



Swansea
University
Prifysgol
Abertawe

Gene regulation in fatty acid pathways in
cyanobacteria exposed to far-red light

Patrick William Colledge
BSc (Hons)

*Submitted to Swansea University in fulfilment of the
requirements
for the Degree of MRes Biosciences*

Swansea University

2022

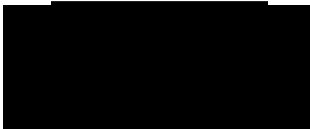
Abstract

Cyanobacteria are photosynthetic microorganisms that can utilise white, far-red or UV light. The organisms adapt their cellular metabolism to their environment by regulating their genetic expression to accumulate or reduce metabolites such as valuable polyunsaturated fatty acids. The production of high-value metabolites whilst using light and CO₂ to grow make these organisms a successful candidate within the biotechnological industry. Far-red light is known to influence cyanobacteria, especially their photosynthetic apparatus. Thylakoid membranes are known to be altered by far-red light photoacclimation (FaRLiP). However, the role of fatty acid synthesis and desaturation during FarLip is little studied. This study examines the gene expression of 33 fatty acid-related genes when the cyanobacterium *Chlorogloeopsis fritschii* PCC 6912 is grown under far-red light compared to white light using RNA-seq data from the NCBI database. The transcriptomic analysis, encompassing a bioinformatic pipeline to process and quantify transcripts, found that only two genes of the 33 examined genes were differentially expressed. The two genes were down-regulated and encoded for a fatty acid desaturase (*fad*) and a *pfaD*/polyketide biosynthesis protein. The results suggest that fatty acid desaturation decreases in favour of saturated fatty acids. Furthermore, the identification of a *pfa* gene cluster may highlight a secondary path of polyunsaturated fatty acids within heterocyst cyanobacteria. This study also highlights the need for greater biological replicates within RNA-seq experiments and increased completeness in genome annotations in biotechnological databases.

University Declarations and Statements

I declare that this work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

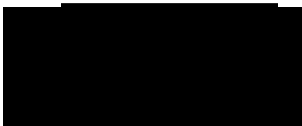
Signed:



Patrick William Colledge
Date 30/08/22

This thesis is the result of my own investigations, except where otherwise stated and that other sources are acknowledged.

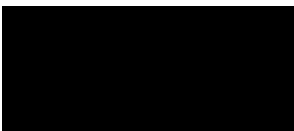
Signed:



Patrick William Colledge
30/08/22

I give consent for the thesis, if accepted to be made available online in the University's Open Access Repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed:



Patrick William Colledge
30/08/22

Statement of Expenditure

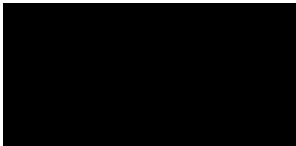
Student name: Patrick Colledge

Student number: [REDACTED]

Project title: Gene regulation in fatty acid pathways in cyanobacteria exposed to far-red light.



I hereby certify that the above information is true and correct to the best of my knowledge



Student signature

Statement of Contributions

Below is a table providing all the contributing roles and contributors involved within each role:

PWC – Patrick William Colledge, CAL – Prof Carole Llewellyn, AOU – Almudena Ortiz-Urquiza, ML – Miguel Lurgi

Contributor Role	Contributor(s)
Conceptualisation	PWC, CAL, AOU
Data Curation	PWC, CAL, ALG-AD research group, Swansea University
Formal Analysis	PWC
Investigation	PWC, AOU
Methodology	PWC
Project Administration	CAL, AOU, ML
Resources	PWC, CAL
Supervision	CAL, AOU, ML
Validation	PWC, AOU
Visualisation	PWC
Writing – Original Draft Preparation	PWC
Writing – Review & Editing	PWC, AOU, CAL


Copy of Ethics Approval

Below is proof that the ethics of this study has been investigated as approved to allow the study to occur:

Student Details

Name: Patrick Colledge
Student Number: [REDACTED]
Level: 7
Course: Biosciences
Project Supervisor: Dr Miguel Lurgi Rivera
Last Updated Date: 26 Aug 2022, 1:48 p.m.
Last Reviewed Date: 26 Aug 2022, 1:55 p.m.
Reviewed by: Miguel Lurgi Rivera

Projects Ethics Assessment Status

Project Title	Status	Approval Number
Cyanobacterial fatty Acid gene regulation within different light conditions	 Completed	[REDACTED]

Risk assessment:

Risk Assessment			
Desk-based study on public data.			
College/ PSU	FSE Biosciences	Assessment Date	03/02/21
Location	Singleton Campus	Assessor	Patrick Colledge
Activity	MRes Project	Review Date (if applicable)	n/a
Associated documents	<ul style="list-style-type: none"> N/A 		

Part 1: Risk Assessment

What are the hazards?	Who might be harmed?	How could they be harmed?	What are you already doing?	S	L	Risk (SxL)	Do you need to do anything else to manage this risk?	S	L	Risk (SxL)	Additional Action Required
Visual Display Equipment	Myself	Eye strain, posture discomfort or misalignment, injury to their hands, arms, neck or back due to overuse and poorly designed workstations with improper supporting chairs.	Breaks taken to reduce possibility, awareness of good posture and use of comfortable seating that supports back and neck posture. Working in an environment where shading prevents reflection and glare of screen.	2	2	4	Review work prevention methods to assess preventing health issues from arising at any sign of injury or symptoms occurring	1	2	2	No
Extended	Myself	Misalignment of posture,	Taking regular	2	2	4	Review recommend	1	2	2	No

What are the hazards?	Who might be harmed?	How could they be harmed?	What are you already doing?	S	L	Risk (SxL)	Do you need to do anything else to manage this risk?	S	L	Risk (SxL)	Additional Action Required
periods of sitting indoors on a computer due to a completely desk-based thesis		degeneration of muscles, decrease in heart health and increase in blood pressure to do the lowered opportunity to exercise. Increased chance of obesity, depression, stress and fatigue due to long periods indoors.	breaks to stretch and move around the room at least once an hour. Eating healthy and incorporating socializing and exercise into longer breaks and during weekly routines outside of work. Eating and relaxing away from the desk space.				actions for the prevention of mental decline and take a break and seek medical attention if signs or symptoms occur.				

Part 2: Actions arising from risk assessment

Actions	Lead	Target Date	Done Yes/No
N/A	N/A	N/A	N/A

Table of Contents

<i>Acknowledgements</i>	1
<i>Tables and Figures</i>	2
<i>Definitions and Abbreviations</i>	4
<i>Introduction</i>	6
Cyanobacteria and their importance	6
What are fatty acids and where are they found in cyanobacteria?.....	7
How are fatty acids produced?	9
How are fatty acids different in cyanobacteria?	11
How is fatty acid synthesis affected by far-red light?	13
Aims and objectives of this study	14
<i>Methodology</i>	15
<i>Chlorogloeopsis fritschii</i> cultures, light treatments and RNA extraction and sequencing.....	15
Transcript quality assessment	16
Genome alignment and transcriptome assembly	17
Transcript quantification	18
Quantification of gene expression and differential expression analysis of fatty acids	18
<i>Results and Discussion</i>	20
<i>Transcript identification and quantification of fatty acid-related genes</i>	20
Differential expression of genes involved in fatty acid synthesis and desaturation	26
<i>Principal Component Analysis of the complete transcriptome and fatty acid-related genes</i>	26
<i>Differential expression of fatty acid-related genes</i>	30
<i>Discussion</i>	35
<i>Conclusion</i>	40
<i>Appendices A – Data from the RNA processing and quality assessment</i>	42
Transcript quality control, identification, and quantification	42
<i>Removal of low-quality reads</i>	42
<i>Assessment of GC Content</i>	47
<i>Transcript trimming and mapping</i>	48
<i>Appendices B – Coding of the data analysis</i>	51
Code within Ubuntu	51
Code within R Studio	75
<i>References</i>	87

Acknowledgements

I'd like to thank Dr Almudena Ortiz-Urquiza and Prof Carole Llewellyn for supervising me through this thesis and helping me through many struggles. Both have taught me invaluable life skills and inspire me to try to conduct research as stellar as they do. A special thanks to Dr Nicholas Delhomme at Umeå Plant Science Centre (UPSCb) for the invaluable training required to complete this thesis. Thank you to the ALG-AD team at Swansea University under Prof Carole Llewellyn for conducting and producing the raw data that allowed to me investigate fatty acid pathway expression.

This thesis would not be complete without the support from three friends, Alexie Jenkins, Calista Collins, and Ellie Evans; I was lucky to live with and experience support for each other like no other. Thank you to my partner, Thomas Evans, for helping me stay focussed, supporting and listening to me even when I stopped making sense.

However, I'd like to thank my mother, Alison Colledge, the most for pushing me through all the hard times with love and strength whilst losing loved ones on the way. To my grandmother, Margaret Baker, thank you for telling me how proud you were of this work even though you aren't here for the submission. To my grandmother, Mary Colledge, who passed away two weeks before submission, you were an inspiration with your strength and created a kind and supportive father for me, thank you.

Finally, Katie Colledge, thank you for all the walks we did that gave me some sort of sanity throughout.

Tables and Figures

Figure / Table number	Title	Page Number
Figure 1.	FAS II Fatty acid synthesis producing acyl-ACP within cyanobacteria. Metabolites are shown in black whilst enzymes and the genes encoding them are shown in blue and red, respectively. Adapted from Taylor (2012) by the addition of enzyme and associated gene names.	10
Figure 2	The comparison of fab genes (encoding for FASII) between <i>Synechococcus elongatus</i> and <i>Escherichia coli</i> . <i>E. coli</i> has two extra genes and a different operonic structure. This may be due to being heterotrophic or due to a greater need of carbon from heterotrophic growth. From Santos-Merino (2018).	12
Figure 3	The cultivation of 3 x 800 mL <i>C. fritschii</i> cultures firstly in white light for six days then moved to far-red light for 24 hours. 50 mL aliquots were taken at the end of each light cultivation.	15
Figure 4	The bioinformatic pipeline to enhance, align and quantify RNA-seq transcripts.	17
Table 1	The RNA-seq sample names, their sample code and link to the NCBI database.	20
Table 2	Genes selected for analysis. The 33 genes were identified within the reference genome and had transcripts aligned. The table contains details about the gene selected from the genome reference as well as further information about the enzyme the gene encodes for within previous literature.	22-25
Figure 5	The variance of expression for each gene within the transcriptome between each sample before VST. Greater variance can be seen in higher expressed genes whilst most genes are densely distributed around the mean standard deviation (red line).	26
Figure 6	The expression variance for each gene within the transcriptome between each sample after VST. The density of genes around the mean (red line) has greatly increased, and the variation in standard deviation (left axis) at all expression levels has decreased.	27
Figure 7	The Principal Component Analysis (PCA) of the quantified transcriptome before DESeq2 analysis to investigate visual differences between control and treatment.	28
Figure 8	The PCA plot of the six samples only using the 33 genes associated with fatty acid synthesis before DESeq2 analysis to investigate visual differences between control and treatment	29
Figure 9	The differential expression cut-off using LFC and FDR of 0.5 and 0.01, respectively, shows a total of 974 genes that are differentially expressed under far-red light presented in a volcano plot.	31
Figure 10	The heatmap produced with calculated Z values (left) for the fatty acid synthesis genes present within the annotated genome alongside the Log2fold change and BH-adjusted p-values (middle).	32

Figure / Table number	Title	Page Number
	Additionally, the biochemical pathway of fatty acid synthesis (right) is presented, showing the genes in red and the enzymes they encode in blue.	
Figure 11	The heatmap produced (left) for the oleic acid synthesis genes alongside their Log2fold and BH-adjusted p-values (middle). Additionally, the biochemical pathway of oleic acid synthesis (right) shows the gene in red and the enzyme they encode in blue.	33
Figure 12	The variance of expression for each gene within the transcriptome between each sample after VST. The density of genes around the mean (red line) has greatly increased, and the variation in standard deviation (left axis) at all expression levels has decreased.	34
Figure 13	Four genes related to polyunsaturated fatty acid desaturase (PUFA-producing polyketide-like synthases biosynthesis) were identified. Their Z-values were calculated and displayed within a heatmap (left). Additionally, their log2fold values and BH-adjusted p-values are shown (right).	35
Figure 14	The number of RNA reads from each biological sample throughout the bioinformatic pipeline. For all samples, the different steps within the process of quality assessment allowed for greater identification of unique reads while reducing duplicate reads.	41
Figure 15	The mean Phred score for each base position within the samples along the bioinformatic pathway.	43
Figure 16	The number of sequences with each Phred score	44
Figure 17	The GC content of the samples along the bioinformatic pathway show a normal deviation throughout the bioinformatic pipeline. Only minor differences are seen between flasks (i.e., biological replicates) and between different stages of the quality assessment process implemented in the pipeline.	46
Figure 18	The Trimmomatic output breakdown shows that >99.9% of both transcripts within each file survived the Trimmomatic analysis.	47
Figure 19	<i>The proportion of reads mapped or unmapped during the STAR transcript alignment. Most transcripts were uniquely mapped during STAR alignment.</i>	48

Definitions and Abbreviations

<u>Abbreviation</u>	<u>Definition</u>
ALA	Alpha-linolenic acid
ARA	Arachidonic acid
AU	Absorbance units
BH-p values	Benjamini-Hochberg adjusted probability value
bp	Base pairs
CLA	Conjugated Linoleum acid
CoA	Coenzyme A
DES	Fatty acid desaturases
DHA	Docosahexaenoic acid
EPA	Eicosapentaenoic acid
FaRLiP	Far red photoacclimation
FAS	Fatty acid synthesis
FFA	Free fatty acid
Fwd	Forward facing strand
GC	Guanine-Cytosine
GLA	Gamma-linolenic acid
KAS	3-ketoacyl-ACP-synthase
LA	Linoleum acid
LED	Light emitting diode
LFC	Log fold change
log	Logarithmic scale
MMP	Maximal mappable prefix
mRNA	Messenger RNA
MUFA	Monounsaturated fatty acid
NCBI	National Centre for Biotechnology Information
OD	Optical Density
PBS	Phycobilisome
PC1 & 2	Principal component 1 & 2
PCA	Principal component analysis
PE	Paired-end
PfaD	Polyunsaturated fatty acid desaturases
PKS	Polyketide synthase
PSI & II	Photosystem I & II
PUFA	Polyunsaturated fatty acid
PUFA	Polyunsaturated fatty acid
Ref	reference
Rev	Reverse facing strand
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequence
ROS	Reactive Oxygen species
rRNA	Ribosomal ribonucleic acid

<u>Abbreviation</u>	<u>Definition</u>
UV	Ultraviolet
VST	Variance stabilising transformation

Introduction

Cyanobacteria are photosynthesising prokaryotes that utilise light and carbon sources to adapt to many different environments. In recent years, interest has grown in the biotechnological applications of cyanobacteria, including in bioremediation, as food supplements and for carbon uptake to reduce of carbon emissions (Zahra *et al.*, 2020). The advantages of using cyanobacteria to produce metabolites include their ease of cultivation compared to land plants, the high diversity of different species and the vast range of molecules they can produce (Wijffels, Kruse & Hellingwerf, 2013; Al-Haj *et al.*, 2016).

Key metabolites within cyanobacteria (proteins, lipids, carbohydrates, and pigments) are useful for producing nutritional food with their biosynthesis affected by changes in light intensity and spectra (Muramatsu & Hihara, 2012; Ho & Bryant, 2019). When cyanobacteria become exposed to the extreme ends of the light spectrum, far-red and UV light, a stress response is triggered due to chlorophyll becoming photosensitised and the light-driven electron transport outpacing the rate of electron consumption during CO₂ fixation. Consequently, this exposure initially leads to the production of reactive oxygen species (ROS) (Latifi, Ruiz & Zhang, 2009). Under photosynthetic stress, cyanobacteria alter their gene expression to produce antioxidant metabolites to alleviate ROS produced and adapt to extreme light spectra (Latifi, Ruiz & Zhang, 2009; Ho & Bryant, 2019). Far-red light ($\lambda = 700\text{-}780\text{nm}$) can cause an increase in high-value compounds, such as the pigment phycoerythrocyanin, whilst reducing the synthesis of other key metabolites by a decrease in expression, limiting growth and energy production (Ho & Bryant, 2019).

Cyanobacteria and their importance

One of the oldest extant life forms on our planet are cyanobacteria which can adapt to varying light, temperature, salinity, and pH. This is due to morphological plasticity and heterogeneity within and between species (Gaysina, Saraf & Singh, 2019). Some cyanobacteria form heterocysts used for nitrogen fixation and contain specific chlorophyll for far-red light absorption, such as the species *Chlorogloeopsis fritschii*, a subsection V cyanobacterium isolated from paddy fields in India (Evans, Foulds & Carr, 1976). *Chlorogloeopsis* can undergo morphological changes, forming four different morphological types of cells due to the change

of light, carbon source and nitrate presence (Evans, Foulds & Carr, 1976). Heterocysts within *C. fritschii* form anaerobic cellular environments and deploy unique metabolic pathways for cellular functions, such as fatty acid synthesis. Furthermore, omega oils have been identified within *C. fritschii* (Kenyon, Rippka & Stanier, 1972). Thus, *C. fritschii* may be able to serve as an alternate source to produce fatty acids, specifically, omega oils, that is more efficient and environmentally beneficial than the current primary source which is aquaculture (Sánchez-Bayo et al., 2020).

Cyanobacteria have been shown to produce a range of different metabolites such as carotenoids, fatty acids, or proteins whilst absorbing CO₂ (Sánchez-Bayo et al., 2020). These products are all high-value compounds which can be extracted from the harvested biomass during downstream processing (Zahra et al., 2020). One interest of cyanobacteria is their production of polyunsaturated fatty acids (PUFAs), which have been shown to have many health benefits when increased in human and animal diets (Kumar et al., 2019). PUFAs produced by cyanobacteria as essential components include eicosapentaenoic acid (20:5; EPA) and docosahexaenoic acid (22:6; DHA) which are often used within supplements and infant formulas (Liu et al., 2021).

The autotrophic growth of cyanobacteria uses light and CO₂ within photosynthesis (Saha & Murray, 2018). During cultivation, the light source that the culture is exposed to can be altered by intensity and wavelength, which has been shown to cause variation within cellular metabolites, specifically photosynthetic compounds such as carotenoids, chlorophyll and phycobiliproteins (Khantoon et al., 2018). This approach might be an effective way to increase the yields of protein and other high-value compounds produced by cyanobacteria.

What are fatty acids and where are they found in cyanobacteria?

Fatty acids (FA), the key constituent of lipids, consist of a hydrocarbon chain with varying length and saturation that end in a carboxylic acid and account for 15% of cellular content within cyanobacteria (Jónasdóttir, 2019). As mono- (FFAs; free fatty acids) or polymers (lipids) within all organisms, fatty acids function in membrane formation, energy storage or exocytosis (Kaczmarzyk & Fulda, 2010; Los & Mironov, 2015).

Fatty acids are present in all forms of life and vary within the hydrocarbon chain. These variations include the number of carbons within the hydrocarbon chain, i.e., chain length, or the number of double bonds between two carbons within the chain, i.e., the degree of saturation. Each fatty acid terminates with a carboxyl acid (Δ) on one side and a methyl carbon (ω) on the other. Desaturation of fatty acids introduces a double bond to the hydrocarbon chain; the position can be annotated from the carboxyl terminus using Δ or from the distal end, the methyl group, using ω . When using either, the carbons are numbered starting at the terminal used in annotation. For oleic acid (18:1; 18 carbon atoms: 1 double bond) with a trans double bond in the middle of the hydrocarbon chain (C9=C10), it could be annotated as *trans- Δ^9* or *trans- ω^9* .

The carboxyl terminus binds to other molecules to produce fatty acid polymers, known as lipids, such as triacylglycerides, glycolipids and phospholipids. For phospholipids, two fatty acids bind by their carboxyl terminus to a glycerol molecule which is subsequently bound to a phosphate molecule. The fatty acids are hydrophobic whilst the phosphate head is hydrophilic; therefore, a lipid bilayer formation occurs. Fatty acid desaturation can increase membrane fluidity to aid organisms in adapting to a change in environment, such as low temperatures (Nalley O'Donnell & Litchman, 2018). Additionally, free fatty acids (FFAs) are often found within organisms, ranging from 6 to 24 carbon atoms long, and are excreted from cells (Kaczmarzyk & Fulda, 2010; Jónasdóttir, 2019). FFAs are also taken up by organisms or recycled from degraded polymers by the enzyme long-chain-fatty-acid CoA ligase, encoded by *aas*, which has been shown to increase fatty acid excretion when its function is impaired via gene knockout (Kaczmarzyk & Fulda, 2010).

The term "microalgae" encompass phyla within different kingdoms. Within cyanobacteria, 15% of the biomass is made of lipids, which is the lowest among the "microalgal phyla" e.g., Chlorophyta, Diatoma, Dinophyta, Haptophyta, Cryptophyta and Ochrophyta (Jónasdóttir, 2019). However, compared to a study investigating seven marine bacterial species, 15% lipid content is seen as relatively high, as lipid content was typically between 4-6% (Brown *et al.*, 1996). The variation in lipid profiles in bacteria occurs further within fatty acid profiles and has been used as a phylogenetic marker for the classification of organisms, including cyanobacteria (Los & Mironov, 2015). Essential fatty acids include α - and γ -linolenic-acid

(18:2 ω^6 and 18:3 ω^3 ; ALA and GLA), and additionally, their C20 derivatives, EPA (20:5 ω^3) and arachidonic acid (20:4 ω^6 ; ARA) have been found within cyanobacteria; *Microcystis aeruginosa* has these four essential fatty acids present, whilst *Chlorogloeopsis fritschii* has been shown to produce only ALA (Kenyon, Rippka & Stanier, 1972; Sharathchandra & Rajashehar, 2011). The desaturase genes present within cyanobacterial species are strongly conserved and used as a phylogenetic marker (Los & Mirinov, 2015). This conservation of fatty acid desaturation genes translates to the fatty acid composition of cyanobacterial species with specific desaturation within different cyanobacterial classes. First proposed by Kenyon, Rippka & Stanier (1972), the fatty acid composition for cyanobacterial species was used to classify the phylum into five distinct groups based on their fatty acid desaturation and has been updated further with the use of genetic data (Murata, Wada & Gombos, 1992; Los & Mirinov, 2015). *Chlorogloeopsis fritschii* is shown to contain only ALA and not GLA, so is placed within group 3 α (Kenyon, Rippka & Stanier, 1972). In comparison, *Synechocystis sp.* PCC 6714 is in group 3 γ due to only containing GLA, whilst *Synechocystis sp.* PCC 6803 is within group 4 due to containing ALA and GLA (Los & Mirinov, 2015).

How are fatty acids produced?

Fatty acid synthesis (FAS) differs in bacteria, including cyanobacteria, and plants, compared to animals, yeast, and filamentous fungi. Whilst animals, yeast and fungi have fatty acid synthase type I (FAS I), plants and bacteria use fatty acid synthase type II (FAS II). FAS I uses one multi-complex enzyme to produce fatty acids. In contrast, FAS II uses multiple single enzymes to complete each reaction to produce acyl-ACP, the precursor of fatty acids (Santos-Merino, Garcillán-Barcia & de la Cruz, 2018). FAS II uses acetyl-CoA to produce acyl-ACP, which is then cycled around the same pathway to add two further carbons each cycle, forming a hydrocarbon chain. This often stops when the molecule has 16 carbons attached, producing palmitic acid (16:0).

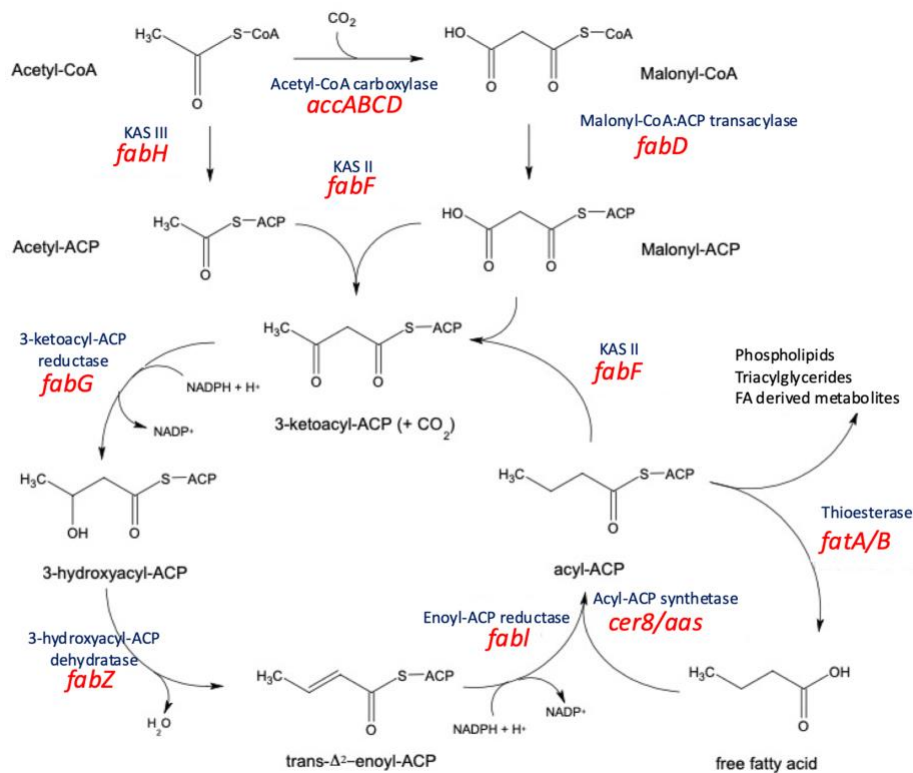


Figure 1. FAS II Fatty acid synthesis producing acyl-ACP within cyanobacteria. Metabolites are shown in black whilst enzymes and the genes encoding them are shown in blue and red, respectively. Adapted from Taylor (2012) by the addition of enzyme and associated gene names.

As seen in Figure 1, acetyl-CoA is formed from pyruvate produced from glycolysis and Coenzyme A (CoA). CoA is produced from the precursors, β-alanine and pantoate, and a series of phosphorylation and de-phosphorylation reactions (Taylor, 2012). The initial step of fatty acid biosynthesis is the carboxylation, via acetyl-CoA carboxylase (EC: 6.4.1.2), of acetyl CoA to form malonyl-CoA; both acetyl- and malonyl-CoA are used to elongate fatty acid chains (Jónasdóttir, 2019). Acetyl-CoA carboxylase is a multienzyme complex containing a biotin carboxyl carrier protein, a biotin carboxylase and the α and β subunits of acetyl-CoA carboxyltransferase. Malonyl- and acetyl-CoA are both turned into their -ACP counterparts by malonyl-CoA:ACP transacylase and 3-Ketoacyl-ACP-synthase III (KAS III), respectively. The two molecules formed are then condensed together by 3-ketoacyl-ACP-synthase II (KAS II) to produce 3-ketobutyryl-ACP and CO₂. Next, the NADPH-dependant ketoacyl-ACP reductase reduces its respective molecule to form 3-hydroxybutyryl-ACP before being dehydrated by

3-hydroxyacyl-ACP dehydrase to produce trans- Δ^2 -butenoyl-ACP. The latter molecule can then be reduced by enoyl-ACP reductase to produce butyryl-ACP (an acyl-ACP product). This product is then able to start the second cycle by being condensed with malonyl-ACP, forming a six-carbon 3-ketoacyl-ACP (3-ketohexanoyl-ACP) by KAS II. This cycle is then able to occur again, adding an additional two carbons from malonyl-ACP each cycle.

Fatty acid desaturation occurs to fatty acid-ACP substrates via fatty acid desaturases (DES) that are all known to be intrinsic membrane proteins within cyanobacteria (Sakamoto & Bryant, 1997). Two classes of polyunsaturated fatty acids (PUFAs) highly attractive in the nutraceutical industry are ω -3 and ω -6. Examples of ω -3 and ω -6 fatty acids that are attractive include ALA, EPA, DHA, linoleum acid (LA), ARA, gamma linoleum acid (GLA) and conjugated linoleum acid (CLA). These PUFAs are synthesised from ALA (C18:3 Δ 9,12,15) and LA (C18:2 Δ 9,15), respectively (Kumar *et al.*, 2019). The first stage in desaturation involves stearic acid (C18:0) being desaturated by Δ 9-Des into oleic acid (C18:1 Δ 9). Oleic acid is then able to be desaturated by Δ 12-DES to produce linoleic acid which in turn can be desaturated by Δ 15-DES to produce α -linolenic acid (Shanab, Hafez & Fouad, 2018).

How are fatty acids different in cyanobacteria?

Autotrophic organisms benefit from producing acetyl-CoA from the reductive pentose phosphate cycle or Calvin-Benson cycle (Liu, Sheng & Curtis III, 2011). Therefore, the main source of carbon for fatty acid synthesis comes from carbon fixation (Jónasdóttir, 2019; Liu, Sheng & Curtis III, 2011). The number of fatty acid synthases (FAS II) varies in bacteria, and the number of FAS II paralogs fluctuates depending on the species.

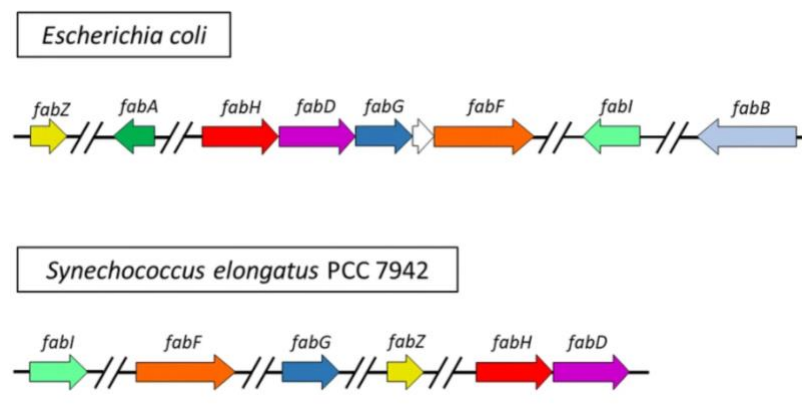


Figure 2. The comparison of *fab* genes (encoding for FASII) between *Synechococcus elongatus* and *Escherichia coli*. *E. coli* has two extra genes and a different operonic structure. This may be due to being heterotrophic or due to a greater need of carbon from heterotrophic growth. From Santos-Merino (2018).

For instance, two more FAS II genes are present for *E. coli*, *fabA* and *fabB*, compared to the cyanobacteria *Synechococcus elongatus* (Figure 2). In *E. coli*, the dehydration of 3-hydroxy-ACP to *trans*-2-enoyl-ACP involves *fabA* alongside *fabZ*, while the condensation of fatty-acyl-ACP with malonyl-ACP is carried out by *fabB* and *fabF* (Santos-Merino *et al.*, 2018). The increased complexity in FAS II genes in *E. coli* may be due to gene duplication or lateral gene flow driven by a greater need for acetyl-CoA. In any case, the presence of more FAS II paralogs does make *E. coli* more genetically complex than cyanobacteria and, therefore, less attractive for genome editing.

The genetic regulation of fatty acid synthesis in cyanobacteria appears simpler. For example, the genetic regulation of FAS II enzymes within *Synechococcus sp.* has shown to depend on *fabH*, and in the cyanobacteria *Synechocystis sp.*, the global regulator *LexA* has been shown to repress *fabD*, *fabH*, *fabF* and *fabG* genes (Kizawa *et al.*, 2017; Kuo & Khosia, 2014).

Unlike most other bacterial classes, some cyanobacteria can produce the PUFAs EPA and DHA through a different pathway than fatty acid desaturation (Okuyama *et al.*, 2006). The enzymes, known as polyunsaturated fatty acid synthases, are encoded by the *pfa* genes and the production of EPA involves the genes *pfaABCD* whilst the production of DHA additionally requires *pfaE* (Allen & Bartlett, 2002). These genes have multiple homologous domains to polyketide synthases (PKSs), so they are often referred to as polyketide-like synthases within

literature (Metz *et al.*, 2001; Okuyama *et al.*, 2006). The *pfa* genes from the proteobacteria *Shewanella baltica* have been expressed in *E. coli* for the recombinant production of EPA and DHA (Amiri-jami & Griffiths, 2010). Nevertheless, microorganisms (e.g. bacteria and microalgae) able to produce EPA and DHA via fatty acid elongation/desaturation and polyunsaturated fatty acid synthases (Okuyama *et al.*, 2006; Kannan, Rao & Nair, 2021) are probable more desirable for the biotechnological industry due to the nutraceutical value of these PUFAs.

How is fatty acid synthesis affected by far-red light?

Within plants, shading increases the amount of far-red exposure in proportion to white light. An increased amount of far-red received by *Arabidopsis* showed reduced expression of fatty acid desaturation whilst increasing the production of fully or partially (1 or 2 double bonds) saturated acids (Arico *et al.*, 2019).

Cyanobacteria grow in niche environments where visible light is less abundant than far-red light, such as deep within caves (Behrendt *et al.*, 2020). They adapt their photosynthetic apparatus by increasing the concentration of chlorophyll *d* and *f* and other far-red light-absorbing compounds such as phycoerythrin. This adaptation is known as far-red light photoacclimation (FaRLip) (Chen *et al.*, 2010). FaRLip alters the physiology of different cyanobacterial species. The greatest change occurs within the photosystem I & II (PS I & II) and the phycobilisome (PBS) which are found within the thylakoid membrane. The subunits of these photosynthetic complexes are replaced with the products of a 21-gene cluster that are specifically expressed within far-red light. The products include a knotless red/far-red phytochrome and two response regulators and produce photosynthetic complexes which are complementary to the incident radiation between 700-750 nm (Gan *et al.*, 2014).

In some cases, FaRLIP also affects cyanobacterial fatty acid pathways and yields. Transcriptional profiling of *C. fritschii* PCC 9212 cultivated in far-red light showed a significant decrease in the expression of genes involved in fatty acid synthesis (Ho & Bryant, 2019). Despite what is known about fatty acid metabolism and how it is affected by far-red light, this information is not comprehensive, particularly within specific fatty acid pathways. An increase or decrease of expression of certain fatty acid synthetase or desaturase genes could cause a flux of specific desired fatty acids or improve fatty acid excretion. This has relevance

from a biotechnological perspective where the increased production of fatty acid is sought for use in nutraceuticals and biofuels (Al-Haj *et al.*, 2016). Furthermore, no change in the fatty acid pathways may be desired to produce an increase in a different high-value compound, such as phycoerythrin pigments, so no negative effect occurs within fatty acid-related functions such as membrane formation or energy storage (MacGregor-Chatwin *et al.*, 2022).

Aims and objectives of this study

This desk-based study aims to investigate the genes associated with fatty acid pathways within the cyanobacteria *C. fritschii* PPC 6912 when exposed to far-red light compared to white light control. This study uses the RNA-seq data from a previous experiment, conducted by a different research team and not the work of this author, where *C. fritschii* PCC 6912 is cultivated under white and far-red light (Llewellyn *et al.*, 2020). The methodology to produce the RNA samples to produce the RNA-seq data set is listed below within the methodology for reproducibility purposes. The raw transcriptomic data obtained from Llewellyn *et al.* (2020) were analysed to improve quality before being aligned and quantified. The quantified transcripts were then investigated to identify differentially expressed genes between the white and far-red samples. Genes associated with fatty acid pathways were identified within the reference genome used to obtain a subset of the transcriptome, allowing for the identification of the genetic expression of specific fatty acid genes within the fatty acid synthesis pathway under white and far-red light.

Methodology

Chlorogloeopsis fritschii cultures, light treatments and RNA extraction and sequencing

As mentioned within the Aims and objectives of the study, the experiment from set-up to obtaining raw-mRNA transcripts was not conducted in this study but instead by Llewellyn *et al.* (2020). The experiment that produced the RNA-seq datasets deposited on NCBI (Bioproject: PRJNA545395) was originally used by Llewellyn *et al.* (2020) to investigate mycosporine-like aromatic amino acids. In summary, three 1-L conical flasks containing 800 mL of *Chlorogloeopsis fritschii* PCC 6912 culture with an initial optical density (OD) of 0.1-0.2 Absorbance Units (AU) at 750nm were cultivated under white LED light (intensity: $100 \mu\text{mol photons m}^{-2} \text{s}^{-2}$) for six days, reaching an OD₇₅₀ of 0.4 AU (Figure 3). 3 X 50 mL aliquots were taken from each flask, which were then used to provide the control sample (white light). Next, the 1 L flasks were cultivated under far-red LED light (710nm, $\sim 18 \mu\text{mol photons m}^{-2} \text{s}^{-2}$) for another 24 h. After the 24 h incubation time, another 50 mL aliquot was taken from each flask to provide for the far-red samples.

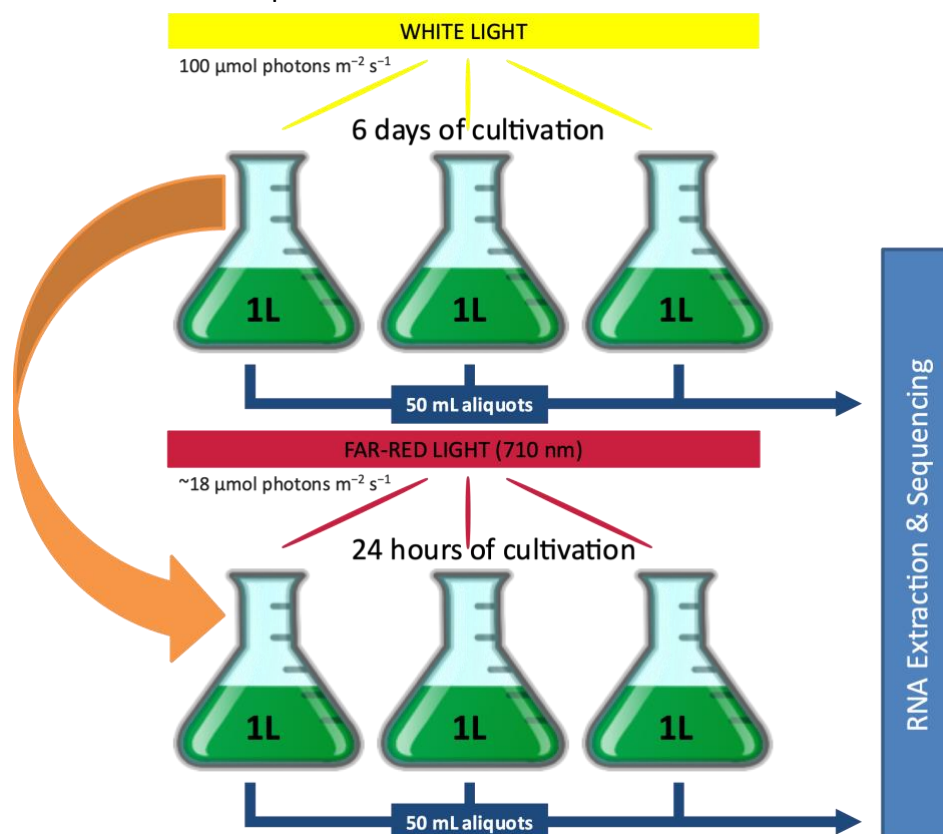


Figure 3. The cultivation of 3 x 800 mL *C. fritschii* cultures firstly in white light for six days then moved to far-red light for 24 hours. 50 mL aliquots were taken at the end of each light cultivation.

The RNA was then extracted from the 50-mL aliquots by centrifugation for 15 min at 4 °C at 3500 rpm, and then concentrated further at 4 °C at 5000 rpm. After supernatant removal, samples were weighed to calculate wet weights before being flash-frozen in liquid nitrogen. Pellets were resuspended in 1mL in Trizol reagent (Thermo Fisher Scientific) and homogenised using 0.5 mm glass beads (VK05) at 6500rpm for 2 x 20 s with a 10 s break. After 5 min, the sample was extracted using 0.2 mL chloroform followed by centrifugation at 12,000xg for 15 min at 4 °C. The upper aqueous phase was mixed with an equal volume of 70% ethanol and applied to a PureLink RNA Mini Kit spin cartridge (Thermo Fisher Scientific) following the manufacturer's instructions with an extra wash step before drying and eluting the extracted RNA in 100 µL RNase-free H₂O. Residual DNA contamination was removed using TURBO DNase (Thermo Fisher Scientific) followed by enzyme removal from the inactivation reagent supplied in TURBO DNA-free Kit (Thermo Fisher Scientific). Total RNA was used as starting material for the generation of sequence-ready libraries. After brief rRNA removal and to ensure an even read distribution for all RNA samples, the final RNA libraries were adjusted to a concentration of 4 nM prior to pooling. A final library concentration of 20 pM was used to sequence libraries on MiSeq platform at the Swansea University Sequencing Facility, generating a total of 99,448,410 reads using multiple V3 2x75 bp PE sequencing runs (Llewellyn *et al.*, 2020).

Transcript quality assessment

Samples were accessed using the SRA Toolkit (NCBI) using the accession files (SRR11810824-SRR11810829) for the far-red and white-light experiments in BioProject PRJNA545395. After all transcript samples were downloaded, a quality assessment and improvement were undertaken with a series of bioinformatic tools running on Ubuntu (Figure 4).

Firstly, FastQC (version 0.11.9; Andrew, 2010) was completed on each FASTQ file, representing each sample, to provide an overview of the quality control of the raw RNA-seq data based on Phred scores and GC content. After the initial FASTQC quality check, SortMeRNA (version 4.3.4; kopylova, Noé & Touzet, 2012) was used to remove any other RNAs other than mRNA. This was completed by first creating an index file with RNA reference strands from archaea, bacteria, and eukaryotic organisms (i.e., 16S and 23S for bacteria and

archaea and 16S and 28S for eukaryotic organisms). Flags used included `-out 2` and `-fastx` to keep the forward and reverse strands separated as required by other tools down the pipeline.

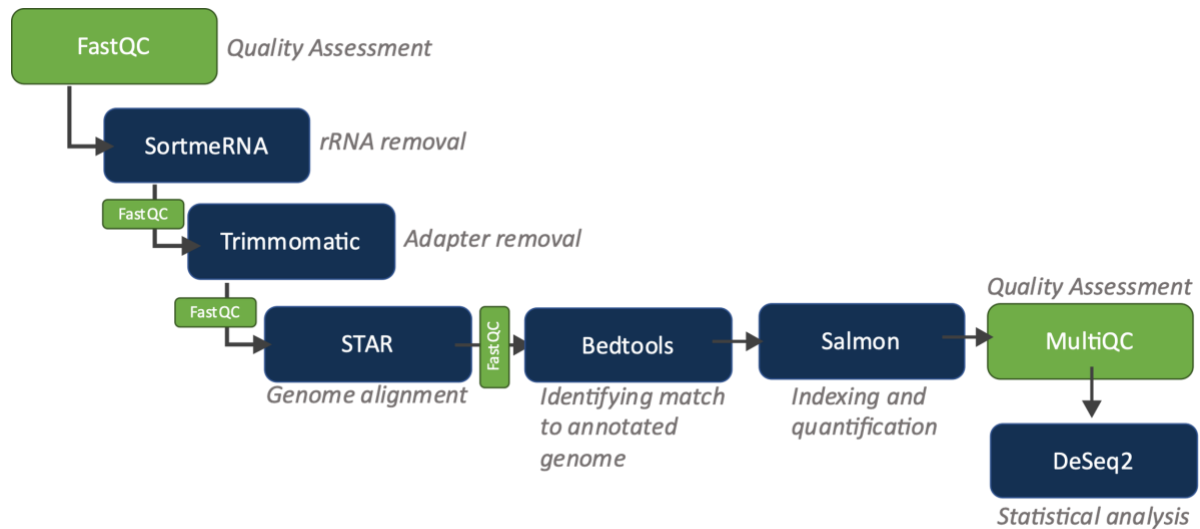


Figure 4. The bioinformatic pipeline to enhance, align and quantify RNA-seq transcripts.

The output files were then analysed with FASTQC before using the paired-end (PE) version of Trimmomatic (version 0.39; Bolger, Lohse & Usadel, 2014) to remove poorly sequenced reads. A sliding length: mean minimum of 5:5 and a minimum length of 50 was used as a threshold to discard poor quality reads. In addition, Trimmomatic was used to trim adaptor sequences and poor-quality bases to improve mappability as the quality of the reads decreases towards their 3' end. The output files were then analysed one more time by FASTQC to see if quality had improved by evaluating the Phred score and mean length of the reads.

Genome alignment and transcriptome assembly

The genome reference file, GCA_000317285.1, sequenced and annotated by Dagan *et al.* (2013), was obtained from NCBI and then converted from .gff to .gtf file format using AGAT (version 0.6.0; Dainat *et al.*, 2021). STAR (v2.7.10; Dobin *et al.*, 2013) was the genome alignment software used. Therefore, the index file for transcript alignment was made from the genome files before the alignment could occur. After indexing, each sample was aligned to the reference genome and the output was set to a BAM "SortedByCoordinate" to make sure the reads stayed within order. The BAM output files were then one more time assessed by FASTQC. Overlap of transcripts with the genome was then investigated using Bedtools

(v2.27.1; Quinlan & Hall, 2010), which found there was overlap and the alignment was successful. Therefore, transcript quantification could take place.

Transcript quantification

Salmon (v1.4.0; Patro *et al.*, 2017) was the biocomputational software used to quantify the transcripts. To create an index file for the salmon quantification, a “gentrome” reference was produced. This was produced by extracting a transcript file from the genome and annotation file and collating it with the genome reference using GFF Utilities (v0.11.7; Pertea & Pertea, 2020). A decoy file was then produced of the gene names. A salmon index was then created using the gentrome and decoy files. For Salmon, BAM files cannot be used so Samtools (v1.13; Danecek *et al.*, 2021) and Bedtools were used to order and convert BAM files from the STAR alignment into FASTQ files. To assure quality, FASTQC was run on the newly produced FASTQ files. Salmon was then used to quantify the transcripts where the flags for GC, positional and sequence bias are used along with a Gibbs Sampler of 100 iterations. After quantification was completed, the output files from all the software used, including FASTQC outputs, were run through MultiQC (v1.11; Ewels *et al.*, 2016), producing a MultiQC file for each sample. The salmon quantification was then used within R Studio for further analysis (R Core Team, 2020).

Quantification of gene expression and differential expression analysis of fatty acids

The Salmon quantification outputs were imported into R Studio 4.0.3 using tximport (v1.18.0; Sonesson, Love & Robinson, 2015). Salmon counts were then used to create a DESeq dataset with the package DESeq2 (1.30.1; Love, Huber & Anders, 2014). A variance stabilising transformation (VST) was carried out on the data within the DESeq2 package because the variance in higher expressed genes is likely to be greater. Thus, highly expressed genes may have a greater influence on the DESeq analysis downstream. To analyse the variance of gene expression, the ranked mean expression of each gene and its standard deviation was calculated using EdgeR and plotted using the package, vsn (v3.32.1, McCarthy, Chen & Smyth, 2012; v3.58.0, Huber *et al.*, 2002).

The genes of interest, such as fatty acid synthases and desaturases, were identified by searching enzymatic names from fatty acid synthesis pathways within the protein list of the

annotated genome from NCBI using the BioCyc database which maps and determines genes within pathways using MetaCyc (v23.5; Caspi *et al.*, 2020) and the Pathway Tools software from PathoLogic (v19.0; Karp *et al.*, 2016). This produced a list of 33 genes related to fatty acids, which were then used to make a table describing their locus tag, gene name, enzyme number, EC number, Protein size (kDa) and reference to the protein size.

To assess the variation within samples as well as the treatment and control, two principal component analyses (PCA) were carried out using the *prcomp* within the stats package from R (v4.0.3; R Core Team, 2020). The first PCA was calculated with all genes present within the quantified transcriptome whilst the second uses the 33 identified fatty acid genes. These PCA plots are mainly for exploratory analysis to visually identify differences between control and treatment samples and provide reasoning for further downstream analysis.

The DESeq results were then extracted with the cutoffs, $lfc = 0.5$ and $padj = 0.01$, suggested by Schurch *et al.* (2016), following the workflow proposed by Love *et al.* (2016). The results extracted were then subset to only contain the genes of interest, which were separated into four categories: fatty acid synthesis (FAS), oleic acid pathway, fatty acid desaturases (*fad*) and polyunsaturated fatty acid desaturases (*pfa*). The z scores, produced from the log2fold divided by the standard error, were then used to create heatmaps for each group and are presented alongside the log2fold value and the Benjamini-Hochberg adjusted p-value (Love, Huber & Anders, 2014). The heatmaps were produced by the Rpackage, *gplots* (v3.1.3; Warnes *et al.*, 2022).

Results

Transcript identification and quantification of fatty acid-related genes

The raw transcripts, from Llewellyn *et al.* (2020), were downloaded from the NCBI database (Table 1) and underwent non-mRNA removal, adapter trimming, genome alignment and transcript quantification by the respective bioinformatic tools mentioned within the Methodology. Further detail about the bioinformatics process can be seen within Appendix A.

Table 1. The RNA-seq sample names, their sample code and link to the NCBI database.

Sample	Sample ID	NCBI Accession Code
White 1	SRR11010824	SRX8362192
White 2	SRR11010825	SRX8362191
White 3	SRR11010826	SRX8362190
Far red 1	SRR11010827	SRX8362189
Far red 2	SRR11010828	SRX8362188
Far red 3	SRR11010829	SRX8362187

As the quality control showed, the mRNA samples were of high quality, and the transcript mappability was high for all samples. The aligned/mapped transcripts were quantified through Salmon using quasi-mapping in two phases (Patro *et al.*, 2017). This step allows for the estimation of transcript abundance and, therefore, for the quantification of gene expression, which is assessed statistically with R studio. To allow analysis to occur specifically of fatty acid-related genes, enzymatic nomenclature was searched within the protein list of the annotated genome. Additionally, the BioCyc genome database collection was used to assign genes to their respective metabolic pathways (Caspi *et al.*, 2020; Karp *et al.*, 2016). This resulted in 33 genes being identified within the annotated genome (Table 2).

The annotated genome used was produced by Dagan *et al.* (2013). However, only 22 of the 33 genes are annotated to gene level, known as completely annotated. The incompletely annotated genes have been annotated to the gene family level, for example, there are ten “fatty acid desaturase” genes. However, their specific role within fatty acid desaturation cannot be identified. The genes that are completely annotated have been investigated by identifying previously researched orthologs within other cyanobacteria and bacteria to

identify their protein size (kDa), which is listed along the species it is distinguished in and the relevant reference used (Table 1).

Table 2. Genes selected for analysis. The 33 genes were identified within the reference genome and had transcripts aligned. The table contains details about the gene selected from the genome reference as well as further information about the enzyme the gene encodes for within previous literature.

Contig Accession ID	Gene ID / Locas Tag	Gene name	Gene length (bp)	Protein Name	EC number	protein size (kDa)	reference	species used in ref
NZ_AJLN01000125.1	gene-UYC_RS0129155	<i>fabF</i>	416	beta-ketoacyl-ACP synthase II	2.3.1.179	43	Moche <i>et al.</i> , 2001	<i>Synechocystis sp.</i>
NZ_AJLN01000116.1	gene-UYC_RS0126220	<i>fabI</i>	258	enoyl-ACP reductase	1.2.1.9	28	Liu <i>et al.</i> , 2011	<i>Burkholderia pseudomallei</i>
NZ_AJLN01000110.1	gene-UYC_RS0123455	<i>fabD</i>	292	ACP S-malonyltransferase	2.3.1.39	32	Serre <i>et al.</i> , 1995	<i>Escherichia coli</i>
NZ_AJLN01000081.1	gene-UYC_RS0114860	<i>fabZ</i>	176	3-hydroxyacyl-ACP dehydratase	4.2.1.59	112	Kimber <i>et al.</i> , 2004	<i>Pseudomonas aeruginosa</i>
NZ_AJLN01000110.1	gene-UYC_RS0123460	<i>fabH</i>	330	ketoacyl-ACP synthase III	2.3.1.180	~60	Khandekar <i>et al.</i> , 2001	<i>Streptococcus pneumoniae</i>
NZ_AJLN01000116.1	gene-UYC_RS0124730	<i>accA</i>	326	acetyl-CoA carboxylase carboxyltransferase subunit alpha	6.4.1.2	35	Ye <i>et al.</i> , 2020	<i>Escherichia coli</i>
NZ_AJLN01000040.1	gene-UYC_RS0103785	<i>accB</i>	170	acetyl-CoA carboxylase biotin carboxyl carrier protein	6.4.1.2	16.7	Choi-Rhee & Cronan, 2003	<i>Escherichia coli</i>

Contig Accession ID	Gene ID / Locas Tag	Gene name	Gene length (bp)	Protein Name	EC number	protein size (kDa)	reference	species used in ref
NZ_AJLN01000074.1	gene-UYC_RS0113435	<i>accC</i>	447	acetyl-CoA carboxylase biotin carboxylase subunit	6.3.4.14	49.4	Choi-Rhee & Cronan, 2003	<i>Escherichia coli</i>
NZ_AJLN01000109.1	gene-UYC_RS0123160	<i>fabG</i> (1)	240	beta-ketoacyl-ACP reductase	1.1.1.100	48	Cross <i>et al.</i> , 2022	<i>Acinetobacter baumannii</i>
NZ_AJLN01000037.1	gene-UYC_RS0102950	<i>fabG</i> (2)	232	3-oxoacyl-ACP reductase	1.1.1.100	48	Cross <i>et al.</i> , 2021	<i>Acinetobacter baumannii</i>
NZ_AJLN01000037.2	gene-UYC_RS0102740	<i>fabG</i> (3)	254	3-oxoacyl-ACP reductase	1.1.1.100	48	Cross <i>et al.</i> , 2022	<i>Acinetobacter baumannii</i>
NZ_AJLN01000109.1	gene-UYC_RS0123225	<i>fabG</i> (4)	242	beta-ketoacyl-ACP reductase	1.1.1.100	48	Cross <i>et al.</i> , 2021	<i>Acinetobacter baumannii</i>
NZ_AJLN01000145.1	gene-UYC_RS0132090		658	long-chain fatty acid--CoA ligase	EC 6.2.1.3			
NZ_AJLN01000116.1	gene-UYC_RS0123270		347	fatty acid desaturase (1)	1.14.19.-			
NZ_AJLN01000015.1	gene-UYC_RS0100770		270	fatty acid desaturase (2)	1.14.19.-			
NZ_AJLN01000116.3	gene-UYC_RS0125315		261	fatty acid desaturase (3)	1.14.19.-			

N/A : lack of detailed annotation

Contig Accession ID	Gene ID / Locas Tag	Gene name	Gene length (bp)	Protein Name	EC number	protein size (kDa)	reference	species used in ref
NZ_AJLN01000116.1	gene-UYC_RS0124270		347	fatty acid desaturase (4)	1.14.19.-	N/A : lack of detailed annotation		
NZ_AJLN01000134.1	gene-UYC_RS0124130		299	fatty acid desaturase (5)	1.14.19.-			
NZ_AJLN01000116.2	gene-UYC_RS0129390		320	fatty acid desaturase (6)	1.14.19.-			
NZ_AJLN01000083.1	gene-UYC_RS0115415		355	fatty acid desaturase (7)	1.14.19.-			
NZ_AJLN01000065.1	gene-UYC_RS0111135		320	fatty acid desaturase (8)	1.14.19.-			
NZ_AJLN01000049.1	gene-UYC_RS0106555		270	fatty acid desaturase (9)	1.14.19.-			
NZ_AJLN01000109.1	gene-UYC_RS0115420		357	fatty acid desaturase (10)	1.14.19.-			
NZ_AJLN01000116.1	gene-UYC_RS0124170	<i>plsY</i>	227	glycerol-3-phosphate 1-O-acyltransferase	2.3.1.15	23	Lu <i>et al.</i> , 2007	<i>Streptococcus pneumoniae</i>
NZ_AJLN01000110.1	gene-UYC_RS0123425	<i>plsC (1)</i>	212	1-acyl-sn-glycerol-3-phosphate acyltransferase	2.3.1.52	27.5	Coleman, 1993	<i>Escherichia coli</i>

Contig Accession ID	Gene ID / Locas Tag	Gene name	Gene length (bp)	Protein Name	EC number	protein size (kDa)	reference	species used in ref
NZ_AJLN01000075.1	gene-UYC_RS0113660	<i>plsC</i> (2)	318	1-acyl-sn-glycerol-3-phosphate acyltransferase	2.3.1.54	27.5	Coleman, 1995	<i>Escherichia coli</i>
NZ_AJLN01000098.1	gene-UYC_RS0118885	<i>plsC</i> (3)	472	1-acyl-sn-glycerol-3-phosphate acyltransferase	2.3.1.53	27.5	Coleman, 1994	<i>Escherichia coli</i>
NZ_AJLN01000116.1	gene-UYC_RS0126710	<i>plsC</i> (4)	240	1-acyl-sn-glycerol-3-phosphate acyltransferase	2.3.1.51	27.5	Coleman, 1992	<i>Escherichia coli</i>
NZ_AJLN01000083.1	gene-UYC_RS0115425	<i>desC</i>	271	sn-1 stearyl-lipid-9-desaturase	1.14.1.9.28	37.2	Murata & Wada, 1995	<i>Syechocystis sp. PCC 6803</i>
NZ_AJLN01000051.1	gene-UYC_RS0108330	<i>pfaD</i>	539	polyunsaturated fatty acid/polyketide biosynthesis protein (1)	2.3.1	59	Metz <i>et al.</i> , 2001	<i>Shewanella sp.</i>
NZ_AJLN01000080.1	gene-UYC_RS0114680	<i>pfaD</i>	557	polyunsaturated fatty acid/polyketide biosynthesis protein (2)	2.3.2	59	Metz <i>et al.</i> , 2002	<i>Shewanella sp.</i>
NZ_AJLN01000097.1	gene-UYC_RS0118545	<i>pfaD</i>	556	polyunsaturated fatty acid/polyketide biosynthesis protein (3)	2.3.3	59	Metz <i>et al.</i> , 2003	<i>Shewanella sp.</i>
NZ_AJLN01000037.1	gene-UYC_RS0102830	<i>pfaD</i>	549	polyunsaturated fatty acid/polyketide biosynthesis protein (4)	2.3.4	59	Metz <i>et al.</i> , 2004	<i>Shewanella sp.</i>

Differential expression of genes involved in fatty acid synthesis and desaturation

Principal Component Analysis of the complete transcriptome and fatty acid-related genes

Before exploratory analysis of multidimensional data, homoscedastic data are required to prevent large scales, such as expression levels, influencing variance. RNA transcript data is known to have a greater variance at higher expressed genes meaning that the data is often heteroscedastic. Therefore, it is important to determine this before multidimensional analysis such as principal component analysis (PCA) occurs.

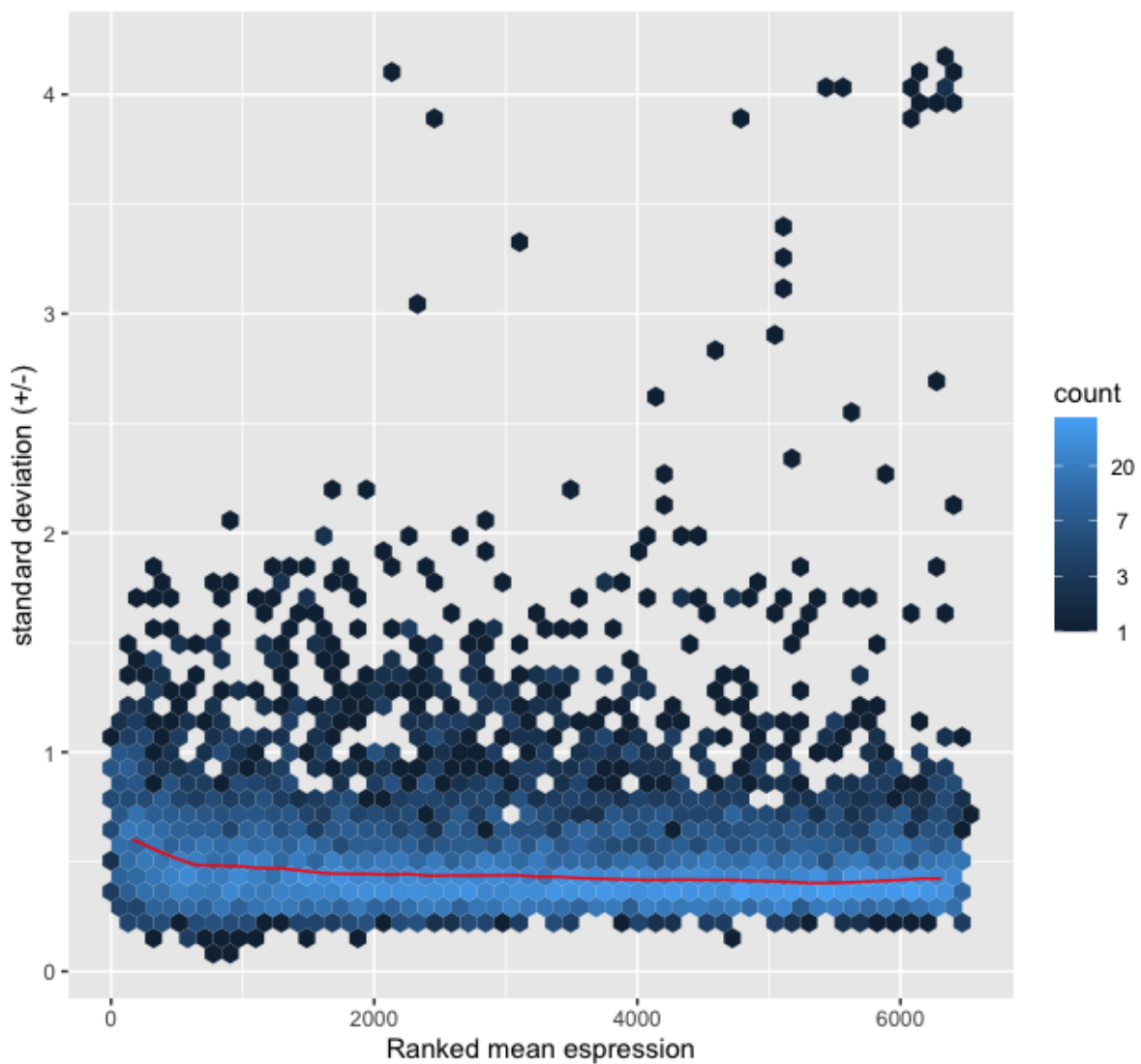


Figure 5. The variance of expression for each gene within the transcriptome between each sample before VST. Greater variance can be seen in higher expressed genes whilst most genes are densely distributed around the mean standard deviation (red line).

As expected, there was a greater variance in genetic expression between the samples in genes that were more highly expressed (Figure 5). Before variance stabilising

transformation (VST), the data is heteroscedastic, and variance is not independent of the gene expression.

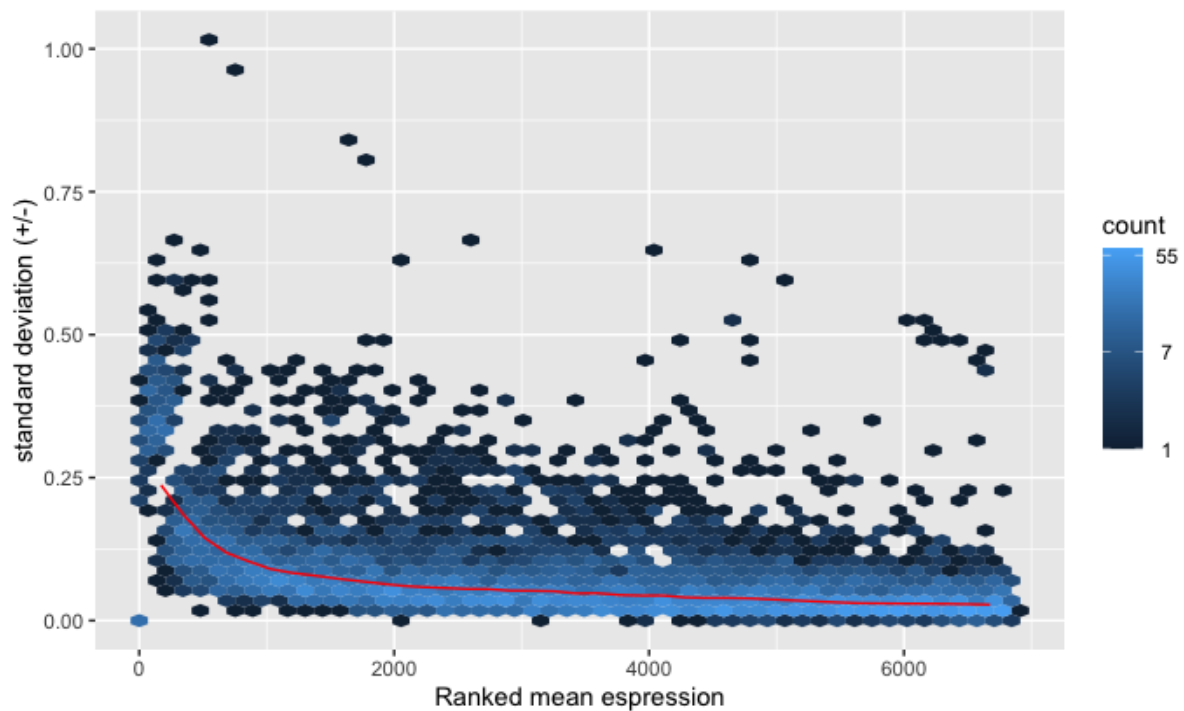


Figure 6. The expression variance for each gene within the transcriptome between each sample after VST. The density of genes around the mean (red line) has greatly increased, and the variation in standard deviation (left axis) at all expression levels has decreased.

VST can prevent an overreliance on highly or lowly expressed genes by normalising the library factor size of each gene and sample, causing the data to become more homoscedastic. This can be seen in Figure 6. As the standard deviation between samples has decreased and genes are more densely populated within the plot, showing less variation and greater independence of variation to expression levels. VST allows for a more reliable PCA to occur on the data.

Multidimensional data can be analysed using a PCA, which determines how the two highest factors affect variance. The first variance, Principal Component 1 (PC1), is the factor that causes the greatest variance in the data, whilst the second variance, Principal Component 2 (PC2), shows the variance determined by the second factor after PC1 is accounted for. As seen in Figure 7, the variance accounted for by PC1 and PC2 totals 98% of all variances within the data, meaning that all other factors share another 2% of the variance.

To understand the variation between samples and between far-red and white light, a PCA was completed on all genes identified and quantified and on the 33-fatty acid-related genes of interest (Figure 7 and 8, respectively). This is to explore the data before statistical analysis through DESeq2 visually.

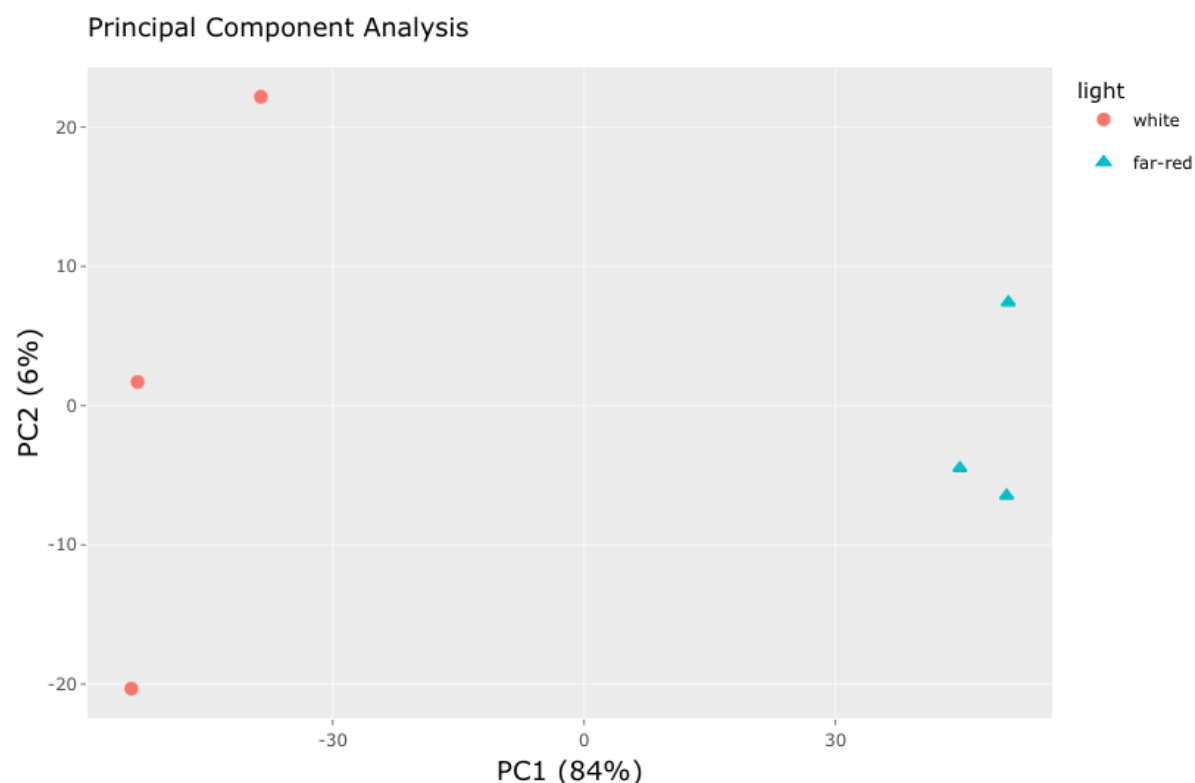


Figure 7. The Principal Component Analysis (PCA) of the quantified transcriptome before DESeq2 analysis to investigate visual differences between control and treatment.

All six samples can be seen in Figure 7, where the three cultures of *C. fritschii* grown under white light are on the left, and the three cultures grown under far-red light are on the right. The horizontal separation along the axis is the first principal component (PC1), accounting for 96% of the variance between the six samples. The idea that the three replicates are close together and the conditionals are separated horizontally suggests that PC1 is the effect of cultivating the cyanobacteria under far-red compared to white light. Thus, visually showing a difference in genetic expression when *C. fritschii* is cultivated under far-red. Figure 7 investigates the whole transcriptome and doesn't specifically investigate the effect on fatty acid-related genes. Therefore, a PCA was carried out on the 33 genes of interest additionally (Figure 8).

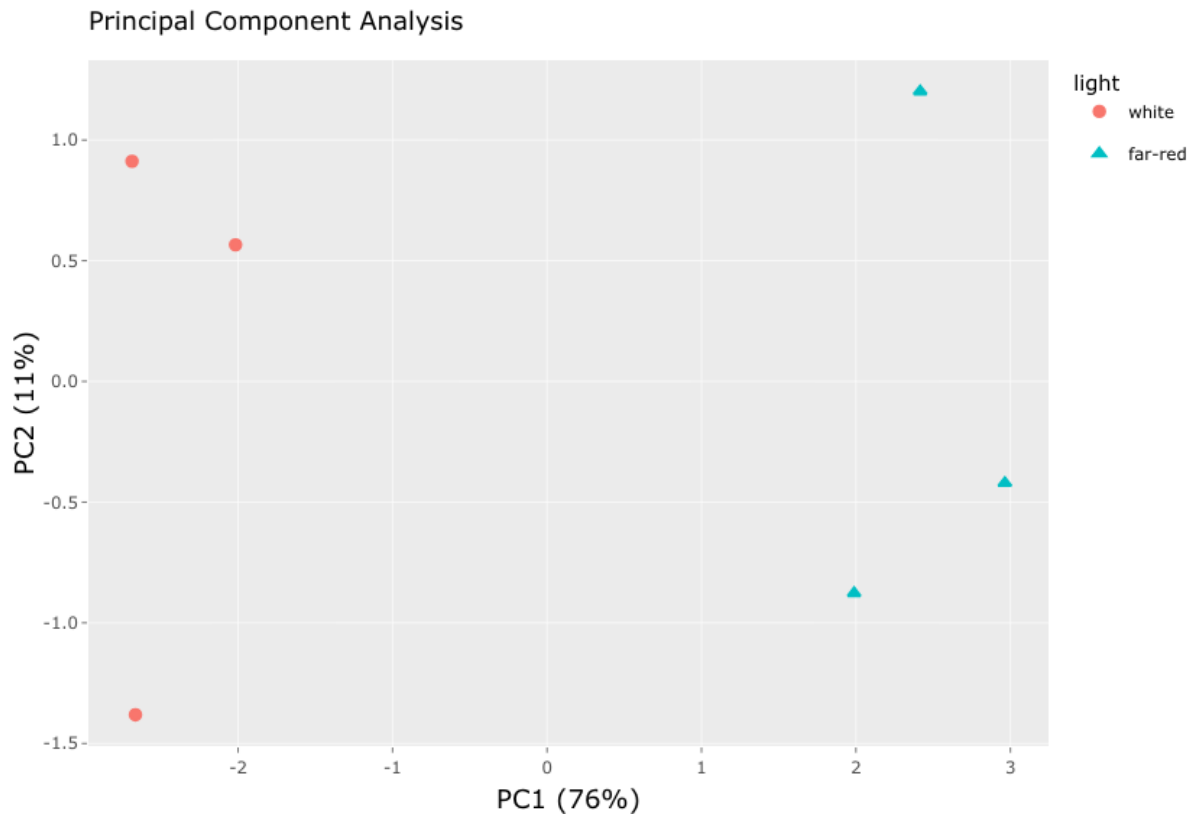


Figure 8. The PCA plot of the six samples only using the 33 genes associated with fatty acid synthesis before DESeq2 analysis to investigate visual differences between control and treatment.

In Figure 8, PC1 explains 76% of the variation within data compared to PC2, which explains 11% of the variation. Compared to Figure 7, PC1 has dropped by 8%, and PC2 has increased by 5%. This shows that when investigating only the fatty acid-related genes, the effect of PC1 decreases, and there is less effect of light on the expression of fatty acid-related genes. An additional comparison to Figure 7 is that the far-red light samples are less clustered together when looking at only fatty acid-related genes. This indicates more variability among samples when only the expression of fatty acid-related genes is examined.

PCA tests can reduce the dimensionality of large datasets whilst preserving variance, allowing analysis to occur on how similar two datasets are (Jolliffe & Cadima, 2016). Both PCA plots, investigating the whole quantified transcriptome and fatty acid-related genes, show significant variation between treatments. However, when all genes within the transcriptome are investigated, there is a higher variation explained by PC1, i.e., the type of light, than in the PCA including only fatty acid-related genes. Further statistical analysis could occur within the PCA analysis. However, this is often not required within bioinformatic investigations, unlike

within ecological investigations, due to it PCA plots only being used as “checkpoints” that there is a visual difference between samples before undergoing the differential expression analysis using bioinformatics software.

Differential expression of fatty acid-related genes

To further investigate whether far-red affects the expression of the investigated fatty acid-related genes, their expression data were processed using DESeq2, which tests for differential expression within a transcriptome. This package fits a glm model following a negative binomial distribution for each gene using the quantified data from Salmon. DESeq2 incorporates sequencing depth for each sample and normalises the expression from each gene using dispersion in expression within samples for each gene and additionally looks at dispersion estimates of genes with similar expression strengths to produce log-fold changes using empirical Bayes shrinkage (Love Huber & Anders, 2014).

To increase reliability of results, Schurch *et al.* (2016) suggestion of a log-fold change (LFC) threshold of 0 ± 0.5 and a false discovery rate (p-value) of 0.01 was used on the DESeq results. From implementing this filter to the whole transcriptome, the change in light conditions (from white to far-red) significantly altered the expression of 974 genes. Up-regulated genes changed from 1658 genes to 323 genes whilst the significantly down-regulated changed from 1680 to 651 (Figure 9). Most genes within the transcriptome, however, were not differentially expressed.

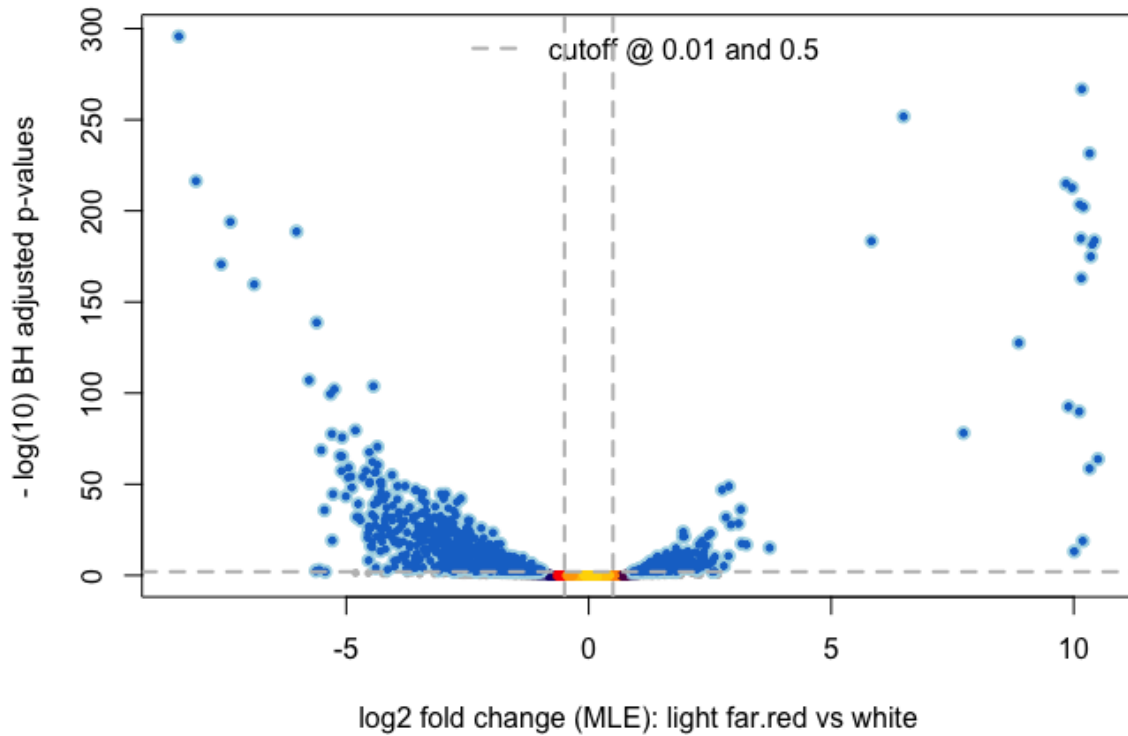


Figure 9. The differential expression cut-off using LFC and FDR of 0.5 and 0.01, respectively, shows a total of 974 genes that are differentially expressed under far-red light presented in a volcano plot.

Using the DESeq2 output, the fatty acid-related genes were then subset to produce heatmaps using the Standard Value (Z Value), which is produced by the Log2fold value divided by the standard error. Alongside, the Log2fold is presented with the Benjamini-Hochberg adjusted p-value (BH p-value). As stated above, the genes are only seen as significantly expressed with a Log2fold of 0 ± 0.5 and a p value < 0.01 following the recommendations by Schurch *et al.* (2016) for datasets with only three replicates.

The first set of fatty acid genes investigated are genes within fatty acid synthesis (fatty acid synthases or FAS) (Figure 10). Within the 12 FAS genes present in the annotated genome, none were found to be differentially expressed. This is due to their p values being >0.01 before the results from the DESeq2 cut-offs were set. Furthermore, BH p-values of 1.0 were produced when the Log2fold was smaller than its standard error.

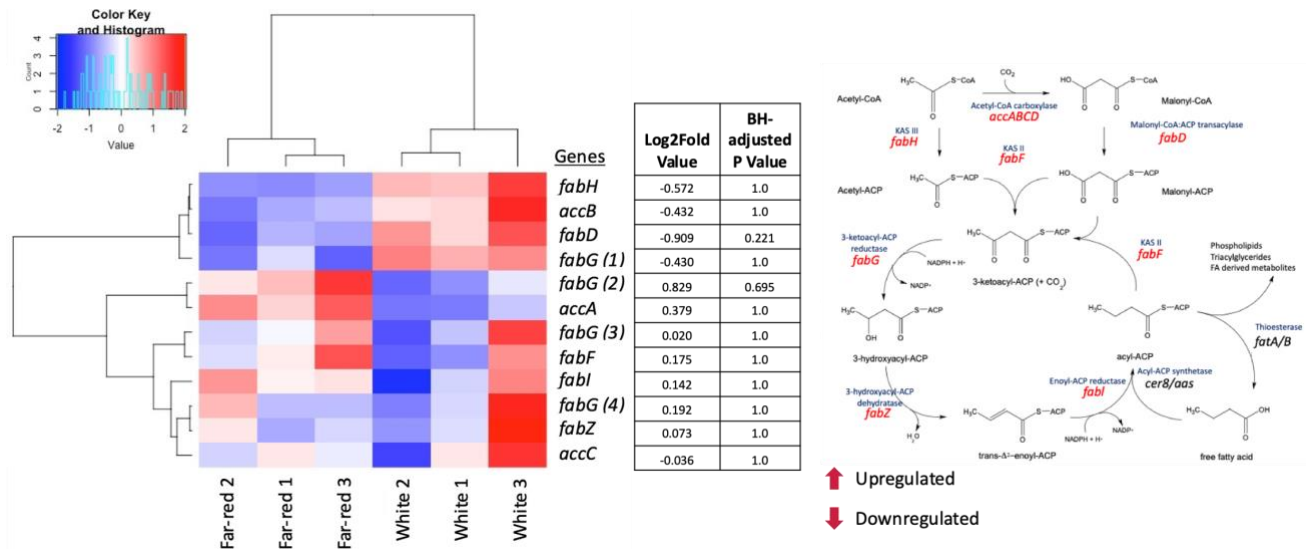


Figure 10. The heatmap produced with calculated Z values (left) for the fatty acid synthesis genes present within the annotated genome alongside the Log2fold change and BH-adjusted p-values (middle). Additionally, the biochemical pathway of fatty acid synthesis (right) is presented, showing the genes in red and the enzymes they encode in blue.

The Z values (heatmap in Figure 10) for the genes associated with fatty acid synthesis indicate that selected genes might appear differentially expressed. However, the observed differences are not statistically significant. Therefore, it is concluded that any variation present is due to experimental variation and not a change in light cultivation.

Six genes of interest, that were annotated completely (Table 2), encode the enzymes for the pathway to oleic acid (18:1). The genes 4x *plsC*, *plsY* and *desC* encode the enzymes glycerol-3-phosphate 1-O-acyltransferase, 1-acyl-sn-glycerol-3-phosphate acyltransferase and sn-1 stearoyl-lipid-9-desaturase, respectively.

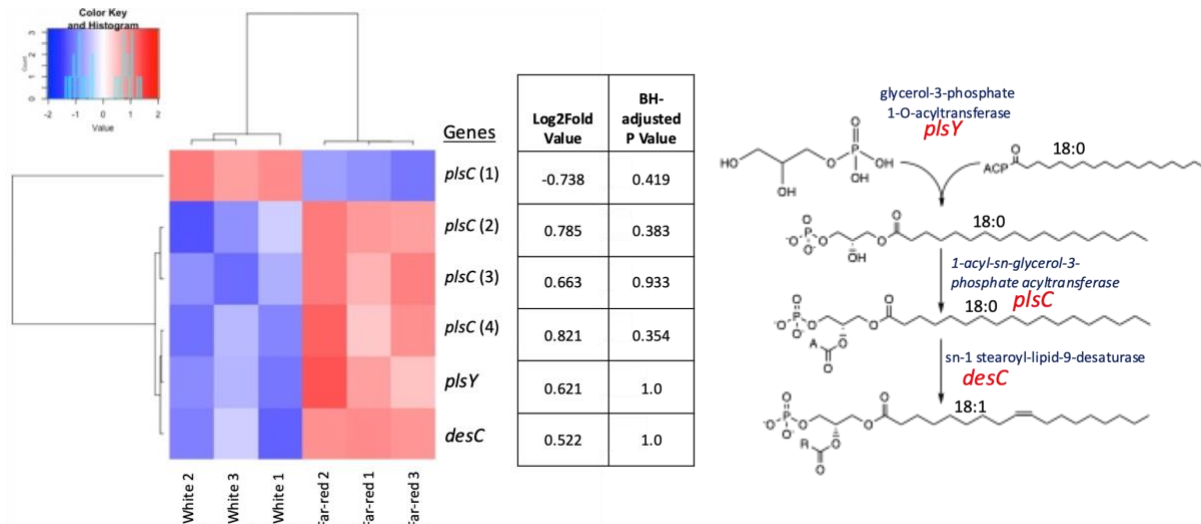


Figure 11. The heatmap produced (left) for the oleic acid synthesis genes alongside their Log2fold and BH-adjusted p-values (middle). Additionally, the biochemical pathway of oleic acid synthesis (right) shows the gene in red and the enzyme they encode in blue.

Out of the six genes within the oleic pathway, four genes that were annotated as *plsC* and to encode for a glycerol-3-phosphate-O-acyltransferase and only one *plsY* and *desC*, respectively (Figure 11). All the genes within the oleate (oleic acid) synthesis pathway are not significantly expressed.

Ten fatty acid desaturase (*fad*) genes were identified in the genome of *C. fritschii*, although there is no information on the specific desaturated fatty acid that they produce (Table 2). Only one *fad* gene, fatty acid desaturase (8), was found to be significantly under-expressed when growing in far-red with a Log2fold value of -1.443 ($p = 1.06 \times 10^{-5}$). The Z values for *fad8* were (far-red 1-3, then, white 1-3) -1.027, -0.671, -0.971 and 1.230, 0.695, 0.751. The log2fold value is reported on a logarithmic scale to base 2, and therefore the log2fold for *fad8*, -1.443, shows that the gene's expression has "decreased" by a multiplicative factor of -0.368, indicating a 2.71-fold decrease in expression.

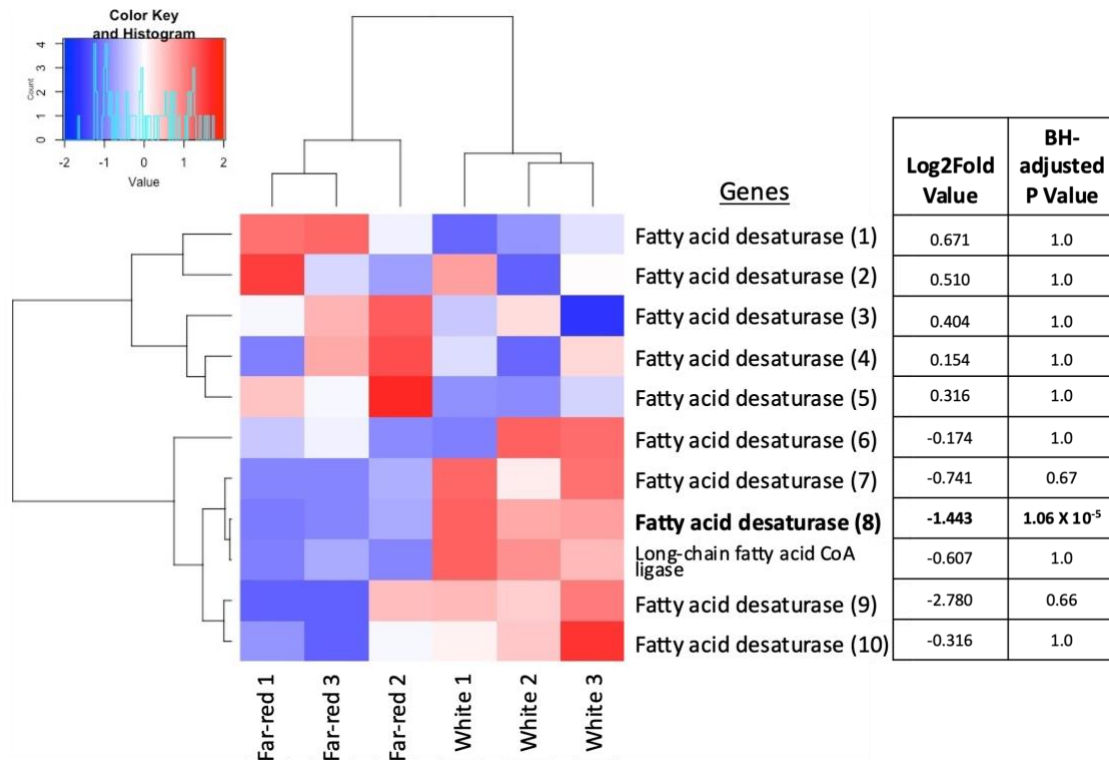


Figure 12. Ten fatty acid desaturases and one CoA ligase gene were identified, and their Z values calculated and displayed within a heatmap (left). Additionally, their Log2fold values and BH-adjusted p-values are shown (right).

The gene encoding long-chain fatty-acid CoA ligase is included in Figure 12. There was no effect on the latter gene's expression when the *C. fritschii* cultures were exposed to far-red light. The long-chain fatty-acid CoA-ligase is involved in recycling and excreting fatty acids; however, there is not much insight about this involvement in specific pathways due to its lack of complete annotation.

The final four genes identified, out of the 33 fatty acid related genes of interest, were related to the synthesis of polyunsaturated fatty acid. All four genes are labelled as the gene *pfad* (Figure 13). As seen in Figure 13, the annotation suggests these four genes may also potentially be related to polyketide biosynthesis.

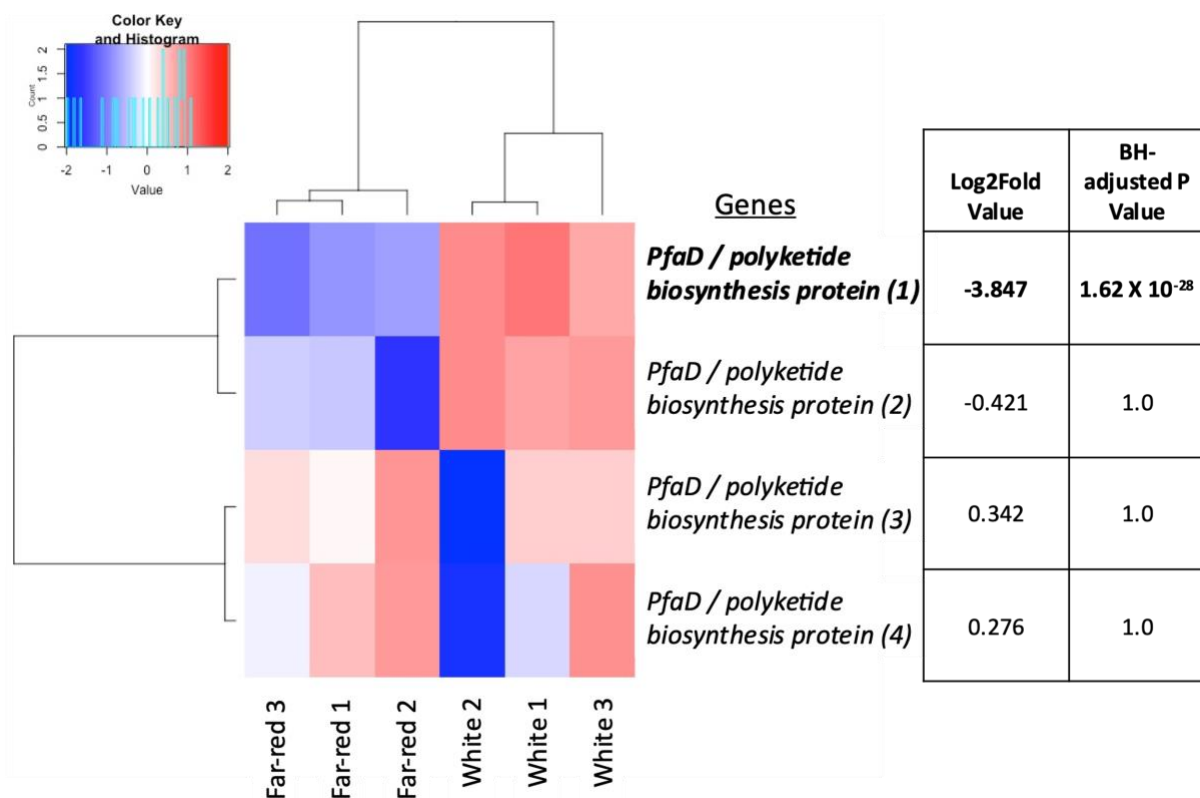


Figure 13. Four genes related to polyunsaturated fatty acid desaturase (PUFA-producing polyketide-like synthases biosynthesis) were identified. Their Z-values were calculated and displayed within a heatmap (left). Additionally, their log2fold values and BH-adjusted p-values are shown (right).

One of the *pfaD* genes, *pfaD1*, is significantly down regulated with a log2fold of -3.847 ($p = 1.62 \times 10^{-28}$), which indicated a 14.40-fold reduction in expression (i.e., a multiplicative factor of -0.069).

Discussion

Fatty acid synthesis is essential in bacteria as FAs are the building blocks of membrane lipids. Cyanobacteria can modify the fatty acid composition of their membranes in response to environmental cues, allowing organisms to survive under undesirable conditions (Roy, Dare & Ibba, 2009; Koba, 2012). Additionally, in cyanobacteria, FA are important components of glycolipids in the thylakoid membranes (Hölzl & Dörmann, 2007).

In this study, I investigated the response of FA-related genes in the cyanobacterium *C. fritschii* PCC6912 under far-red light and white light (control). Few investigations into far-red light on

cyanobacteria have looked at fatty acid metabolism. My results show that far-red light influences fatty acid-related pathways, specifically the desaturation of fatty acids.

FaRLiP allows cyanobacteria to grow within environments that are exposed to far-red light instead of visible light, such as caves or within aquatic environments (Behrendt *et al.*, 2020). The change of energy input for the cyanobacteria leads to an altered metabolism to sustain the required metabolites for cellular function (Ho & Bryant, 2019). As cyanobacteria possess many unique valuable compounds, including FA like EPA and DHA, it is vital for research to investigate how metabolic pathways can be altered to improve industrial production of such molecules such as fatty acids. Most genes within the analysis were not differentially expressed for multiple reasons, including functional genes that are rarely affected by environmental factors. We have found that 974 genes were differentially expressed when *Chlorogloeopsis fritschii* PCC6912 was exposed to far-red light. However, only two genes related to fatty acid synthesis pathways were seen as differentially expressed – *fad(8)* and *pfaD(1)* both being downregulated.

The FASII pathway-related genes (i.e., FA synthases) in *C. fritschii* PCC6912 were not differentially expressed. However, other studies found a decrease in expression. Ho & Bryant (2019) found *accBC*, *fabD*, *fabF*, *fabG*, three *fabH*, *fabI* and *fabZ* to be downregulated when *C. fritschii* PCC 9212 was exposed to far-red light whilst *accB*, *fabD*, and two each of *fabF* and *fabH* were found to be downregulated in *Acaryochloris marino* cultivated under far-red light (Hernández-Prieto *et al.*, 2018). For the gene encoding fatty-acid CoA-ligase, Hernández-Prieto *et al.* (2018) found an increased expression whilst Ho & Bryant (2019) found no change in expression, as I did. These discrepancies and similarities in FASII genes highlight metabolic differences between cyanobacterial species and strains.

Oleic acid is the precursor fatty acid for all other desaturated fatty acids, meaning a decrease in the in expression of genes involved in oleic acid synthesis may cause a decrease in the amount of desaturated fatty acids in the fatty acid composition of the cyanobacteria. No change in expression was seen for the six genes identified involved in the oleic acid pathway meaning that the effect of far-red light cultivation might not cause changes in the production of oleic acid. An unaltered production of oleic acid under far-red may result in no changes in

the profiles of *C. fritschii* desaturated fatty acids. However, this cannot be concluded from transcriptomic research without further information such as metabolic analysis. Conversely, *Acaryochloris marino* cultivated under far-red and white light found the oleic acid-related gene, *desC*, to be down regulated. The study identified 23 genes related to fatty acid metabolism in which six of the seven differentially expressed genes showed a decreased expression (Hernández-Prieto *et al.*, 2018). Two of the down-regulated genes encoded Δ 9-desaturases, which produce oleic acid from stearic acid.

Changes in the expression of fatty acid desaturases (*fad*) can influence the amount of mono- (e.g., oleic acid) and polyunsaturated fatty acids (MUFAs and PUFAs, respectively) in *C. fritschii* as these enzymes introduce a double bond in the fatty acid hydrocarbon chain. This study shows that far-red light can cause *fad* genes to be downregulated, specifically fatty acid desaturase (8). This decrease in expression of the *fad8* gene would result in a decreased production of the unsaturated fatty acid it produces. Unfortunately, without more information on the function of this gene or additional metabolic or functional analyses, it is impossible to further speculate on putative changes in the profiles of *C. fritschii* unsaturated fatty acids under far-red light. In a similar FaRLiP experiment within *C. fritschii* PCC 9212, a decrease in transcript abundance occurred for five genes related to fatty desaturation (Ho & Bryant, 2019). Three of the down-regulated genes within fatty acid desaturase were only annotated as “fatty acid desaturases” whilst the other two encoded for Δ 9-desaturase and Δ 15-desaturase; the latter producing the polyunsaturated fatty acid ALA (α -linolenic acid). Furthermore, Ho & Bryant (2019) suggested that when cyanobacteria exhibit FaRLiP, metabolic power is preserved by the decrease in gene expression which would explain some of the results found by this study.

The *pfa* gene family encode for polyketide-like synthases that are multidomain proteins which catalyse condensation, keto-reduction, dehydration, and enoyl-reduction reactions to produce PUFAs (polyunsaturated fatty acids). *pfaD* encodes for an enoyl-reductase whilst the dehydratase domains within *pfaC* that show homology to the fatty acid synthase genes *fabA* and *fabZ* (Oyola-Robles *et al.*, 2014). Bacteria containing the *pfa* gene cluster either contain four genes, *pfaA*, *B*, *C*, and *D*, or five, *pfaA*, *B*, *C*, *D* and *E*. Most bacteria only require *pfaABCD* to produce the PUFA, EPA and the additional *pfaE* gene allows the organism to produce DHA

(Allen & Barlett, 2002). One of the four *pfaD* genes identified in *C. fritschii* PCC6912 was found to be downregulated whilst the other three were not differentially expressed. A decrease expression of the *pfaD* gene will cause a decrease in the production of PUFAs, specifically EPA. However, cannot be concluded that there will be an effect on DHA production due to no identification of the gene *pfaE* within the genome. This result is concurrent to the *pfaD* and *pfaB* genes that are downregulated within the supplementary information of Ho & Bryant (2019). The annotation within this study provides only four *pfaD* genes and no other *pfa* genes (i.e., *pfaA*, *B* or *C*), which is assumed to be due to poor annotation. It is plausible that the four *pfaD* genes found in *C. fritschii* represent the four genes present in the cluster *pfaABCD* and required to produce EPA. This is not able to be concluded without further genomic, phylogenetic and domain analyses on these four *pfa* genes found in *C. fritschii*. Research into fatty acid pathways often neglects the *pfa* gene family encoding polyunsaturated fatty acid desaturases (Oyola-Robles *et al.*, 2013). This may be because the *pfa* gene family is seen as a novel pathway of fatty acid desaturation mostly present within deep-sea bacteria, which provides the ability to desaturate fatty acids anaerobically (Metz *et al.*, 2002; Oyola-Robles *et al.*, 2013). However, the presence of *pfa* genes in the coccolithophore *Emiliania huxleyi* as well as *C. fritschii* suggests the *pfa* genes may be more widespread in other marine microbes than first thought (Yoshida *et al.*, 2016).

My RNA-seq analysis conclude that far-red light influences fatty acid-related pathways, specifically the desaturation of fatty acids. The downregulation of two genes, *pfaD1* and *fad8*, that produce unsaturated fatty acids within different pathways suggests that there is a reduction in the amount of unsaturated fatty acids. This is likely due to the restructuring of the thylakoid membrane for the photoacclimation to far-red light (FaRLiP), as cyanobacteria grown under far-red light have more layers of thylakoid membrane with a reduced distance between each layer (Li *et al.*, 2016; Majumder *et al.*, 2017; MacGregor-Chatwin *et al.*, 2022).

Whilst this study has successfully identified a difference in fatty acid related gene expression when *Chlorogloeopsis fritschii* was cultivated under far-red light, it must be considered that there are further improvements that could be done within future studies. By only using transcriptomics, the study assumes that the gene expression has led directly to the production of the protein it expresses. This occurs often, however, in some cases, the gene

expression doesn't correlate with protein abundance or protein translation. The use of proteomics to identify enzyme abundance and, also, the use of metabolomics to quantify the reactants and products can provide a more detailed "metabolic picture" of fatty acid production by *C. fritschii* PCC6912. Therefore, this is suggested for any future experiments.

Furthermore, this study experienced issues with gene annotation which were unable to be resolved with time and resources available to this research. The lack of complete annotation is common for many cyanobacterial species. However, the latter is slowly being improved with the aid of genome databases such as CyanoBase, which incorporates genomes for 376 species (Fujisawa *et al.*, 2017; Pathania *et al.*, 2022). When investigating genome annotation within the bacterial tree of life, Lobb *et al.* (2020) found that annotation completeness ranged from 17% to 98% within different species from >27,000 examined genomes. A similar study involving eukaryotic algae showed that genome assembly quality, annotation quality and genome completeness had declined over time (Hanschen & Starkenburg, 2020). Increased research in long-read sequencing outputs, scaffolding and gene prediction technology, as well as the implementation of additional genome completeness analyses using software such as BUSCO might translate into better quality assembly and annotation (Seppey, Manni & Zdobnov, 2019; Hanschen & Starkenburg, 2020).

Another consideration is our experimental replication. It will be argued that the experimental replicates within this study are not "real replicates" due to the cultivation, and RNA extraction occurring alongside each other within the same lab space and the experiment not being repeated at a different time. Furthermore, the number of replicates are low which may have affected the output of the differentially expressed genes, as too much variability between the samples can result in non-statistical differences inferred by DESeq. The fact that our study shows fewer differentially expressed genes may be due to the experimental set-up, which includes only three replicates per treatment. Schurch *et al.* (2016) showed within a 48-replicate yeast study that only 20-40% of the differentially expressed genes were detected when only three replicates were used within the analysis. As shown by Lamarre *et al.* (2018), four replicates can significantly improve detection of differentially expressed genes compared to using three, or to include a public dataset that has a similar experimental design to the dataset initially being used. Thus, with a greater number of replicates, more genes might have

been identified as differentially expressed within the fatty acid pathways (Stark, Grzelak & Hadfield, 2019). Nevertheless, our core results are concurrent with the two studies on FaRLiP acclimation of cyanobacteria, which found at least two *fad* genes to be downregulated and the majority of both fatty acid and non-fatty acid-related genes having no differential expression (Hernández-Prieto *et al.*, 2018; Ho & Bryant, 2019).

Conclusion

Overall, the results above show that the bioinformatic process that was used to investigate genetic expression of fatty acid pathways was successful in removing unwanted adapters and low-quality transcripts whilst successfully aligning most of the transcripts to the reference genome and quantifying the gene expression through transcript quantification. Additionally, we found that cultivating *C. fritschii* under far-red light caused a differential expression of 974 genes – 323 upregulated and 651 downregulated. Furthermore, 33 fatty acid-related genes were identified and only 2 genes, *fad(8)* and *pfaD(1)*, were shown to have a significantly differential expression when the *C. fritschii* cultures were exposed to far-red light. The main pathways, fatty acid synthesis and oleic acid synthesis, that were identified showed no significant change in genetic expression, suggesting that there is little difference in fatty acid production when switching from white light to far-red light cultivation. However, FaRLiP appears to influence the unsaturation of fatty acids by downregulating the expression of *fad(8)* and *pfaD(1)*. While *fad(8)* showed a 2.71-fold decrease when *C. fritschii* PCC6912 was grown under far-red, *pfaD(1)*, functioning in a separate pathway of FA desaturation to the *fad* gene family, has a 14.40-fold reduction in expression when grown under far-red light. The decrease of a fatty acid desaturase is expected to cause an increase in saturation in lipid membranes which may increase membrane fluidity. This suggests that there is a general decrease within fatty acid desaturation within *C. fritschii* PCC6912 when exposed to far-red light which is likely explained by the remodelling of the thylakoid membranes to incorporate photosystems able to utilise far-red light. In addition, a decrease of a fatty acid desaturase activity is expected to cause an increase in saturation in lipid membranes which may increase membrane fluidity in *C. fritschii*. While FARLiP is beneficial to increase the production of certain cyanobacterial compounds (e.g., phycobiliproteins and MAAs), it isn't an appropriate approach to obtain/increase the yield of unsaturated fatty acids from cyanobacteria.

Appendix A – Data from the RNA processing and quality assessment

Transcript quality control, identification, and quantification

Quality control was performed on the raw reads to assess sequence quality, GC content, the presence of adaptors and/or sequencing errors (Conesa *et al.*, 2016). Also, raw RNA reads are processed through multiple bioinformatic programmes, including, SortMeRNA (Kopylova, Noé & Touzet, 2012) and Trimmomatic (Bolger, Lohse & Usadel, 2014) to remove non-messenger RNAs and any poor-quality reads, respectively. These processes create a higher quality data set, and improve mappability (i.e., the number of transcripts overlapping the genome) and transcript quantification via STAR and Salmon, respectively (Dobin *et al.*, 2013; Patro *et al.*, 2017).



Figure 14. The number of RNA reads from each biological sample throughout the bioinformatic pipeline. For all samples, the different steps within the process of quality assessment allowed for greater identification of unique reads while reducing duplicate reads.

Removal of low-quality reads

Low-quality reads were identified by Phred scores and GC content using FastQC (Andrew, 2010) and MultiQC (Ewels *et al.*, 2016). Within all samples, the mean Phred score at each bp position and mean Phred score for each RNA transcript were investigated. Additionally, an in-depth look at the Trimmomatic output and STAR alignment is shown in Figure 18 & 19.

It is predicted that the number of reads will decrease along the bioinformatic pathway due to the programs removing non-mRNA transcripts and low quality and unmapped. As expected along the pathway, there was a decrease in the overall number of reads, specifically after SortMeRNA and STAR (Figure 14).

When comparing the combined number of forward and reverse transcripts within the raw transcripts and final filtered transcripts for white 1 and far-red 1, white 1 loses 767,315 unique and 1,695,573 duplicate reads whilst far-red 1 loses 1,781,819 unique and 3,282,169 duplicate reads. This means that in white 1, 11.6% (unique) and 51.7% (duplicate) of the reads removed occurred in SortMeRNA and 85.9% (unique) and 48.1% (duplicate) within STAR. Similarly, white 2 and 3 had 15.5% and 18.9% (unique) and 29.7% and 61.6% (duplicate) of reads removed within SortMeRNA, respectively. While via STAR, white 2 and 3 had 83.5% and 78.4% (unique) and 70% and 37.7% (duplicate) of reads removed.

For far-red 1, 14.5% (unique) and 73.5% (duplicate) of the reads removed occurred in SortMeRNA and 83.8% (unique) and 26.1% (duplicate) within STAR. Likewise, far-red 2 and 3 had 14.3% and 38.6% (unique) and 73.4% and 76% (duplicate) of reads removed within SortMeRNA, respectively. During STAR, far-red 2 and 3 had 83.9% and 59.2% (unique) and 26.2% and 23.6% (duplicate) of reads removed (more information in Appendix A , Table 3).

Overall, the variation between samples overlaps and more unique reads were lost within SortMeRNA, and more duplicate reads were lost within STAR. The combined STAR output identified that some of the duplicates are uniquely mapped, increasing the overall unique reads within this combined output. Therefore, it is concluded that the drop in reads has improved the quality and mappability of the transcripts, which mainly occurred by the removal of non-mRNAs and poorly mapped transcripts (kopylova, Noé & Touzet, 2012; Dobin *et al.*, 2013).

Phred quality scores (Figure 15) provide a greater insight into the quality of Illumina sequenced data by identifying how accurately the nucleotide for each base position is identified (Ewing *et al.*, 1998). All samples have a Phred score above 30, meaning that the probability of successfully identifying the correct nucleotide is >99.9%. Figure 15 shows that throughout all samples, the reverse strand has a lower Phred score. This is likely due to the longer time that these samples have spent on the sequencing plate, and fluorescent markers degrading during this time (Ozsolak & Milos, 2011).

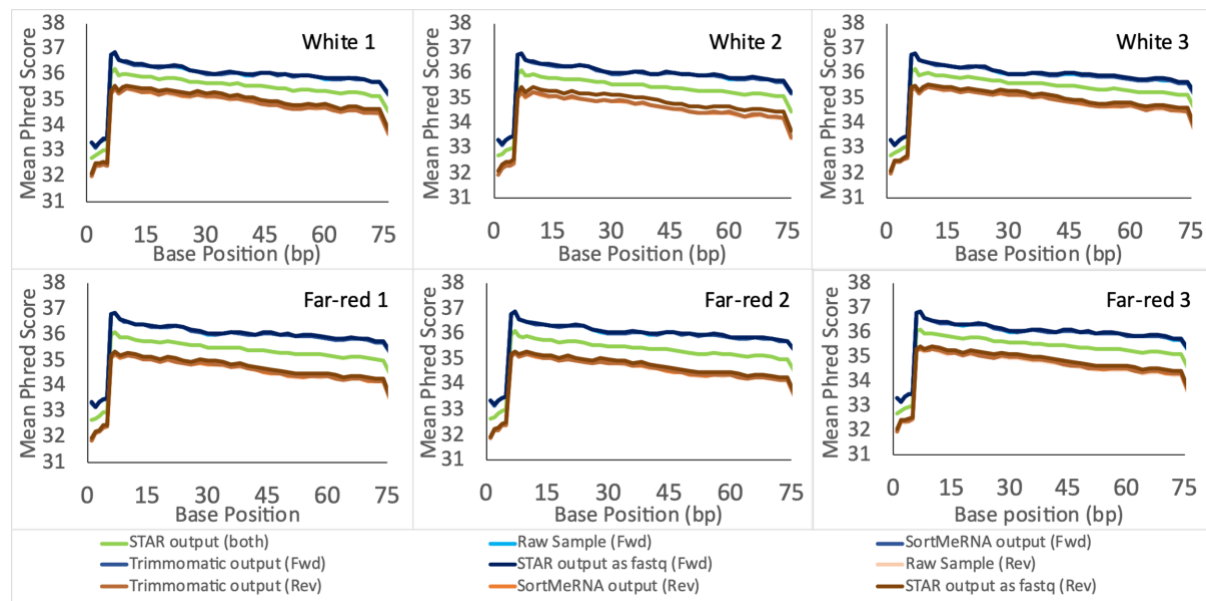


Figure 15. The mean Phred score for each base position within the samples along the bioinformatic pathway.

Within white 1, the average Phred score within the raw forward and reverse transcripts are 35.76 and 34.61, respectively. This is similar to white 2 and 3, which have average Phred scores of 35.73 and 35.70 (forward) and 34.38 and 34.63 (reverse), respectively. Far-red 1 has a mean Phred score within the raw forward and reverse transcripts of 35.76 and 34.34, respectively. This is comparable to far-red 2 and 3, which have average Phred scores of 35.77 and 35.77 (forward) and 34.35 and 34.46 (reverse), respectively.

The mean Phred score for each base position increases in all samples throughout the quality control pipeline. After the quality assessment, the mean forward and reverse transcripts reach 35.78 and 34.76, respectively, in white 1 reads. Again, white 2 and 3 follow the same trend, with the forward transcripts reaching 35.75 and 35.73 and the reverse transcripts reaching 34.60 and 34.75, respectively.

Quality-assessed far-red samples also follow the same increasing pattern. For instance, far-red 1 exhibits a mean Phred score of 35.78 and 34.4 for the final forward and reverse transcripts, respectively. Similarly, far-red 2 and 3 reach 35.79 and 35.80 for the forward transcripts, and 34.47 and 34.60, respectively, for the reverse transcripts. (More information within Appendix A, Table 4 – 9.).

Overall, the implemented bioinformatic pipeline improved the quality of all the analysed strands, although a decrease in Phred scores, and therefore, quality, was seen in reverse strands, but also at the first 15 base positions of both reverse and forward-facing strands. A decrease in Phred scores at the beginning of the read is likely to be due to random hexamer primers present from the initial start of the sequencing occurring on each strand, causing a lower nucleotide frequency. Hansen, Brenner & Dudoit (2010) analysed RNA-seq experiments and found that the nucleotide frequencies (Phred score) varied a lot in the first 13 base positions, after that, any observed variation became independent of position.

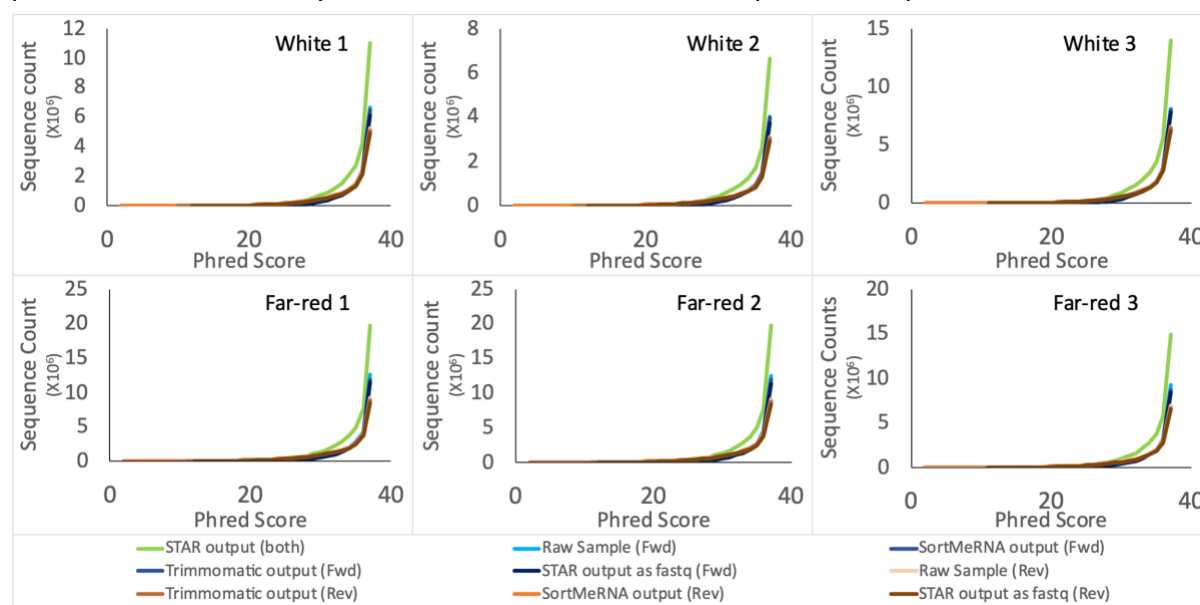


Figure 16. The number of sequences with each Phred score.

The number of sequences with their respective Phred scores is presented in Figure 16. These data depict that most of the identified sequences throughout the bioinformatic pipeline display a Phred score >30. For example, the raw forward and reverse transcripts of white 3 have 570,222 (3.3%) and 1,954,179 (11%) transcripts with Phred scores below 30 and 17,183,299 (96.7%) and 15,816,370 (89%) above 30, respectively. This is similar to the other

white samples, as white 1 and 2 have 92.8% and 92.0% of forward and reverse transcripts with Phred scores above 30, respectively. For the raw reverse and forward transcripts of far-red 3, 599,377 (3.1%) and 2,410,314 (12.5%) transcripts had Phred scores below 30 and 18,730,525 (96.9%) and 16,917,012 (87.5%) with a Phred score above 30. Once again, this is similar within far-red 1 and 2, with 91.7% and 91.5% of forward and reverse transcripts having Phred scores above 30.

An increase in Phred scores for all transcripts is seen after the samples have been processed through SortMeRNA, Trimmomatic and STAR. For white 3 forward and reverse strands, 520,187 (3.1%) and 1,711,324 (10.2%) transcripts had Phred scores below 30 and 16,237,054 (96.9%) and 15,045,917 (89.8%) above 30, respectively. white 1 and 2 follow this trend, with now 93.4% and 92.19% of reverse and forward transcripts having Phred scores above 30.

For the forward and reverse strands of far-red 3 after SortMeRNA, Trimmomatic and STAR have 514,688 (2.9%) and 2,045,715 (11.6%) below 30 and 17,125,673 (97.1%) and 15,594,643 (88.4%) above 30, respectively. This pattern is seen also within far-red 1 and 2, where 92.18% and 92.23% of the forward and reverse transcripts have Phred scored above 30. In general, reverse strands have a higher quantity of transcripts with Phred scores below 30, highlighting the drop in quality seen within reverse strand sequencing across all samples (More information within Appendix A, Table 10 – 15.).

The method of RNA extraction and purification, as outlined in Llewellyn *et al.* (2020), shows that all ribosomal RNA (rRNA) is supposed to be removed. However, the drop in the number of transcripts present in each data set after running SortMeRNA indicated the presence of rRNA transcripts within the raw data. The importance of using SortMeRNA is due to rRNA not always being fully removed from RNA samples no matter which rRNA depletion method is used, which can leave up to 20% of the RNA sample being rRNA (Stark, Grzelak & Hadfield, 2019). Thus, it is highly important to incorporate a bioinformatic step able to identify rRNA and filter out the metatranscriptomic samples of any rRNA from the species but also from sample contamination (Kopylova, Noé & Touzet, 2012).

Assessment of GC Content

GC content is an organism-specific feature, but these levels need to stay homogenous across samples of the same experiment (Conesa *et al.*, 2016). Figure 17, derived from the FastQC output, shows the distribution of the sequences' GC content (%). For all samples, white and far-red light, the GC content of the transcripts shows a normal distribution, with most of the sequences having a GC content of 50% (Figure 17). There is minor deviation between samples which is expected due to each sample having slightly different sequencing profiles and expression levels.

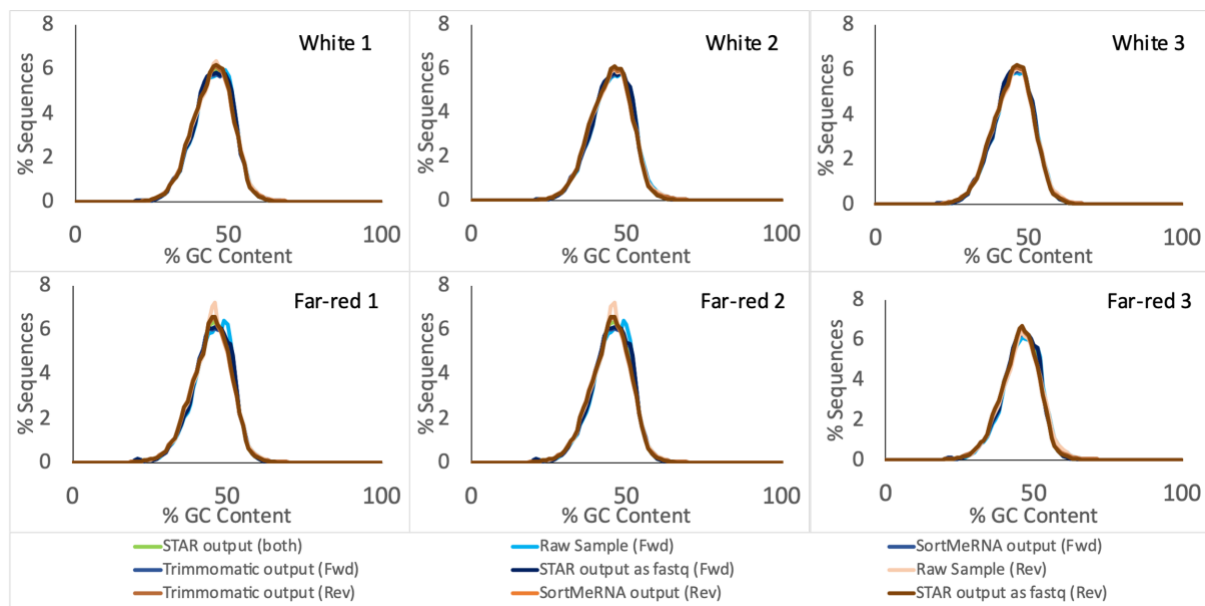


Figure 17. The GC content of the samples along the bioinformatic pathway show a normal deviation throughout the bioinformatic pipeline. Only minor differences are seen between flasks (i.e., biological replicates) and between different stages of the quality assessment process implemented in the pipeline.

Removal of transcripts by SortMeRNA, Trimmomatic and STAR can also cause minor variations in GC content across the processed data set. If the removal of transcripts had been too intense or included a GC bias, a more obvious change in the GC content would have been seen. Therefore, the lack of major variations in GC content throughout the bioinformatic pipeline shows successful transcript filtering, which in turn will increase transcript mappability and quantification. Therefore, we can conclude that our data show no GC bias during rRNA removal, trimming or alignment (Andrews 2010; Ewels *et al.*, 2016).

Transcript trimming and mapping

Overall, most transcripts were conserved by Trimmomatic and mapped via STAR. Trimmomatic investigates the strands for low-quality reads and bases and adapters still present. As shown previously, the reverse strand can have lower quality due to the increased length of time for sequencing, allowing more degradation to occur. Therefore, some of the reverse strands are not of great quality in comparison to their respective forward-facing transcript. During Trimmomatic, a minimum filter length was used to prevent aggressive trimming, which has been shown to impact expression estimates due to short reads produced from trimming (Williams *et al.*, 2016).

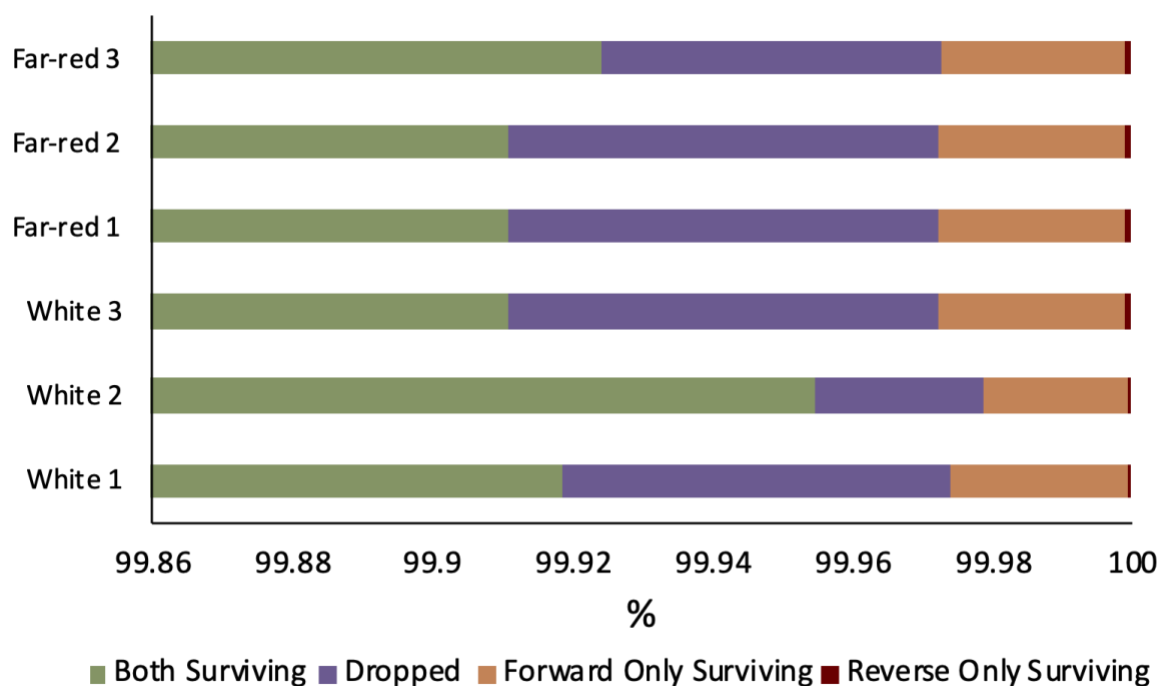


Figure 18. The Trimmomatic output breakdown shows that >99.9% of both transcripts within each file survived the Trimmomatic analysis.

All samples examined had >99.9% of reads untrimmed, therefore, the Trimmomatic is not seen as aggressive and, therefore, shouldn't have affected the gene expression and only increased the quality of transcript reads. It must be noted that newer research recommends not to incorporate trimming algorithms into their bioinformatic pipelines and that this method may become redundant for transcript expression studies. Liao & Shi (2019) have shown that some pieces of alignment software, such as STAR, are able to remove the adapters

by its soft clipping procedures, allowing for a better alignment of low-quality reads and more accurate expression profiles.

As seen in Figure 18, >99.9% of transcripts survived after Trimmomatic, whilst <0.06% were dropped from both forward and reverse transcripts. On average, a higher percentage of transcripts survived in the white light samples, however, the difference is negligible.

STAR is a bioinformatic program used to align transcripts to a reference genome by mapping the transcripts out along the genome and allowing the quantification of transcripts of specific genes. By finding multiple potential matches for each strand to the reference genome, STAR maps the transcript to the gene that has the highest match to the transcript. If the transcript does not find an exact match, STAR finds where the transcript has the closest match to one gene, the maximal mappable prefix (MMP). The unmapped part of the transcript is spliced and then aligned to the MMP related to itself. If the transcript is stitched to two genes, this is then reported as so within the output as mapped to multiple loci (Figure 19).

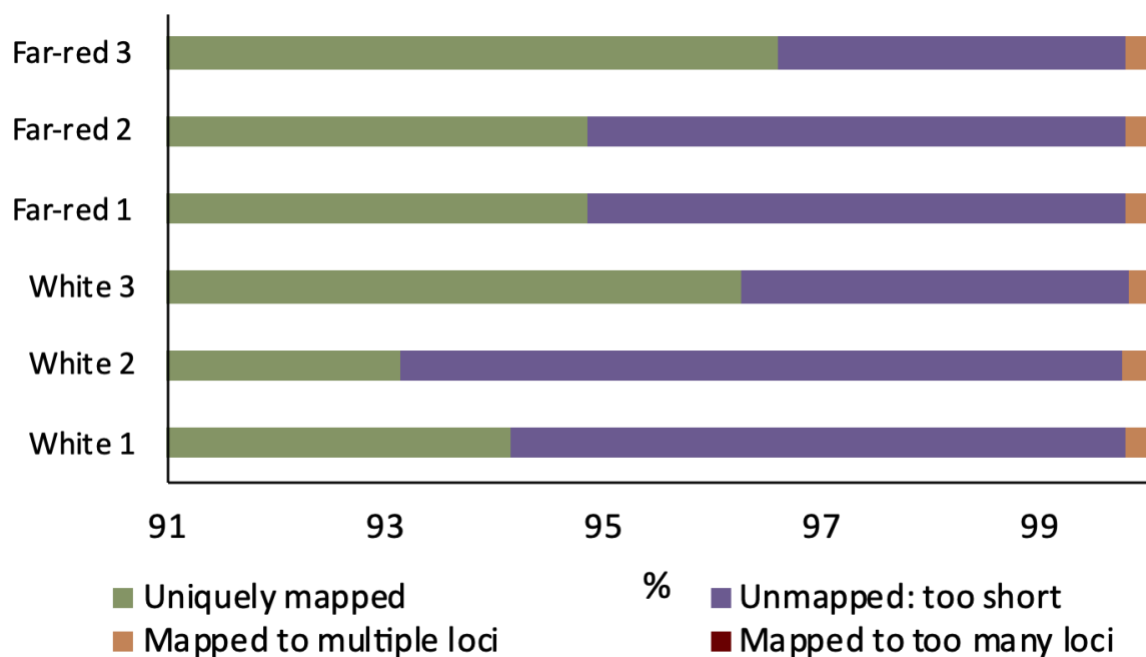


Figure 19. The proportion of reads mapped or unmapped during the STAR transcript alignment. Most transcripts were uniquely mapped during STAR alignment.

Over 90% of the transcripts in all six samples were uniquely mapped to a gene (Figure 19), meaning that the likelihood of a transcript being aligned to the wrong gene is far less as there

are no other matches within the genome that cause the transcript to be mapped to multiple genes. Under 6% of transcripts could not be mapped. This is potentially due to these transcripts not being long enough - the average length of the transcripts is only 76-bp long - or it could be due to the transcripts being long enough but not aligning well enough to specific regions of the reference genome to be mapped. Due to this being a small percentage of the transcripts, it is not a cause of concern. The percentage of transcripts that are mapped to multiple or too many gene loci is <0.1% suggesting that the number of misaligned transcripts is minimal.

STAR has been proven to be an extremely fast alignment tool that is 180 times faster than GSNAP and MapSplice programs (Dobin *et al.*, 2013). However, STAR may produce false exon junctions within the output compared to the more conservative alignment, such as MapSplice (Engström *et al.*, 2013). This is not an issue for the data within this study as the percentage of transcripts mapped to multiple loci is <0.1% in all samples.

Appendix B – Coding of the data analysis

Code within Ubuntu

```
#!/usr/bin/env bash
set -euxo pipefail
set -o erretrace
```

```
#####
```

```
    #Download fast- and MultiQC#
```

```
#####
```

```
conda create -n fastqc
conda activate fastq
conda install -c bioconda fastqc
conda deactivate fastqc
```

```
conda create -n multiqc
conda activate multiqc
conda install -c bioconda multiqc
conda deactivate
```

```
#####
```

```
    #Download Analysis Programs#
```

```
#####
```

```
# Curl
sudo apt install curl
```

```
# SortMeRNA
conda create -n sortmerna -c conda-forge -c bioconda sortmerna
```

```
# Trimmomatic
conda create -n trimmomatic
conda activate trimmomatic
conda install -c bioconda trimmomatic
conda deactivate
```

```
# AGAT
conda create -n agat -c conda-forge -c bioconda agat
```

```
# STAR
wget https://github.com/alexdobin/STAR/archive/2.7.9a.tar.gz
tar -xzf 2.7.9a.tar.gz
cd ~/STAR-2.7.9a/source/
make STAR
cd
```

```
# Bedtools
sudo apt-get install bedtools

# GFFRead
cd /some/build/dir
git clone https://github.com/gpertea/gffread
cd gffread
make release

# Samtools
conda create -n samtools
conda activate samtools
conda install -c bioconda samtools
conda deactivate

# Salmon
conda create -n salmon
conda activate salmon
conda install -c bioconda salmon
conda deactivate
```

```
#####
```

```
#creating file structure#
```

```
#####
```

```
cd
mkdir /home/patrick/cyano/

cd /home/patrick/cyano/
mkdir /home/patrick/cyano/FR_reads/
mkdir /home/patrick/cyano/Sortmerna/
mkdir /home/patrick/cyano/trimmomatic/
mkdir /home/patrick/cyano/QC/
mkdir /home/patrick/cyano/STAR/
mkdir /home/patrick/cyano/genome/
mkdir /home/patrick/cyano/Salmon/
mkdir /home/patrick/cyano/Bedtools/
mkdir /home/patrick/cyano/salmon_index/
```

```
cd /home/patrick/cyano/Sortmerna/
mkdir SRR11810824
mkdir SRR11810825
mkdir SRR11810826
mkdir SRR11810827
mkdir SRR11810828
mkdir SRR11810829
```

```
cp \
```

```
-r \  
/home/patrick/cyano/Sortmerna/* \  
/home/patrick/cyano/trimmomatic/
```

```
cp \  
-r \  
/home/patrick/cyano/Sortmerna/* \  
/home/patrick/cyano/QC/
```

```
cp \  
-r \  
/home/patrick/cyano/Sortmerna/* \  
/home/patrick/cyano/STAR/
```

```
cp \  
-r \  
/home/patrick/cyano/Sortmerna/* \  
/home/patrick/cyano/Salmon/
```

```
cp \  
-r \  
/home/patrick/cyano/Sortmerna/* \  
/home/patrick/cyano/Bedtools/
```

```
# adding multiqc file to qc directory
```

```
mkdir /home/patrick/cyano/QC/SRR11810824/MultiQC  
mkdir /home/patrick/cyano/QC/SRR11810825/MultiQC  
mkdir /home/patrick/cyano/QC/SRR11810826/MultiQC  
mkdir /home/patrick/cyano/QC/SRR11810827/MultiQC  
mkdir /home/patrick/cyano/QC/SRR11810828/MultiQC  
mkdir /home/patrick/cyano/QC/SRR11810829/MultiQC
```

```
mkdir /home/patrick/cyano/QC/SRR11810829/Logs  
mkdir /home/patrick/cyano/QC/SRR11810828/Logs  
mkdir /home/patrick/cyano/QC/SRR11810827/Logs  
mkdir /home/patrick/cyano/QC/SRR11810826/Logs  
mkdir /home/patrick/cyano/QC/SRR11810825/Logs  
mkdir /home/patrick/cyano/QC/SRR11810824/Logs
```

```
#####  
#Download RNA sequences#  
#####
```

```
#!/usr/bin/env bash
```

```

curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/025/SRR11810825/SRR11810825_1.fastq.gz -o
SRR11810825_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/025/SRR11810825/SRR11810825_2.fastq.gz -o
SRR11810825_2.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/024/SRR11810824/SRR11810824_1.fastq.gz -o
SRR11810824_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/024/SRR11810824/SRR11810824_2.fastq.gz -o
SRR11810824_2.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/027/SRR11810827/SRR11810827_1.fastq.gz -o
SRR11810827_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/027/SRR11810827/SRR11810827_2.fastq.gz -o
SRR11810827_2.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/028/SRR11810828/SRR11810828_1.fastq.gz -o
SRR11810828_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/028/SRR11810828/SRR11810828_2.fastq.gz -o
SRR11810828_2.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/029/SRR11810829/SRR11810829_1.fastq.gz -o
SRR11810829_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/029/SRR11810829/SRR11810829_2.fastq.gz -o
SRR11810829_2.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/026/SRR11810826/SRR11810826_1.fastq.gz -o
SRR11810826_1.fastq.gz
curl -L
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR118/026/SRR11810826/SRR11810826_2.fastq.gz -o
SRR11810826_2.fastq.gz

mv ~/SRR11810829*.fastq.gz ~/cyano/FR_reads/

#####
#QC#
#####
conda activate fastqc

fastqc \
/home/patrick/cyano/FR_reads/SRR11810824*.fastq.gz \
-o /home/patrick/cyano/QC/SRR11810824/

```

```
fastqc \  
/home/patrick/cyano/FR_reads/SRR11810825*.fastq.gz \  
-o /home/patrick/cyano/QC/SRR11810825/
```

```
fastqc \  
/home/patrick/cyano/FR_reads/SRR11810826*.fastq.gz \  
-o /home/patrick/cyano/QC/SRR11810826/
```

```
fastqc \  
/home/patrick/cyano/FR_reads/SRR11810827*.fastq.gz \  
-o /home/patrick/cyano/QC/SRR11810827/
```

```
fastqc \  
/home/patrick/cyano/FR_reads/SRR11810828*.fastq.gz \  
-o /home/patrick/cyano/QC/SRR11810828/
```

```
fastqc \  
/home/patrick/cyano/FR_reads/SRR11810829*.fastq.gz \  
-o /home/patrick/cyano/QC/SRR11810829/
```

```
conda deactivate
```

```
#####  
#SortMeRNA#  
#####
```

```
conda activate sortmerna
```

```
indexdb_rna --ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-16s-  
id90.fasta,/home/patrick/cyano/Sortmerna/index/silva-bac-16s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/silva-bac-23s-  
id98.fasta,/home/patrick/cyano/Sortmerna/index/silva-bac-23s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/silva-arc-16s-  
id95.fasta,/home/patrick/cyano/Sortmerna/index/silva-arc-16s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/silva-arc-23s-  
id98.fasta,/home/patrick/cyano/Sortmerna/index/silva-arc-23s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/silva-euk-18s-  
id95.fasta,/home/patrick/cyano/Sortmerna/index/silva-euk-18s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/silva-euk-28s-  
id98.fasta,/home/patrick/cyano/Sortmerna/index/silva-euk-28s:\  
/home/patrick/cyano/Sortmerna/rna_refs/rfam-5s-database-  
id98.fasta,/home/patrick/cyano/Sortmerna/index/rfam-5s-db:\  
/home/patrick/cyano/Sortmerna/rna_refs/rfam-5.8s-database-  
id98.fasta,/home/patrick/cyano/Sortmerna/index/rfam-5.8s-db
```



```

sortmerna \
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5.8s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-16s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-16s-id90.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-18s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-28s-id98.fasta \
--reads ~/cyano/FR_reads/SRR11810824_1.fastq.gz \
--reads ~/cyano/FR_reads/SRR11810824_2.fastq.gz \
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \
--workdir /home/patrick/cyano/Sortmerna/ \
--out2 \
--fastx \
--threads 8 \
--paired_in \
--kvdb ~/cyano/Sortmerna/SRR11810824/SRR11810824_kvdb \
--other ~/cyano/Sortmerna/SRR11810824/SRR11810824_sortmerna |&
tee ~/cyano/Sortmerna/SRR11810824/SRR11810824_sortmerna.log

```

```

cp /home/patrick/cyano/Sortmerna/SRR11810824/SRR11810824_sortmerna.log
/home/patrick/cyano/QC/Logs/SRR11810824/SRR11810824_sortmerna.log

```

```

sortmerna \
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5.8s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-16s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-16s-id90.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-18s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-28s-id98.fasta \
--reads ~/cyano/FR_reads/SRR11810825_1.fastq.gz \
--reads ~/cyano/FR_reads/SRR11810825_2.fastq.gz \
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \
--workdir /home/patrick/cyano/Sortmerna/ \
--out2 \
--fastx \
--threads 8 \
--paired_in \
--kvdb ~/cyano/Sortmerna/SRR11810825/SRR11810825_kvdb \
--other ~/cyano/Sortmerna/SRR11810825/SRR11810825_sortmerna |&
tee ~/cyano/Sortmerna/SRR11810825/SRR11810825_sortmerna.log

```

```

cp /home/patrick/cyano/Sortmerna/SRR11810825/SRR11810825_sortmerna.log
/home/patrick/cyano/QC/Logs/SRR11810825/SRR11810825_sortmerna.log

```

```

sortmerna \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/rfam-5.8s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/rfam-5s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-arc-16s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-arc-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-bac-16s-id90.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-bac-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-euk-18s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-euk-28s-id98.fasta \
--reads ~/cyano/FR_reads/SRR11810826_1.fastq.gz \
--reads ~/cyano/FR_reads/SRR11810826_2.fastq.gz \
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \
--workdir /home/patrick/cyano/Sortmerna/ \
--out2 \
--fastx \
--threads 8 \
--paired_in \
--kvdb ~/cyano/Sortmerna/SRR11810826/SRR11810826_kvdb \
--other ~/cyano/Sortmerna/SRR11810826/SRR11810826_sortmerna |&
tee ~/cyano/Sortmerna/SRR11810826/SRR11810826_sortmerna.log

```

```

cp /home/patrick/cyano/Sortmerna/SRR11810826/SRR11810826_sortmerna.log
/home/patrick/cyano/QC/SRR11810826/SRR11810826_sortmerna.log

```

```

sortmerna \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/rfam-5.8s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/rfam-5s-database-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-arc-16s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-arc-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-bac-16s-id90.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-bac-23s-id98.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-euk-18s-id95.fasta \
--ref /home/patrick/cyano/Sortmerna/rrna_refs/silva-euk-28s-id98.fasta \
--reads ~/cyano/FR_reads/SRR11810827_1.fastq.gz \
--reads ~/cyano/FR_reads/SRR11810827_2.fastq.gz \
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \
--workdir /home/patrick/cyano/Sortmerna/ \
--out2 \
--fastx \
--threads 8 \
--paired_in \
--kvdb ~/cyano/Sortmerna/SRR11810827/SRR11810827_kvdb \
--other ~/cyano/Sortmerna/SRR11810827/SRR11810827_sortmerna |&
tee ~/cyano/Sortmerna/SRR11810827/SRR11810827_sortmerna.log

```

```
cp /home/patrick/cyano/Sortmerna/SRR11810827/SRR11810827_sortmerna.log
/home/patrick/cyano/QC/Logs/SRR11810827/SRR11810827_sortmerna.log
```

```
sortmerna \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5.8s-database-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5s-database-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-16s-id95.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-23s-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-16s-id90.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-23s-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-18s-id95.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-28s-id98.fasta \  
--reads ~/cyano/FR_reads/SRR11810828_1.fastq.gz \  
--reads ~/cyano/FR_reads/SRR11810828_2.fastq.gz \  
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \  
--workdir /home/patrick/cyano/Sortmerna/ \  
--out2 \  
--fastx \  
--threads 8 \  
--paired_in \  
--kvdb ~/cyano/Sortmerna/SRR11810828/SRR11810828_kvdb \  
--other ~/cyano/Sortmerna/SRR11810828/SRR11810828_sortmerna | &  
tee ~/cyano/Sortmerna/SRR11810828/SRR11810828_sortmerna.log
```

```
cp /home/patrick/cyano/Sortmerna/SRR11810828/SRR11810828_sortmerna.log
/home/patrick/cyano/QC/Logs/SRR11810828/SRR11810828_sortmerna.log
```

#27469.pts-1.patrick-All-Series running sortmerna

```
sortmerna \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5.8s-database-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/rfam-5s-database-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-16s-id95.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-arc-23s-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-16s-id90.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-bac-23s-id98.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-18s-id95.fasta \  
--ref /home/patrick/cyano/Sortmerna/rna_refs/silva-euk-28s-id98.fasta \  
--reads ~/cyano/FR_reads/SRR11810829_1.fastq.gz \  
--reads ~/cyano/FR_reads/SRR11810829_2.fastq.gz \  
--idx-dir /home/patrick/cyano/Sortmerna/index2/ \  
--workdir /home/patrick/cyano/Sortmerna/ \  
--out2 \  
--fastx \  
--threads 8 \  
--paired_in \  

```

```
--kvdb ~/cyano/Sortmerna/SRR11810829/SRR11810829_kvdb \  
--other ~/cyano/Sortmerna/SRR11810829/SRR11810829_sortmerna | &  
tee ~/cyano/Sortmerna/SRR11810829/SRR11810829_sortmerna.log
```

```
cp /home/patrick/cyano/Sortmerna/SRR11810829/SRR11810829_sortmerna.log  
/home/patrick/cyano/QC/Logs/SRR11810829/SRR11810829_sortmerna.log
```

```
conda deactivate  
#####  
#QC#  
#####  
conda activate fastqc
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810824/SRR11810824*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810824/
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810825/SRR11810825*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810825/
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810826/SRR11810826*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810826/
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810827/SRR11810827*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810827/
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810828/SRR11810828*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810828/
```

```
fastqc \  
/home/patrick/cyano/Sortmerna/SRR11810829/SRR11810829*.fq.gz \  
-o /home/patrick/cyano/QC/SRR11810829/
```

```
conda deactivate  
#####  
#TRIMMOMATIC#  
#####  
conda activate trimmomatic
```

```
trimmomatic PE \  
-threads 8 \  
/home/patrick/cyano/Sortmerna/SRR11810824/SRR11810824_sortmerna_fwd.fq.gz
```

```

/home/patrick/cyano/Sortmerna/SRR11810824/SRR11810824_sortmerna_rev.fq.gz

/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_trimmomatic_
1.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_unpaired_1.fq
.gz \

/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_trimmomatic_
2.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_unpaired_2.fq
.gz \
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \
SLIDINGWINDOW:5:5 \
#5 is seen as plenty for the sliding length and mean minimum
MINLEN:50 |&
tee /home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_trimmomatic.log

cp /home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_trimmomatic.log
/home/patrick/cyano/QC/Logs/SRR11810824/SRR11810824_trimmomatic.log

trimmomatic PE \
-threads 8 \
/home/patrick/cyano/Sortmerna/SRR11810825/SRR11810825_sortmerna_fwd.fq.gz
/home/patrick/cyano/Sortmerna/SRR11810825/SRR11810825_sortmerna_rev.fq.gz

/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_trimmomatic_
1.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_unpaired_1.fq
.gz \

/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_trimmomatic_
2.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_unpaired_2.fq
.gz \
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \
SLIDINGWINDOW:5:5 \
#5 is seen as plenty for the sliding length and mean minimum
MINLEN:50 |&
tee /home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_trimmomatic.log

cp /home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_trimmomatic.log
/home/patrick/cyano/QC/Logs/SRR11810825/SRR11810825_trimmomatic.log

```

```
trimmomatic PE \  
-threads 8 \  
/home/patrick/cyano/Sortmerna/SRR11810826/SRR11810826_sortmerna_fwd.fq.gz  
/home/patrick/cyano/Sortmerna/SRR11810826/SRR11810826_sortmerna_rev.fq.gz  
  
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_trimmomatic_  
1.fq.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_unpaired_1.fq  
.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_trimmomatic_  
2.fq.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_unpaired_2.fq  
.gz \  
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \  
SLIDINGWINDOW:5:5 \  
#5 is seen as plenty for the sliding length and mean minimum  
MINLEN:50 |&  
tee /home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_trimmomatic.log  
  
cp /home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_trimmomatic.log  
/home/patrick/cyano/QC/Logs/SRR11810826/SRR11810826_trimmomatic.log
```

```
trimmomatic PE \  
-threads 8 \  
/home/patrick/cyano/Sortmerna/SRR11810827/SRR11810827_sortmerna_fwd.fq.gz  
/home/patrick/cyano/Sortmerna/SRR11810827/SRR11810827_sortmerna_rev.fq.gz  
  
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_trimmomatic_  
1.fq.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_unpaired_1.fq  
.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_trimmomatic_  
2.fq.gz \  
  
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_unpaired_2.fq  
.gz \  
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \  
SLIDINGWINDOW:5:5 \  
#5 is seen as plenty for the sliding length and mean minimum  
MINLEN:50 |&
```

```

tee /home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_trimmomatic.log

cp /home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_trimmomatic.log
/home/patrick/cyano/QC/Logs/SRR11810827/SRR11810827_trimmomatic.log

trimmomatic PE \
-threads 8 \
/home/patrick/cyano/Sortmerna/SRR11810828/SRR11810828_sortmerna_fwd.fq.gz
/home/patrick/cyano/Sortmerna/SRR11810828/SRR11810828_sortmerna_rev.fq.gz

/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_trimmomatic_
1.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_unpaired_1.fq
.gz \

/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_trimmomatic_
2.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_unpaired_2.fq
.gz \
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \
SLIDINGWINDOW:5:5 \
#5 is seen as plenty for the sliding length and mean minimum
MINLEN:50 |&
tee /home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_trimmomatic.log

cp /home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_trimmomatic.log
/home/patrick/cyano/QC/Logs/SRR11810828/SRR11810828_trimmomatic.log

trimmomatic PE \
-threads 8 \
/home/patrick/cyano/Sortmerna/SRR11810829/SRR11810829_sortmerna_fwd.fq.gz
/home/patrick/cyano/Sortmerna/SRR11810829/SRR11810829_sortmerna_rev.fq.gz

/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_trimmomatic_
1.fq.gz \

/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_unpaired_1.fq
.gz \

/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_trimmomatic_
2.fq.gz \

```

```
/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_unpaired_2.fq.gz \
ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE.fa:5:30:10 \
SLIDINGWINDOW:5:5 \
#5 is seen as plenty for the sliding length and mean minimum
MINLEN:50 |&
tee /home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_trimmomatic.log
```

```
cp /home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_trimmomatic.log
/home/patrick/cyano/QC/Logs/SRR11810829/SRR11810829_trimmomatic.log
```

```
conda deactivate
#####
#QC#
#####
conda activate fastqc
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810824/ \
/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_trimmomatic_
*.fq.gz
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810825/ \
/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_trimmomatic_
*.fq.gz
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810826/ \
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_trimmomatic_
*.fq.gz
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810827/ \
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_trimmomatic_
*.fq.gz
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810828/ \
/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_trimmomatic_
*.fq.gz
```

```
fastqc \
-o /home/patrick/cyano/QC/SRR11810829/ \
```



```
/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_trimmomatic_*.fq.gz
```

```
conda deactivate
```

```
#####
```

```
#Sequence Alignment#
```

```
#####
```

```
#convert genome annotation gff file to gtf
```

```
conda activate agat
```

```
agat_convert_sp_gff2gtf.pl \
```

```
--gff /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \
```

```
-o /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gtf
```

```
conda deactivate
```

```
#####
```

```
#STAR transcript alignment#
```

```
#####
```

```
cd ~/STAR-2.7.9a/source/
```

```
STAR \
```

```
--runThreadN 2 \
```

```
--runMode genomeGenerate \
```

```
--genomeDir /home/patrick/cyano/genome/index/ \
```

```
--genomeFastaFiles
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.fna \
```

```
--sjdbGTFfile
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gtf \
```

```
--sjdbOverhang 74 \
```

```
--genomeSAindexNbases 11
```

```
STAR \
```

```
--runThreadN 2 \
```

```
--genomeDir \
```

```
/home/patrick/cyano/genome/index/ \
```

```
--readFilesIn
```

```
/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_trimmomatic_1.fq.gz
```

```
/home/patrick/cyano/trimmomatic/SRR11810829/SRR11810829_sortmerna_trimmomatic_2.fq.gz \
```

```
--readFilesCommand gunzip -c \
```

```
--outSAMtype BAM SortedByCoordinate \
```

```
--limitBAMsortRAM 1159145820 \
```

```
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810829/
```

```
cp /home/patrick/cyano/STAR/SRR11810829/Log.final.out
/home/patrick/cyano/QC/SRR11810829/Log.final.out
```

```
STAR \
--runThreadN 2 \
--genomeDir \
/home/patrick/cyano/genome/index/ \
--readFilesIn
/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_trimmomatic_
1.fq.gz
/home/patrick/cyano/trimmomatic/SRR11810828/SRR11810828_sortmerna_trimmomatic_
2.fq.gz \
--readFilesCommand gunzip -c \
--outSAMtype BAM SortedByCoordinate \
--limitBAMsortRAM 1159145820 \
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810828/
```

```
cp /home/patrick/cyano/STAR/SRR11810828/Log.final.out
/home/patrick/cyano/QC/SRR11810828/Log.final.out
```

```
STAR \
--runThreadN 2 \
--genomeDir \
/home/patrick/cyano/genome/index/ \
--readFilesIn
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_trimmomatic_
1.fq.gz
/home/patrick/cyano/trimmomatic/SRR11810827/SRR11810827_sortmerna_trimmomatic_
2.fq.gz \
--readFilesCommand gunzip -c \
--outSAMtype BAM SortedByCoordinate \
--limitBAMsortRAM 1159145820 \
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810827/
```

```
cp /home/patrick/cyano/STAR/SRR11810827/Log.final.out
/home/patrick/cyano/QC/SRR11810827/Log.final.out
```

```
STAR \
--runThreadN 2 \
--genomeDir \
/home/patrick/cyano/genome/index/ \
--readFilesIn
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_trimmomatic_
1.fq.gz
```

```
/home/patrick/cyano/trimmomatic/SRR11810826/SRR11810826_sortmerna_trimmomatic_2.fq.gz \  
--readFilesCommand gunzip -c \  
--outSAMtype BAM SortedByCoordinate \  
--limitBAMsortRAM 1159145820 \  
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810826/
```

```
cp /home/patrick/cyano/STAR/SRR11810826/Log.final.out  
/home/patrick/cyano/QC/SRR11810826/Log.final.out
```

```
STAR \  
--runThreadN 2 \  
--genomeDir \  
/home/patrick/cyano/genome/index/ \  
--readFilesIn  
/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_trimmomatic_1.fq.gz  
/home/patrick/cyano/trimmomatic/SRR11810825/SRR11810825_sortmerna_trimmomatic_2.fq.gz \  
--readFilesCommand gunzip -c \  
--outSAMtype BAM SortedByCoordinate \  
--limitBAMsortRAM 1159145820 \  
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810825/
```

```
cp /home/patrick/cyano/STAR/SRR11810825/Log.final.out  
/home/patrick/cyano/QC/SRR11810825/Log.final.out
```

```
STAR \  
--runThreadN 2 \  
--genomeDir \  
/home/patrick/cyano/genome/index/ \  
--readFilesIn  
/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_trimmomatic_1.fq.gz  
/home/patrick/cyano/trimmomatic/SRR11810824/SRR11810824_sortmerna_trimmomatic_2.fq.gz \  
--readFilesCommand gunzip -c \  
--outSAMtype BAM SortedByCoordinate \  
--limitBAMsortRAM 1159145820 \  
--outFileNamePrefix /home/patrick/cyano/STAR/SRR11810824/
```

```
cp /home/patrick/cyano/STAR/SRR11810824/Log.final.out  
/home/patrick/cyano/QC/SRR11810824/Log.final.out
```

```

#####
#QC#
#####
conda activate fastqc

fastqc \
-o /home/patrick/cyano/QC/SRR11810824/ \
/home/patrick/cyano/STAR/SRR11810824/Aligned.sortedByCoord.out.bam

fastqc \
-o /home/patrick/cyano/QC/SRR11810825/ \
/home/patrick/cyano/STAR/SRR11810825/Aligned.sortedByCoord.out.bam

fastqc \
-o /home/patrick/cyano/QC/SRR11810826/ \
/home/patrick/cyano/STAR/SRR11810826/Aligned.sortedByCoord.out.bam

fastqc \
-o /home/patrick/cyano/QC/SRR11810827/ \
/home/patrick/cyano/STAR/SRR11810827/Aligned.sortedByCoord.out.bam

fastqc \
-o /home/patrick/cyano/QC/SRR11810828/ \
/home/patrick/cyano/STAR/SRR11810828/Aligned.sortedByCoord.out.bam

fastqc \
-o /home/patrick/cyano/QC/SRR11810829/ \
/home/patrick/cyano/STAR/SRR11810829/Aligned.sortedByCoord.out.bam

conda deactivate
#####
#STAR Alignment assessment#
#####
#investigate overlapp

bedtools intersect \
-a /home/patrick/cyano/STAR/SRR11810829/Aligned.sortedByCoord.out.bam \
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \
-bed > /home/patrick/cyano/Bedtools/SR11810829/output.txt

bedtools intersect \
-a /home/patrick/cyano/STAR/SRR11810828/Aligned.sortedByCoord.out.bam \
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \
-bed > /home/patrick/cyano/Bedtools/SR11810828/output.txt

bedtools intersect \
-a /home/patrick/cyano/STAR/SRR11810827/Aligned.sortedByCoord.out.bam \

```

```
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \  
-bed > /home/patrick/cyano/Bedtools/SR11810827/output.txt
```

```
bedtools intersect \  
-a /home/patrick/cyano/STAR/SRR11810826/Aligned.sortedByCoord.out.bam \  
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \  
-bed > /home/patrick/cyano/Bedtools/SR11810826/output.txt
```

```
bedtools intersect \  
-a /home/patrick/cyano/STAR/SRR11810825/Aligned.sortedByCoord.out.bam \  
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \  
-bed > /home/patrick/cyano/Bedtools/SR11810825/output.txt
```

```
bedtools intersect \  
-a /home/patrick/cyano/STAR/SRR11810824/Aligned.sortedByCoord.out.bam \  
-b /home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff \  
-bed > /home/patrick/cyano/Bedtools/SR11810824/output.txt
```

```
#####
```

```
#Salmon preprocessing#
```

```
#####
```

```
#creating transcript file from genome and gff annotation
```

```
gffread -x transcript.fasta -g
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.fna
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.gff
```

```
# FASTA index file
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.fna.fai
```

```
created.
```

```
#producing the decoy index, firstly extracting the gene names from the genome
```

```
grep "^>" <
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.fna | cut -d " "
```

```
-f 1 > /home/patrick/cyano/genome/decoys.txt
```

```
sed -i.bak -e 's/>//g' /home/patrick/cyano/genome/decoys.txt
```

```
#attaching the genome and transcriptome together to produce the gentrome to use for the  
salmon index
```

```
cat /home/patrick/cyano/genome/transcript.fasta
```

```
/home/patrick/cyano/genome/GCF_000317285.1_ChIPCC6912_1.0_genomic.fna >
```

```
/home/patrick/cyano/genome/gentrome.fa.gz
```

```
#creating the salmon index
```

```
conda activate salmon
```

```
salmon index -t /home/patrick/cyano/genome/gentrome.fa.gz -d  
/home/patrick/cyano/genome/decoys.txt -p 12 -i  
/home/patrick/cyano/Salmon/Index/salmon_index --gencode
```

```
conda deactivate
```

```
# converting Bam file to fasta :)
```

```
#sort the Bam file with samtools  
conda activate samtools
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810829/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810829/Aligned.sortedByCoord.out.bam
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810828/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810828/Aligned.sortedByCoord.out.bam
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810827/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810827/Aligned.sortedByCoord.out.bam
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810826/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810826/Aligned.sortedByCoord.out.bam
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810825/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810825/Aligned.sortedByCoord.out.bam
```

```
samtools sort -n -o /home/patrick/cyano/Salmon/SRR11810824/aln.qsort.bam  
/home/patrick/cyano/STAR/SRR11810824/Aligned.sortedByCoord.out.bam
```

```
conda deactivate
```

```
#Bam to fastq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810829/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810829/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810829/aln.end2.fq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810828/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810828/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810828/aln.end2.fq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810827/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810827/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810827/aln.end2.fq
```

```
-fq /home/patrick/cyano/Salmon/SRR11810827/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810827/aln.end2.fq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810826/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810826/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810826/aln.end2.fq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810825/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810825/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810825/aln.end2.fq
```

```
bedtools bamtofastq -i /home/patrick/cyano/Salmon/SRR11810824/aln.qsort.bam \  
-fq /home/patrick/cyano/Salmon/SRR11810824/aln.end1.fq \  
-fq2 /home/patrick/cyano/Salmon/SRR11810824/aln.end2.fq
```

```
#####
```

```
#FastQC#
```

```
#####
```

```
conda activate fastqc
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810824/ \  
/home/patrick/cyano/Salmon/SRR11810824/aln.end*.fq
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810825/ \  
/home/patrick/cyano/Salmon/SRR11810825/aln.end*.fq
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810826/ \  
/home/patrick/cyano/Salmon/SRR11810826/aln.end*.fq
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810827/ \  
/home/patrick/cyano/Salmon/SRR11810827/aln.end*.fq
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810828/ \  
/home/patrick/cyano/Salmon/SRR11810828/aln.end*.fq
```

```
fastqc \  
-o /home/patrick/cyano/QC/SRR11810829/ \  
/home/patrick/cyano/Salmon/SRR11810829/aln.end*.fq
```

```
conda deactivate
```

```
#####
```

```
#Salmon Transcript Quantification#
```

```
#####
```

```
conda activate salmon
```

```
salmon quant \
```

```
-i /home/patrick/cyano/Salmon/Index/salmon_index \
```

```
-l A \
```

```
-1 /home/patrick/cyano/Salmon/SRR11810829/aln.end1.fq \
```

```
-2 /home/patrick/cyano/Salmon/SRR11810829/aln.end2.fq \
```

```
--validateMappings \
```

```
--dumpEq \
```

```
--numGibbsSamples 100 \
```

```
--seqBias \
```

```
--gcBias \
```

```
--posBias \
```

```
-o /home/patrick/cyano/Salmon/SRR11810829/
```

```
cp /home/patrick/cyano/Salmon/SRR11810829/logs/salmon_quant.log
```

```
/home/patrick/QC/SRR11810829/salmon_quant.log
```

```
salmon quant \
```

```
-i /home/patrick/cyano/Salmon/Index/salmon_index \
```

```
-l A \
```

```
-1 /home/patrick/cyano/Salmon/SRR11810828/aln.end1.fq \
```

```
-2 /home/patrick/cyano/Salmon/SRR11810828/aln.end2.fq \
```

```
--validateMappings \
```

```
--dumpEq \
```

```
--numGibbsSamples 100 \
```

```
--seqBias \
```

```
--gcBias \
```

```
--posBias \
```

```
-o /home/patrick/cyano/Salmon/SRR11810828/
```

```
cp /home/patrick/cyano/Salmon/SRR11810828/logs/salmon_quant.log
```

```
/home/patrick/QC/SRR11810828/salmon_quant.log
```

```
salmon quant \
```

```
-i /home/patrick/cyano/Salmon/Index/salmon_index \
```

```
-l A \
```

```
-1 /home/patrick/cyano/Salmon/SRR11810827/aln.end1.fq \
```

```
-2 /home/patrick/cyano/Salmon/SRR11810827/aln.end2.fq \
```

```
--validateMappings \
```

```
--dumpEq \
```

```
--numGibbsSamples 100 \
```

```
--seqBias \
```

```
--gcBias \
```



```
--posBias \  
-o /home/patrick/cyano/Salmon/SRR11810827/
```

```
cp /home/patrick/cyano/Salmon/SRR11810827/logs/salmon_quant.log  
/home/patrick/QC/SRR11810827/salmon_quant.log
```

```
salmon quant \  
-i /home/patrick/cyano/Salmon/Index/salmon_index \  
-l A \  
-1 /home/patrick/cyano/Salmon/SRR11810826/aln.end1.fq \  
-2 /home/patrick/cyano/Salmon/SRR11810826/aln.end2.fq \  
--validateMappings \  
--dumpEq \  
--numGibbsSamples 100 \  
--seqBias \  
--gcBias \  
--posBias \  
-o /home/patrick/cyano/Salmon/SRR11810826/
```

```
cp /home/patrick/cyano/Salmon/SRR11810826/logs/salmon_quant.log  
/home/patrick/QC/SRR11810826/salmon_quant.log
```

```
salmon quant \  
-i /home/patrick/cyano/Salmon/Index/salmon_index \  
-l A \  
-1 /home/patrick/cyano/Salmon/SRR11810825/aln.end1.fq \  
-2 /home/patrick/cyano/Salmon/SRR11810825/aln.end2.fq \  
--validateMappings \  
--dumpEq \  
--numGibbsSamples 100 \  
--seqBias \  
--gcBias \  
--posBias \  
-o /home/patrick/cyano/Salmon/SRR11810825/
```

```
cp /home/patrick/cyano/Salmon/SRR11810825/logs/salmon_quant.log  
/home/patrick/QC/SRR11810825/salmon_quant.log
```

```
salmon quant \  
-i /home/patrick/cyano/Salmon/Index/salmon_index \  
-l A \  
-1 /home/patrick/cyano/Salmon/SRR11810824/aln.end1.fq \  
-2 /home/patrick/cyano/Salmon/SRR11810824/aln.end2.fq \  
--validateMappings \  
--dumpEq \  
--numGibbsSamples 100 \  
--seqBias \  
-o /home/patrick/cyano/Salmon/SRR11810824/
```

```

--gcBias \
--posBias \
-o /home/patrick/cyano/Salmon/SRR11810824/

cp /home/patrick/cyano/Salmon/SRR11810824/logs/salmon_quant.log
/home/patrick/QC/SRR11810824/salmon_quant.log

conda deactivate

#####
#QC#
#####

# unable to be complete on .sf files however .sf files can be used in multiqc

cp /home/patrick/cyano/Salmon/SRR11810824/quant.sf
/home/patrick/cyano/QC/SRR11810824/

cp /home/patrick/cyano/Salmon/SRR11810825/quant.sf
/home/patrick/cyano/QC/SRR11810825/

cp /home/patrick/cyano/Salmon/SRR11810826/quant.sf
/home/patrick/cyano/QC/SRR11810826/

cp /home/patrick/cyano/Salmon/SRR11810827/quant.sf
/home/patrick/cyano/QC/SRR11810827/

cp /home/patrick/cyano/Salmon/SRR11810828/quant.sf
/home/patrick/cyano/QC/SRR11810828/

cp /home/patrick/cyano/Salmon/SRR11810829/quant.sf
/home/patrick/cyano/QC/SRR11810829/

#####
#MultiQC#
#####

conda activate multiqc

multiqc \
-o /home/patrick/cyano/QC/SRR11810824/MultiQC/ \
/home/patrick/cyano/QC/SRR11810824/

multiqc \
-o /home/patrick/cyano/QC/SRR11810825/MultiQC/ \
/home/patrick/cyano/QC/SRR11810825/

```

```
multiqc \  
-o /home/patrick/cyano/QC/SRR11810826/MultiQC/ \  
/home/patrick/cyano/QC/SRR11810826/
```

```
multiqc \  
-o /home/patrick/cyano/QC/SRR11810827/MultiQC/ \  
/home/patrick/cyano/QC/SRR11810827/
```

```
multiqc \  
-o /home/patrick/cyano/QC/SRR11810828/MultiQC/ \  
/home/patrick/cyano/QC/SRR11810828/
```

```
multiqc \  
-o /home/patrick/cyano/QC/SRR11810829/MultiQC/ \  
/home/patrick/cyano/QC/SRR11810829/
```

```
conda deactivate  
#####  
#Data Quality control complete#  
#####
```

Code within R Studio

```
# download correct packages
#deseq2
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DESeq2")

#edgeR

BiocManager::install("edgeR")

#here
install.packages("here")

#LSD
install.packages("LSD")

#RColourBrewer
install.packages("RColorBrewer")

#summarised experiment

BiocManager::install("SummarizedExperiment")

#Tidyverse
install.packages("tidyverse")

#hexbin
install.packages("hexbin")

#tximport
BiocManager::install("tximport")

# vsn
BiocManager::install("vsn")

#Rsubread
BiocManager::install("Rsubread")

#pheatmap
install.packages("pheatmap")

#VennDiagram
install.packages("VennDiagram")
```

```

#textshape
install.packages("textshape")

#gplots
install.packages("gplots")

#dplyr
install.packages("dplyr")

suppressPackageStartupMessages({
  library(DESeq2)
  library(edgeR)
  library(here)
  library(LSD)
  library(SummarizedExperiment)
  library(tidyverse)
  library(tximport)
  library(vsn)
  library(Rsubread)
  library(hexbin)
  library(pheatmap)
  library(RColorBrewer)
  library(VennDiagram)
  library(readr)
  library(tidyr)
  library(dplyr)
  library(textshape)
  library(gplots)
  library(plotly)
  library(hyperSpec)
})

#creating line plotting

"line_plot" <- function(dds=ds_se,vst=vsd,gene_id=gene_id){
  message(paste("Plotting",gene_id))
  sel <- grepl(gene_id,rownames(vsd))
  stopifnot(sum(sel)==1)

  p <- ggplot(bind_cols(as.data.frame(colData(ds_se)),
    data.frame(value=vsd[sel,])),
    aes(x=MDay,y=value,col=MGenotype,group=MGenotype)) +
  geom_point() + geom_smooth() +
  scale_y_continuous(name="VST expression") +
  ggtitle(label=paste("Expression for: ",gene_id))
}

```

```

suppressMessages(suppressWarnings(plot(p)))
return(NULL)
}
##### extracting results script #####
mar <- par("mar")

#Extracting deseq results from deseq dataset script:
"ExtractResults" <- function(dds,vst,contrast,
                             padj=0.01,lfc=0.5,
                             plot=TRUE,verbose=TRUE,

export=TRUE,default_dir=here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/DE/")),
                             default_prefix="DE-",
                             labels=colnames(dds),
                             sample_sel=1:ncol(dds),
                             expression_cutoff=0,
                             debug=FALSE,filter=c("median",NULL),...){

# get the filter
if(!is.null(match.arg(filter))){
  filter <- rowMedians(counts(dds,normalized=TRUE))
  message("Using the median normalized counts as default, set filter=NULL to revert to
using the mean")
}

# validation
if(length(contrast)==1){
  res <- results(dds,name=contrast,filter = filter,lfcThreshold=lfc,alpha=padj)
} else {
  res <- results(dds,contrast=contrast,filter = filter,lfcThreshold=lfc,alpha=padj)
}

stopifnot(length(sample_sel)==ncol(vst))

if(plot){
  par(mar=c(5,5,5,5))
  volcanoPlot(res)
  par(mar=mar)
}

# a look at independent filtering
if(plot){
  plot(metadata(res)$filterNumRej,
        type="b", ylab="number of rejections",
        xlab="quantiles of filter")
}

```

```

lines(metadata(res)$lo.fit, col="red")
abline(v=metadata(res)$filterTheta)
}

if(verbose){
  message(sprintf("The independent filtering cutoff is %s, removing %s of the data",
    round(metadata(res)$filterThreshold,digits=5),
    names(metadata(res)$filterThreshold)))

  max.theta <-
metadata(res)$filterNumRej[which.max(metadata(res)$filterNumRej$numRej),"theta"]
  message(sprintf("The independent filtering maximises for %s %% of the data,
corresponding to a base mean expression of %s (library-size normalised read)",
    round(max.theta*100,digits=5),
    round(quantile(counts(dds,normalized=TRUE),probs=max.theta),digits=5)))
}

if(plot){

qtl.exp=quantile(counts(dds,normalized=TRUE),probs=metadata(res)$filterNumRej$theta)
dat <- data.frame(thetas=metadata(res)$filterNumRej$theta,
  qtl.exp=qtl.exp,
  number.degs=sapply(lapply(qtl.exp,function(qe){
    res$padj <= padj & abs(res$log2FoldChange) >= lfc &
    ! is.na(res$padj) & res$baseMean >= qe
  })),sum))
if(debug){
  plot(ggplot(dat,aes(x=thetas,y=qtl.exp)) +
    geom_line() + geom_point() +
    scale_x_continuous("quantiles of expression") +
    scale_y_continuous("base mean expression") +
    geom_hline(yintercept=expression_cutoff,
      linetype="dotted",col="red"))

  p <- ggplot(dat,aes(x=thetas,y=qtl.exp)) +
    geom_line() + geom_point() +
    scale_x_continuous("quantiles of expression") +
    scale_y_log10("base mean expression") +
    geom_hline(yintercept=expression_cutoff,
      linetype="dotted",col="red")
  suppressMessages(suppressWarnings(plot(p)))

  plot(ggplot(dat,aes(x=thetas,y=number.degs)) +
    geom_line() + geom_point() +
    geom_hline(yintercept=dat$number.degs[1],linetype="dashed") +
    scale_x_continuous("quantiles of expression") +
    scale_y_continuous("Number of DE genes"))
}

```

```

plot(ggplot(dat,aes(x=thetas,y=number.degs[1] - number.degs),aes()) +
  geom_line() + geom_point() +
  scale_x_continuous("quantiles of expression") +
  scale_y_continuous("Cumulative number of DE genes"))

plot(ggplot(data.frame(x=dat$thetas[-1],
  y=diff(dat$number.degs[1] - dat$number.degs)),aes(x,y)) +
  geom_line() + geom_point() +
  scale_x_continuous("quantiles of expression") +
  scale_y_continuous("Number of DE genes per interval"))

plot(ggplot(data.frame(x=dat$ctl.exp[-1],
  y=diff(dat$number.degs[1] - dat$number.degs)),aes(x,y)) +
  geom_line() + geom_point() +
  scale_x_continuous("base mean of expression") +
  scale_y_continuous("Number of DE genes per interval"))

p <- ggplot(data.frame(x=dat$ctl.exp[-1],
  y=diff(dat$number.degs[1] - dat$number.degs)),aes(x,y)) +
  geom_line() + geom_point() +
  scale_x_log10("base mean of expression") +
  scale_y_continuous("Number of DE genes per interval") +
  geom_vline(xintercept=expression_cutoff,
  linetype="dotted",col="red")
suppressMessages(suppressWarnings(plot(p)))
}
}

sel <- res$padj <= padj & abs(res$log2FoldChange) >= lfc & ! is.na(res$padj) &
res$baseMean >= expression_cutoff

if(verbose){
  message(sprintf(paste(
    ifelse(sum(sel)==1,
      "There is %s gene that is DE",
      "There are %s genes that are DE"),
    "with the following parameters: FDR <= %s, |log2FC| >= %s, base mean expression >
%s"),
    sum(sel),padj,
    lfc,expression_cutoff))
}

# proceed only if there are DE genes
if(sum(sel) > 0){
  val <- rowSums(vst[sel,sample_sel,drop=FALSE])==0
  if (sum(val) >0){

```



```

warning(sprintf(paste(
  ifelse(sum(val)==1,
    "There is %s DE gene that has",
    "There are %s DE genes that have"),
  "no vst expression in the selected samples"),sum(val)))
sel[sel][val] <- FALSE
}

if(export){
  if(!dir.exists(default_dir)){
    dir.create(default_dir,showWarnings=FALSE,recursive=TRUE,mode="0771")
  }
  write.csv(res,file=file.path(default_dir,paste0(default_prefix,"results.csv")))
  write.csv(res[sel,],file.path(default_dir,paste0(default_prefix,"genes.csv")))
}
if(plot & sum(sel)>1){
  heatmap.2(t(scale(t(vst[sel,sample_sel]))),
    distfun = pearson.dist,
    hclustfun = function(X){hclust(X,method="ward.D2")},
    trace="none",col=hpal,labRow = FALSE,
    labCol=labels[sample_sel],...
  )
}
}
return(list(all=rownames(res[sel,]),
  up=rownames(res[sel & res$log2FoldChange > 0,]),
  dn=rownames(res[sel & res$log2FoldChange < 0,])))
}

##### Data Analysis #####

# import metatdata

meta <- read_tsv(file.path("~/Documents/Univeristy/MRes/Research Project/BioInf/",
"Metadata.txt"),col_types=cols(.default=col_factor()))

#test metadata
meta$light
levels(meta$light)
as.integer(meta$light)
#metadata shows that it is read as 6 samples that are either light or far-red

# importing Salmon data
#setting file location

```

```

salmonfiles <- file.path("~/Documents/Univeristy/MRes/Research Project/Biolnf/",
"salmon", meta$sample,"quant.sf")
names(salmonfiles) <- meta$sample
stopifnot(all(file.exists(salmonfiles)))

#importing data
sg <- tximport(files=salmonfiles,type="salmon",txOut=TRUE)

#rounding estimated count to integers
counts_salmon <- round(sg$counts)
rownames(counts_salmon) <- sub("\\.1$", "",rownames(counts_salmon))

# sequencing depth
dat <- tibble(x=colnames(counts_salmon),y=colSums(counts_salmon)) %>%
  bind_cols(meta)

ggplot(dat,aes(x,y,fill=light)) + geom_col() +
  scale_y_continuous(name="reads") +
  theme(axis.text.x=element_text(angle=90,size=10),axis.title.x=element_blank())

#creating the model for DESeq2
meta <- meta %>% column_to_rownames("sample")
meta
#the sample names are now rownames instead of in the first column.

stopifnot(all(colnames(counts_salmon) == rownames(meta)))
#both colnames and rownames match.

# creating DESeq
ds_se <- DESeqDataSetFromTximport(txi=sg, colData=meta ,
  design = ~ light)

#creating dgelist for edgeR
genetable <- data.frame(gene.id = rownames(counts_salmon),
  stringsAsFactors = FALSE)
stopifnot(all(rownames(meta) == colnames(counts_salmon)))
dge <- DGEList(counts = counts_salmon,
  samples = meta,
  genes = genetable)
names(dge)

#averaget transcript lengths are then determined to identify offsets which are added
manually to the dge
avetxlengths <- sg$length
rownames(avetxlengths) <- sub("\\.1$", "",rownames(avetxlengths))

```

```

#checking gene names concur with salmon could dataset
stopifnot(all(rownames(avetxlengths) == rownames(counts_salmon)))
stopifnot(all(colnames(avetxlengths) == colnames(counts_salmon)))
#they do concur

avetxlengths <- avetxlengths/exp(rowMeans(log(avetxlengths)))

offsets <- log(calcNormFactors(counts_salmon/avetxlengths)) +
  log(colSums(counts_salmon/avetxlengths))
dge <- scaleOffset(dge, t(t(log(avetxlengths)) + offsets))
names(dge)

#analysing the data (look at standard deviation of genes at different expression levels)
meanSdPlot(log(assay(ds_se)[rowSums(assay(ds_se))>30,]), ylab = "standard deviation (+/-
)", xlab = "Ranked mean espression")
#This shows that a higher mean expression causes a greater sd more often than in lower
expressed genes.
#This can affect data during the analysis as the data is heteroskedastic and therefore
during PCA there may be
#greater reliance on higher expressed genes.
#to reduce this we can use variance stabilizing transformation (VST) which will transform
the results to have
# an even variance of expression no matter the expression, making the data
homoskedastic.
#This transforms the variance to be independent of the mean.

vsd <- DESeq2::vst(ds_se)
vst <- assay(vsd)
vst <- vst - min(vst)

meanSdPlot(log(assay(vsd)[rowSums(assay(vsd))>0,]), ylab = "standard deviation (+/-)", xlab
= "Ranked mean espression")
#now there is more variety throughout.

#PCA Plot of all genes

pc <- prcomp(t(vst))
percent <- round(summary(pc)$importance[2,]*100)
#define number of variables in the experiment
nvar=1
nlevel=nlevels(dds$light)

pc.dat <- bind_cols(PC1=pc$x[,1],
  PC2=pc$x[,2],
  as.data.frame(colData(dds)))

p <- ggplot(pc.dat,aes(x=PC1,y=PC2,col=light,shape=light,text=light)) +

```

```

geom_point(size=2) +
ggtitle("Principal Component Analysis",subtitle="variance stabilized counts")

p

#PCA graph made

ggplotly(p) %>% layout(xaxis=list(title=paste("PC1 (",percent[1],"%",sep="")),
  yaxis=list(title=paste("PC2 (",percent[2],"%",sep="")))

# PCA if fatty acid genes #####

FA_vst <- vst[unlist(all_genes), ]

FA_pc <- prcomp(t(FA_vst))

FA_percent <- round(summary(FA_pc)$importance[2,]*100)

FA_pc.dat <- bind_cols(PC1=FA_pc$x[,1],
  PC2=FA_pc$x[,2],
  as.data.frame(colData(dds)))

FA_p <- ggplot(FA_pc.dat,aes(x=PC1,y=PC2,col=light,shape=light,text=light)) +
  geom_point(size=2) +
  ggtitle("Principal Component Analysis of Fatty Acid genes",subtitle="variance stabilized
counts")

FA_p

ggplotly(FA_p) %>% layout(xaxis=list(title=paste("PC1 (",FA_percent[1],"%",sep="")),
  yaxis=list(title=paste("PC2 (",FA_percent[2],"%",sep="")))

#saving data for expression analysis _if needed_
rownames(ds_se) <- sub("\\.1","",rownames(ds_se))
dir.create(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/","DE"),showWarnings=FALSE)
save(ds_se,dge,vsd,file=file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/","DE","differentialExpressionObjects.rda"))

#' * Helpers
source(here("R/volcanoPlot.R"))

#' * Graphics
pal <- brewer.pal(8,"Dark2")

```

```

#looking at dispersion estimates
ds_se <- DESeq2::DESeq(ds_se)
#plotting
DESeq2::plotDispEsts(ds_se)
#majority do not vary in dispersion and data is accepted.

#extracting results

ExtractResults(dds = ds_se, vst = vst, contrast = "light_far.red_vs_white")

DE.results <- read.csv("~/Documents/Univeristy/MRes/Research Project/BioInf/DE/DE-
results.csv")
DE.results <- DE.results %>% column_to_rownames("X")

FAS_RES <- data.frame(DE.results[FAS,])
view(FAS_RES)
write.csv(FAS_RES,file=file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/DE/", "FAS_results.csv"))

FAD_RES <- data.frame(DE.results[FAD,])
view(FAD_RES)
write.csv(FAS_RES,file=file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/DE/", "FAD_results.csv"))

oleate_RES <- data.frame(DE.results[oleate,])
view(oleate_RES)
write.csv(oleate_RES,file=file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/DE/", "oleate_results.csv"))

PfaD_RES <- data.frame(DE.results[PfaD,])
view(PfaD_RES)
write.csv(PfaD_RES,file=file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/DE/", "PfaD_results.csv"))

# volcano plot to identify DEgenes
volcanoPlot(resSchurch)
#yellow shows there is a high density of genes that arent DE, this is good.

# producing heatmap
mat <- assay(vsd)[head(order(resSchurch$padj), 30), ]
mat <- mat - rowMeans(mat)
df <- as.data.frame(colData(vsd)[, c("light")])
rownames(df) <- colnames(mat)
pheatmap(mat, annotation_col = df)

```

```

#Importing genes of interest

FAS <- read_lines(here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/", "FAS_genes.txt")))
FAD <- read_lines(here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/", "FAD_genes.txt")))
oleate <- read_lines(here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/", "oleate_synth.txt")))
PfaD <- read_lines(here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/", "PfaD_genes.txt")))
allgenes <- read_lines(here(file.path("~/Documents/Univeristy/MRes/Research
Project/BioInf/", "all_genes.txt")))

# creating colour palette
hpal <- colorRampPalette(c("blue", "white", "red"))(100)

#producing heatmaps
zscore <- t(scale(t(vst[rowSums(vst)>0,])))

FAS_Heatmap <- heatmap.2(t(scale(t(vst)))[FAS,],
  distfun=pearson.dist,
  hclustfun=function(X){hclust(X,method="ward.D2")},
  breaks=seq(min(zscore),max(zscore),length.out=length(hpal)+1),
  trace = "none",
  labCol = conds,
  col=hpal, margins = c(10,12))

FAD_Heatmap <- heatmap.2(t(scale(t(vst)))[FAD,],
  distfun=pearson.dist,
  hclustfun=function(X){hclust(X,method="ward.D2")},
  breaks=seq(min(zscore),max(zscore),length.out=length(hpal)+1),
  trace = "none",
  labCol = conds,
  col=hpal, margins = c(6,15))

oleate_Heatmap <- heatmap.2(t(scale(t(vst)))[oleate,],
  distfun=pearson.dist,
  hclustfun=function(X){hclust(X,method="ward.D2")},
  breaks=seq(min(zscore),max(zscore),length.out=length(hpal)+1),
  trace = "none",
  labCol = conds,
  col=hpal, margins = c(6,15))

PfaD_Heatmap <- heatmap.2(t(scale(t(vst)))[PfaD,],
  distfun=pearson.dist,
  hclustfun=function(X){hclust(X,method="ward.D2")},

```

```
breaks=seq(min(zscore),max(zscore),length.out=length(hpal)+1),  
trace = "none",  
labCol = conds,  
cexRow = 1.5,  
col=hpal, margins = c(6,15))
```

```
all_Heatmap <- heatmap.2(t(scale(t(vst)))[allgenes,],  
distfun=pearson.dist,  
hclustfun=function(X){hclust(X,method="ward.D2")},  
breaks=seq(min(zscore),max(zscore),length.out=length(hpal)+1),  
labRow = allgenes,trace = "none",  
labCol = conds,  
col=hpal, margins = c(6,10))
```

References

- Al-Haj, L., Lui, Y. T., Abed, R. M., Gomaa, M. A., & Purton, S. (2016). Cyanobacteria as chassis for industrial biotechnology: progress and prospects. *Life*, 6(4), 42.
- Allen, E. E., & Bartlett, D. H. (2002). Structure and regulation of the omega-3 polyunsaturated fatty acid synthase genes from the deep-sea bacterium *Photobacterium profundum* strain SS9. The GenBank accession numbers for the sequences reported in this paper are AF409100 and AF467805. *Microbiology*, 148(6), 1903-1913.
- Amiri-Jami, M., & Griffiths, M. W. (2010). Recombinant production of omega-3 fatty acids in *Escherichia coli* using a gene cluster isolated from *Shewanella baltica* MAC1. *Journal of Applied Microbiology*, 109(6), 1897-1905.
- Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Arico, D., Legris, M., Castro, L., Garcia, C. F., Laino, A., Casal, J. J., & Mazzella, M. A. (2019). Neighbour signals perceived by phytochrome B increase thermotolerance in *Arabidopsis*. *Plant, Cell & Environment*, 42(9), 2554-2566.
- Behrendt, L., Trampe, E.L., Nord, N.B., Nguyen, J., Kuhl, M., Lonco, D., Nyarko, A., Dhinojwala, A., Hershey, O.S. & Barton, H. (2020). Life in the dark: far-red absorbing cyanobacteria extend photic zones deep into terrestrial caves. *Environmental Microbiology*, 22(3), 952-963.
- Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
- Brown, M. R., Barrett, S. M., Volkman, J. K., Nearhos, S. P., Nell, J. A., & Allan, G. L. (1996). Biochemical composition of new yeasts and bacteria evaluated as food for bivalve aquaculture. *Aquaculture*, 143(3-4), 341-360.
- Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P. & Karp, P.D. (2020). The MetaCyc database of metabolic pathways and enzymes—a 2019 update. *Nucleic Acids Research*, 48(D1), D445-D453.
- Chen, M., Schliep, R. D., Willows, Z. L., Cai, B. A., Neilan & Scheer, H. (2010). A red-shifted chlorophyll. *Science*, 329, 1318–1319.
- Choi-Rhee, E., & Cronan, J. E. (2003). The biotin carboxylase-biotin carboxyl carrier protein complex of *Escherichia coli* acetyl-CoA carboxylase. *Journal of Biological Chemistry*, 278(33), 30806-30812.

Coleman, J. (1992). Characterization of the Escherichia coli gene for 1-acyl-sn-glycerol-3-phosphate acyltransferase (plsC). *Molecular and General Genetics MGG*, 232(2), 295-303.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 1-19.

Cross, E. M., Adams, F. G., Waters, J. K., Aragão, D., Eijkelkamp, B. A., & Forwood, J. K. (2021). Insights into *Acinetobacter baumannii* fatty acid synthesis 3-oxoacyl-ACP reductases. *Scientific Reports*, 11(1), 1-16.

Dagan, T., Roettger, M., Stucken, K., Landan, G., Koch, R., Major, P., Gould, S.B., Goremykin, V.V., Rippka, R., Tandeau de Marsac, N. & Gugger, M. (2013). Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biology and Evolution*, 5(1), 31-44.

Dainat, J., Hereñú, D., & Pucholt, P. (2021). NBISweden/AGAT: AGAT-v0.6.0 (v0.6.0). *Zenodo*. <https://doi.org/10.5281/zenodo.4637977>

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), giab008.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21.

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rättsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R. & Bertone, P. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 10(12), 1185-1191.

Evans, E. H., Foulds, I., & Carr, N. G. (1976). Environmental conditions and morphological variation in the blue-green alga *Chlorogloea fritschii*. *Microbiology*, 92(1), 147-155

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047-3048.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8(3), 175-185.

Fujisawa, T., Narikawa, R., Maeda, S.I., Watanabe, S., Kanesaki, Y., Kobayashi, K., Nomata, J., Hanaoka, M., Watanabe, M., Ehira, S. & Suzuki, E. (2017). CyanoBase: a large-scale update on its 20th anniversary. *Nucleic Acids Research*, 45(D1), D551-D554.

- Gan, F., Shen, G., & Bryant, D. A. (2014). Occurrence of far-red light photoacclimation (FaRLiP) in diverse cyanobacteria. *Life*, *5*(1), 4-24.
- Gan, F., Zhang, S., Rockwell, N. C., Martin, S. S., Lagarias, J. C., & Bryant, D. A. (2014). Extensive remodeling of a cyanobacterial photosynthetic apparatus in far-red light. *Science*, *345*(6202), 1312-1317.
- Gaysina, L. A., Saraf, A., & Singh, P. (2019). Cyanobacteria in diverse habitats. In *Cyanobacteria* (pp. 1-28). Academic Press.
- Hanschen, E. R., & Starkenburg, S. R. (2020). The state of algal genome quality and diversity. *Algal Research*, *50*, 101968.
- Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, *38*(12), e131-e131.
- Hernández-Prieto, M. A., Li, Y., Postier, B. L., Blankenship, R. E., & Chen, M. (2018). Far-red light promotes biofilm formation in the cyanobacterium *Acaryochloris marina*. *Environmental Microbiology*, *20*(2), 535-545.
- Ho, M. Y., & Bryant, D. A. (2019). Global transcriptional profiling of the cyanobacterium *Chlorogloeopsis fritschii* PCC 9212 in far-red light: insights into the regulation of chlorophyll d synthesis. *Frontiers in Microbiology*, *10*, 465.
- Hölzl, G., & Dörmann, P. (2007). Structure and function of glycoacyl lipids in plants and bacteria. *Progress in Lipid Research*, *46*(5), 225-243.
- Huber, W., Von Heydebreck, A., Sülthmann, H., Poustka, A. & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, *18*(suppl_1), S96-S104.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.
- Jónasdóttir, S. H. (2019). Fatty acid profiles and production in marine phytoplankton. *Marine Drugs*, *17*(3), 151.
- Kaczmarzyk, D., & Fulda, M. (2010). Fatty acid activation in cyanobacteria mediated by acyl-acyl carrier protein synthetase enables fatty acid recycling. *Plant Physiology*, *152*(3), 1598-1610.
- Kannan, N., Rao, A. S., & Nair, A. (2021). Microbial production of omega-3 fatty acids: an overview. *Journal of Applied Microbiology*, *131*(5), 2114-2130.

Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P. & Spaulding, A. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 17(5), 877-890.

Kenyon, C. N., Rippka, R., & Stanier, R. Y. (1972). Fatty acid composition and physiological properties of some filamentous blue-green algae. *Archiv für Mikrobiologie*, 83(3), 216-236.

Khandekar, S.S., Gentry, D.R., Van Aller, G.S., Warren, P., Xiang, H., Silverman, C., Doyle, M.L., Chambers, P.A., Konstantinidis, A.K., Brandt, M. & Daines, R.A. (2001). Identification, Substrate Specificity, and Inhibition of the *Streptococcus pneumoniae* β -Ketoacyl-Acyl Carrier Protein Synthase III (FabH). *Journal of Biological Chemistry*, 276(32), 30024-30030.

Khatoon, H., Leong, L. K., Rahman, N. A., Mian, S., Begum, H., Banerjee, S., & Endut, A. (2018). Effects of different light source and media on growth and production of phycobiliprotein from freshwater cyanobacteria. *Bioresource Technology*, 249, 652-658.

Kimber, M. S., Martin, F., Lu, Y., Houston, S., Vedadi, M., Dharamsi, A., Fiebig, K. M., Schmid, M., & Rock, C. O. (2004). The structure of (3R)-hydroxyacyl-acyl carrier protein dehydratase (FabZ) from *Pseudomonas aeruginosa*. *Journal of Biological Chemistry*, 279(50), 52593-52602.

Kizawa, A., Kawahara, A., Takashima, K., Takimura, Y., Nishiyama, Y., & Hihara, Y. (2017). The LexA transcription factor regulates fatty acid biosynthetic genes in the cyanobacterium *Synechocystis sp.* PCC 6803. *The Plant Journal*, 92(2), 189-198.

Koga, Y. (2012). Thermal adaptation of the archaeal and bacterial lipid membranes. *Archaea*, 2012, 789652.

Konishi, T. (2015). Principal component analysis for designed experiments. *BMC Bioinformatics*, 16(18), 1-9.

Kopylova, E., Noé, L. & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24), 3211-3217.

Kumar, B. R., Deviram, G., Mathimani, T., Duc, P. A., & Pugazhendhi, A. (2019). Microalgae as rich source of polyunsaturated fatty acids. *Biocatalysis and Agricultural Biotechnology*, 17, 583-588.

Kuo, J., & Khosla, C. (2014). The initiation ketosynthase (FabH) is the sole rate-limiting enzyme of the fatty acid synthase of *Synechococcus sp.* PCC 7002. *Metabolic Engineering*, 22, 53-59.

- Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M. & Maza, E. (2018). Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. *Frontiers in Plant Science*, *9*, 108.
- Latifi, A., Ruiz, M., & Zhang, C. C. (2009). Oxidative stress in cyanobacteria. *FEMS Microbiology Reviews*, *33*(2), 258-278.
- Li, Y., Lin, Y., Garvey, C.J., Birch, D., Corkery, R.W., Loughlin, P.C., Scheer, H., Willows, R.D. & Chen, M. (2016). Characterization of red-shifted phycobilisomes isolated from the chlorophyll f-containing cyanobacterium *Halomicronema hongdechloris*. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, *1857*(1), 107-114.
- Liao, Y., & Shi, W. (2019). Read trimming is not required for mapping and quantification of RNA-seq reads. *BioRxiv*, 833962.
- Liu, D., Liberton, M., Hendry, J. I., Aminian-Dehkordi, J., Maranas, C. D., & Pakrasi, H. B. (2021). Engineering biology approaches for food and nutrient production by cyanobacteria. *Current Opinion in Biotechnology*, *67*, 1-6.
- Liu, N., Cummings, J. E., England, K., Slayden, R. A., & Tonge, P. J. (2011). Mechanism and inhibition of the FabI enoyl-ACP reductase from *Burkholderia pseudomallei*. *Journal of Antimicrobial Chemotherapy*, *66*(3), 564-573.
- Liu, X., Sheng, J., & Curtiss III, R. (2011). Fatty acid production in genetically modified cyanobacteria. *Proceedings of the National Academy of Sciences*, *108*(17), 6899-6904.
- Llewellyn, C.A., Greig, C., Silkina, A., Kultschar, B., Hitchings, M.D. & Farnham, G. (2020). Mycosporine-like amino acid and aromatic amino acid transcriptome response to UV and far-red light in the cyanobacterium *Chlorogloeopsis fritschii* PCC 6912. *Scientific Reports*, *10*(1), 1-13.
- Llewellyn, C.A., Greig, C., Silkina, A., Kultschar, B., Hitchings, M.D. & Farnham, G. (2020). Mycosporine-like amino acid and aromatic amino acid transcriptome response to UV and far-red light in the cyanobacterium *Chlorogloeopsis fritschii* PCC 6912. *Scientific Reports*, *10*(1), 1-13.
- Los, D. A., & Mironov, K. S. (2015). Modes of fatty acid desaturation in cyanobacteria: an update. *Life*, *5*(1), 554-567.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1-21.

- Lu, Y. J., Zhang, F., Grimes, K. D., Lee, R. E., & Rock, C. O. (2007). Topology and active site of PIsY: the bacterial acylphosphate: glycerol-3-phosphate acyltransferase. *Journal of Biological Chemistry*, *282*(15), 11339-11346.
- MacGregor-Chatwin, C., Nürnberg, D.J., Jackson, P.J., Vasilev, C., Hitchcock, A., Ho, M.Y., Shen, G., Gisriel, C.J., Wood, W.H., Mahbub, M. & Selinger, V.M. (2022). Changes in supramolecular organization of cyanobacterial thylakoid membrane complexes in response to far-red light photoacclimation. *Science Advances*, *8*(6), eabj4437.
- Majumder, E. L. W., Wolf, B. M., Liu, H., Berg, R. H., Timlin, J. A., Chen, M., & Blankenship, R. E. (2017). Subcellular pigment distribution is altered under far-red light acclimation in cyanobacteria that contain chlorophyll f. *Photosynthesis Research*, *134*(2), 183-192.
- McCarthy, D.J., Chen, Y., & Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*, 4288-4297.
- Metz, J.G., Roessler, P., Facciotti, D., Levering, C., Dittrich, F., Lassner, M., Valentine, R., Lardizabal, K., Domergue, F., Yamada, A. & Yazawa, K. (2001). Production of polyunsaturated fatty acids by polyketide synthases in both prokaryotes and eukaryotes. *Science*, *293*(5528), 290-293.
- Moche, M., Dehesh, K., Edwards, P., & Lindqvist, Y. (2001). The crystal structure of β -ketoacyl-acyl carrier protein synthase II from *Synechocystis sp.* at 1.54 Å resolution and its relationship to other condensing enzymes. *Journal of Molecular Biology*, *305*(3), 491-503.
- Muramatsu, M., & Hihara, Y. (2012). Acclimation to high-light conditions in cyanobacteria: from gene expression to physiological responses. *Journal of Plant Research*, *125*(1), 11-39.
- Murata, N., & Wada, H. (1995). Acyl-lipid desaturases and their importance in the tolerance and acclimatization to cold of cyanobacteria. *Biochemical Journal*, *308*(Pt 1), 1.
- Murata, N., Wada, H., & Gombos, Z. (1992). Modes of fatty-acid desaturation in cyanobacteria. *Plant and Cell Physiology*, *33*(7), 933-941.
- Nalley, J. O., O'Donnell, D. R., & Litchman, E. (2018). Temperature effects on growth rates and fatty acid content in freshwater algae and cyanobacteria. *Algal Research*, *35*, 500-507.
- Okuyama, H., Orikasa, Y., Nishida, T., Watanabe, K., & Morita, N. (2007). Bacterial genes responsible for the biosynthesis of eicosapentaenoic and docosahexaenoic acids and their heterologous expression. *Applied and Environmental Microbiology*, *73*(3), 665-670.

Oyola-Robles, D., Gay, D.C., Trujillo, U., Sánchez-Parés, J.M., Bermúdez, M.L., Rivera-Díaz, M., Carballeira, N.M. & Baerga-Ortiz, A. (2013). Identification of novel protein domains required for the expression of an active dehydratase fragment from a polyunsaturated fatty acid synthase. *Protein Science*, 22(7), 954-963.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2), 87-98.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417-419.

Pertea, G. & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research*, 9.

Quinlan, A.R. & Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Roy, H., Dare, K., & Ibbá, M. (2009). Adaptation of the bacterial membrane to changing environments using aminoacylated phospholipids. *Molecular Microbiology*, 71(3), 547-550.

Saha, S. K. & Murray, P. (2018). Exploitation of Microalgae Species for Nutraceutical Purposes: Cultivation Aspects. *Fermentation*, 4, 46.

Sakamoto, T., & Bryant, D. A. (1997). Temperature-regulated mRNA accumulation and stabilization for fatty acid desaturase genes in the cyanobacterium *Synechococcus sp.* strain PCC 7002. *Molecular Microbiology*, 23(6), 1281-1292.

Sánchez-Bayo, A., Morales, V., Rodríguez, R., Vicente, G., & Bautista, L. F. (2020). Cultivation of microalgae and cyanobacteria: effect of operating conditions on growth and biomass composition. *Molecules*, 25(12), 2834.

Santos-Merino, M., Garcillán-Barcia, M. P., & de la Cruz, F. (2018). Engineering the fatty acid synthesis pathway in *Synechococcus elongatus* PCC 7942 improves omega-3 fatty acid production. *Biotechnology for Biofuels*, 11(1), 1-13.

Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T. & Blaxter, M. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA*, 22(6), 839-851.

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. In *Gene prediction* (pp. 227-245). Humana, New York, NY.

Serre, L., Verbree, E. C., Dauter, Z., Stuitje, A. R., & Derewenda, Z. S. (1995). The *Escherichia coli* Malonyl-CoA: Acyl Carrier Protein Transacylase at 1.5-Å Resolution: Crystal structure of a fatty acid synthase component. *Journal of Biological Chemistry*, 270(22), 12961-12964.

Shanab, S. M., Hafez, R. M., & Fouad, A. S. (2018). A review on algae and plants as potential source of arachidonic acid. *Journal of Advanced Research*, 11, 3-13.

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1), 1-18.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631-656.

Taylor, G. (2012). Fatty Acid Metabolism in Cyanobacteria. Doctor of Philosophy Thesis, University of Exeter.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M. & Venables, B. (2022). gplots: Various R Programming Tools for Plotting Data. R package version 3.1.3. <https://CRAN.R-project.org/package=gplots>.

Wijffels, R. H., Kruse, O., & Hellingwerf, K. J. (2013). Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Current Opinion in Biotechnology*, 24(3), 405-413.

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, 17(1), 1-13.

Ye, Y., Nikovics, K., To, A., Lepiniec, L., Fedosejevs, E.T., Van Doren, S.R., Baud, S. & Thelen, J.J., (2020). Docking of acetyl-CoA carboxylase to the plastid envelope membrane attenuates fatty acid production in plants. *Nature Communications*, 11(1), 1-14.

Yoshida, K., Hashimoto, M., Hori, R., Adachi, T., Okuyama, H., Orikasa, Y., Nagamine, T., Shimizu, S., Ueno, A. & Morita, N. (2016). Bacterial long-chain polyunsaturated fatty acids: their biosynthetic genes, functions, and practical use. *Marine Drugs*, 14(5), 94.

Zahra, Z., Choo, D. H., Lee, H., & Parveen, A. (2020). Cyanobacteria: review of current potentials and applications. *Environments*, 7(2), 13.