–32, Jan-Feb 2010.

[73] Q. Hu, M. Sommerfeld, E. Jarvis, M. Ghirardi, M. Posewitz, M. Seibert, and A. Darzins. Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *The Plant Journal*, 54(4):621–39, May 2008. doi: 10.1111/j.1365-313X.2008.03492.x.

[74] S. A. Stuart, M. P. Davey, J. S. Dennis, I. Horst, C. J. Howe, D. J. Lea-Smith, and A. G. Smith. Biodiesel from algae: challenges and prospects. *Current Opinion in Biotechnology*, 21(3):277–286, 2010.

[75] F. X. Malcata. Microalgae and biofuels: a promising partnership? *Trends in Biotechnology*, 29(11):542–9, Nov 2011. doi: 10.1016/j.tibtech.2011.05.005.

[76] P. M. Schenk, S. R. Thomas-Hall, E. Stephens, U. C. Marx, J. H. Mussgnug, C. Posten, O. Kruse, and B. Hankamer. Second generation biofuels: High-efficiency microalgae for biodiesel production. *BioEnergy Research*, 1(1):20–43, Mar 2008.

[77] M. Hannon, J. Gimpel, M. Tran, B. Rasala, and S. Mayfield. Biofuels from algae: challenges and potential. *Biofuels*, 1(5):763–784, Sep 2010.

[78] P. T. Pienkos and A. Darzins. The promise and challenges of microalgal-derived biofuels. *Biofuels, Bioproducts and Biorefining*, 3(4):431–440, May 2009. doi: 10.1002/bbb.159.

[79] A. Singh, P. S. Nigam, and J. D. Murphy. Mechanism and challenges in commercialisation of algal biofuels. *Bioresource Technology*, 102(1):26–34, Jan 2011. doi: 10.1016/j.biortech.2010.06.057.

[80] R. H. Wijffels and M. J. Barbosa. An outlook on microalgal biofuels. *Science*, 329 (5993):796–9, Aug 2010. doi: 10.1126/science.1189003.

[81] P. D. Gressler, T. R. Bjerk, Schneider R. C., M. P. Souza, E. A. Lobo, A. L. Zappe, V. A. Corbellini, and M. S. Moraes. Cultivation of Desmodesmus subspicastus in a tubular photobioreactor for bioremediation and microalgae oil production. *Environmental Technology*, 35(1-4):209–19, Jan-Feb 2014.

[82] C. Fuentes-Grünewald, E. Garcés, E. Alacid, N. Sampedro, S. Rossi, and J. Camp. Improvement of lipid production in the marine strains Alexandrium minutum and Heterosigma akashiwo by utilizing abiotic parameters. *Journal of Industrial Microbiology and Biotechnology*, 39(1):207–16, Jan-Feb 2012.

[83] Randor Radakovits, Robert E. Jinkerson, Susan I. Fuerstenberg, Hongseok Tae, Robert E. Settlage, Jeffrey L. Boore, and Matthew C. Posewitz. Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana. *Nature Communications*, 3, Feb 2012.

[84] D. Shiran, I. Khozin, Y. M. Heimer, and Z. Cohen. Biosynthesis of eicosapentaenoic acid in the microalga Porphyridium cruentum. I: The use of externally supplied fatty acids. *Lipids*, 31(12):1277–82, Dec 1996.

[85] C. Somerville, J. Browse, J. G. Jaworski, and J. B. Ohlrogge. pages 456–527, 2000.

[86] Liliana Rodolfi, Graziella Chini Zittelli, Niccolò Bassi, Giulia Padovani, Natascia Biondi, Gimena Bonini, and Mario R. Tredici. Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnology and Bioengineering*, 102(1):100–112, 2009.

[87] Randor Radakovits, Robert E. Jinkerson, Susan I. Fuerstenberg, Hongseok Tae, Robert E. Settlage, Jeffrey L. Boore, and Matthew C. Posewitz. Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana. *Nature Communications*, 3, 2012.

[88] Hamid Rismani-Yazdi, Berat Z. Haznedaroglu, Carol Hsin, and Jordan Peccia. Transcriptomic analysis of the oleaginous microalga Neochloris oleoabundans reveals metabolic insights into triacylglyceride accumulation. *Biotechnology for Biofuels*, 5(74), 2012.

[89] N. M. Courchesne, A. Parisien, B. Wang, and C. Q. Lan. Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches. *Journal of Biotechnology*, 141(1-2):31–41, Apr 2009. doi: 10.1016/j.jbiotec.2009.02.018.

[90] C. Adams, V. Godfrey, B. Wahlen, L. Seefeldt, and B. Bugbee. Understanding precision nitrogen stress to optimize the growth and lipid content tradeoff in oleaginous green microalgae. *Bioresource Technology*, 131:188–94, Mar 2013. doi: 10.1016/j.biortech.2012.12.143.

[91] Akira Satoh, Kyonosuke Ichii, Mitsufumi Matsumoto, Chihiro Kubota, Michiko Nemoto, Masayoshi Tanaka, Tomoko Yoshino, Tadashi Matsunaga, and Tsuyoshi Tanaka. A process design and productivity evaluation for oil production by indoor mass cultivation of a marine diatom, Fistulifera sp. JPCC DA0580. *Bioresource Technology*, 137:132–8, Jun 2013.

[92] Masaki Muto, Yorikane Fukuda, Michiko Nemoto, Tomoko Yoshino, Tadashi Matsunaga, and Tsuyoshi Tanaka. Establishment of a Genetic Transformation System for the Marine Pennate Diatom Fistulifera sp. Strain JPCC DA0580—A High Triglyceride Producer. *Marine Biotechnology*, 15(1):48–55, Feb 2013.

[93] Daisuke Nojima, Tomoko Yoshino, Yoshiaki Maeda, Masayoshi Tanaka, Michiko Nemoto, and Tsuyoshi Tanaka. Proteomics Analysis of Oil Body-Associated Proteins in the Oleaginous Diatom. *Journal of Proteome Research*, 12(11):5293–301, Nov 2013.

[94] Q. Hu, M. Sommerfield, E. Jarvis, M. Ghirardi, M. Posewitz, M. Seibert, and A. Darzins. Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *The Plant Journal*, 54(4):621–39, May 2008. doi: 10.1111/j.1365-313X.2008.03492.x.

[95] R. Radakovits, R. E. Jinkerson, A. Darzins, and M. C. Posewitz. Genetic engineering of algae for enhanced biofuel production. *Eukaryotic Cell*, 9(4):486–501, Apr 2010. doi: 10.1128/EC.00364-09.

[96] I. Khozin-Goldberg and Z. Cohen. Unraveling algal lipid metabolism: Recent advances in gene identification. *Biochimie*, 93(1):91–100, Jan 2011. doi: 10.1016/j.biochi.2010.07.020.

[97] M. Miyahara, M. Aoi, N. Inoue-Kashino, Y. Kashino, and K. Ifuku. Highly efficient transformation of the diatom Phaeodactylum tricornutum by multi-pulse electroporation. *Bioscience, Biotechnology, and Biochemistry*, 77(4):874–876, 2013.

[98] K. E. Apt, P. G. Kroth-Pancic, and A. R. Grossman. Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Molecular Genetics and Genomics*, 252 (5):572–579, Oct 1996.

[99] C. Bowlder, A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, C. Martens, F. Maumus, and R. P. Otillar. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–244, Nov 2008.

[100] Z. K. Yang, Y. F. Niu, Y. H. Ma, J. Xue, M. H. Zhang, W. D. Yang, J. S. Liu, S. H. Lu, Y. Guan, and H. Y. Li. Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation. *Biotechnology for Biofuels*, 6(1):67, May 2013.

[101] A. V. Klepikova, M. D. Logacheva, S. E. Dmitriev, and A. A. Penin. RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation. *BMC Genomics*, 16(466), June 2015. doi: 10.1186/s12864-015-1688-9.

[102] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.r-project.org`.

[103] Charles D. Warden, Yate-Ching Yuan, and Xiwei Wu. Optimal Calculation of RNA-Seq Fold-Change Values . *International Journal of Computational Bioinformatics and In Silico Modeling*, 2(6):285–292, 2013.

[104] K. Yamada and K. Tomii. Revisiting amino acid substitution matrices for identifying distantly related proteins. *Bioinformatics*, 30(3):317–25, Feb 2014.

[105] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, Apr 1999.

[106] M. Lynch, M. O'Hely, B. Walsh, and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4), Dec 2001.

[107] W. Qian and J. Zhang. Genomic evidence for adaptation by gene duplication. *Genome Research*, 24(8):1356–62, Aug 2014. doi: 10.1101/gr.172098.114.

[108] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind. The role of lineage-specific gene family expansion in the evolution of eukaryotes.

[109] C. Vogel and C. Chothia. Protein family expansions and biological complexity. *PLoS Computational Biology*, 2(5):e48, May 2006. doi: 10.1371/journal.pcbi.0020048.

[110] S. Ohno. *Evolution by gene duplication*. Springer-Verlag, 1970.

[111] G. Wu, D. E. Hufnagel, A. K. Denton, and S. H. Shiu. Retained duplicate genes in green alga Chlamydomonas reinhardtii tend to be stress responsive and experience frequent response gains. *BMC Genomics*, 16:149, Mar 2015. doi: 10.1186/s12864-015-1335-5.

[112] T. Tanaka, Y. Fukuda, T. Yoshino, Y. Maeda, M. Muto, M. Mitsufumi, S. Mayama, and T. Matsunaga. High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, Fistulifera sp. strain JPCC DA0580. *Photosynthesis Research*, 109(1-3):223–9, Sep 2011.

[113] Tsuyoshi Tanaka, Y. Maeda, A. Veluchamy, Michihiro Tanaka, H. Abida, E. Maréchal, C. Bowler, M. Muto, Y. Sunaga, Masayoshi Tanaka, T. Yoshino, T. Taniguchi, Y. Fukuda, M. Nemoto, M. Matsumoto, P. S. Wong, S. Aburatani, and W. Fujibuchi. Oil Accumulation by the Oleaginous Diatom Fistulifera solaris as Revealed by the Genome and Transcriptome . *The Plant Cell*, 27:162–176, Jan 2015.

[114] I. Khozen-Goldberg and Z. Cohen. Unraveling algal lipid metabolism: Recent advances in gene identification. *Biochimie*, 93(1):91–100, Jan-Feb 2011.

[115] R. V. Lenth. *lsmeans: Least-Squares Means*, 2014.

[116] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[117] M. D. Allen, J. A. Del Campo, J. Kropat, and S. S. Merchant. FEA1, FEA2, and FRE1, encoding two homologous secreted proteins and a candidate ferrireductase, are expressed coordinately with FOX1 and FTR1 in iron-deficient Chlamydomonas reinhardtii. *Eukaryotic Cell*, 6(10):1841–1852, 2007.

[118] H. A. Dailey and T. A. Dailey. Protoporphyrinogen oxidase of Myxococcus xanthus. Expression, purification, and characterization of the cloned enzyme. *Journal of Biological Chemistry*, 271(15):8714–8, Apr 1996.

[119] T. Ishikawa and S. Shigeoka. Recent advances in ascorbate biosynthesis and the physiological significance of ascorbate peroxidase in photosynthesizing organisms. *Bioscience, Biotechnology, and Biochemistry*, 72(5):1143–54, May 2008.

[120] S. W. Lee, H. W. Lee, H. J. Chung, Y. A. Kim, Y. J. Kim, Y. Hahn, J. H. Chung, and Y. S. Park. Identification of the genes encoding enzymes involved in the early biosynthetic pathway of pteridines in Synechocystis sp. PCC 6803. *FEMS Microbiology Letters*, 176(1):169–79, Jul 1999.

[121] S. Pohl, H. M. Mitchison, A. Kohlschütter, O. Van Diggelen, T. Braulke, and O. Storch. Increased expression of lysosomal acid phosphatase in CLN3-defective cells and mouse brain tissue. *Journal of Neurochemistry*, 103(6):2177–88, Dec 2007.

[122] A. Beranek, G. Rechberger, H. Knauer, H. Wolinski, S. D. Kohlwein, and R. Leber. Identification of a cardiolipin-specific phospholipase encoded by the gene CLD1 (YGR110W) in yeast. *Journal of Biological Chemistry*, 284(17):11572–8, Apr 2009.

[123] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, 8 (2), 2012.

[124] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*, 4(144):44–57, 2009.

[125] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(144):1–13, 2009.

[126] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–43, Apr 2013. doi: 10.1093/bioinformatics.

[127] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A Novel Signaling Pathway Impact Analysis (SPIA). *Bioinformatics*, 25(144):75–82, 2009.

[128] Seon-Young Kim and David J Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, 6(144), 2005.

[129] Gabriel F. Berrix, John E. Beaver, Can Cenik, Murat Tasan, and Frederick P. Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25(22): 3043–3044, 2009.

[130] Q. Zheng and X. Wang. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, 36, 2008.

[131] J. Gillis, M. Mistry, and P. Pavlidis. Gene function analysis in complex data sets using ErmineJ. *Nature Protoc*, 5:1148–1159, 2010.

[132] Weijun Luo, Michael Friedman, Kerby Shedden, Kurt Hankenson, and Peter Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(161), 2009.

[133] Z. Hu, J. Mellor, J. Wu, and C. DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5(17), Feb 2004. doi: 10.1186/1471-2105-5-17.

[134] T. Yamada, I. Letunic, S. Okuda, M. Kanehisa, and P. Bork. iPath2.0: interactive pathway explorer. *Nucleic Acids Research*, 39:W412–W415, May 2011. doi: 10.1093/nar/gkr313.

[135] Charlotte Maia. *mecdf: Multivariate Empirical Cumulative Distribution Functions*, 2011.

[136] Jitao David Zhang and Stefan Wiemann. KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. *Bioinformatics*, 25(11):1470–1471, 2009.

[137] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[138] T. J. Ettema, H. Ahmed, A. C. Geerling, J. van der Oost, and B. Siebers. The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) of Sulfolobus solfataricus: a key-enzyme of the semi-phosphorylative branch of the Entner-Doudoroff pathway. *Extremophiles*, 12(1):75–88, Jan 2008.

[139] E. J. Grasso, M. B. Scalambro, and R. O. Calderón. Differential response of the urothelial V-ATPase activity to the lipid environment. *Cell Biochemistry and Biophysics*, 61(1):157–168, September 2011.

[140] Roberta Croce and Herbert van Amerongen. Light-harvesting in photosystem I. *Photosynthesis Research*, May 2013.

[141] J. Doebley and L. Lukens. Transcriptional regulators and the evolution of plant form. *The Plant Cell*, 10(7):1075–1082, Jul 1998. doi: 10.1105/tpc.10.7.1075.

[142] D. Tautz. Evolution of transcriptional regulation. *Current Opinion in Genetics and Development*, 10(5):575–9, Oct 2000. doi: 10.1016/S0959-437X(00)00130-1.

[143] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96:4285–4288, April 1999.

[144] Y. Kobayashi and D. Weigel. Move on up, it's time for change–mobile signals controlling photoperiod-dependent flowering . *Genes and Development*, 21(19): 2371–84, October 2007. doi: 10.1101/gad.1589007.

[145] R. Borner, G. Kampmann, J. Chandler, R. Gleissner, E. Wisman, K. Apel, and S. Melzer. A MADS domain gene involved in the transition to flowering in Arabidopsis. *The Plant Journal*, 24(5):591–9, Dec 2000. doi: 10.1046/j.1365-313x.2000.00906.x.

[146] H. Lee, S. S. Suh, E. Park, E. Cho, J. H. Ahn, S. G. Kim, J. S. Lee, Y. M. Kwon, and I. Lee. The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in Arabidopsis. *Genes and Development*, 14(18):2366–76, Sep 2000. doi: 10.1101/gad.813600.

[147] M. Abe, Y. Kobayashi, S. Yamamoto, Y. Daimon, A. Yamaguchi, Y. Ikeda, H. Ichinoki, M. Notaguchi, K. Goto, and T. Araki. FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science*, 309(5737):1052–6, Aug 2005. doi: 10.1126/science.1115983.

[148] P. A. Wigge, M. C. Kim, K. E. Jaeger, W. Busch, M. Schmid, J. U. Lohmann, and D. Weigel. Integration of spatial and temporal information during floral induction in Arabidopsis. *Science*, 309(5737):1056–9, Aug 2005. doi: 10.1126/science.1114358.

[149] S. D. Michaels and R. M. Amasino. FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell*, 11(5): 949–56, May 1999.

[150] U. Hartmann, S. Höhmann, K. Nettesheim, E. Wisman, H. Saedler, and P. Huijser. Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. *The Plant Journal*, 21(4):351–60, Feb 2000. doi: 10.1046/j.1365-313x.2000.00682.x.

[151] M. A. Blázquez, L. N. Soowal, I. Lee, and D. Weigel. LEAFY expression and flower initiation in Arabidopsis. *Development*, 124(19):3835–44, Oct 1997.

[152] F. D. Hempel, D. Weigel, M. A. Mandel, G. Ditta, P. C. Zambryski, L. J. Feldman, and M. F. Yanofsky. Floral determination and expression of floral regulatory genes in Arabidopsis. *Development*, 124(19):3845–53, Oct 1997.

[153] M. A. Busch, Bomblies K., and D. Weigel. Activation of a floral homeotic gene in Arabidopsis. *Science*, 285(5427):585–7, Jul 1999. doi: 10.1126/science.285.5427.585.

[154] S. A. Kempin, B. Savidge, and M. F. Yanofsky. Molecular basis of the cauliflower phenotype in Arabidopsis. *Science*, 267(5197):522–5, Jan 1995.

[155] C. Smaczniak, R. G. Immink, J. M. Muiño, R. Blanvillain, M. Busscher, J. Busscher-Lange, Q. D. Dinh, S. Liu, A. H. Westphal, S. Boeren, F. Parcy, L. Xu, C. C. Carles, G. C. Angenent, and K. Kaufmann. Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proceedings of the National Academy of Sciences of the United States of America*, 109(5):1560–65, Jan 2012. doi: 10.1073/pnas.1112871109.

[156] G. H. Cardon, S. Höhmann, K. Nettesheim, H. Saedler, and P. Huijser. Functional analysis of the Arabidopsis thaliana SBP-box gene SPL3: a novel gene involved in the floral transition. *The Plant Journal*, 12(4):367–77, Aug 1997. doi: 10.1046/j.1365-313X.1997.12020367.x.

[157] S. Schwarz, A. V. Grande, N. Bujdoso, H. Saedler, and P. Huijser. The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in Arabidopsis. *Plant Molecular Biology*, 67(1-2):183–95, May 2008. doi: 10.1007/s11103-008-9310-z.

[158] J. W. Wang, B. Czech, and D. Weigel. miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana . *Cell*, 138(4):738–49, Aug 2009. doi: 10.1016/j.cell.2009.06.014.

[159] A. Yamaguchi, M. F. Wu, L. Yang, G. Wu, R. S. Poethig, and D. Wagner. The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. *Developmental Cell*, 17(2):268–78, Aug 2009. doi: 10.1016/j.devcel.2009.06.007.

[160] M. A. Mandel, C. Gustafson-Brown, B. Savidge, and M. F. Yanofsky. Molecular characterization of the Arabidopsis floral homeotic gene APETALA1. *Nature*, 360(6401):273–7, Nov 1992. doi: 10.1038/360273a0.

[161] M. Ng and M. F. Yanofsky. Activation of the Arabidopsis B class homeotic genes by APETALA1. *Plant Cell*, 13(4):739–53, Apr 2001.

[162] R. D. Fin, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. The Pfam Protein Families Database. *Nucleic Acids Research*, (42):D222–D230, 2014 .

[163] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919, Nov 2001.

[164] Sarah K. Kummerfeld and Sarah A. Teichmann. DBD: a transcription factor prediction database. *Nucl. Acids Res.*, 34(suppl_1):D74–81, 2006. doi: 10.1093/nar. URL `http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D74`.

[165] E. Rayko, F. Maumus, U. Maheswari, K. Jabbari, and C. Bowler. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytologist*, 188(1):52–66, Oct 2010. doi: 10.1111/j.1469-8137.2010.03371.x.

[166] P. Pérez-Rodríguez, D. M. Riaño Pachón, L. G. G. Corrêa, S. A. Rensing, B. Kersten, and B. Mueller-Roeber. PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research*, 38:D822–D827, October 2010. doi: 10.1093/nar.

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

[168] X. Yi, Z. Du, and Z. Su. PlantGSEA: a Gene Set Enrichment Analysis toolkit for plant community. *Nucleic Acids Research*, 41(W1):98–103, Jul 2013. doi: 10.1093/nar.

[169] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PloS One*, 6(7):e21800, Jul 2011. doi: 10.1371/journal.pone.0021800.

[170] J. Hu, D. Wang, J. Li, G. Jing, K. Ning, and J. Xu. Genome-wide identification of transcription factors and transcription-factor binding sites in oleaginous microalgae Nannochloropsis. *Scientific Reports*, 4(5454), Jun 2014. doi: 10.1038/srep05454.

[171] N. Focks and C. Benning. Wrinkled1: a novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. *Plant Physiology*, 118(1):91–101, Sep 1998.

[172] S. Baud, S. Wuillème, A. To, C. Rochat, and L. Lepiniec. Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in Arabidopsis. *The Plant Journal*, 60(6):933–947, Dec 2009. doi: 10.1111/j.1365-313X.2009.04011.x.

[173] A. Cernac and C. Benning. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in Arabidopsis. *The Plant Journal*, 40(4):575–585, Nov 2004.

[174] L. Nover, K. Bharti, P. Döring, S. K. Mishra, A. Ganguli, and K. D. Scharf. Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? *Cell Stress Chaperones*, 6:177–189, 2001.

[175] K. D. Scharf, T. Berberich, I. Ebersberger, and L. Nover. The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochimica et Biophysica Acta*, 1819:104–119, 2012.

[176] I. Pérez-Salamó, C. Papdi, G. Rigó, L. Zsigmond, B. Vilela, V. Lumbreras, I. Nagy, B. Horváth, M. Domoki, Z. Darula, K. Medzihradszky, L. Bögre, C. Koncz, and L. Szabados. The heat shock factor A4A confers salt tolerance and is regulated by oxidative stress and the mitogen-activated protein kinases MPK3 and MPK6. *Plant Physiology*, 165(1):319–334, May 2014.

[177] T. Phillips. Genetic Signaling: Transcription Factor Cascades and Segmentation. *Nature Education*, 1(1):200, 2008.

[178] D. Greenbaum, R. Jansen, and M. Gerstein. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*, 18 (4), Oct 2002. doi: 10.1093/bioinformatics/18.4.585.

[179] C. Robert and M. Watson. Errors in RNA-Seq quantification affect genes of

relevance to human disease. *Genome Biology*, 16(177), Jun 2015. doi: 10.1186/s13059-015-0734-x.

[180] J. M. Franco-Zorrilla, P. Cubas, J. A. Jarillo, B. Fernández-Calvín, J. Salinas, and J. M. Martínez-Zapater. AtREM1, a Member of a New Family of B3 Domain-Containing Genes, Is Preferentially Expressed in Reproductive Meristems. *Plant Physiology*, 128(2):418–427, Feb 2002.

[181] O. Mantegazza, V. Gregis, M. A. Mendes, P. Morandini, M. Alves-Ferreira, C. M. Patreze, S. M. Nardeli, M. M. Kater, and L. Colombo. Analysis of the arabidopsis REM gene family predicts functions during flower development. *Annals of Botany*, 114(7):1507–15, Nov 2014. doi: 10.1093/aob.

[182] Y. Oono, M. Seki, M. Satou, K. Iida, K. Akiyama, T. Sakurai, M. Fujita, K. Yamaguchi-Shinozaki, and K. Shinozaku. Monitoring expression profiles of Arabidopsis genes during cold acclimation and deacclimation using DNA microarrays. *Functional & Integrative Genomics*, 6(14), Feb 2006. doi: 10.1007/s10142-005-0014-z.

[183] K. Ohashi-Ito and D. C. Bergmann. Arabidopsis FAMA Controls the Final Proliferation/Differentiation Switch during Stomatal Development. *American Society of Plant Biologists*, 18(10):2493–2505, Oct 2006.

[184] Y. Zhang, S. Schwarz, H. Saedler, and P. Huijser. SPL8, a local regulator in a subset of gibberellin-mediated developmental processes in Arabidopsis. *Plant Molecular Biology*, 63(3):429–439, Feb 2007. doi: 10.1007/s11103-006-9099-6.

[185] S. A. Jorgensen and J. C. Preston. Differential SPL gene expression patterns reveal candidate genes underlying flowering time and architectural differences in Mimulus and Arabidopsis. *Molecular Phylogenetics and Evolution*, 73:129–39, Apr 2014. doi: 10.1016/j.ympev.2014.01.029.

[186] M. K. Chen, W. H. Hsu, P. F. Lee, M. Thiruvengadam, H. I. Chen, and C. H. Yang. The MADS box ge ne, FOREVER YOUNG FLOWER, acts as a repressor controlling floral organ senescence and abscission in Arabidopsis. *The Plant Journal*, 68 (1):168–85, Oct 2011. doi: 10.1111/j.1365-313X.2011.04677.x.

[187] W. H. Chen, P. F. Li, M. K. Chen, Y. I. Lee, and C. H. Yang. FOREVER YOUNG FLOWER Negatively Regulates Ethylene Response DNA-binding Factors (EDFs), by Activating An Ethylene-Responsive Factor (ERF), to Control Arabidopsis Floral Organ Senescence and Abscission. *Plant Physiology*, 10, Jun 2015.

[188] D. Spies and C. Ciaudo. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Computational and Structural Biotechnology Journal*, 13:469–477, Aug 2015. doi: 10.1016/j.csbj.2015.08.004.

[189] W. R. Pearson and M. L. Sierk. The limits of protein sequence comparison? *Current Opinion in Structural Biology*, 15(3):254–260, Jun 2005. doi: 10.1016/j.sbi.2005.05.005.

[190] P. Myllynen, M. Kummu, and E. Sieppi. ABCB1 and ABCG2 expression in the placenta and fetus: an interspecies comparison. *Expert Opinion on Drug Metabolism and Toxicology*, 6(11):1385–98, Nov 2010. doi: 10.1517/17425255.2010.514264.

[191] Y. Liang, Y. Maeda, Y. Sunaga, M. Muto, M. Matsumoto, T. Yoshino, and T. Tanaka. Biosynthesis of Polyunsaturated Fatty Acids in the Oleaginous Marine Diatom Fistulifera sp. Strain JPCC DA0580. *Marine drugs*, 11(12):5008–23, Dec 2013.

[192] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39:D691–7, Jan 2011. doi: 10.1093/nar.

[193] J. Koussa, A. Chaiboonchoe, and K. Salehi-Ashtiani. Computational Approaches for Microalgal Biofuel Optimization: A Review. *BioMed Research International*, 2014(2014), Sep 2014. doi: 10.1155/2014/649453.

[194] S. M. Dittami, D. Scornet, J. Petit, B. Ségurens, C. Silva, E. Corre, M. Dondrup, K. Glatting, R. König, L. Sterck, P. Rouzé, Y. Van de Peer, J. M. Cock, C. Boyen, and T. Tonon. Global expression analysis of the brown alga Ectocarpus siliculosus (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biology*, 10:R66, Jun 2009. doi: 10.1186/gb-2009-10-6-r66.

[195] Q. Jiang, L. Zhao, J. Dai, and Q. Wu. Analysis of autophagy genes in microalgae: Chlorella as a potential model to study mechanism of autophagy. *PloS One*, 7(7):e41826, 2012. doi: 10.1371/journal.pone.0041826.

[196] M. Gargouri, J. Park, F. O. Holguin, M. Kim, H. Wang, R. Deshpande, Y. Shachar-Hill, L. M. Hicks, and D. R. Gang. Identification of regulatory network hubs that

control lipid metabolism in Chlamydomonas reinhardtii. *Journal of Experimental Botany*, 66(20), May 2015. doi: 10.1093/jxb/erv217.

[197] M. Iwai, K. Ikeda, M. Shimojima, and H. Ohta. Enhancement of extraplastidic oil synthesis in Chlamydomonas reinhardtii using a type-2 diacylglycerol acyltransferase with a phosphorus starvation–inducible promoter. *Plant Biotechnology Journal*, 12(6):808–819, Aug 2014. doi: 10.1111/pbi.12210.

[198] T. Cakmak, P. Augun, Y. E. Demiray, A. D. Ozkan, Z. Elibol, and T. Tekinay. Differential effects of nitrogen and sulfur deprivation on growth and biodiesel feedstock production of Chlamydomonas reinhardtii. *Biotechnology and Bioengineering*, 109(8):1947–1957, Aug 2012. doi: 10.1002/bit.24474.

[199] V. H. Work, R. Radakovits, R. E. Jinkerson, J. E. Meuser, L. G. Elliott, D. J. Vinyard, L. M. L. Laurens, G. C. Dismukes, and M. C. Posewitz. Increased Lipid Accumulation in the Chlamydomonas reinhardtiista7-10 Starchless Isoamylase Mutant and Increased Carbohydrate Synthesis in Complemented Strains. *Eukaryotic Cell*, 9(8):1251–1261, Aug 2010. doi: 10.1128/EC.00075-10.

[200] J. Zhang, Q. Hao, J. Xu, W. Yin, L. Song, L. Xu, X. Guo, C. Fan, Y. Chen, J. Ruan, S. Hao, Y. Li, R. R. Wang, and Z. Hu. Overexpression of the soybean transcription factor GmDof4 significantly enhances the lipid content of Chlorella ellipsoidea. *Biotechnology for Biofuels*, 7:128, Sep 2014. doi: 10.1186/s13068-014-0128-4.