Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biology

Arshad Kadevalappil Ajilan

Investigating changes in lipidomics of *Alb-3b* mutants as compared to wild type of *Phaeodactylum tricornutum*

Master's thesis in Master of Science in Ocean Resources
(Biochemistry and Biotechnology)
Supervisor: Atle Magnar Bones
Co-supervisor: Per Winge

January 2022

NTNU
Norwegian University of
Science and Technology

Arshad Kadevalappil Ajilan

Investigating changes in lipidomics of *Alb-3b* mutants as compared to wild type of *Phaeodactylum tricornutum*

Master's thesis in Master of Science in Ocean Resources (Biochemistry and Biotechnology)
Supervisor: Atle Magnar Bones
Co-supervisor: Per Winge
January 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biology

NTNU
Norwegian University of
Science and Technology

1 Acknowledgement

I am profoundly grateful to the individuals whose support, guidance, and contributions have been instrumental in completing this master's thesis. Their unwavering commitment and expertise have significantly enriched the journey of research and learning that led to the culmination of this work. First and foremost, I extend my heartfelt appreciation to Atle M. Bones, whose role as my supervisor has been indispensable. I thank him for encouraging me to follow what I wanted, giving me a great deal of research independence, and being there to provide his insightful guidance and mentorship whenever I felt stuck. His support has not only shaped this thesis but has also fostered my growth as a researcher. I am equally thankful to Per Winge for his invaluable contributions as a co-supervisor. His expertise, constructive feedback, and unwavering support have helped me improve my knowledge base and have played a crucial role in refining the scope and quality of this research. Special gratitude is extended to Ralph Kissen, whose exceptional support has been a cornerstone of this project. From training with equipment and standard protocols to meticulously overseeing my lab work, his dedication, and expertise have been critical throughout all the lab work in this project. His willingness to assist in every aspect of this endeavor has made a significant difference. I am also indebted to Marianne Nymark, Marthe Caroline Grønbech Hafskjold, and Frederike Sasse for their assistance in this project. Their willingness to share their expertise, clarify doubts, and provide informational support has been crucial in navigating through the major stages that this project has gone through. Furthermore, I thank Tore Brembu for his valuable assistance in providing technical information on lab protocols and other relevant information for the BODIPY staining experiments using Flow cytometry. His guidance and help have been necessary to ensure the smooth execution of the research. Furthermore, I wish to thank Felicity Jayne Ashcroft for providing training with fluorescence microscopy and Astrid Bjørkøy for her time for the detailed explanation of using the confocal laser scanning microscope. I also want to express my sincere appreciation to Maren Brænden Mathisen and Elise Marlene Sørhøy for their generosity in sharing their lab work knowledge and being willing to help whenever needed. Their assistance was crucial for executing the pulse amplitude modulation(PAM) experiments and getting used to other aspects of the required lab work part of this project. Furthermore, I wish to express my appreciation to all other members of the cell, molecular biology, and genomics group(CMBG), whose inputs have supported me along this journey. Last but not least, I am deeply grateful to my Mom, sister, and brother for their patience and struggles to ensure my well-being.

2 Abbreviations

- Alb3b: Albino 3b mutant
- ACP: Acyl carrier protein
- API: Application programming interface
- ANOVA: Analysis of variance
- BODIPY: Boron-dipyrromethene
- Chl: Chlorophyll
- CLSM: Confocal laser scanning microscopy
- CNRQ: Calibrated normalized relative quantities
- CoA: Coenzyme A
- CRISPR: Clustered regularly interspaced short palindromic repeats
- dNTPs: Deoxynucleotide triphosphates
- Dd: Diadinoxanthin
- DES: De-epoxidation state
- Dt: Diatoxanthin
- EDA: Exploratory data analysis
- ESI: Electron spray ionization
- ETR: Electron transport rate
- FITC: Fluorescein isothiocyanate
- FSC: Forward scattering
- Fx: Fucoxanthin
- GC-MS: Gas chromatography coupled mass spectrometry
- GFP: Green fluorescent protein
- G3P: Glycerol-3-phosphate
- HPLC: High-performance liquid chromatography
- HL: High light
- LD: Lipid droplet
- LL: Low light
- MALDI: Matrix-assisted laser desorption ionization
- ML: Medium light
- NPQ: Non-photochemical quenching
- NRMSE: Normalized Root Mean Squared Error
- OLS: Ordinary least squares
- PAR: Photosynthetic active radiation
- PCA: Principal component analysis
- PAM: Pulse amplitude modulation
- PPFD: Photosynthetic photon flux density
- PS II: Photosystem II
- rETR: Relative electron transport rate
- RFU: Relative fluorescence units
- SSC: Side scattering
- WT: Wild type

2.1 Lipids and Lipid metabolism associated enzymes

- ACC: Acetyl-coA-carboxylase
- ATS1 & ATS2: Arabidopsis seed gene 1& 2
- CDP: Cytidine diphosphate
- CDIP: CDP-DAG inositol 3-phosphatidyl transferase
- CDS: CDP-DAG synthase
- DAG: Diacylglycerol
- DGAT: Diacylglycerol acyltransferase
- DGK: Diacylglycerol kinase
- DGTA: Diacylglycerol trimethyl-β-alanine
- DGD: Digalactosyldiacylglycerol synthase
- DGDG: Digalactocsyl diacylglycerol
- DHA: Docosahexaenoic acid
- EAR: Enoyl-ACP reductase
- EPA: Eicosapentaenoic acid
- FA: Fatty acid
- GPAT: Glycerol-3-phosphate acyltransferase
- HAD: Hydroxyacyl-ACP dehydrate
- KAR: ketoacyl-ACP reductase
- KAS: ketoacyl-ACP synthase
- LACS: Long-chain acyl-CoA synthetase
- LPAAT: Lysophosphatidic acid acyltransferase
- MAT: Malonyl-CoA ACP transacylase
- MGD: Monogalactosyldiacylglycerol synthase
- MGDG: Monogalactosyl diacylglycerol
- PC: Phosphatidylcholine
- PAP: Phosphatidate phosphatase
- PE: Phosphatidylethanolamine
- PI: Phosphatidylinositol
- PIPLC: Phosphoinositide phospholipase C
- PGP: Phosphatidylglycerophosphatase
- PGPS: Phosphatidylglycerol phosphate synthase
- PSD: Phosphatidylserine decarboxylase
- PSS: Phosphatidylserine synthase
- PUFA: Polyunsaturated fatty acid
- SQD2: Sulfoquinovosyl transferase
- SQDG: Sulphoquinovosyl diacylglycerol
- TAG: Triacylglycerol
- TE: Thioesterase

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

3 Abstract

This master thesis project aimed to investigate the impact of a mutation in the Albino3b insertase gene on the lipid profile and other associated parameters of *Phaeodactylum tricornutum*, a model marine eukaryotic species of diatom. Understanding the influences of particular proteins and enzymes on the lipid profile can give insights into the lipid metabolism in diatoms and shed light on approaches for using them as potential candidates for sustainable production of various lipid-based products of industrial importance like Biodiesel, nutraceuticals, or as an alternative source of PUFAs in aquaculture feed, among various other applications. The ALBINO-3B knock-out mutation has been proven to have resulted in reduced levels of photosynthetic pigments in the light-harvesting complexes of the cells, and also have been reported to alter the photosynthetic and growth parameters.

The project commenced with developing a standardized data analysis pipeline for a tandem mass spectrometry (MS-MS) lipid profiling dataset using Python. Leveraging this pipeline, a detailed lipidomic comparison was conducted to reveal substantial variations in different lipid classes and fatty acid compositions between the wild-type and 3 different *Alb3b* mutant lines(*Alb3b*-14,16, 19). This was followed by light-stress experiments, wherein the cell lines were treated under three light conditions(Low light(LL), Medium light(ML), and High light(HL)), followed by experimental cell and molecular biology techniques to understand the underlying mechanisms for the variations observed and to compare results with previous research.

The results indicated differences between the mutants and the WT and among the different mutant lines in how the different lipid classes and the fatty acid compositions in certain lipid classes change between these cell lines when compared between LL and ML. Variations were also observed in the changes in photosynthetic and photoprotective parameters, with the mutants having reduced chlorophyll pigmentation, increased non-photochemical quenching, and better photoacclimation properties than the wild type. Additionally, changes were observed in Lipid droplet (LD) structure, particularly in their number within the cell, between cell lines, and between light conditions. Furthermore, The expression levels of certain enzymes involved in phospholipid and fatty acid metabolism between the cell lines were detected to be deferentially regulated under the three different light conditions using quantitative PCR, during the lab work part of the project. The project was concluded with interpretations of the results obtained and possible explanations and predictions about the same based on findings from the literature search. The stress responses in lipid metabolism are predicted to be different between the WT and the *Alb3b*-mutants possibly because of the changes observed in photo physiologies between them.

4 Sammendrag

Denne masteroppgaven hadde som mål å undersøke virkningen av en mutasjon i Albino3b insertase-genet på lipidprofilen og andre tilknyttede parametere til *Phaeodactylum tricornutum*, en modellorganisme for marine eukaryoter av diatomer. Å forstå innflytelsen til spesifikke proteiner og enzymer på lipidprofilen kan gi innsikt i lipidmetabolismen hos diatomer og kaste lys over tilnærminger for å bruke dem som potensielle kandidater for bærekraftig produksjon av ulike lipidbaserte produkter av industriell betydning som biodiesel, næringsmidler eller som en alternativ kilde til PUFAer i akvakulturfor.

ALBINO-3B knock-out mutasjonen har vist seg å ha resultert i reduserte nivåer av fotosyntetiske pigmenter i lysoppsamlingskompleksene til cellene, og det har også blitt rapportert at den endrer de fotosyntetiske og vekstparametrene.

Prosjektet startet med å utvikle en standardisert dataanalysepipeline for et tandem massespektrometri (MS-MS) lipidprofileringsdatasett ved hjelp av Python. Ved hjelp av denne pipelinen ble det gjennomført en detaljert lipidomisk sammenligning for å avdekke betydelige variasjoner i forskjellige lipidklasser og fettsammensetninger mellom villtypen og 3 ulike *Alb3b*-mutantlinjer (*Alb3b*-14,16, 19). Dette ble etterfulgt av lysstressforsøk, der cellelinjene ble behandlet under tre lysforhold (Lavt lys (LL), Medium lys (ML) og Høyt lys (HL)), etterfulgt av eksperimentelle celle- og molekylærbiologiske teknikker for å forstå de underliggende mekanismene for de observerte variasjonene og sammenligne resultatene med tidligere forskning.

Resultatene indikerte forskjeller mellom mutantene og WT og blant de ulike mutantlinjene i hvordan de ulike lipidklassene og fettsammensetningene i visse lipidklasser endres mellom disse cellelinjene ved sammenligning mellom LL og ML. Variasjoner ble også observert i endringene i fotosyntetiske og fotobeskyttende parametere, der mutantene hadde redusert klorofyllpigmentering, økt ikke-fotokjemisk sluking og bedre fotoakklimeringsegenskaper enn villtypen. Endringer ble også observert i lipiddråpe (LD) -struktur, spesielt i antallet innenfor cellen, mellom cellelinjer og mellom lysforhold. Videre ble uttrykksnivåene til visse enzymer involvert i fosfolipid- og fettsyremetabolisme mellom cellelinjene påvist å være differensielt regulert under de tre forskjellige lysforholdene ved bruk av kvantitativ PCR, under labarbeidet i prosjektet. Prosjektet ble avsluttet med tolkninger av de oppnådde resultatene og mulige forklaringer og spådommer om det samme basert på funn fra litteratursøket. Stressresponsene i lipidmetabolismen antas å være forskjellige mellom WT og *Alb3b*-mutanter, muligens på grunn av endringene som er observert i fotofysiologiene mellom dem.

5 Introduction

Diatoms are organisms of great significance in diverse research fields owing to their contribution to primary production, characteristic evolutionary biology, ecological interactions, and biotechnological applications. Their contribution of about 20% to the global primary productivity marks their relevance in the earth system(Malviya et al., 2016). This is almost equivalent to the organic carbon production of all the rainforests combined(Malviya et al., 2016). Organic matter production to this scale indicates their ecological relevance as a crucial primary producer in the ocean food chain. Furthermore, diatoms are also involved in other complex interactions in marine ecosystems, including their well-known dominance in algal blooms during events of nutrient enrichment and the toxic threats posed by some of the strains to

other aquatic life(Armbrust, 2009). In addition to being major carbon fixing agents, their silicate uptake and metabolism to construct and maintain their siliceous cell walls or frustules make them important regulators of past and present-day ocean bio-geochemistry(Armbrust, 2009). Although the exact estimates of global diatom species diversity are hard to determine, they are reputed as the most diverse group of phytoplankton, accounting for about 12000 to 30000 species(Malviya et al., 2016). In addition to this fact, their development through secondary endosymbiosis gives them great importance in evolutionary research. The increased accretion of molecular sequence data because of technological development has immensely supported increasing interest in diatom phylogenetic analysis (Williams, 2007).

Apart from being a crucial ecological actor, diatoms also serve as feedstock for biotechnological applications that have the potential to support both human and planet welfare. This includes biofuel production, human and animal nutraceuticals, and pharmaceuticals, the development of bio-active compounds, the development of nanotechnological materials, and wastewater treatment(Bozarth et al., 2009).

Widespread research has been and is being carried out with diatoms to harness their potential for these applications. One group of this research area focuses on the lipidomics of diatoms as the different lipids in diatoms or their associated fatty acids can serve several beneficial products. The lipid production in diatoms could potentially be exploited for developing sustainable alternatives for various commercial and industrial purposes. The lipid content in the diatom cell comprises different classes each of which has different roles in cell metabolism and applications in the industry.

Triacylglycerols(TAG), a class of neutral lipids that forms the major carbon and energy storage compound in diatoms, are efficiently accumulated in significant amounts by the cells accounting for about 15-25% of the overall dry biomass. TAG and its constituent fatty acids with various chain lengths can be utilized for many applications mentioned above For instance, most past research on diatoms for biotechnological applications has been focused on the production of the long-chain polyunsaturated fatty acids, eicosapentaenoic acid (EPA) and docosahexaenoic acid(DHA)(Lebeau and Robert, 2003), which are well described for their nutritional benefits, especially in supporting cardiovascular health and brain development(Yi et al., 2017). However, other short or medium-chained fatty acids in the TAG content have other relevant uses. For example, the short or medium-chained fatty acids (C14-C18) in TAG can be chemically processed to produce Fatty acid methyl esters(FAME), commercially called bio-diesel, a sustainable fossil-fuel alternative. Furthermore, some of the C16 fatty acids found in diatoms were shown to have antibacterial properties, thus opening new possibilities in drug development(Yi et al., 2017).

Several studies on TAG accumulation in diatoms have been executed using environmental manipulations, the majority of which involve stress conditions like nutrient starvation or non-optimal light levels, which are environmentally germane growth stressors. Reduction of growth due to the various stressors like nutrient or phosphorous starvation or high light intensities is clearly shown to induce TAG accumulation that will help the cell to reserve its carbon and energy to support itself during recovery from the particular stressors. However, the cell's metabolic responses vary depending on the type of stress introduced, and the TAG accumulation in cells is affected by metabolic pathways other than just the fatty acid synthesis pathways(Yi et al., 2017). This could include pathways associated with amino acid metabolism, photosynthesis, and metabolic pathways of other lipid classesGe et al., 2014.

A comprehensive understanding of TAG accumulation at the system level needs to be obtained to better manipulate the culture environment or create lipid production-optimal strains through mutations or genetic engineering for supporting sustainable production strategies. One way to approach this is by

studying the lipid profile changes in mutants with knock-out mutations in genes associated with metabolic pathways that directly or indirectly affect cell growth.

5.1 Aim of the project

This master project aims to analyze and compare the lipidomics data obtained through MS-MS of the lipid samples from a wild type of *P. tricornutum* and a particular knock-out mutant called Alb-3b of the same species. The project progressed through the following phases before completion:

- Developing a data analysis pipeline with exploratory data analysis, visualizations, and hypothesis testing using functions and packages in Python 3.8 using the JupyterLab API.
- Acclimating the mutants and WT to the previous light experiment conditions used for mass spectrometry to quantify LD accumulation using neutral lipid staining and flow cytometry, and to measure various photosynthetic and photo-protective parameters using pulse amplitude modulation.
- Using confocal laser scanning microscopy (CLSM) to observe the fluorescence-stained LDs and compare between and among the WT and the mutants under different light conditions.
- Conducting real-time polymerase chain reaction(q-PCR) for differential expression analysis of selected lipid metabolism genes.

5.2 Wild type of *Phaeodactylum tricornutum*

Phaeodactylum tricornutum is a model species of diatom extensively used in biotechnology research. Selection of the species as a model for research can be based on several factors such as:

- High growth rates and ease of culturing
- Availability of whole genome sequence
- The ability of the species to grow without silicified frustule formation

These factors also make it a potential candidate for applications like biofuel production, recombinant protein expression, and silicon nanofabrication. The wild type of the species was obtained from

5.3 *Alb3b*-Mutants

The *Alb3b* mutant strains of *Phaeodactylum tricornutum* are knock-out mutants in which the functionality of the ALBINO3B insertase protein has been lost(Nymark et al., 2019). The ALBINO3B protein is required by the cells for the insertion of the fucoxanthin-chlorophyll binding protein into the thylakoid membrane(Nymark et al., 2019). Although plants and green algae possess *Alb3b* proteins with known functionality, phylogenetic analysis led to the categorization of the *Alb3b* proteins in diatoms as a distinct group. This makes the prediction of *Alb3b* proteins in diatoms based on characterization of the same in plants or green algae less reliable(Nymark et al., 2019).

The *Alb3b* knock-out strains were developed through a biolistic transformation using pKS diaCas9-sgRNA plasmid for CRISPR cas9-based knockout (Nymark et al., 2019). The original gene editing experiment targeted both the Alb3a and *Alb3b* paralogs of the protein but succeeded only in creating mutants

with edit in the *Alb3b* gene. The mutants, with large insertions in the 5' end of the gene, have been shown to have significantly reduced(75%) fucoxanthin-chlorophyll a/c-binding proteins, reduced pigmentation, and non-photochemical quenching, truncated light-harvesting antennae and changes in photosynthetic saturation light levels(Nymark et al., 2019). However, the knock-out mutations were shown not to affect the levels of diadinoxanthin or diatoxanthins, the carotenoids involved in the xanthophyll cycle for photo-protection(Nymark et al., 2019).

Three lines of knock-out mutants where the *Alb3b* gene has been disrupted through the CRISPR-cas9 mechanism were used in the previous mass spectrometry experiment on which analysis for this thesis work is based and the flow cytometry and imaging experiments involved in the project. These are:*Alb3b*-14, *Alb3b*-16 and *Alb3b*-19.

5.4 Findings from the previous study conducted on the mutants to understand the role of *Alb3b* in Diatoms

The previous study by Nymark et al., 2019 aimed at investigating the functionality and molecular mechanisms of ALBINO3B insertase protein in the assembly and integration of the light harvest complex proteins and how it affects other crucial metabolic processes including growth, photosynthesis, and photoprotection. As mentioned above, knock-out mutants for the ALBINO3B gene were developed with *P.tricornutum* and were subjected to varying light conditions during the experiment. The study involved a comprehensive experimental setup combining genetic, biochemical, and imaging techniques, to provide detailed insights into the functional role of ALBINO3B in diatoms. This multidisciplinary experimental approach used various experimental techniques, including CRISPR/Cas9-based genome editing, spectral analyses, photosynthetic performance analyses, complementation studies, transmission electron microscopy, and HPLC pigment analysis to obtain a global view of the impact of ALBINO3B on diatom photosynthesis. A brief description of the main components of the experimental setup is as follows:

CRISPR/Cas9-based Genome Editing:

- Design of guide RNA targeting the ALBINO3B gene and induction of double-strand breaks using the Cas9 endonuclease.
- Mutation screening and confirmation via PCR and sequencing.

Analyses of photophysiological parameters

- Determination of changes in light energy absorption and energy transfer efficiency in mutant strains.
- Insights into alterations in pigments' contributions to the reaction center of photosystem II (PSII).
- Light-saturation curves of photosynthesis based on oxygen evolution.
- Variable in vivo Chl a fluorescence measurement.
- Assessment of photosynthetic parameters: respiration, Pmax, and Es.

Complementation Studies:

- Introduction of a modified ALBINO3B gene to restore the wild-type phenotype.

- Confirmation of ALBINO3B's essential role in maintaining normal light-harvesting antenna structure and function.

Transmission Electron Microscopy:

- Visualization of thylakoid architecture in mutant strains.
- Insights into structural alterations resulting from the absence of ALBINO3B.

HPLC Pigment Analysis:

- Determination of changes in pigment content in mutant strains.
- Uncovering shifts in the composition of pigments associated with ALBINO3B deficiency.

The above analyses were performed for cells exposed or acclimated to different light conditions for different exposure times. The light conditions employed are:

- LL: 35µmol photons m^{-2} s^{-1}
- Medium-light: 200µmol photons $m^{-2}s^{-1}$
- HL: 480µmol photons m^{-2} s^{-1}

This comprehensive evaluation of wild-type and mutant strains yielded some noteworthy results, As mentioned above, various parameters were scrutinized to discern the impact of ALBINO3B deficiency on the light-harvesting antenna complex in diatoms.

HPLC pigment analysis revealed a substantial reduction in Chl c and Fx content in mutant strains, suggestive of an altered light-harvesting antenna complex. Spectral analyses unveiled pronounced disparities in the in vivo fluorescence excitation spectra, indicating diminished energy transfer from Chl c and Fx to the PSII reaction center in mutants. This pigment reduction is confirmed by the fact that the absorption spectra differences aligned with alterations observed in in vivo fluorescence excitation spectra.

Notably, the pulse amplitude modulation measurements to study photosynthetic performance by Fv/Fm values in this study indicated a high photosynthetic activity in mutant lines as compared to the wild-type. This is also supported by the observed higher maximum electron transport rate (rETR $_{max}$)and light saturation intensity(E_k) in the mutants in the initial phase of light experiments. Further investigation into this discrepancy by analysis of light saturation curves of photosynthesis revealed the mutants to have a higher P $_{max}$ (maximum photosynthetic rate) and E $_s$(light saturation index) and a lower maximum light utilization coefficient (α). These results point towards a light-harvesting antenna with reduced cross-sectional area in the mutants and avoidance of interpreting the results for improved photosynthetic performance in the mutants.

The mutant strains initially had lower NPQ levels compared to the wild type at higher light intensities(>400 µmol photons m^{-2} s^{-1}), but over time, the differences in NPQ capacity between the mutant strains and the wild type decreased. However, the levels of the photo-protective pigment, that are diatoxanthin(Dt) and diadinoxanthin(Dd), involved in the xanthophyll cycle, appeared not to be affected in the knock-out mutants. Moreover, the deepoxidation state(DES) indices were even found to be higher in the mutants than in the wild-type.

Transmission electron microscopy depicted structural transformations in mutant chloroplasts, characterized by a reduced number of thylakoid membranes per chloroplast.

Complementation studies, however, showcased a recovery of the wild-type phenotype in complemented mutants, affirming the pivotal role of ALBINO3B in maintaining normal pigment and LHCF content. Collectively, these findings underscore the intricate influence of ALBINO3B on the structural and functional aspects of the light-harvesting antenna complex, shedding light on its indispensable role in diatom photosynthesis.

5.5 Lipid profile in diatoms

Lipids constitute a major group of biomolecules in all living organisms, performing crucial functions like energy storage, transport, cellular structure, and signal transduction(Guo et al., 2020). Hydrophobic or amphiphilic molecules are synthesized from fatty acid units of different chain lengths(Guo et al., 2020). Lipids in all living organisms including algae and diatoms can be broadly classified into polar lipids and non-polar lipids(Manning, 2022). Non-polar lipids mainly serve as energy reserve molecules whereas polar lipids are mostly involved in other functions like structural roles in various cell membranes(Manning, 2022).

Diatoms contain various lipids belonging to these different classes with a variety of fatty acid chain lengths present in each of the classes to form the overall lipid profile of the cell. The lipid metabolic pathways are altered by the organism based on environmental cues like stress, resulting in a varied lipid profile adapted to the new conditions(Maeda et al., 2017. The exact functions and sub-cellular localization of most of these lipids are of great research interest especially because of the complex membrane system that emerged from secondary endosymbiosis(Maeda et al., 2017). However, genes associated with enzymes for synthesizing these lipids and those involved in the remodeling of lipid metabolic pathways under stress conditions have been discovered in diatoms(Levitan et al., 2015, Sayanova et al., 2017). The major classes of lipids, their functions, and molecular structures are as follows:

- **Neutral lipids**: The neutral lipids are mostly in the form of triacylglycerol(TAG), wherein three fatty acid molecules attach to a glycerol backbone. These molecules serve as the major carbon and energy storage source, especially under stress conditions (Manning, 2022). Diatom cells concentrate their neutral lipids in special organelles called lipid droplets(Maeda et al., 2017).

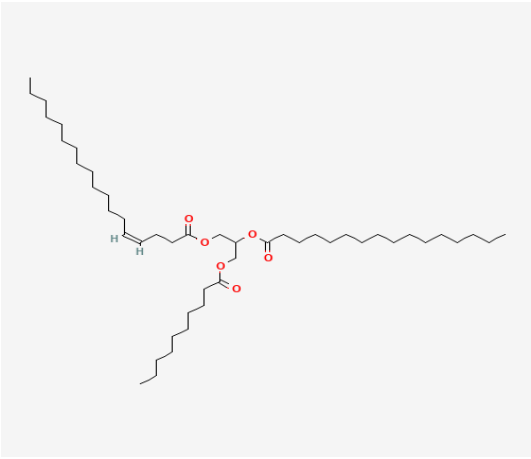
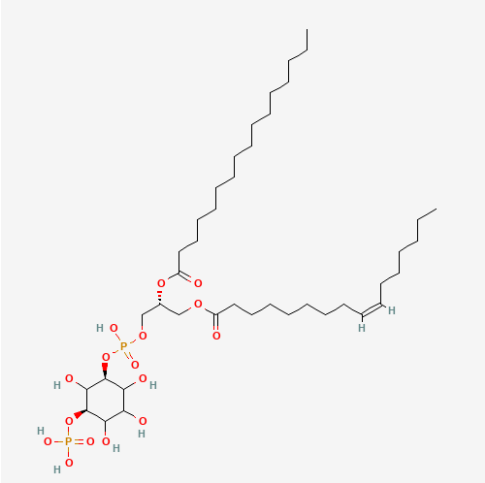


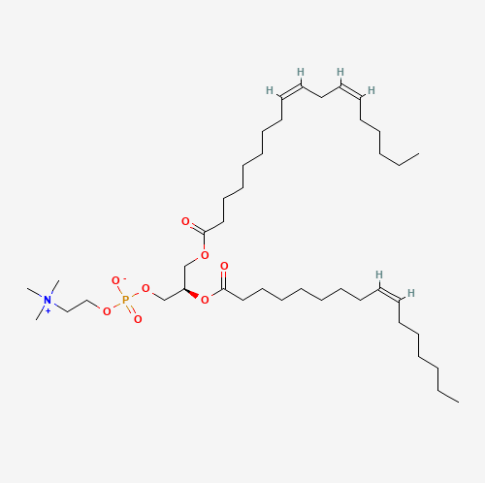
Figure 1: Chemical structure of the neutral lipid, TAG obtained from PubChem.

- **Phospholipids**: The glycerophospholipids, a major group of polar lipids predominantly act as membrane lipids and form most parts of the cell membranes and endoplasmic reticulum(Tanaka et al., 2022). These include molecules with a phosphate group-containing hydrophilic head and a hydrophilic part comprised of fatty acid chains attached to a glycerol backbone. Phosphatidylcholine (PC), Phosphatidylinositol (PI), phosphatidylglycerol (PG), and phosphatidylethanolamine (PE) are the main phospholipids detected in diatoms(Manning, 2022).

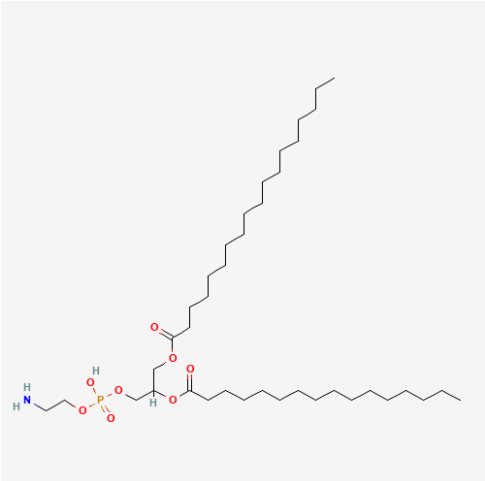
(a) Phosphatidylinositol



(b) Phosphatidylcholine



(c) Phosphatidylethanolamine



(d) Phosphatidylglycerol

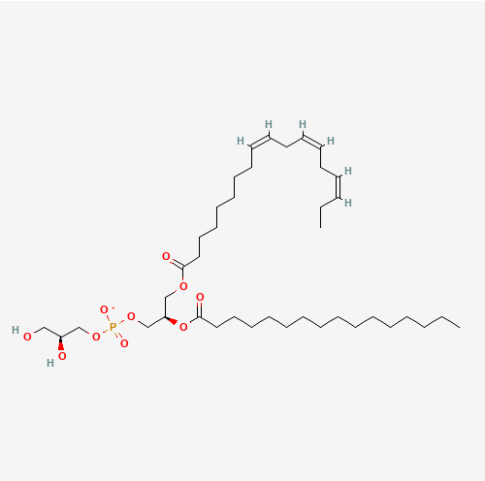
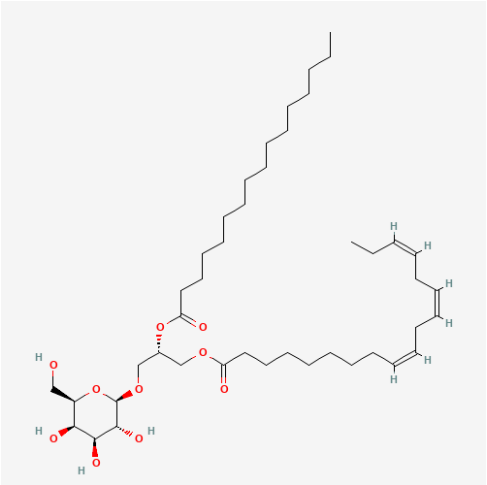


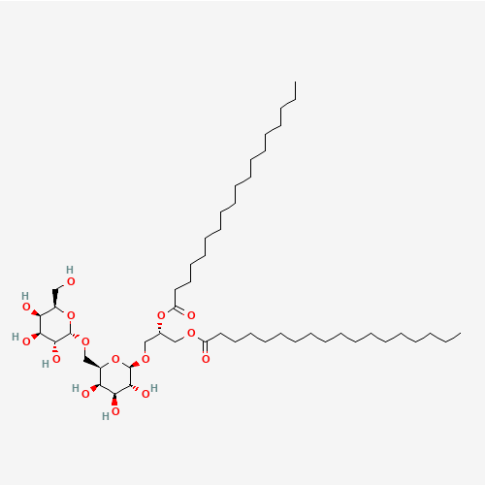
Figure 2: Chemical structures of the major types of phospholipids obtained from PubChem.

- **Glycolipids**: These lipids are another major group of poplar lipids which are carbohydrate-containing lipids. A subgroup of the glycolipids, called galactolipids, wherein the sugar group is galactose, comprises a crucial fraction of the diatom lipid profile(Manning, 2022). This fraction involves species such as sulfoquinovosyl diacylglycerol (SQDG) monogalactosyl diacylglycerol (MGDG), and galactosyl diacylglycerol (DGDG)(Manning, 2022), which predominates the thylakoid membranes within the diatom plastids along with the phospholipids PG and PC(Tanaka et al., 2022). They are involved in photosynthesis, membrane organization, and photoprotection functions of the cell(Tanaka et al., 2022).

(a) Monogalactosyldiacylglycerol



(b) Digalactosyldiacylglycerol



(c) Sulphoquinovosyldiacylglycerol

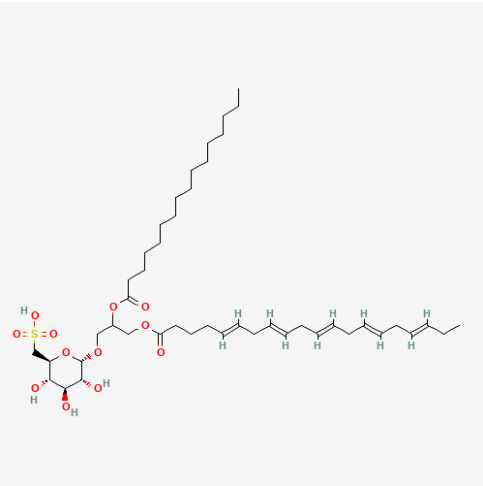


Figure 3: Chemical structures of the major types of glycolipids obtained from PubChem.

Betaine lipids and sterols form two other groups of lipids detected in diatom, Even though betaine lipids like the Diacylglycerylhydroxymethyl- N, N, N - trimethyl-β-alanine (DGTA) and sterols like the brassicasterol has been detected in lipid droplets (Lupette et al., 2019) and the genes associated with synthesis and remodeling of both these groups has been identified in diatoms the exact metabolic pathways in diatoms still need to be delineated(Tanaka et al., 2022).

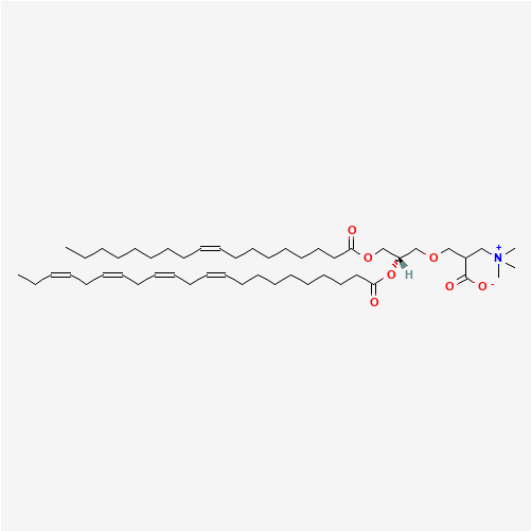


Figure 4: Chemical structure of the betaine lipid, Diacylglycerylhydroxymethyl- N, N, N - trimethyl--alanine from PubChem.

5.5.1 Denovo fatty acid synthesis, elongation and desaturation reactions in diatoms

The common building block for all these lipids is the fatty acids(Tanaka et al., 2022). These are produced by the cells through an energy-dependent fatty acid synthesis pathway in the plastid starting with Acetyl-CoA molecules(Tanaka et al., 2022). The reaction is initiated by the conversion of acetyl-CoA to Malonyl-coenzyme A (CoA) catalyzed by acetyl-CoA carboxylase (ACC). This ATP-dependent step sets the stage for the subsequent fatty acid synthesis (FAS) pathway(Tanaka et al., 2022).

The type II fatty acid synthesis pathway is observed in diatom cells wherein multiple mono-functional enzymes are involved in different processes(Apt et al., 2002). According to the information on putative pathways of fatty acid synthesis from Tanaka et al., 2015, the Malonyl-coA produced in the first reaction is converted to Malonyl-acyl carrier protein (ACP) through the action of Malonyl-CoA: ACP transacylase. The Malonyl-ACP acts as a carbon donor molecule to elongate the acyl chain by 2 carbons during each cycle of the acyl elongation pathway in the plastid. This pathway involves multiple enzymatic domains that catalyze sequential reactions, including condensation, reduction, dehydration, and reduction, to elongate the fatty acid chain. This characteristic cyclic reaction results in the elongation of the acyl chain of the fatty acid precursor called Acyl-ACP(Tanaka et al., 2022).

A detailed view of the FAS pathway is illustrated in the figure 5.

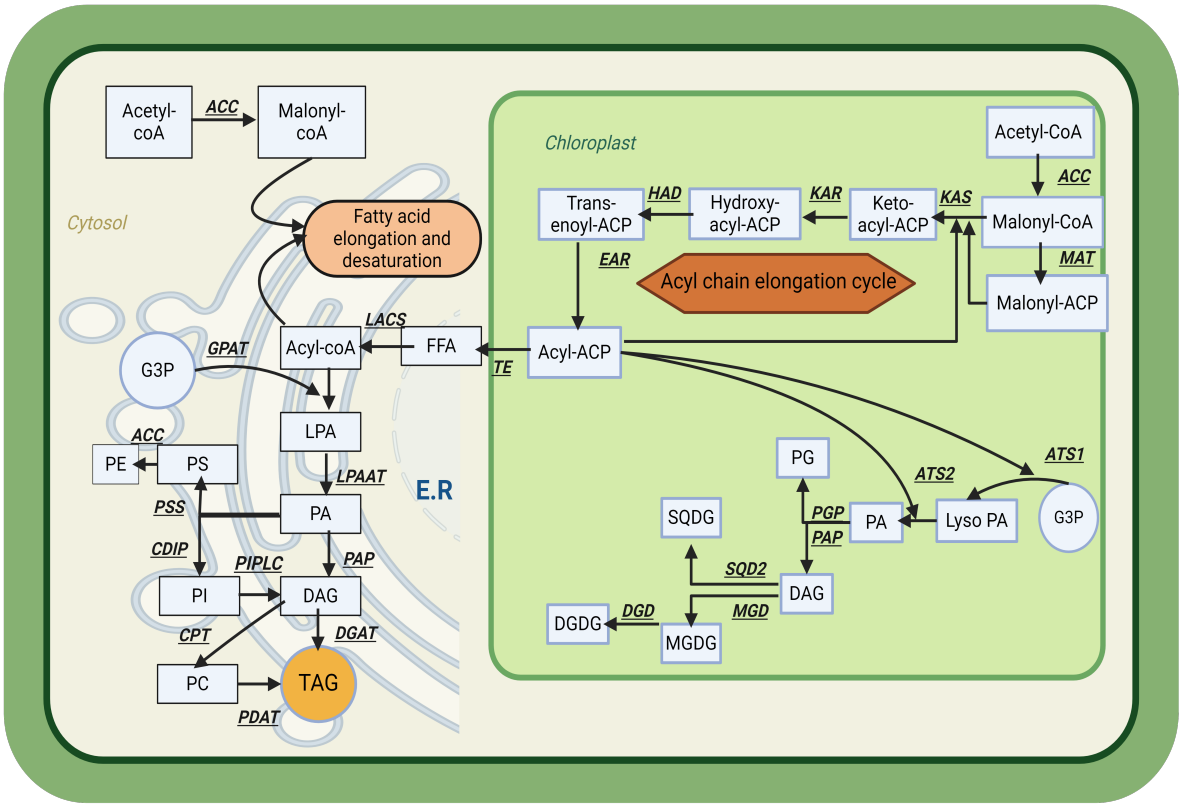


Figure 5: An overview of the lipid metabolic pathways in diatoms cells created using information from Tanaka et al., 2022. The figure depicts the acyl chain elongation cycle leading to the production of Acyl-ACP, an important intermediate for synthesizing glycerolipids, within the chloroplast. Also, the synthesis of the important thylakoid membrane lipids like the glycolipids(MGDG, DGDG) sulpholipids (SQDG), and PG, are shown within the chloroplast. TAG synthesis through the Kennedy pathway using Acyl-coA and glycerol triphosphate(G3P) and phospholipid remodeling is depicted in the Endoplasmic reticulum. The figure was created using Biorender.

Abbreviations of the lipids and enzymes in Figure 5 involved in different steps are as follows: ACC: Acetyl-coA-carboxylase, ATS1 ATS2: Arabidopsis seed gene 1 2, CDP: Cytidine diphosphate, CDIP: CDP-DAG inositol 3-phosphatidyl transferase, CDS: CDP-DAG synthase, DAG: Diacylglycerol, DGAT: Diacylglycerol acyltransferase, DGK: Diacylglycerol kinase, DGTA: Diacylglycerol trimethyl-β-alanine, DGD: Digalactosyldiacylglycerol synthase, DGDG: Digalactocsyl diacylglycerol, DHA: Docosahexaenoic acid, EAR: Enoyl-ACP reductase, EPA: Eicosapentaenoic acid, FA: Fatty acid, GPAT: Glycerol-3-phosphate acyltransferase, HAD: Hydroxyacyl-ACP dehydrate, KAR: ketoacyl-ACP reductase, KAS: ketoacyl-ACP synthase, LACS: Long-chain acyl-CoA synthetase, LPAAT: Lysophosphatidic acid acyltransferase, MAT: Malonyl-CoA ACP transacylase, MGD: Monogalactosyldiacylglycerol synthase, MGDG: Monogalactosyl diacylglycerol, PC: Phosphatidylcholine, PAP: Phosphatidate phosphatase, PE: Phosphatidylethanolamine, PI: Phosphatidylinositol, PIPLC: Phosphoinositide phospholipase C, PGP: Phos-

phatidylglycerophosphatase, PGPS: Phosphatidylglycerol phosphate synthase, PSD: Phosphatidylserine decarboxylase, PSS: Phosphatidylserine synthase, PUFA: Polyunsaturated fatty acid, SQD2: Sulfoquinovosyl transferase, SQDG: Sulphoquinovosyl diacylglycerol, TAG: Triacylglycerol, TE: Thioesterase

Thioesterase enzymes act upon this Acyl-ACP to produce free fatty acids by hydrolysis(Hao et al., 2018). The free fatty acid molecules thus produced get converted through enzymatic reaction into Acyl-coA, which along with the glycerol 3 phosphate molecules from glycolysis are used in different pathways for denovo synthesis of both neutral lipids, that is TAG, and the various phospholipids(Tanaka et al., 2022). TAG could be also synthesized in an Acyl-coA-independent manner, wherein the polar lipids like phospholipids or glycolipids, undergo remodeling into TAG, especially under stress conditions like nutrient starvation(Abida et al., 2015, Maréchal and Lupette, 2020). the galactolipids(MGDG, DGDG, and SQDG) are synthesized from another set of pathways that uses the acyl-ACP intermediate and glycerol 3 phosphate as the feed stock(Tanaka et al., 2022).

The thioesterase does not consume all the acyl-ACP as a fraction is transported out of the plastids and undergoes further elongation and desaturation in the endoplasmic reticulum to produce the long chain or very long chain and mono or polyunsaturated fatty acid molecules like the EPA and DHA(Tanaka et al., 2022).

Although most of the elongation and desaturation reactions occur in the ER, some of these types of reactions are also known to occur within the chloroplast(Sayanova et al., 2017). The ER-mediated desaturation and elongation processes are integral for synthesizing intricate lipids characterized by specific fatty acid compositions. These pathways are crucial for diatoms, as the resulting PUFAs play pivotal roles in modulating membrane fluidity, participating in signaling pathways, and mediating stress responses within diatom cells(Montecillo-Aguado et al., 2023). Desaturation involves the introduction of double bonds into the fatty acid chain, leading to the formation of unsaturated fatty acids(J. M. Lee et al., 2016). Enzymes known as desaturases catalyze these reactions by inserting double bonds at specific positions in the fatty acid chain(J. M. Lee et al., 2016). Elongation processes in the ER involve the addition of two-carbon units to the fatty acid chain, leading to the synthesis of longer-chain fatty acids(Wang et al., 2023). Enzymes called elongases are responsible for extending the fatty acid chain length(Wang et al., 2023). The carbon donor molecule for the elongation step is the same Malonyl-coA as in the FAS pathway within the chloroplast(Tanaka et al., 2022). However, the Malonyl-coA molecules for elongation reactions in the ER are produced from Acetyl-coA in the cytosol by the action of cytosolic ACC enzymes(Tanaka et al., 2022).

There are two crucial elongation and desaturation pathways occurring in the diatom ER, namely the $\omega 6$ and $\omega 3$ pathways(Arao and Yamada, 1994), which involve specific elongase and desaturase enzymes as indicated in the figure. The $\omega 3$ desaturase enzyme acts as a crossover enzyme between these two pathways by using products from ω 6 pathway to form products of the $\omega 3$ pathway(Sayanova et al., 2017). The detailed process is illustrated in figure 6.

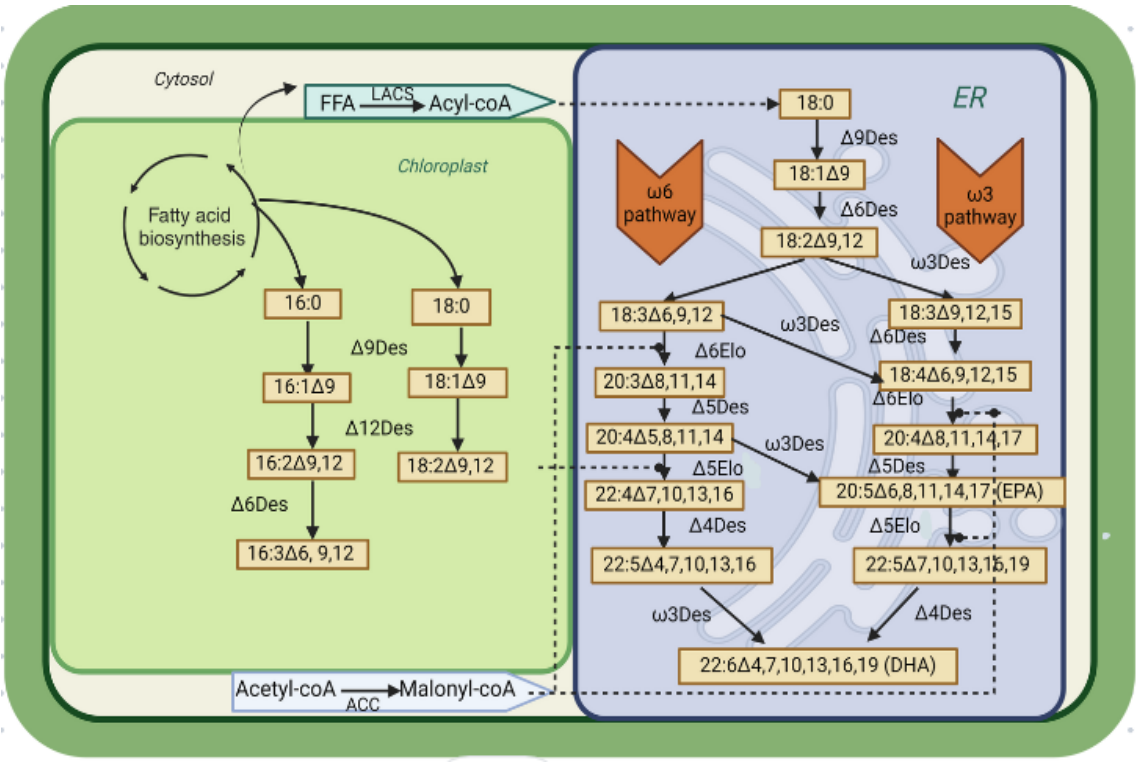


Figure 6: An overview of the fatty acid elongation and desaturation reaction pathways within diatom cells based on information from Tanaka et al., 2022. The figure shows only the fatty acids with 16:0, and 18:0 chains as the starting substrates obtained from the fatty acid synthesis depicted in figure 5. This is just to show the two most important desaturation pathways that occur in the ER, namely the $\omega 3$ and $\omega 6$ pathways leading to the production of relevant PUFAs.Similar to acyl chain elongation reactions, the carbon donor in the elongation reactions, catalyzed by elongases, is the Malonyl-coA.However, this is obtained from cytosolic acetyl-coA, unlike the plastidic ones in acyl chain elongation cycles.Des stands for desaturase and Elo stands for elongase. The figure was created using Biorender.

Changes in the activity of various enzymes involved in the FAS and the desaturation pathways, with different acyl chain preferences, will affect the fatty acid chain lengths and levels of unsaturation in the final lipid profile of the cells(Tanaka et al., 2022,Haslam et al., 2020). Some of the most important ones among these include:

- TE (Thioesterase): The enzyme involved in the conversion of the acyl-ACP intermediate in the acyl elongation process in the FAS pathway into free fatty acids thereby resulting in the termination of chain elongation in chloroplast. The specificity of this enzyme towards a particular acyl chain length could thus potentially affect the final fatty acid profile.

- GPAT (glycerol 3 phosphate acyl transferase): One of the starting enzymes involved in both TAG and Phospholipid synthesis pathways, which converts glycerol 3 phosphate into Lysophosphatidic acid, the precursor of phosphatidic acid from which the phospholipids or TAG can be synthesized through different pathways.

- LPAAT (Lysophosphatidic acid acyl transferase): The enzyme that converts the lysophosphatidic acid into phosphatidic acid.

- DGAT (Diacylglycerol acyl transferase: The final enzyme in the TAG synthesis pathway that adds the final acyl chain to Diacylglycerol (DAG) to form TAG.
- PDAT (Phospholipid: diacylglycerol acyl transferase): One of the main enzymes in the phospholipid remodeling pathway or the acyl-coA independent TAG synthesis that transfers acyl chains from phospholipids like PC into DAG to from TAG

5.6 Lipidomics using mass spectrometry

Metabolomics involves the qualitative and quantitative analyses of metabolites in a biological sample and includes the analyses of proteins or peptides (proteomics or peptidomics) and lipids (lipidomics)(Wu et al., 2020). Mass spectrometry is a powerful tool that has enabled high throughput metabolomic analyses through advanced molecular characterization and quantification approaches. The role of mass spectrometry is considered inevitable in the field of lipidomics(Wu et al., 2020). Mass spectrometry involves the detection of molecules that differ in their mass, charge, shape, and size after ionizing the constituent molecules in the gas phase and separating them based on their mass-to-charge (m/z) ratio. Contemporary mass spectrometry methods combine it with chromatographic separation techniques to increase the resolution of the analysis. There exist different variants of mass spectrometry based on the type of chromatography with which it is combined, the method used for the ionization of molecules, and the technique employed for molecular separation based on mass-to-charge ratio and detection.

Common chromatography techniques that are integrated with MS include:

- High performance liquid chromatography(HPLC)
- Gas chromatography
- Thin layer chromatography(TLC)

The previous experiment conducted by Nymark et al., 2019 isolated lipids from all the above-mentioned mutant strains and wild types which were exposed to two different light-level treatments, that is LL(35µmol photons m^{-2} s^{-1}) and ML(200µmol photons m^{-2} s^{-1}). The isolated lipid samples were characterized using tandem mass spectrometry at the Institut National de la Recherche Agronomique, Université Grenoble Alpes, in Grenoble, France. Generally, the glycerolipids are separated and eluted with HPLC/TLC, and the Fatty acids with GC-FID before MS-MS. The positioning of the FAs on the glycerol moiety is determined by radiolocalization. Both lipid class data and Fatty acid composition data from this experiment were used to base this thesis work.

The general workflow in MS-MS coupled with GC is depicted in figure 7

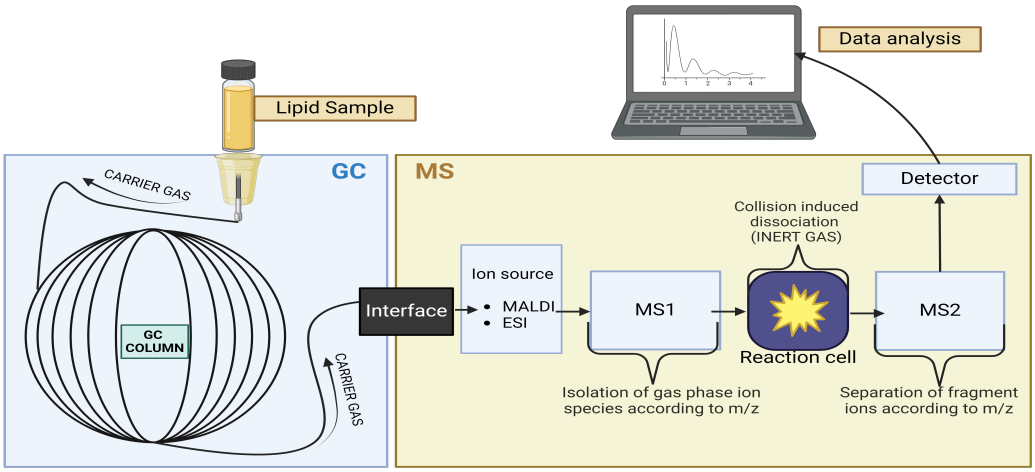


Figure 7: General workflow in gas chromatography coupled mass spectrometry for Fatty acid analysis. The lipid samples with different components are injected after vaporization into the GC COLUMN along with the carrier gas or mobile phase of chromatography. The movement of the different components through the column differs based on their mass thus changing their time to traverse the column. Consequently, the different components in the lipid sample are separated and adsorbed onto the stationary phase at different points in the column, and then later on eluted to enter the MS phase for further separation and detection based on the m/z ratio. The MS phase in lipid analysis usually involves tandem mass spectrometry(MS-MS) to achieve high resolution by 2 consecutive MS steps separated by a collision-induced fragmentation step. The final separated components are then detected by a detector that generates the final chromatogram based on the detected electrical signals. The figure was created using Biorender.

There are two main methods which are employed in MS for gas-phase ionization of the samples after elution from the GC column:

- Electron spray ionization: The biological sample in the liquid phase is forced through an orifice with an electric field across it resulting in the subsequent breakdown into smaller particles with charge surrounding it and vaporization of the molecules into charged ions.
- Matrix-assisted laser desorption ionization (MALDI): The biological sample is in a solid state by intercalating it in a suitable solid matrix. This is then subjected to exposure by a focused laser that causes the desorption and ionization of the molecules.

The separation of ions based on mass and charge can be done through one of the following techniques

- Time-of-Flight: The ionized molecules are separated based on their velocities when accelerated through a specific voltage(Sinha and Mann, 2020). Different velocities result in different times for traversing the fixed trajectory which is a measure of mass-to-charge ratio and can be estimated based on the arrival times at the detector(Sinha and Mann, 2020).

- Orbitrap: Ions are differentiated based on the variations in oscillation frequencies of the molecules as they move along a central metal spindle after being trapped in the Orbitrap(Sinha and Mann, 2020).

Tandem mass spectrometry(MS-MS) is a variant of mass spectrometric analysis that offers a much higher resolution in ion detection by combining with subsequent mass spectrometry events. Here the separated particles in the first MS are further disintegrated into smaller particles by collision with inert gases like nitrogen in a chamber(Sinha and Mann, 2020). The smaller particles are then separated in the second MS event.

Chromatography, MS, and radiolocalization analysis of lipid samples can yield two types of data;

- Lipid class data: This quantifies the amounts of different classes of lipids present in the samples in a standard unit, usually nanomolar levels.
- Fatty acid composition and position data: This is a much more resolved view of lipidome that indicates quantities of different lipids within each class that vary in the carbon chain length and position(for example, on the glycerol backbone for glycerolipids) of the constituent fatty acids.

5.7 Flow cytometry

Flow cytometry, a dynamic technology, swiftly analyzes individual cells or particles as they move through a buffered salt-based solution, exposed to lasers (McKinnon, 2018). Each particle undergoes scrutiny for the visible light scatter and various fluorescence parameters, offering valuable insights into their unique characteristics(McKinnon, 2018). This versatile technology finds applications across diverse scientific disciplines, including immunology, virology, molecular biology, cancer biology, and infectious disease monitoring(McKinnon, 2018). Flow cytometry also has immense use in algal research as it could be used to measure cell counts within cultures and also for analyzing algal cells stained with different fluorochromes. For instance, propidium iodide staining of algal cells and their subsequent detection using appropriate detectors in the flow cytometer helps determine the extent of cell death. Another relevant use is the measurement of gene expression by measurement of fluorescence from fluorescent tags like the green fluorescent protein(GFP) that is attached to the gene under study. Flow cytometry also finds application in algal lipid studies specifically. Lipids within cells can be stained using dyes such as Nile red or BODIPY with characteristic excitation-emission spectra. Emission detection measurement with defined gates of the flow cytometer under specific detectors will aid in measuring the lipid content within cells.

As stated, measurement of visible light scatter is a crucial aspect of flow cytometry. There are two important visible light scatter measurement parameters measured by separate detectors, namely:

- Forward scatter (FSC): The measurement of scattered light in the forward direction which is indicative of the relative size of the particles or cells causing the scattering(McKinnon, 2018)
- Side scatter (SSC): The measurement of scattering in the perpendicular direction, which indicates the cell complexity or granularity(McKinnon, 2018)

Apart from these two detectors, the flow cytometer will also be equipped with other fluorescence detectors. These detectors can be of different types like photo-multiplier tubes or photo-diodes(McKinnon,

2018). Some of the advanced flow cytometers with improved sensitivity employ avalanche photo-diodes for fluorescence detection(McKinnon, 2018).

In addition to lasers and detectors, filters form an essential actor in the instrumentation. These include both excitation filters for controlling the wavelength of the laser beam used for exciting the flurochromes and the emission filters to allow only specific wavelengths to pass towards a detector or for deflecting specific wavelengths to respective detectors(McKinnon, 2018). Most flow cytometers used at present have precisely positioned dichroic filters and bandpass filters that steer and filter the emitted light, facilitating the detection and measurement of individual fluorochromes(McKinnon, 2018).

The use of multiple lasers, filters, and detectors in the instruments coupled with advanced software systems for data interpretation aids in the effective assessment of several parameters associated with the studied samples(McKinnon, 2018).

5.8 Autofluorescence measurements

The inherent ability of different biological molecules including chlorophyll, called native fluorescence or autofluorescence, can be utilized in cell biology studies for monitoring and assessment of the condition of the biological samples. One of the applications for this is the establishment of algal growth curves through autofluorescence measurements. The excitation of algal samples using wavelengths that specifically target fluorescence emissions from the chlorophyll molecules is a commonly used approach to create growth curves as these measurements are directly proportional to the fraction of active cells in the culture samples. This can be considered a non-invasive and insightful method for assessing the physiological dynamics of algal cultures as the excitation wavelength used, around 365 nm, is less prone to cause cell damage and the detected wavelengths, around 680 nm, is a reliable marker for cellular activity, eliminating the need for external dyes or intrusive sampling(Rost, 1999,García-Plazaola et al., 2015). This can be achieved using specialized equipment such as flow cytometers or spectrophotometers. The resulting growth curve, characterized by changes in autofluorescence over time, provides a comprehensive view of algal growth phases, including lag, exponential, stationary, and decline. This data can then be used to draw information about cellar content, metabolic activity, and growth dynamics to compare between cell lines or treatment conditions. This approach could prove to be useful in the comparative study between the wild type and *Alb3b* mutant cell lines of *P.tricornutum*. A conspicuous change in the autofluorescence growth curve is expected in this case as the previous study by Nymark et al., 2019 has proved a difference in pigment concentrations, photosynthetic activity, and growth pattern between the cell lines.

5.9 Pulse amplitude modulation(PAM) fluorometry for measuring photosynthetic efficiency

Apart from having applications in growth measurements and assessment of cell health, Autofluorescence from pigments can be utilized to evaluate the photosynthetic efficiency of diatoms. This is possible because autofluorescence emission is one among four techniques that the phototrophic cell uses to de-excite the pigments, particularly chlorophyll molecules in the photosynthetic reaction centers(Consalvey et al., 2005). The other three mechanisms by which these molecules, which are excited by light, get back to their stable state are as follows:

- **Thermal Dissipation:** In this process, energy is dissipated in the form of heat to a random neighboring molecule through the molecular motion of the excited chlorophyll molecules(Consalvey et al., 2005).

- **Energy Transfer:** In this process, energy is simply transferred to another chlorophyll molecule in the vicinity of the excited chlorophyll molecule, causing the excitation of an electron in the latter to a higher energy state(Consalvey et al., 2005). This also facilitates energy distribution within the photosynthetic system.

- **Photochemical Reaction:** This crucial step involves utilizing absorbed energy to drive a photochemical reaction(Consalvey et al., 2005). The excited electron, released from the chlorophyll molecule, initiates photosynthesis by participating in chemical reactions that convert light energy into chemical energy.

Additionally, a part of the excess energy acquired from the irradiance could be dissipated as heat through a mechanism called non-photochemical quenching(NPQ). This involves specific carotenoid pigments called diadinoxanthin and diatoxanthin which are inter-converted between each other in the xanthophyll cycle, the underlying reaction process of NPQ(Jahns and Holzwarth, 2012).

The strengths at which all these processes of dealing with the acquired light energy, including autofluorescence, are affected by each other. Therefore, the measurement of certain autofluorescence parameters helps us assess the levels of other mechanisms and their associated processes(Consalvey et al., 2005). The photosynthetic efficiency is usually determined by calculating the light utilization efficiency and electron transport rate(Consalvey et al., 2005). Pulse amplitude modulation fluorometry is one of the techniques in phyto-biological and algal research that utilizes fluorescence measurements from chlorophyll to calculate the photosynthetic efficiency and the levels of the other light-associated processes.

PAM employs short light pulses that can induce chlorophyll autofluorescence but not photochemical reactions, thus enabling the differentiation of fluorescence emissions from actinic light, which in turn diversifies its applications(Consalvey et al., 2005). Different measures of autofluorescence can be obtained from PAM. Some of the crucial ones among these are:

- **Fo (Minimum Fluorescence Yield):** The minimum fluorescence yield in the cells adapted to dark conditions, where all reaction centers are open. This means that there is no incident light and thereby, no electron transfer into the PSII reaction centers.

- **Fm (Maximum Fluorescence Yield):** The maximum fluorescence yield in the cells adapted to dark conditions after a saturating light pulse, that is intense enough to close all the reaction centers. This means that all the PSII reaction centers receive electrons from a donor in the light-driven electron transport chain, leading to the maximum fluorescence yield.

- **Fv (Variable Fluorescence):** It represents a measure of the range of fluorescence that can be induced in the dark-adapted state and is calculated as the difference between Fm and Fo.

- **F' (Fluorescence Yield in the Light-Adapted State):** The measure of fluorescence yield in the presence of actinic light, reflecting the extent to which PSII is closed under ambient light conditions.

- **Fm' (Maximum Fluorescence in the Light-Adapted State):** The maximum fluorescence yield in the cells adapted particular light levels after a saturating light pulse, indicating the maximum potential quantum yield of PSII under ambient light conditions.
- **Fq' (Quenched Fluorescence by photo-chemistry):** The amount of fluorescence suppressed as the light energy gets utilized for photochemical reactions. This can be calculated as the difference between Fm' and F'.

These measured parameters can then be used to calculate the light utilization efficiency and electron transport rate, which reflects the photosynthetic state of the cells. The maximum light utilization efficiency is calculated as the ratio of variable fluorescence and maximum fluorescence:

$$F_v/F_m = \frac{F_m\text{-}F_0}{F_m} \tag{1}$$

The light utilization efficiency at a particular actinic light level is calculated as the ratio of Photochemistry-quenched fluorescence and maximum light-adapted fluorescence:

$$\phi PSII = \frac{F'_m\text{-}F}{F'_m} = \frac{F'_q}{F'_m} \tag{2}$$

The electron transport rate is calculated as a function of the light utilization efficiency and the level of irradiance:

$$ETR = \phi PSII \times \frac{PPFD}{2} \times A \tag{3}$$

Alternatively, the intensity of Photosynthetic active radiation(PAR) measured in mol photons m^{-2} nm^{-1} s^{-1}) can be used instead of *PPFD*. Furthermore the calculation of *ETR* without the absorbance coefficient A yields the relative electron transport rate(*rETR)*(Consalvey et al., 2005))

5.10 Confocal laser scanning microscopy

Confocal laser scanning microscopy(CLSM) is one of the major advancements in the field of optical microscopy that helps achieve comparatively much higher resolution than conventional wide-field fluorescence microscopy. Although the resolution from CLSM is lower than electron microscopy, it has benefits like less complex sample preparation techniques and the ability to accommodate live three-dimensional imaging(Canette and Briandet, 2014).

As the name suggests, CLSM involves the acquisition of an image from just the focal plane and uses a laser to scan the specimen. Unlike conventional fluorescence microscopy, where the entire specimen is flooded with illumination, CLSM employs localized point-to-point excitation using a focused laser and subsequent detection of excitation (Canette and Briandet, 2014). The specimen, which is usually stained with a fluorescent dye, is illuminated with a laser of an ideal wavelength and at user-defined laser power intensities to a diffraction-limited spot on the specimen using a lens. This allows for a high energy density at the focused point followed by the excitation of the fluorescent probe. The excited signals traverse back through the lens and then are passed through a pinhole before detection which serves to eliminate the out-of-focus light and thereby increase the resolution(Elliott, 2019). The in-focus emissions from the excited

spot are subsequently detected by a photo-multiplier tube or a photo-diode that acts as a transducer to convert the photon signal to an electrical signal to be displayed as a pixel with specific characteristics. This process is repeated as the laser is scanned over the specimen point by point using two rapidly moving perpendicular scanning mirrors(Elliott, 2019).This results in a pixel-by-pixel acquisition of the emitted light from the specimen which can then be reconstructed for the final high-resolution image. The focal plane can be changed to a different depth by adjusting the lens position(Elliott, 2019). This change in focal point depth is used to shift the scanning in the z direction and can be used to obtain multiple optical sections of the specimen to form a Z-stack(Elliott, 2019). The acquired Z-stacks can then be processed to form 3D images of the specimen(Elliott, 2019).

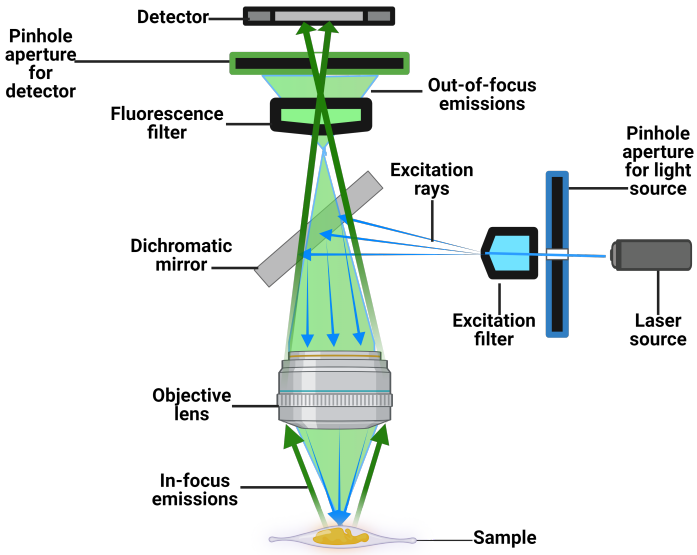


Figure 8: General working principle and components of confocal laser scanning microscope showing the various components like laser, scanning mirror, filters, and lenses. The blue light rays indicate the filtered light rays from the laser to excite the fluorescently labeled samples and the green light rays indicate the fluorescent emission from the dye that stained the sample. it also shows that both the excitation and emission rays pass through the same objective lens and a pinhole aperture that reduces the out-of-focus signals before they reach a detector. Figures were created using Biorender.

The benefits put forth by CLSM such as non-invasive optical sectioning and 3D imaging, improved contrast and resolution, reduced background noise and artifacts, and the ability to observe specimens in different experimental conditions like nutrient or chemical stress, make it one of the ideal imaging alternatives for algal cells under various treatment conditions. This can be used to observe structural variations in the morphology of cells, organelles, or other cellular components under stress conditions. This also includes studying variations in lipid droplet structure and distribution in cells using lipid-specific

stains like Nile red or BODIPY.

5.11 BODIPY staining

Fluorescent dyes that specifically stain lipids are advantageous for lipid measurement in microalgae because they offer a rapid and inexpensive analysis tool to measure neutral lipid content, avoiding time-consuming and costly gravimetric analysis(Rumin et al., 2015). They also allow for high-throughput screening of potential oleaginous microalgae to identify promising sources for commercial biofuel production. Two of the widely used fluorescent dyes for lipid staining are Nile red and BODIPY(Rumin et al., 2015).

Although Nile red (9-diethylamino-5H-benzo[a]phenoxazine-5-one) has been used for a wide range of lipid analysis studies it has certain limitations like reduced solubility and fluorescence in water, interference with chlorophyll, subpar photo stability levels, and permeation challenges for some microalgae species(Rumin et al., 2015). Recent studies have shown that BODIPY 505/515 is a better marker than Nile red for visualizing neutral lipid content in fluorescence microscopy studies(Rumin et al., 2015).

Boron dipyrromethene (commonly called BODIPY) is a class of compounds that has its absorption spectrum in the UV region and exhibits strong emission peaks(Rumin et al., 2015). These compounds can be subjected to chemical modifications to form fluorescent dyes with varying excitation and emission properties to be used in the imaging of different sub-cellular components. Additionally, BODIPY compounds are advantageous to be used in environmental manipulation experiments as they are tolerant to pH variations and polarity in the treatment conditions(Rumin et al., 2015). Furthermore, the non-destructive nature of these compounds allows the cells to be used for further analysis after BODIPY treatment(Rumin et al., 2015).

BODIPY 505/515 (4,4-difluro-1,3,5,7-tetramethyl-4-bora-3a,4a-diaza-s-indacene) is one of the commonly used variants of BODIPY in lipid analysis. This can be excited with a blue laser in the range of 450 to 490 nm and lead to a sharp emission in the green region in the range of 515 to 530 nm(Rumin et al., 2015). The optimal staining conditions may vary depending on the microalgae species, dye concentration, cell concentration, temperature, and incubation duration(Rumin et al., 2015). However, previous research has proved that using a permeation solvent like Dimethyl sulfoxide (DMSO) or glycerol will significantly improve the staining efficiency(Rumin et al., 2015). One of the research that attempted to delineate the ideal staining conditions for microalgae established that the optimal staining concentration be 0.067μg μL^{-1}at 1.10^{6} cells mL^{-1}, temperature at 25 °C, and incubation time of 10 minutes(Rumin et al., 2015). However, these values were calculated for other species of microalgae than *P.tricornutum*, and the optimum cell concentrations are species-specific(Rumin et al., 2015).

5.12 Real-time Polymerase chain reaction or quantitative PCR

Real-time or Quantitative polymerase Chain reaction, abbreviated as q-PCR, is a very commonly used molecular biology technique for detecting and quantifying particular target sequences in genetic material (Králík and Ricchi, 2017). This attribute makes it a widely applied tool for studying gene expression profiles (Bustin et al., 2009). This technique is based on the traditional polymerase Chain Reaction and involves all the basic steps involved in a typical PCR cycle:

- Denaturation: the samples are heated to a high temperature of about 95°C so that the double-stranded DNA becomes denatured to generate single-stranded DNA molecules
- Annealing: The reaction temperature is brought down to about 50 to 60 °C so that the primers, which are oligonucleotide sequences designed complementary to the target sequence, bind to the single-stranded DNA molecules.
- Elongation: The reaction temperature is increased to about 72°C so that A suitable thermo-stable DNA polymerase enzyme elongates the DNA strands at the regions where the primers have annealed using dNTPs.

This cycle is repeated several times until several millions of copies of the target sequence, the amplicon, are obtained. q-PCR differs from traditional PCR in that measurement of the amplicon is made after every cycle using fluorescence. This can be done either by using non-specific fluorescent dyes or specific oligonucleotide fluorescent probes. The higher specificity of the probes to detect just the desired product makes it more advantageous than dye-based detection as the problem of detecting non-target products is avoided(Kubista et al., 2006). These fluorescent dyes or probes used in q-PCR specifically bind to the double-stranded DNA and only fluoresce upon binding. Therefore the amount of fluorescence measured after each PCR cycle indicates the amount of amplified product in the sample after a particular number of cycleKrálík and Ricchi, 2017).

Differential gene expression studies using q-PCR are based on a parameter called the quantification cycle value or C_q value. It can be defined as the number of cycles in q-PCR after which the fluorescent emission from the desired product becomes detectable and distinguishable from the background(Králík and Ricchi, 2017). Since more of the initial amount of desired product leads to the product getting amplified to levels above this threshold earlier, the Cq value is inversely proportional to the amount of the studied product or gene in the sample(Králík and Ricchi, 2017).

The absolute quantities of the expressed gene under study can be calculated by measuring serial dilutions of the samples with known concentrations of the studied gene to make a calibration curve, and then using it to determine unknown concentrations(Yang and Rothman, 2004). However, it is also possible to get comparative gene expression profiles by measuring unknown samples against a control to determine expression as fold-change compared to the control(Bustin et al., 2009).

Differential gene expression studies using q-PCR usually involve extensive sample preparation before the PCR reaction cycles. This includes:

- RNA isolation: Purifying just the RNA fraction from the samples by using reagents, buffers, and filtration steps to remove proteins, fats, DNA, and other biological materials. It is usually done by using standard kits.
- Nanodrop assessment: Measuring the concentration of nucleic acid material present in the isolated samples by spectrophotometry(García-Alegría et al., 2020). The purity of the samples is estimated by measuring the absorbance ratio at 260nm(for nucleic acids) to 280nm(for proteins), and at 260nm to 230nm (for other contaminants like phenol)(García-Alegría et al., 2020).
- RNA integrity testing: Calculating the RNA integrity numbers, ranging from 1 to 10 of isolated RNA samples to determine RNA degradation levels and thereby the suitability of the samples for

q-PCR. This was earlier done using gel electrophoresis and subsequent visualization using ethidium bromide. It is now commonly done using a BioAnalyzer using special chips, fluorescent dyes, and an algorithm to form an electropherogram based on fluorescence intensities from different-sized RNA molecules(Puchta et al., 2020).

- cDNA synthesis: Synthesizing complementary DNA strands from isolated RNA using reverse transcription using an appropriate RT enzyme(Kuang et al., 2018). This is also done using standard kits and employing thermal cycles as in PCR. A primer mix and/or oligo-dT primers are used to anneal to the RNA molecules to initiate reverse transcription(Kuang et al., 2018).

5.13 Data analysis pipeline development

A data analysis pipeline is a series of data processing and analysis steps organized sequentially to transform raw data into meaningful insights. It can also be defined as a workflow that streamlines the entire data analysis process, making it more efficient, reproducible, and scalable. Such a pipeline needed to be developed for the MS-MS results dataset, so that it can be applied to various data from each mutant and wild type to make the analysis fast and comparable. However, the plethora of options available for incorporation into the pipeline, including basic plots, advanced graphs, statistical tests, Outlier imputation techniques, Null value imputation methods, and Machine learning or Deep learning algorithms, among others, could challenge the development process. However, domain expertise in the field of diatom lipidomics, statistical knowledge, and a comprehensive understanding of the dataset to be studied could lead to better decision making resulting in a pipeline that yields deeper insights. The pipeline typically includes various stages, each serving a specific purpose in the analysis process. The key components of a data analysis pipeline may include:

- Data Collection: The original or raw dataset from the MS-MS analysis can be used as the source to gather the required data to carry out a step further in the pipeline. For example, gathering PI levels in *Alb3b* 14 mutants under high light levels and low light levels.

- Data Cleaning and Pre-processing: The missing values and the outliers need to be imputed with scientifically reasonable values before analysis or testing for the results to be more reliable. The raw data is then transformed into an appropriate format for analysis and testing.

- Exploratory Data Analysis (EDA): An initial exploratory analysis is conducted through plots or graphs to understand the basic characteristics of the data, identify patterns and anomalies, and generate hypotheses.

- Feature Engineering or decomposition: Too many features, which in this case are lipid types, can make the analysis complex and the results hard to understand. This can be solved through this step where either new features that explain an adequate level of data variance are generated or some of the original features with less importance are eliminated.

- Modeling: Applying statistical or machine learning models or tests to the processed data for prediction, classification, clustering, or other analytical tasks. This can also involve visualization of values given by the different cell lines in the models to understand and compare their behavior.

- Validation: Assessing the performance of the models and ensuring that they generalize well to new, unseen data or checking whether the requirements for certain statistical analysis steps are met in the gathered and processed dataset.

- Interpretation: Interpreting the results of the analysis and drawing conclusions based on the insights gained.

A data analysis pipeline was developed for processing and exploring the MS-MS results dataset from the light experiments using mutants conducted by Nymark et al., 2019. This included the following main parts among others:

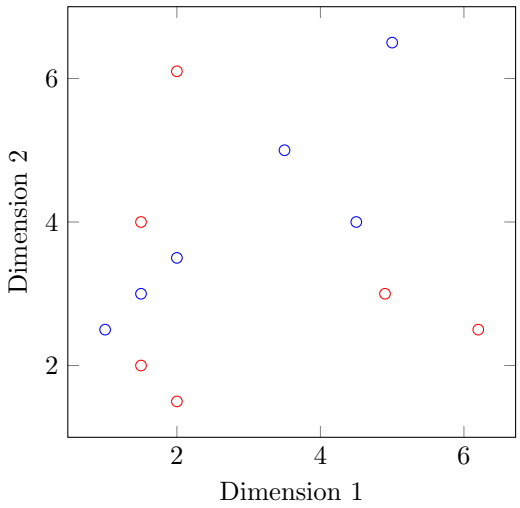
- principal component analysis: It is one of the most common features or dimensionality reduction methods in which datasets with a high number of features or variables are reduced into fewer new sets of variables called principal components(Jolliffe and Cadima, 2016). These principal components are linear functions of the original variables and will retain the maximum possible variance observed in the original dataset successively(Jolliffe and Cadima, 2016).

- Outlier imputation: A dedicated function was defined that takes into account all the similar sample types, that belong to the same cell line and same light conditions, calculates the inter-quartile range, and imputes all the values that are outside the following range (Q1-1.5xIQR) to (Q3+1.5XIQR) where Q1 is the first quartile, Q3 is the third quartile and IQR is the inter-quartile range. The substituted value in the place of the outliers was the mean of all the values belonging to the same sample type.

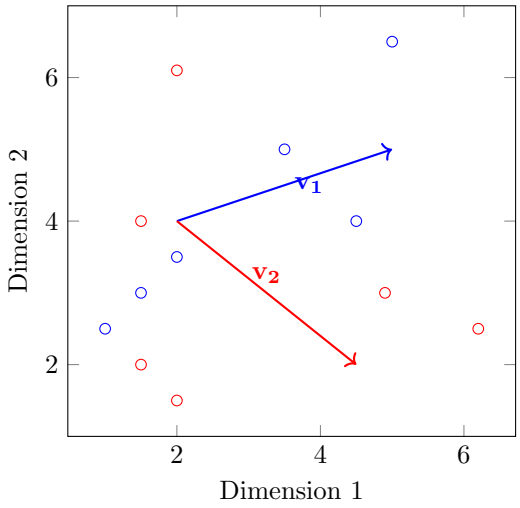
- Welch's T-test: This is similar to the student-T-test except that it does not assume equal variances between the samples compared during the calculation of test statistics and p-values.

5.14 Feature decomposition and extraction using principal component analysis

(a) Original Data Points showing no discernible variation or clustering between the data points from two groups



(b) Calculating the orthogonal PCA Vectors in the original feature space



(c) Transformed Data presented in the first two principal components indicating variation between the two groups of data points

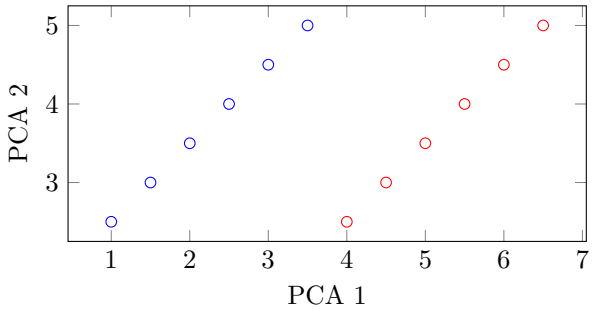


Figure 9: Illustration of different steps in principle component analysis. The figure was created using LaTeX.

Conducting a principal component analysis before exploratory data analysis or predictive analytics is a widely used procedure in most data science project workflows to help reform a complex dataset, usually with a redundant number of variables or features, to a simple and more insightful one with a new set of calculated and uncorrelated features or vectors called as the principal components(Shlens, 2014).

Principal component analysis involves the assessment of the reformed dataset to obtain otherwise clouded information. These components are defined to hold the most variance of the original dataset, with principal component 1 possessing the highest variance followed by principal component 2, and so on. Therefore, PCA contributes to noise reduction by emphasizing components with the highest variance, assumed to represent the underlying structure of the data. This non-parametric technique operates without making assumptions about the data distribution or structure, offering flexibility and applicability across diverse datasets including the MS-MS lipidomics dataset.

These newly established features will allow the generation of more insightful visualizations, presenting relevant information, which can otherwise be vague when presented with the original features. This is because PCA facilitates the representation of high-dimensional data in a lower-dimensional space(Shlens, 2014). Through this plotting based on principal components, intricate patterns, and relationships become more accessible for observation and interpretation(Shlens, 2014). Additionally, this dataset decomposition process reduces the computational power demands required for approaches like machine learning or deep learning. Thus it helps in the management of large volumes of data with minimal loss of the essential trends or patterns. Furthermore, PCA serves as a powerful feature extraction method, enabling the identification of significant features within the data.

The main steps in PCA can be loosely defined as finding new vectors in the original feature space of data points and plotting the same data points in the new vector space to observe for variation in the data. This is illustrated in figure 9

A standard principal component analysis involves the following steps:

5.14.1 Data Standardization

Standardize the dataset to have zero mean (μ) and unit variance (σ) for each variable. This is also called data scaling, where the data points are substituted with corresponding z-values.

$$Z_{ij} = \frac{X_{ij}\text{-}\mu_j}{\sigma_j}$$

Where Z_{ij} is the standardized value for variable j in observation i, X_{ij} is the original value, μ_j is the mean, and σ_j is the standard deviation.

5.14.2 Covariance Matrix

The covariance matrix is a symmetric matrix that provides the covariance values between the different features of the given dataset. In the case of the MS-MS data set this would be the covariance values between different lipid types. It should be noted that the individual samples whose measurements were considered for covariance matrix formation might come from multiple populations, which in this case would be the different cell lines. For instance, the PCA might be done for checking for separate clustering of WT and one of the *Alb3b* mutant lines.

Compute the covariance matrix ($\sum$) of the standardized or scaled dataset.

$$\sum^{(\mathrm{X,Y})} = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(\mathbf{X}_i\text{-}\bar{\mathbf{X}})(\mathbf{Y}_i\text{-}\bar{\mathbf{Y}})$$

Where n is the number of observations, X and Y are two individual features $\mathbf{X}_i$, and $\mathbf{Y}_I$ is the i^{th} standardized measurement or observation for the variable X and Y respectively, and $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ is the mean vectors for the respective variables.

The calculated covariance matrix will have the following format:

$$\sum^{(\mathrm{Z})} = \begin{bmatrix} \mathrm{Cov}(\mathrm{Z}_{\text{lipid 1}}\ ,\ Z_{\text{lipid 1}}) & \mathrm{Cov}(\mathrm{Z}_{\text{lipid 1}}\ ,\ Z_{\text{lipid 2}}) & \dots & \mathrm{Cov}(\mathrm{Z}_{\text{lipid 1}}\ ,\ Z_{\text{lipid n}}) \\ \mathrm{Cov}(\mathrm{Z}_{\text{lipid 2}}\ ,\ Z_{\text{lipid 1}}) & \mathrm{Cov}(\mathrm{Z}_{\text{lipid 2}}\ ,\ Z_{\text{lipid 2}}) & \dots & \mathrm{Cov}(\mathrm{Z}_{\text{lipid 2}}\ ,\ Z_{\text{lipid n}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathrm{Z}_{\text{lipid n}}\ ,\ Z_{\text{lipid 1}}) & \mathrm{Cov}(\mathrm{Z}_{\text{lipid n}}\ ,\ Z_{\text{lipid 2}}) & \dots & \mathrm{Cov}(\mathrm{Z}_{\text{lipid n}}\ ,\ Z_{\text{lipid n}}) \end{bmatrix}$$

Where Z represents the collective population from which the measurements are taken.

It should be noted here that the values along the main diagonal of the matrix are the variance measures for each lipid type for the particular cell line. The number of lipid types, represented as n depends on the dataset under analysis. For the lipid class dataset, it will be 10 incorporating the phospholipids(PI, PC, PG, PE), glycolipids(MGDG, DGDG, SQDG), betaine lipids(DGTA), and neutral lipids(DAG, TAG).

5.14.3 Eigenvalue decomposition

Eigenvectors are non-zero vectors that maintain their direction unchanged after the application of a linear transformation(Libretexts, 2023). The linear transformations applied to a vector can be represented as matrices, which when multiplied by the vector give an image or a transformed version of the eigenvector(Dan Margalit, n.d.). The scale by which the matrix transforms an eigenvector is referred to as the eigenvalue(Dan Margalit, n.d.). That is, for a given matrix A, a unit vector v and a corresponding eigenvalue of λ :

$$Av = \lambda v$$

For the calculated covariance matrix for the MS-MS data, the equation will become:

$$\sum^{(\mathrm{Z})}\mathbf{v} = \lambda\mathbf{v}$$

Where λ is an eigenvalue and $\mathbf{v}$ is the corresponding eigenvector.

Eigenvalue decomposition represents the crucial step in PCA(Raschka, 2015). This is a factorization step to determine the eigenvectors and eigenvalues of the covariance matrix $\Sigma^{(Z)}$. By definition, the principal component or axes are the eigenvectors of the covariance matrix of the dataset with the corresponding eigenvalues representing the extent of variance covered by the eigenvectors(Raschka, 2015). Thus, selecting a subset of the top eigenvectors, with the highest eigenvalues allows for a transformation of the dataset into a new coordinate system. This process enables a reduction in the dimensionality of the dataset while retaining the essential information contained in the most significant directions of variability. Perform eigendecomposition on$\sum^{(\mathrm{z})}$ to obtain eigenvalues (λ) and corresponding eigenvectors ($\mathbf{v}$).

5.14.4 Selection of Principal Components

Sort the eigenvalues in descending order and choose the top k eigenvectors based on the desired number of principal components.

5.14.5 Projection

Project the standardized data onto the selected principal components to obtain the transformed dataset.

$$\mathbf{T} = \mathbf{ZV}$$

Where **T** is the matrix of transformed data, **Z** is the standardized data matrix, and **V** is the matrix of selected eigenvectors.

For the MS-MS data, the projection was performed using both the top two and three principal components for the mutant lines studied. This was then used to create Two and three-dimensional graphs respectively to observe the clustering of data points in the new coordinate system.

5.14.6 Calculation of Loading Scores

The loading scores for the different variables are calculated based on the angle of rotation of the various principle components relative to the axes that originally represent the variables(David T. Harvey, n.d. The loading scores provide insights into the importance of each variable in contributing to the principal components and are a commonly used technique for feature extraction.

5.15 Statistical tests

5.15.1 Levene's Test

Assessment of homogeneity of variances between the groups or samples to be compared can help in deciding the best option for hypothesis testing. Though this can be achieved through plotting options like the "residuals versus fits plot", it can only indicate high or conspicuous differences. Levene's test evaluates the homogeneity variances between samples and is often employed before analysis of variance (ANOVA) or independent T-tests to ensure the samples are in accord with the homogeneity of variances assumption(Gastwirth et al., 2009). The test statistic is based on the absolute deviations of the observations from the group mean(Gastwirth et al., 2009).

$$F = \frac{(N\text{-}k)}{(k\text{-}1)} \times \frac{\sum_{i=1}^{k} n_i (Z_{i\cdot}\text{-}Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij}\text{-}\bar{X}_{i\cdot})^2} \tag{4}$$

where:

W is the Levene's test F-statistic,

N is the total number of observations,

k is the number of groups,

n_i is the number of observations in the i-th group,

$Z_{i\cdot}$ is the mean of the absolute deviations of observations from the group mean in the i-th group,

$Z_{\cdot\cdot}$ is the grand mean of the absolute deviations,

X_{ij} is the j-th observation in the i-th group,

$\bar{X}_{i\cdot}$ is the mean of observations in the i-th group.

5.15.2 Shapiro Wilk's test

Most of the hypothesis testing including ANOVA and T-test assumes that the population under study is normally distributed. This assumption can be tested using basic distribution plotting. But this can potentially involve biases and assessing a large number of samples can be time-consuming. The Shapiro-Wilk test is a statistical test that provides a quantification of how normally distributed the population is(**shapiro**). It is normally used to supplement normality plots before other statistical tests(Shapiro and Wilk, 1965). However, it is sensitive to departures from normality in the distribution tailsShapiro et al., 1968. The test statistic is based on the covariances between the sorted sample values and the expected values under normality(Shapiro and Wilk, 1965).

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_{(i)}\text{-}\bar{X})^2} \tag{5}$$

where:

W is the Shapiro-Wilk test statistic,

n is the sample size,

a_i are constants derived from the covariance matrix of the order statistics,

$X_{(i)}$ is the i-th order statistic,

$\bar{X}$ is the sample mean

5.15.3 T-test

The comparison of two independent groups, for example, cells treated in high light and low light separately, can be executed using a T-test. There are two options available here, namely,

- Student-T-test
- Welch's-T-test

The T-statistic calculations, for the MS-MS dataset, in both tests utilize the following formula:

$$t = \frac{\bar{X}_{lipid1}\text{-}\bar{X}_{lipid2}}{\sqrt{\frac{s^2_{lipid1}}{n_{lipid1}} + \frac{s^2_{lipid2}}{n_{lipid2}}}} \tag{6}$$

where:

t is the Welch's t-test statistic,

$\bar{X}_{lipid1}, \bar{X}_{lipid2}$ are the sample means,

$s^2_{lipid1}, s^2_{lipid2}$ are the sample variances,

n_{lipid1}, n_{lipid2} are the sample sizes.

Both tests assume the existence of normality in the compared groups. However, these tests differ in the fact that the student T-test assumes equality of variances between the compared groups, Whereas the Welch T-test is a modification of the traditional Student's t-test that accommodates unequal variances(Lu and Yuan, 2010). This is achieved by calculating the degrees of freedom(df) in a more detailed manner,

taking individual variances and sample sizes into consideration, unlike the pooled variance used for calculating df in Student's T-test.

This can be represented by the following equations:

df in Student's T-test is calculated as:

$$\mathbf{df} = \mathbf{n1}+\mathbf{n2} - 2$$

where **n1** and **n2** are sample sizes of the independent samples considered.

df in Welch's T-test is calculated as:

$$df = \frac{\left(\frac{s^2_{lipid1}}{n_{lipid1}} + \frac{s^2_{lipid2}}{n_{lipid2}}\right)^2}{\frac{\left(\frac{s^2_{lipid1}}{n_{lipid1}}\right)^2}{n_{lipid1}\text{-}1} + \frac{\left(\frac{s^2_{lipid2}}{n_{lipid2}}\right)^2}{n_{lipid2}\text{-}1}} = \frac{\left(\frac{s^2_{lipid1}}{n_{lipid1}} + \frac{s^2_{lipid2}}{n_{lipid2}}\right)^2}{\frac{\left(\frac{s^2_{lipid1}}{n_{lipid1}}\right)^2}{n_{lipid1}\text{-}1} + \frac{\left(\frac{s^2_{lipid2}}{n_{lipid2}}\right)^2}{n_{lipid2}\text{-}1}}$$

The results from the Levenes test can be used to decide the selection of either of these tests for particular comparisons.

5.16 Statistical plotting

Data visualization plays an inevitable role in exploratory data analysis as it allows the user to get an overview of the entire data set or the results from the statistical analyses of the dataset. A wide range of visualization options are available to plot the data both in two and three-dimensional space under the Cartesian coordinate system. This includes general application plots like bar plots, scatter plots, pie charts, violin plots, box plots, and line plots, among others to some specific plots like the scree plot and loading plots used in principal component analyses. It is possible to make custom plots from scratch or through modification of the plots mentioned above using Python or R programming. In this master thesis, the results of PCA were visualized using Scree plots and Biplots and the results of statistical modeling were visualized using custom-made T-statistic v/s P-value plots.

5.16.1 Scree plot

Scree plots can be bar plots or line plots that represent the amount of variance explained by each of the calculated principal components from PCA. These plots are useful in deciding the number of principal components to be used in further analysis including visualizations. These plots usually have the percentage or fraction of explained variance on the y-axis and the principal components on the x-axis.

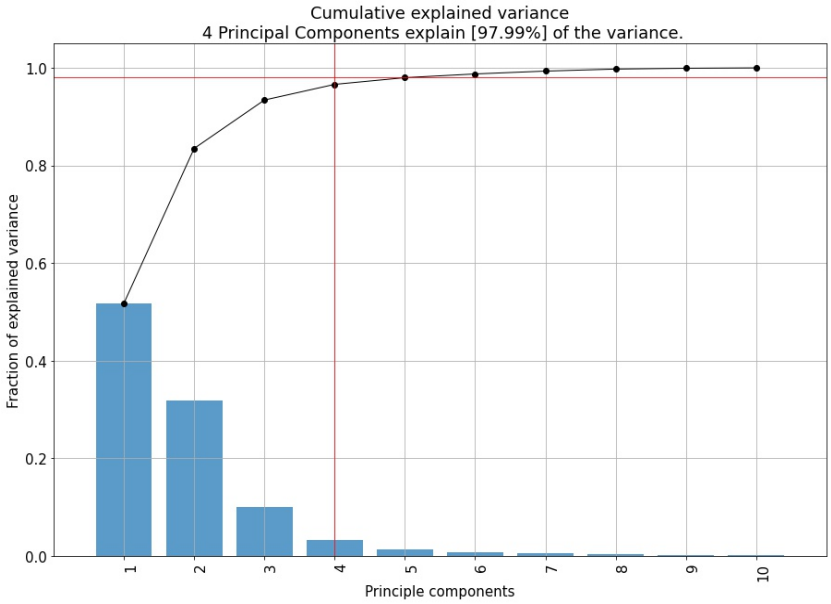


Figure 10: An example scree plot created in Python using the 'plot' function from 'pca' package in Python demonstrates how many components are required to achieve a total explain variance of more than 95%.

5.16.2 Loadings plot

Loading plots are graphical visualizations of the influences or loading scores of the different variables in the original dataset toward the principal components. These plots consist of arrows that represent the original variables in a two or three-dimensional cartesian system with the axes being the principal components. The starting point of all the arrows will have all the coordinates zero and the length of the arrow indicates the loading score(Medium, 2018). The direction in which the arrow points indicates the direction in which the respective variable causes variance and the angle between the arrows is a measure of the correlation between variables(Medium, 2018). For example, if two arrows lie close to each other at a very small acute angle, those two variables probably exhibit a positive correlation. If these are perpendicular to each other, the two variables probably have no significant correlation. In the final case, if the two arrows are at a large obtuse angle they are probably negatively correlated(Medium, 2018).

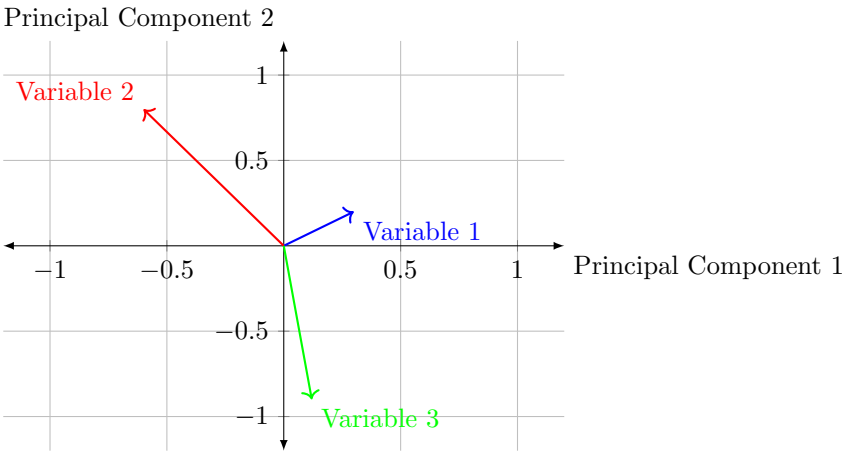


Figure 11: An example loading plot with two principal components and three variables. Each variable is represented by a vector with a specific color. The length of the vector is proportional to the loading score of the vector and the direction of the arrow indicates the direction in which the variable causes variance created using LaTeX.

5.16.3 Biplots

Biplots are a combination of a Loading plot(Figure 11 and a PCA scatterplot (Figure 9c). Thereby, it indicates both how the different data points are distributed and how much the different variables contribute to the variation between different data points in the space of the principal components. Figure 12 represents an example of a biplot with three principal components and four different clusters of data points varying in the values of ten different variables

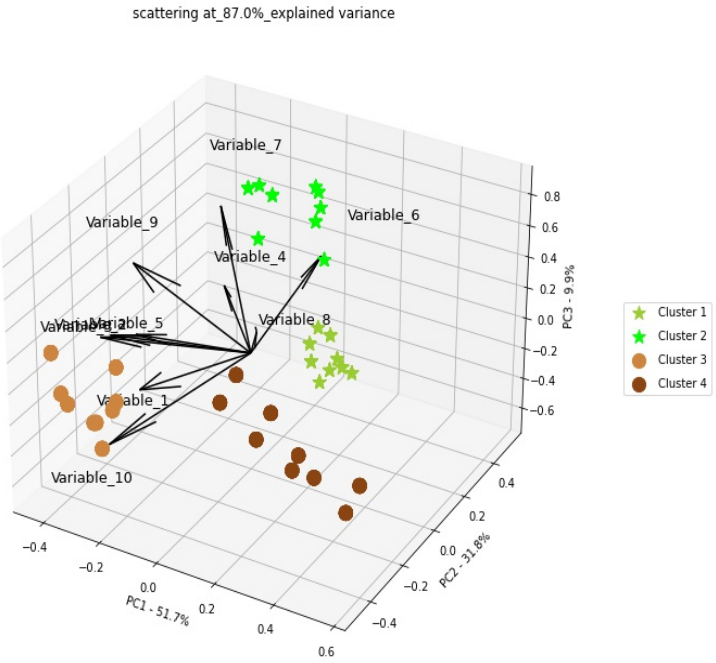


Figure 12: An example three-dimensional biplot created using Python using various functions from the 'Matplotlib' package. This represents a combination of the PCA scatter plot and loadings plot.

6 Materials and Method

6.1 Data analysis pipeline development

Exploratory data analysis and inferential statistics on the MS-MS lipidomics data set were performed using Python(version 3.8) in the Jupyter Notebook API from the Anaconda software package version 3.0. Various Python libraries and associated functions were used during the process. Some of the core libraries employed for the project are:

Table 1: Python Packages used for Data Analysis Pipeline Development and their Purposes

Python Packages	**Purpose**
Pandas	Data manipulation
Numpy	Numerical operations
Matplotlib	Data visualization
Seaborn	Statistical data visualization
Scikit-learn	Machine learning tools
Scipy	Scientific computing

Two main operations were performed on the MS-MS results data set:

- principal component analysis
- Modelling based on inferential statistics

Principal component analyses were performed on the filtered versions of the main data set containing data from samples of one of the mutant lines and the Wild-type. The decision to filter data like this was based on several trial runs of PCA with different combinations of samples in the filtered dataset.

The data pre-processing steps before PCA included:

- **Standard scaling of the data**: Performed using the 'scale' function from the 'pre-processing' package of Sci-kit Learn to standardize the data as explained in section 5.14.1 .
- **Outlier and zero value imputation**: Performed using custom python functions. The outlier imputer function detects outliers based on the IQR rule and imputes them with the mean of the measurements from a cell line under particular light conditions. The zero value imputer replaces the zero or null values with a random number less than 0.01% of the total value of measurements from a cell line under particular light conditions.

The steps in PCA including Covariance matrix calculations, eigenvalue decomposition, and loading score calculations were all performed using the 'PCA' function from the 'decomposition' package of Sci-kit Learn.

The inferential statistics and the modeling based on it were also preceded by the outlier and zero value imputation using custom functions as done for PCA. The statistical analysis involved performing

Figure 13: Different components of the data analysis pipeline developed for analyzing the MS-MS data. The original dataset(green) passes separately through two processes(yellow): PCA and statistical modeling. Both processes have different steps arranged in order from top to bottom and connected by solid downward arrows. Steps in principle components are shown in the orange blocks, while steps in the statistical modeling are in the red blocks. The blue blocks connected to corresponding steps by dashed arrows indicate the Python function and the package used(in parentheses) to execute the step. The flow chart was created using LateX.

Levene's test and Shapiro Wilk's test to assess the homogeneity of variances and Normality respectively. These were executed using the 'Levene' and 'Shapiro' functions from the 'stats' package of Scipy. A custom function was made to perform a T-test that will first extract the data for a specified cell line in both light conditions and store them in separate objects. Then it performs a logarithmic transformation of these data to improve the normality before performing the T-test. The performed test will be either the standard student's T-test or Welch's T-test based on the p-values from Levene's test that indicate equality of variances. That is if the p-values of Levene's test for the data from a cell line in two different light conditions are less than 0.05 the T-test function will perform a Welch's T-test. In the opposite scenario, it will be a Student's T-test. The actual statistical testing steps for calculating the T-statistic and p-values for both the T-test variants were done using the 'ttest_ind' function from the 'stats' package of Scipy. The results from all the T-tests performed were finally represented together in a scatter plot with the T-statistic on the x-axis and p-values on the y-axis to better understand the nature of changes between light conditions for particular cell lines. In these graphs, the points to the right of the vertical line at x=0 indicate an increase, and those toward the left indicate a decrease. The points below the horizontal line at y=0.05 indicate that the change is significant.

All the visualizations including scree plots and biplots were made from scratch using Matplotlib and Seaborn.

The workflow followed during the project is represented in the figure 13

6.2 Sample acquisition and maintenance

- Axenic cultures of three mutant lines (*Alb3b*14.8, 16.7, 19.7) were obtained from Marianne Nymark.
- The cultures were split into multiple stock solutions and were acclimated for 2 weeks under controlled conditions: 15°C $\pm 2, 100\,\mu M$ light, and agitation at 150 rpm.
- Different stock cultures, from the same parent stock, of the various cell lines including the wild type were used for different experiments.

6.3 Experimental setup

6.4 Autofluorescence measurements

- Three treatment conditions were established: High Light (HL) with 700 µmol photons m^{-2} s^{-1}, Medium light(ML) with 200 µmol photons m^{-2} s^{-1} intensity and Low Light (LL) with 35 µmol photons m^{-2} s^{-1} intensity in the same room were the stock cultures were acclimatized.
- Cells were cultured in 12-well plates with three biological replicates for each cell line and three blanks for background subtraction.
- The Cytation 5 plate reader from BioTek was used for the measurement and the following protocol was set for running the plate reader:
  - Orbital shake for 0.07 ms.
  - Temperature set point of 22° Celsius.

- Fluorescence measurements with a excitation wavelength of 480 nm and detection wavelength of 680 nm corresponding to the autofluorescence levels of chlorophyll molecules.

- Measurements were done for 7 consecutive days at a specific time to acquire the growth curves.

6.5 Lipid Measurements with BODIPY using flow cytometry

6.5.1 Experimental setup

- The same light conditions as described in section 6.4 were used for the flow cytometry experiments.
- cultures were maintained at an approximate volume of 40 ml in tissue culture flasks with three biological replicates per cell line(WT, *Alb3b*-14,16,19) for each treatment condition(HL,ML, and LL). That is a total of 12 cultures in each of the three light conditions.
- These cultures were acclimated for 2 weeks with media being refreshed every 2^{nd} day.
- After the acclimation period of 2 weeks, the cell counts were estimated using flow cytometry,

6.5.2 BODIPY staining

- Stock solutions of BODIPY were made using DMSO as the solvent. The stock solutions had a concentration of 5mmol and were stored under refrigeration in dark conditions.
- The estimated cell counts of the cultures acclimated to treatment conditions were used to estimate the volume required to acquire 5 ml solution with 1 million cells per ml using the following formula:

$$\mathbf{V}(in\ mL) = \frac{\mathbf{5}}{\mathbf{M}}$$

  Where V is the volume of culture to be diluted to 5 ml and M is the measured cell concentration
- If the measured concentration was less than 1 million cells per ml, the volume of the culture containing 1 million cells was centrifuged and the cells were re-suspended in 1 ml of F/2 media. The following formula was used for estimating this volume:

$$\mathbf{V}(in\ mL) = \frac{\mathbf{10^6}}{\mathbf{M}}$$

  Where V is the volume of culture to be centrifuged and M is the measured cell concentration
- 5 ml solutions with 1 million cells per ml of all 24 cultures were prepared using the above methods and were used for staining.
- For staining, 1 µL of the BODIPY stain was added to 1 ml of the prepared solution, containing around 1 million cells per ml.
- Three technical replicates were used from the prepared 5 ml of solution of each of the biological replicates.
- The stained cells were then incubated in the dark for 10 minutes before measurement using the flow cytometer. The timing of staining for different samples was timed to achieve exactly 10 minutes of dark incubation for all the samples before measurement.
- A negative control, with 1 ml F/2 medium containing 1 million unstained cells was used for each set of biological replicates for a cell line in a particular light treatment.

6.5.3 Flow Cytometer Operation

- The NovoCyte 2000 instrument from Agilent was used for flow cytometry.
- The median fluorescence intensity values of the peak obtained in the FITC-GFP channel were measured for each technical replicate.
- The final measurement of lipid quantity was calculated as the mean of measured median values of technical replicates for each of the three biological replicates.

Samples from the same experimental setup for flow cytometry were used for CLSM, PAM, and q-PCR.

6.6 Structural Observation of Lipid Droplets Using CLSM

The staining protocol for BODIPY used for imaging using CLSM is the same as that for flow cytometry. That is, specific volumes of the cultures were diluted to obtain a final concentration of 1 million cells per ml and 1 ml of the dilution was stained with 1 µL of BODIPY. However, for imaging, only 1 ml of dilution was prepared from the cultures as the volume required on a microscopic slide is just 6 µL. Thus, for each of the cultures 1ml of dilution containing 1 million cells was prepared by the previously mentioned method and from each of this dilution 3 microscopic slides were prepared with 6 µL on each of them.

6.6.1 Slide preparation

- Standard light microscopy glass slides of size 25 x 75 mm were used for imaging
- A 0.12 mm thick imaging spacer from Grace Bio-labs was stuck onto the center position of the slide.
- 6 µL of the sample was pipetted to the center of the imaging spacer.
- A glass coverslip of 24 x 24 mm was stuck onto the top of the imaging spacer to form a thin layer of the sample between the cover slip and the glass slide.

6.6.2 CLSM operation

- The Leica SP8 confocal laser scanning microscope was used for the imaging
- The LasX software, compatible with the specific instrument, was used for controlling the imaging settings and for image acquisition
- Samples were observed under the 60× objective of the microscope using the 488 nm laser at a power of 1.35 and gain set to 400 V.

6.7 Quantitative PCR(q-PCR)

Genes associated with lipid metabolism in diatoms were selected for the q-PCR based on results from other light experiments, where these genes appeared to have changed expression under different light conditions.

The genes selected for differential gene expression analysis in the wild-type and mutant cell lines under different light treatments are presented in Table 2

Table 2: Information about genes used for differential expression study using q-PCR obtained from NCBI GenBank database

Gene ID	**Accession Number**	**Protein Title**
PHATRDRAFT_37652	XM_002181731.1	FADB (Malonyl-CoA:ACP transacylase)
PHATRDRAFT_48423	XM_002182796.1	PTD12 (Precursor of desaturase omega-6 desaturase)
PHATRDRAFT_20508	XM_002180392.1	ELO6b_2 (Elongase delta 6 elongase)
PHATRDRAFT_50443	XM_002185338.1	Predicted protein (Fatty acid desaturase with Chl targeting motif)
PHATRDRAFT_10068	XM_002177895.1	FABI (Enoyl-ACP reductase)
PHATRDRAFT_20143	XM_002179910.1	ACS1 (long chain acyl-CoA synthetase)
PHATRDRAFT_bd765	XM_002176380.1	Predicted protein (Acyl-CoA thioesterase)
PHATRDRAFT_54756	XM_002181952.1	CDS1 (Phosphatidate cytidylyltransferase)
PHATRDRAFT_42683	XM_002177125.1	PLC_delta (Phospholipase C isoform delta)
PHATRDRAFT_bd976	XM_002176456.1	Predicted protein (Acetyl-CoA Carboxylase)
PHATRDRAFT_41570	XM_002185462.1	PTD15 (Precursor of Omega 3 desaturase)

In addition to these selected genes, two reference genes were selected for normalizing the Cq values during data analysis of the q-PCR results. These are:

- RPS5/30S ribosomal protein S5 encoding gene (Phatr2_42848): This nucleus localized gene was found to be not responsive to different light treatments based on DNA microarray analysis in a previous study(Valle et al., 2014).
- Putative hiv-1 rev binding protein (Phatr2_42776): This gene was determined as non-responsive to high-light treatment using microarray analysis in a previous study(Nymark et al., 2009).

6.7.1 Primer design

Primers required for q-PCR were designed using the NCBI BLAST tool for 11 selected genes associated with lipid metabolism as presented in the supplementary table 5 in the appendix.

6.7.2 Cell harvesting

The cells treated under each light treatment for two weeks were extracted 2 or 3 days after the last refreshing by vacuum filtration using the 0.65µm Durapore membrane filters. The cells were then separated from the filters by centrifugation(5000xg for 1 min) and discarding the supernatant.The cells were then frozen by keeping them in liquid nitrogen(-196°C) and were finally stored at -80°C until RNA isolation was performed.

6.7.3 RNA isolation

Isolation of RNA from the samples treated under different light conditions was done using the RNeasy plant mini kit. This is a standardized kit for isolating RNA from plant and filamentous fungi samples and is based on special spin columns. These columns have silica-based membranes with affinity to RNA

molecules. This particular binding ability is combined with centrifugation for drawing out the RNA fraction in the sample and subsequently eluting them into a solution with RNAase-free water.

As mentioned these kits are designed for extracting plant and fungal RNA. Therefore, a special tissue lysis procedure was followed before the spin column RNA extraction. The 'Lysis and homogenization of fatty acid tissues using the tissuelyser II' Protocol From The "Qiazol Handbook For efficient lysis of fatty tissues and all other types of tissue before RNA purification"(Qiagen, 2021) was followed for this purpose. The RNA pellets obtained at the end of this particular protocol were re-suspended in 100 µL of RNAase-free water and were used as starting material for the "RNA cleanup" protocol in the RNeasy Mini Handbook from Qiagen(Qiagen, 2023). The optional procedure for on-column DNA digestion was applied to all the samples for which the RNAse-free DNAase set provided in the RNeasy mini kit and the corresponding protocol was used.

The isolated RNA was then stored in a frozen state at -80°C.

6.7.4 Nanodrop assessment

The concentration of the isolated RNA and its purity were assessed using the Nanodrop One instrument from Thermofisher Scientific, which is a micro volume UV-Vis spectrophotometer to calculate absorbance for nucleic acids at 230nm. These values are then used to calculate standard ratios to other absorbance wavelengths, particularly 230nm for carbohydrates or compounds that may reside from reagents used during RNA isolation like Phenol, and 280nm for proteins. The measurements were obtained using 1 µL of the sample and RNAase-free water was used for blanking the instrument before measurements.

6.7.5 Bioanalyzer

The Agilent Bioanalyzer 2100 was used for calculating RNA integrity numbers from all the isolated samples. For this purpose, the Agilent RNA 6000 nanokit and the associated standard protocols were used. This procedure included the use of micro-fluidic chips into which a mix of the provided gel matrix and the RNA dye concentrate was added for loading the RNA samples and a standard RNA ladder for reference. All the RNA samples and the ladder were also added with a marker RNA provided in the kit with a known length of 25 nt as a control.

6.7.6 Complementary DNA (cDNA)synthesis

QuantiTect Reverse Transcription Kit and the associated standard protocol from QIAGEN were used for the cDNA synthesis from isolated RNA samples. This involved the following main steps:

- Incubation with a gDNA wipe-out buffer: An incubation process at 42°C for removing genomic DNA contaminants in the sample to avoid errors during q-PCR for 2 minutes.
- Reverse transcription: The process of synthesizing DNA strands from RNA strands by incubation of the samples with a master mix containing the provided Quantiscript reverse transcriptase, primer mix, and an optimization buffer with dNTPs at 42°C for 15 minutes followed by inactivation at 95°C for 3 minutes.

The Quantiscript reverse transcriptase carries out both RNA-dependent DNA polymerase activity and a hybrid-dependent exoribonuclease or RNase H activity to degrade RNA bound to DNA in hybrid

molecules. The primer mix provided is a mix of oligo-dT primers that can bind to poly-A tails of processed mRNAs and random primers that could bind to random complementary sequences in the RNA strands. Along with the cDNA synthesis reaction negative controls(RT-ve) were set up for all the samples in which a master mix with all the components except the reverse transcriptase was used to avoid cDNA synthesis. The volume corresponding to 1µg and 0.5µg of each of the RNA samples were used for cDNA synthesis and the RT-ve controls, respectively. The RT-ve reactions are done to check for genomic DNA contamination later during the q-PCR reactions.

After the reaction process was completed the samples(supposedly having cDNA) and the negative control were diluted 5 times and stored at -80°C.

6.7.7 Running real-time PCR

The Roche LightCycler 480 system was used for running the q-PCR reactions on the mentioned genes and reference genes on specific 96 well plates designed for this equipment(480 Multiwell Plate 96). The LightCycler 480 SYBR Green I Master kit, specifically prepared for q-PCR reactions on this instrument was used for all the samples. This master kit includes a Master mix containing the saturating dye called SYBR green, dNTP mix, and a thermostable DNA polymerase called the Taq polymerase, mixed with reaction buffers and magnesium chloride to optimize the reaction. Additionally, PCR-grade water is provided to adjust the volume of the final reaction mix. The final reaction mix is prepared with the components from the kit and the primers required for the studied gene. 15 µL of this final reaction mix is then added to the 5 µL one of the cDNA samples to form the final reaction volume of 20µL.

6.7.8 q-PCR Data analysis

The data from the Lightcycler 480 system was processed using the LinRegPCR software(version 2021.2) for the calculation of the PCR efficiencies and Cq values. This data was then used to calculate the corrected normalized relative quantity (CNRQ) to determine the relative gene expression levels using the qBase+ software(version 3.4) from Biogazelle. These values were then transformed by taking the logarithm to the base 2 and then subjected to two-way ANOVA and posthoc tests using Python for determining significant upregulations or downregulations.

7 Results

The results are divided into two parts:

- **Insights from the exploratory data analysis(EDA) of MS-MS dataset:** This includes, in the first part, results from PCA that compares each of the mutant cell lines individually with the wild type under both LL and ML conditions. In the second part, the results from the different statistical tests mentioned in the Material and Methods section are included.
- **Results from the lab work**: This includes results from Flow cytometry, Autofluorescence Growth curves, PAM, CLSM, and q-PCR for each of the mutant cell line and the wild type under LL, ML, and HL conditions.

7.1 Results from EDA of the MS-MS dataset

7.1.1 PCA results comparing Alb14 with Wild type

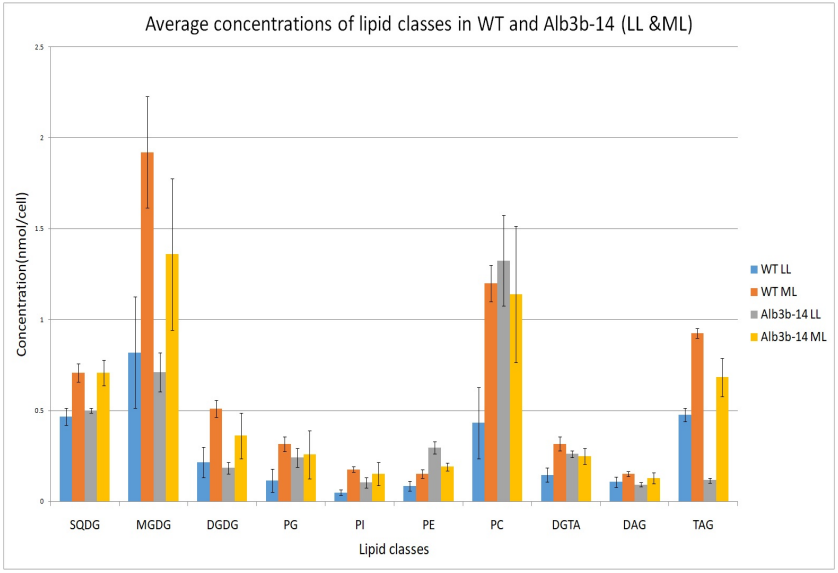
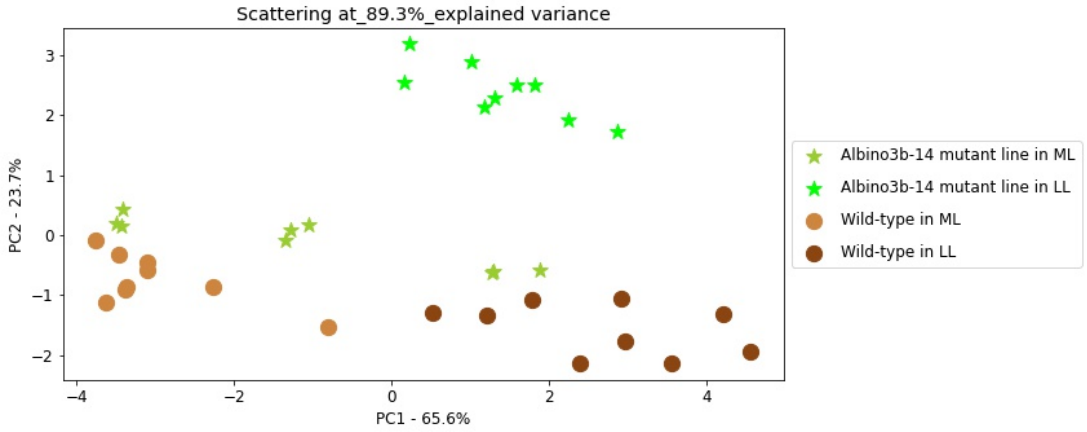


Figure 14: Bar plot showing average concentrations in of different lipid classes in *Alb3b*-14 cell line and Wild type, under LL and ML conditions, in nanomolar level per cell. *Alb3b*-14 behaves differently concerning change in some lipid classes, especially TAG, than the other two mutant lines. Values are mean from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

(a) 2D PCA scatter plot(*Alb3b*-14 & WT)



(b) 3D PCA scatterplot(Alb13b-4 & WT)

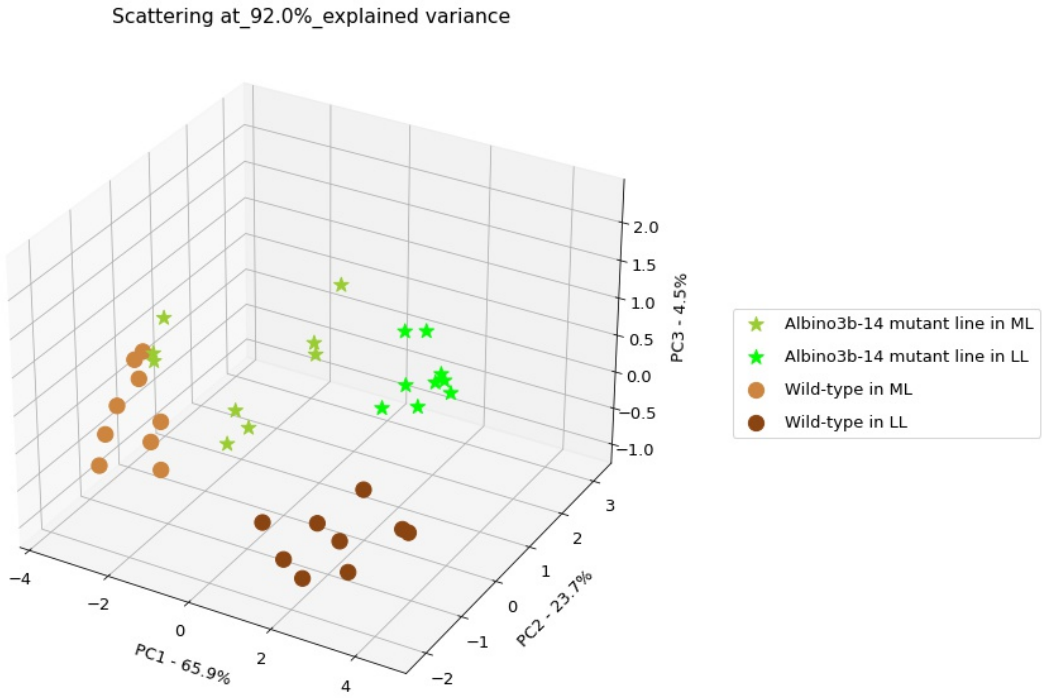


Figure 15: 2D(a) and 3D(b) PCA scatterplots for comparison between the WT and ALB3b-14 under LL and ML. Both plots are included to show how much difference can be observed in the differentiation of clusters formed by different samples when the explained variance is increased by 3% from 89 to 92% from 2 to 3 principle components. The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The scatter plots indicate a clear differentiation or clustering of three sample groups that are, Alb14 under LL, wild type under ML, and wild type under LL. The Alb14 cell line under ML can be observed as dispersed with a major fraction of its cluster spread into the wild type under ML. Although some of these samples can be seen as a part of the wild-type LL cluster in the 2D scatter plot, the 3D scatter plot with a much higher representation of variance does not show the same.

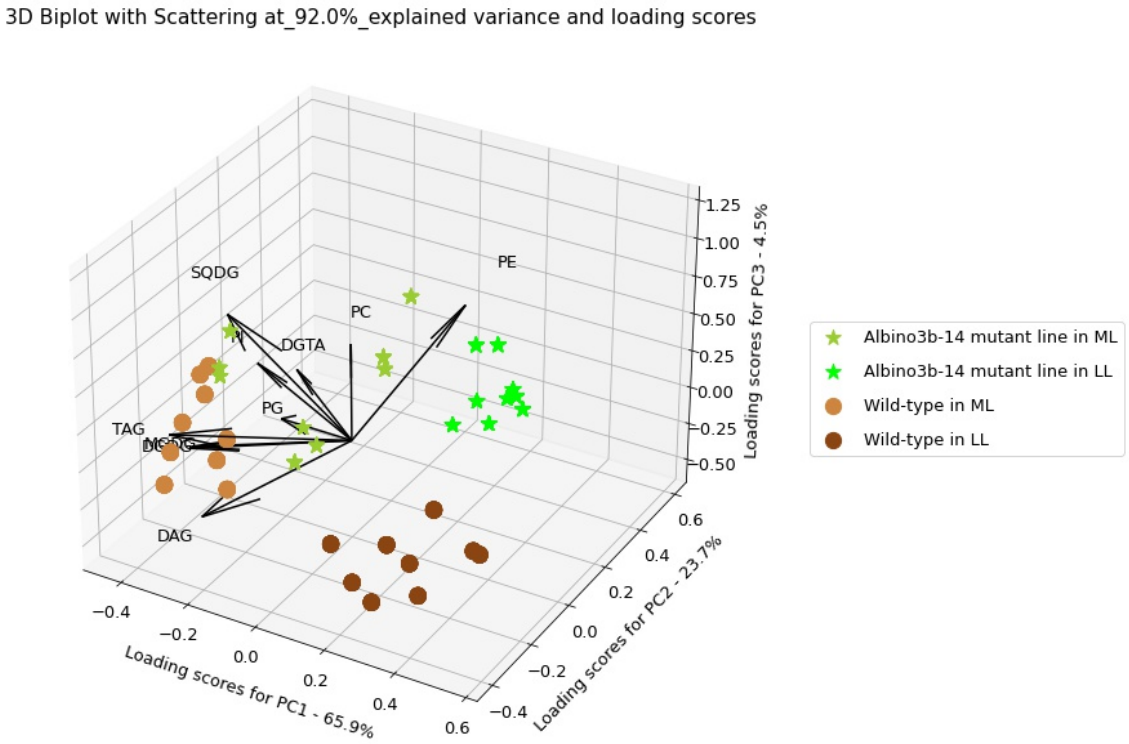


Figure 16: 3D biplot for comparing *Alb3b*-14 and WT samples in LL and ML. The explanation for interpreting the biplot is explained in section 5.16.3. The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The 3D biplot in figure 16 shows that the variance between the cluster of wild-type samples under LL and that of wild-type under ML is mainly across the first principle components and mainly caused by the phospholipid PG and PI, and the sulfolipid SQDG. Another notable observation is that the two cell lines, Alb14, and the Wild-type, are varied mainly across the second principle component and are mainly due to the neutral lipids TAG and DAG and the glycolipids MGDG and DGDG in one direction and the Phospholipids' PE and PC in the other direction. Another observation from the biplot is the possibility of a strong positive correlation between TAG, MGDG, and DGDG. Furthermore, the probable existence of a strong negative correlation between these three and two of the phospholipids, that is PC and PE can be observed. However, whether these correlations exist in both the cell lines separately or between them when the values of the lipid classes change under different light levels or if both of these are true, is hard to delineate from the biplot as it was made by considering all the Alb14 and wild-type samples together. Nevertheless, a cross-verification of these correlations with the bar plot showing the average concentrations(Figure 14) shows that the negative correlation between PE and TAG, the two lipids that contribute the most in differentiating Alb14 from wild-type, probably exist both separately in the two cell-types and between them. Additionally, this negative correlation appears to be much more pronounced in the Alb14, with a reduction PE by around 47% causing an approximately 6-fold increase in TAG. A similar effect can be also observed for PI, however, the high extent of error bars could lead to wrong inferences.

7.1.2 PCA results comparing Alb16 and Alb19 individually with Wild type

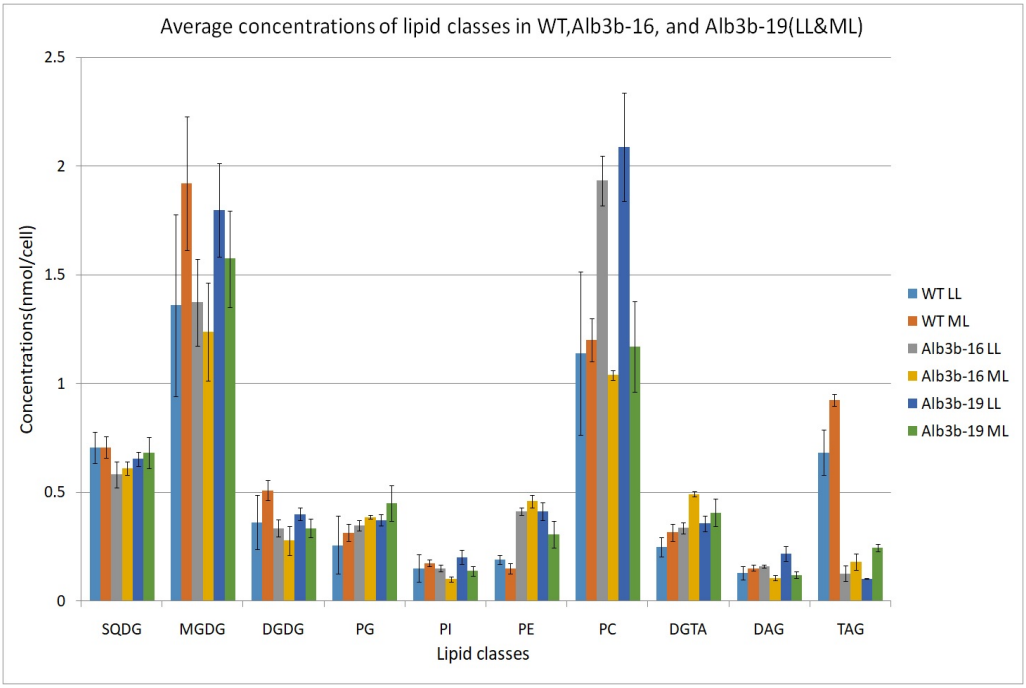
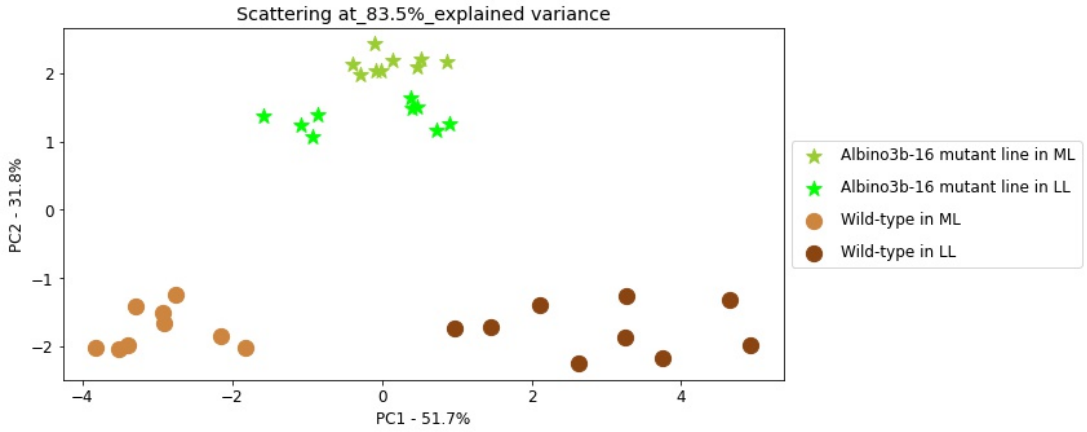


Figure 17: Average concentrations of different lipid classes in the wild-type, *Alb3b*-16 and *Alb3b*-19 cell lines in both LL and ML levels in nmol/cell. The values are mean from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The *Alb3b*-16 and *Alb3b*-19 mutants are compared to WT separately from the *Alb3b*-14 mutants as they for clusters in almost the same fashion in the PCA comparisons with the WT compared to that of the *Alb3b*-14. Additionally, the *Alb3b*-14 mutants also show similar levels of TAG under ML to WT, whereas the *Alb3b*-16 and 19 lines show comparatively very low levels of the same in ML compared to WT.

Unlike the *Alb3b*14 cell line, the*Alb3b*16 indicates 4 clusters with obvious distinction in the PCA scatterplots(Figure 18 encompassing both the cell lines, that is*Alb3b*16 and Wild-type, in two different light conditions. Although the clusters for Alb16 under LL and ML appear close to each other in the 2D scatter plot (Figure 18a), adding more variance by the third principle component (Figure 18b) shows that they are separated considerably along this component.

(a) 2D PCA scatter plot(Alb16 & WT)



(b) 3D PCA scatter plot(Alb16 & WT)

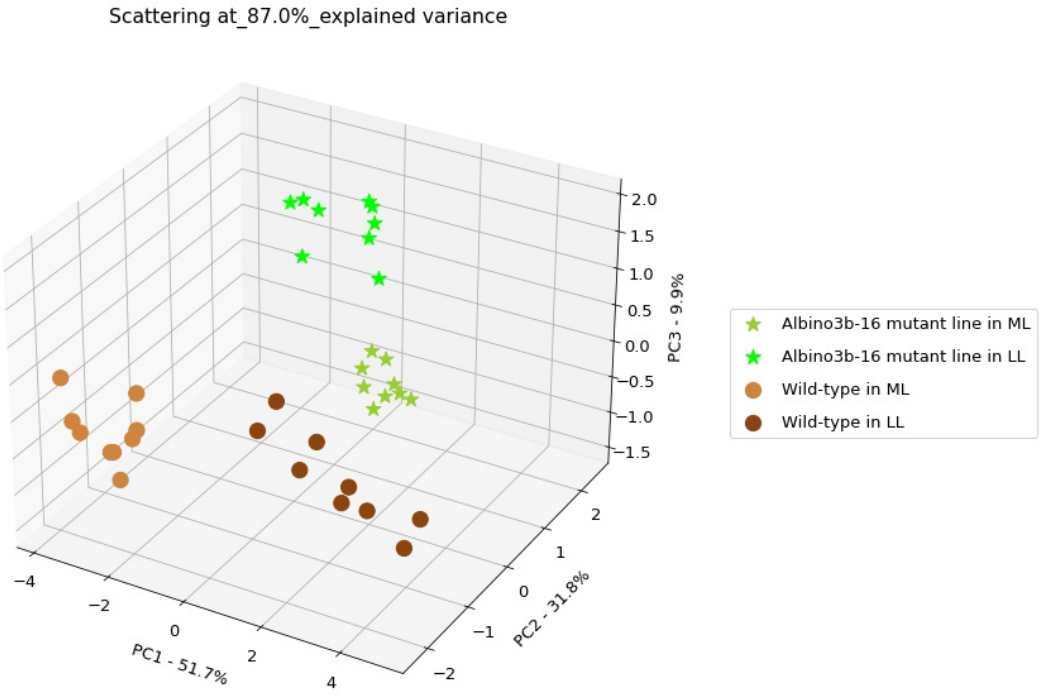


Figure 18: 2D(a) and 3D(b) PCA scatterplots for comparison between the WT and ALB3b-16 under LL and ML. a 3,5% increase in explained variance from 2 to 3 principle components can seen to be causing a notable difference in the clustering of the LL and ML samples of *Alb3b*-16 along the Z-axis represented by PC3.The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

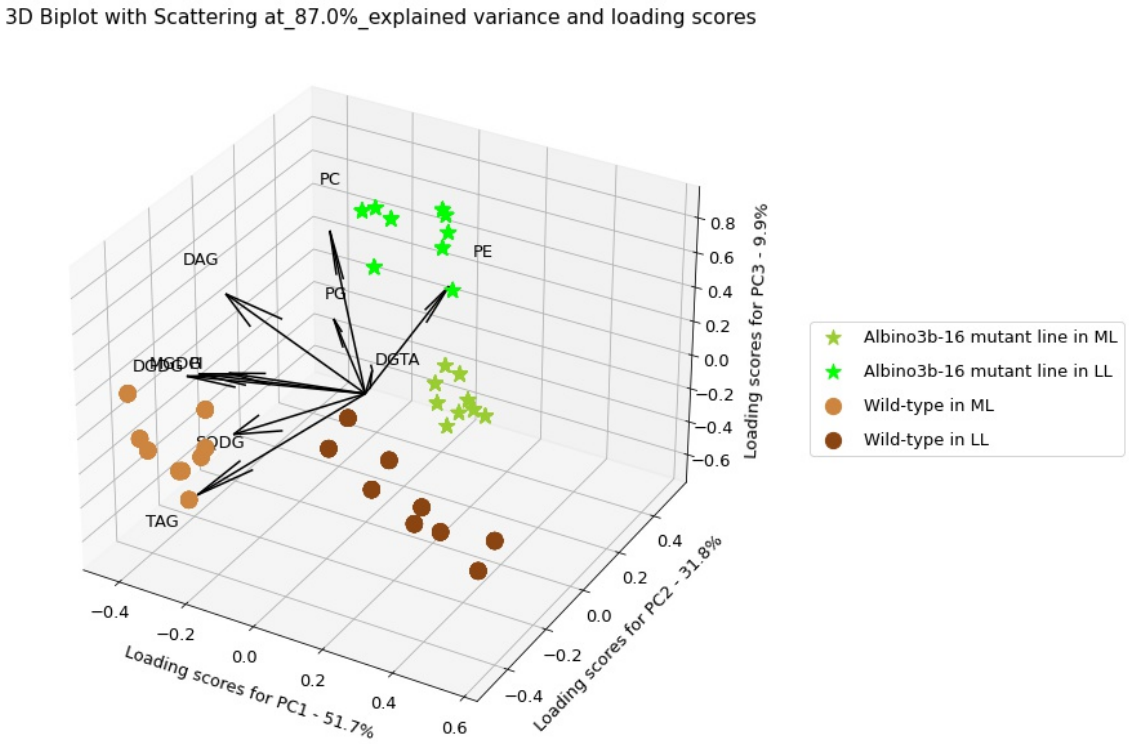
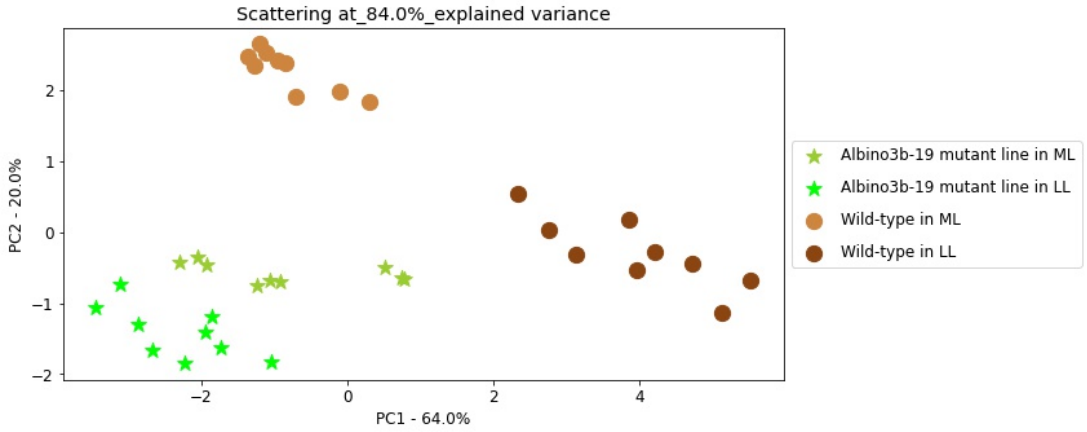


Figure 19: 3D biplot for comparing *Alb3b*-16 and WT samples in LL and ML. The explanation for interpreting the biplot is explained in section 5.16.3.The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The biplot for the comparative analysis of Alb16 with the wild type has similarities and differences to that of Alb14(Figure 16). The main similarity is that the cell lines are mainly differentiated here by the levels of TAG, MGDG, and DGDG in one direction and PE and PC in the other direction. Another similarity is the possibility of a negative correlation between the Phospholipids(PE, PC, and PG in this case) and TAG. However, unlike the Alba4, the Alb 16 cell line under ML forms a non-dispersed cluster separate from the wild-type under ML with a major contribution of variance from the glycolipids MGDG AND SQDG. Additionally, the Alb16 samples under ML vary from the same in LL along the third principle component, with major loading contributions from PE and PC. This is not observed clearly in the Alb14. Furthermore, the possibility of a strong positive correlation between the glycolipids and TAG is not as clearly visible as in the previous biplot(Figure 16)

(a) 2D scatter plot(Alb19)



(b) 3D scatterplot(Alb16)

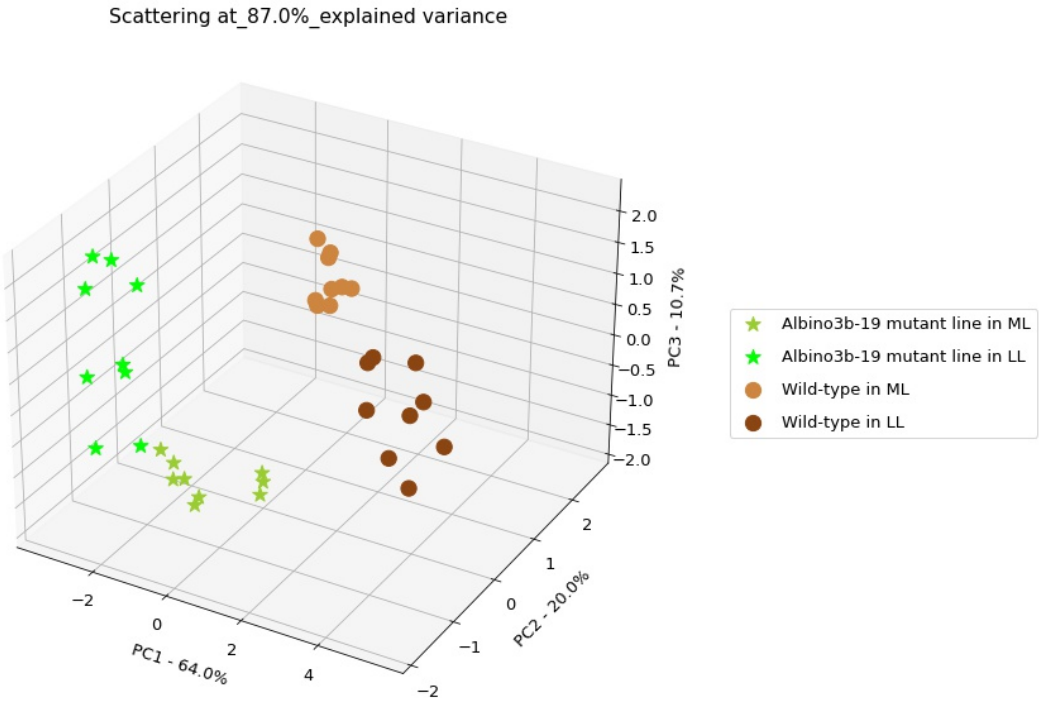


Figure 20: 2D(a) and 3D(b) PCA scatterplots for comparison between the WT and ALB3b-19 under LL and ML. A 3% increase in explained variance from 2 to 3 principle components causes a difference in the clustering of the LL and ML samples of *Alb3b*-19 along the Z-axis represented by PC3.The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The scattering observed for the Alb19 samples compared to wild-type in the principle components space is similar to that of Alb16 to a great extent, except for the dispersion of Alb19 samples under LL along the third principle component(Figure 20b).

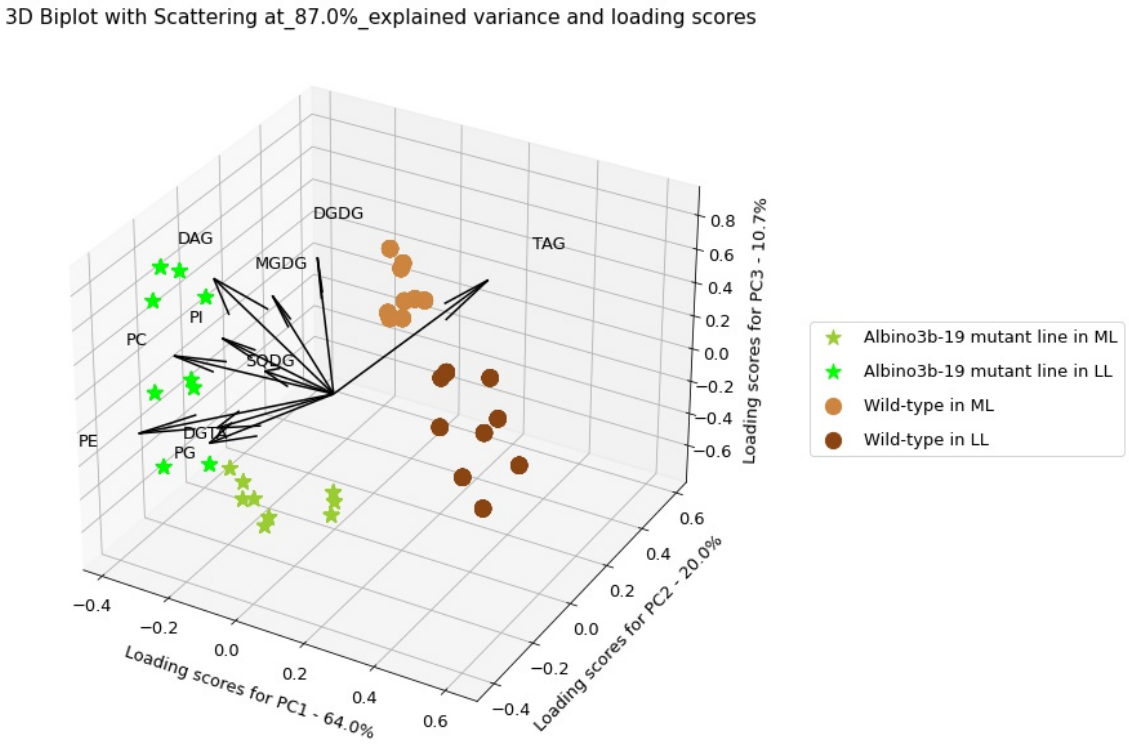


Figure 21: 3D biplot for comparing *Alb3b*-19 and WT samples in LL and ML.The explanation for interpreting the biplot is explained in section 5.16.3. The data points include measurements from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The Observations from the biplot comparing the Alb19 cell line with the wild-type(Figure 21) are also the same as that of Alb16(Figure 19) except that the glycolipids(MGDG and SQDG) do not appear to contribute much to the differentiation of the cell lines, but instead contribute to the variance of each of the cell lines in LL levels from the ML levels.

7.2 Results from statistical tests

Levene's test and Shapiro Wilk's test were performed on data from each lipid class for each cell line. The graphs representing the results from Shapiro Wilk's test show that some lipids are not normally distributed for all the cell lines, as their p-values are below the horizontal line at y equal to 0.05(Supplementary figure 14). To alleviate the impact of this violation of the normal distribution assumption on the T-test results a log transformation was applied to all the data before the T-test was performed. Furthermore, the graphs showing the results of the Levene's test indicate that the *Alb3b*-14 and *Alb3b*- 16 cell lines had some lipid classes in which the ML and LL samples deviated from the equal variance assumption of the T-test (Supplementary figure 13). Therefore, to improve the accuracy Welch's T-test was applied instead of the student's T-test for all the lipid classes for all the cell lines.

The results from Welch's T-test are represented in a scatter plot with the T-statistic on the x-axis and the P value on the y-axis. The vertical line represents x = 0, which is a T-statistic equal to 0. Therefore, all the points on the right to this line indicate an increase and all the points to the left indicate a decrease. Similarly, a horizontal dotted line is placed at y = 0.05. This is to show that all the points below the line indicate a significant change, and those above do not.

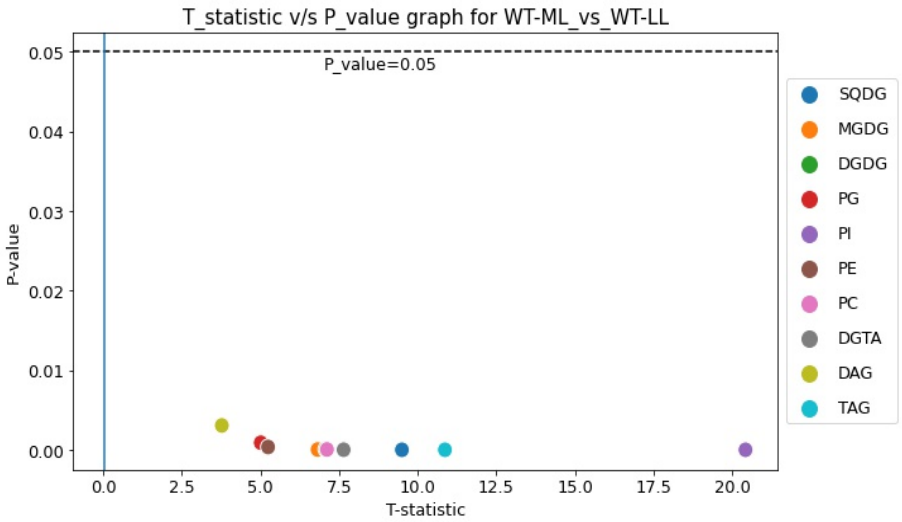


Figure 22: Results for T-test comparing Wild-type cell samples acclimated in ML and LL conditions. All the lipid classes are significantly higher in the ML samples than in LL for the WT. The samples compared included measurements from 3 technical replicates for every 3 biological replicates of WT in one of the studied light conditions.

Graph number one shows the results of the T-test comparing the amounts of lipid classes in the *Alb3b*-14 cell line in LL and the same cell line in ML. It can be observed that there is a significantly increased amount of the neutral lipid, TAG, the sulfolipids, SQDG, and the galactolipids, MGDG and DGDG in the *Alb3b*-14 cell line under ML as compared to the same cell line under LL, with the Neutral lipid showing the highest increase. Additionally, the amount of the phospholipid PE is significantly lower in the sample under ML compared to LL. On the other hand, the same graphical representation comparing the wild-type cell lines under ML and LL shows obvious differences. In the latter, it can be observed that in the ML condition, all the lipid classes show a significant increase compared to the cell line under LL conditions, with the highest increase being for the phospholipid PI.

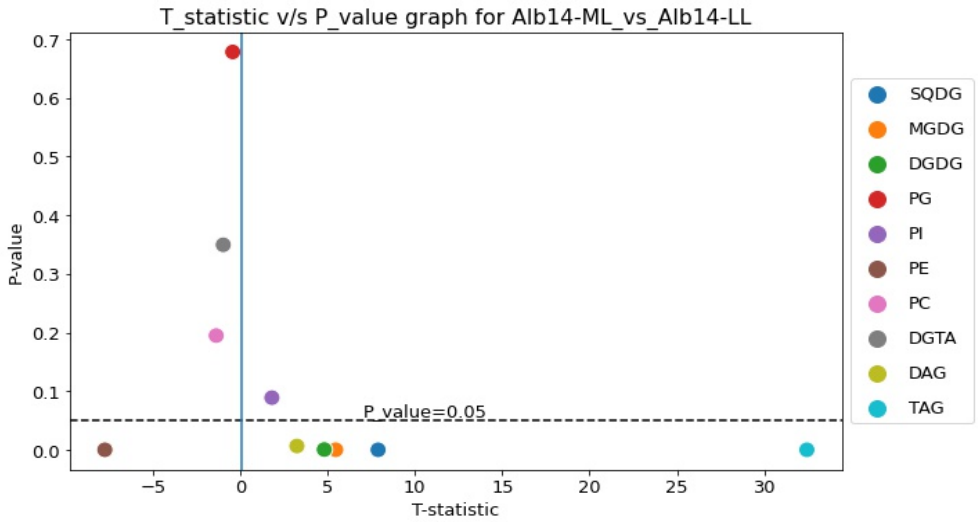


Figure 23: Results for T-test comparing *Alb3b*-14 cell samples acclimated to ML and LL treatments. The neutral lipid, TAG, and glycolipids, SQDG, MGDG, and DGDG are significantly higher in ML, while the phospholipid PE is significantly low. The samples compared included measurements from 3 technical replicates for every 3 biological replicates of *Alb3b*-14 in one of the studied light conditions.

The T-test results comparing the *Alb3b*- 16 and *Alb3b*- 19 cell lines in the two light conditions also show differences in the comparative amount of lipid classes in the two conditions as compared to that of the wild-type. Although the pattern of this difference shown by *Alb3b*-16 and *Alb3b*- 19 cell line from the wild type are almost the same they both differ from the pattern shown by *Alb3b*- 14. In both the *Alb3b*-16 and *Alb3b*- 19 cell lines only one of the glycolipids, that is DGDG, is significantly changing. But unlike in the *Alb3b*- 14, this particular glycolipid is decreasing in both the cell lines. Another similarity observed between *Alb3b*-16 and *Alb3b*-19 cell lines is the significantly increased amount of betaine lipid (DGTA) in the ML samples compared to the LL samples, which are not observed in *Alb3b*- 14 samples. Although these similarities exist between *Alb3b*- 16 and *Alb3b*- 19 they differ in terms of the comparative amounts of phospholipids in the different light conditions. Specifically, for *Alb3b*- 16 only one of the phospholipids, that is PC, is significantly decreasing, and the phospholipid PI is significantly increasing in contrast to the *Alb3b*- 14 cell line. However, for the *Alb3b*- 19 cell line, all the phospholipids are significantly decreasing with the greatest increase for PC.

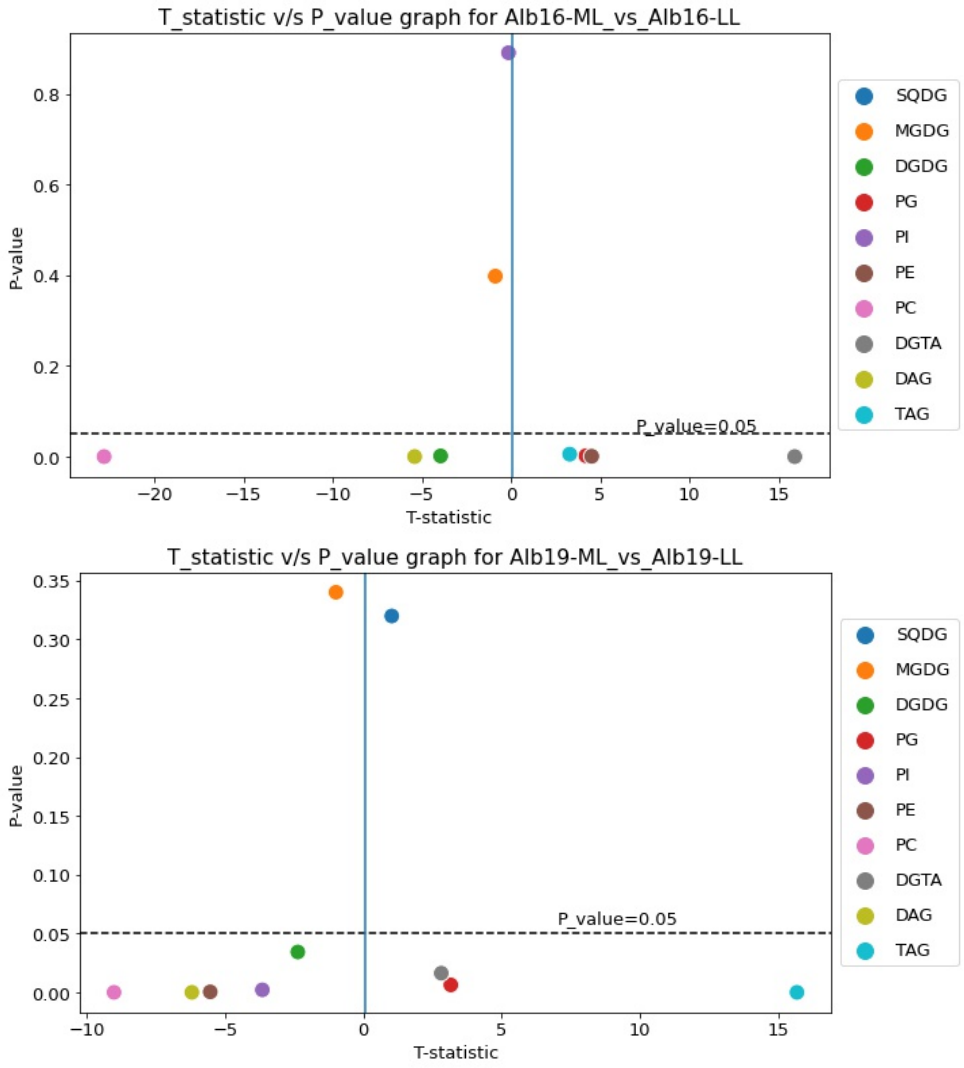


Figure 24: Results for T-test comparing *Alb3b*-16(top) and *Alb3b*-19(bottom) cell samples acclimated in ML and LL. The results are different from that of *Alb3b*-14(Figure 23). Although the TAG levels are significantly higher in ML, the glycolipids are not, with one of them(DGDG) being significantly low. The phospholipid, PE is significantly high in *Alb3b*-16, but low in *Alb3b*-19, while PC is significantly low in both. The samples compared included measurements from 3 technical replicates for every 3 biological replicates of one of the studied cell lines in one of the light conditions.

The MS-MS data indicating the fatty acid composition in each lipid class was also passed through the same data analysis pipeline. This particular dataset shows that the fatty acid composition of different lipid classes involves forms with either 2 or 3 fatty acids attached to a glycerol moiety. The chain length of these fatty acids ranged from 14 (C14) to 22(22C) carbon atoms with varying levels of double bonds or desaturation at different positions. For ease of explanation, the fatty acids with less than 20C length are mentioned as medium chain and those with more than 20C are mentioned as long chain fatty acids.

Similar to the data for lipid classes the fatty acid chain length composition data was also subjected to PCA and Welch's T-test. This was an extensive analysis as each lipid class had a different number of fatty acid combinations and needed to be split into separate data frames before passing through the pipeline. Therefore, each of these 10 different datasets, corresponding to 10 different lipid classes for each 4 cell

lines, thus accounted for a total of 10 different data frames. Consequently, the PCA for these 40 different data frames, comparing mutant lines with the wild type, generated a high number of PCA scatter plots, scree plots, and biplots. Although clear clustering of the wild-type samples and the analyzed *Alb3b*-3b mutant lines in both the light conditions were observed for each of the lipid classes, the high number of fatty acid combinations made the biplots complex and hard to understand. However, no discernible pattern, where specific fatty acid composition causes variation between clusters was not observed for any of the lipids.

Additionally, the comparative analysis of the results from the T-tests between the wild type and all the mutant lines did not indicate any pronounced difference patterns for any of the lipid classes except for the TAG. The T-test results for changes in concentrations of fatty acids in the TAG fraction in the mutants between ML and LL indicate that the amount of most of the TAG with medium-chain fatty acids and low levels of desaturation has significantly increased and most of the TAG with higher chain fatty acids and a higher level of desaturation is significantly decreased under the ML condition as compared to LL conditions. In contrast, an opposite scenario is observed in the wild type wherein under ML conditions most of the TAG with medium-chain fatty acids and lower levels of desaturation is significantly decreased and most of the TAG with long-chain fatty acids and higher levels of desaturation are significantly increased. It can also be observed that both the increment in most medium-chain fatty acids and the decrement in most of the long-chain fatty acids in *Alb3b*-14 are comparatively low as indicated by the low values of T-statistic. However, in the *Alb3b*-16 and *Alb3b*-19 mutants, the increment in some of the medium-chain fatty acids is considerably higher than in the *Alb3b*-14 cell line.

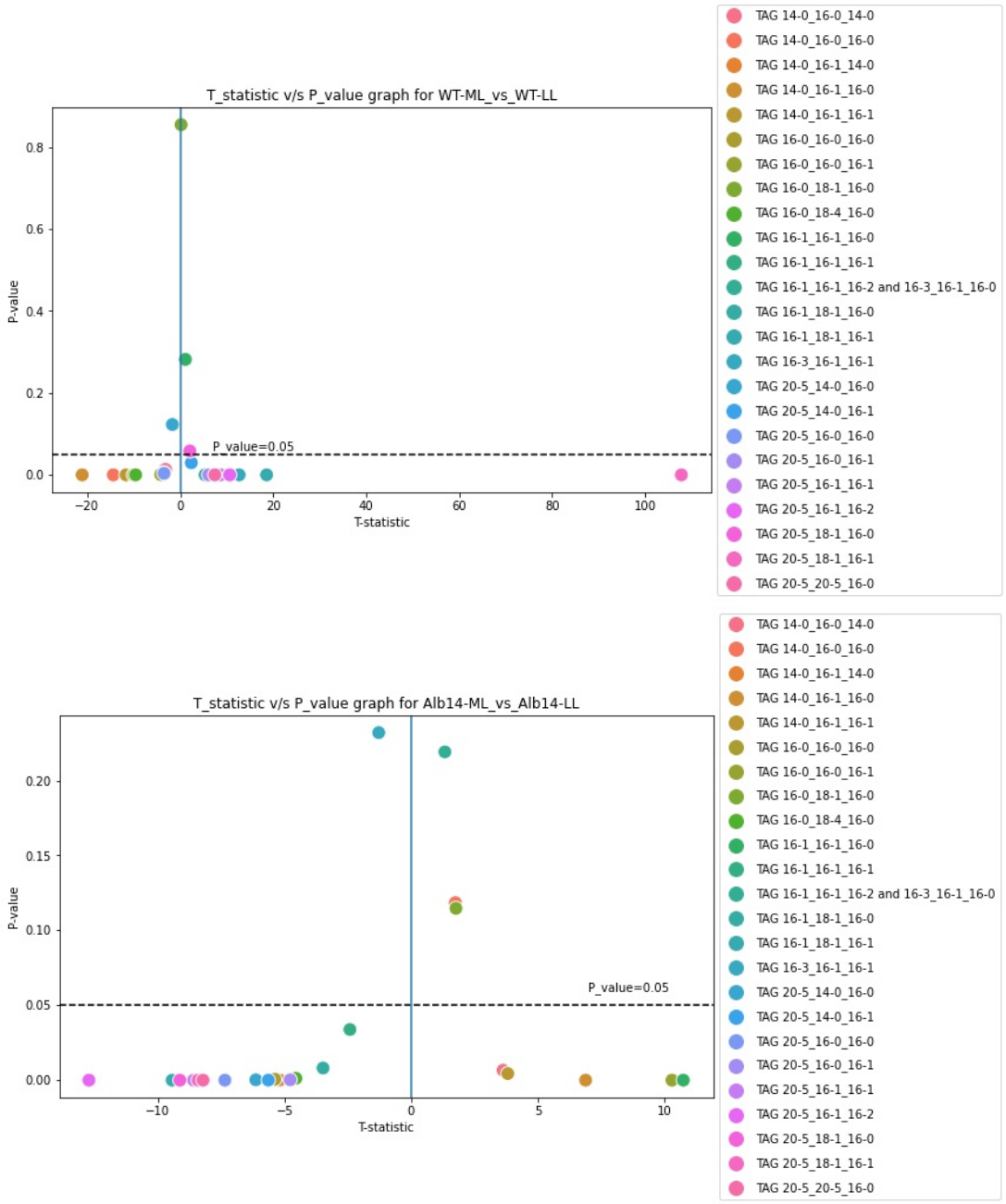


Figure 25: Results for T-test comparing concentrations of different fatty acid compositions of TAG in WT(top) and *Alb3b*-14(bottom) mutants acclimated to LL and ML conditions. The blue, violet, and pink shades represent compositions with at least one long-chain PUFA (EPA(20:5)), while the red, yellow, and green represent compositions with medium-chain, saturated, or monounsaturated FAs. An opposite trend can be observed between the WT and *Alb3b*-14, with WT having significantly higher PUFAs and lower saturated or monounsaturated FAs in general in ML compared to LL, while in the *Alb3b*-14 it is the opposite scenario. The samples compared included measurements from 3 technical replicates for every 3 biological replicates of one of the studied cell lines in one of the light conditions.

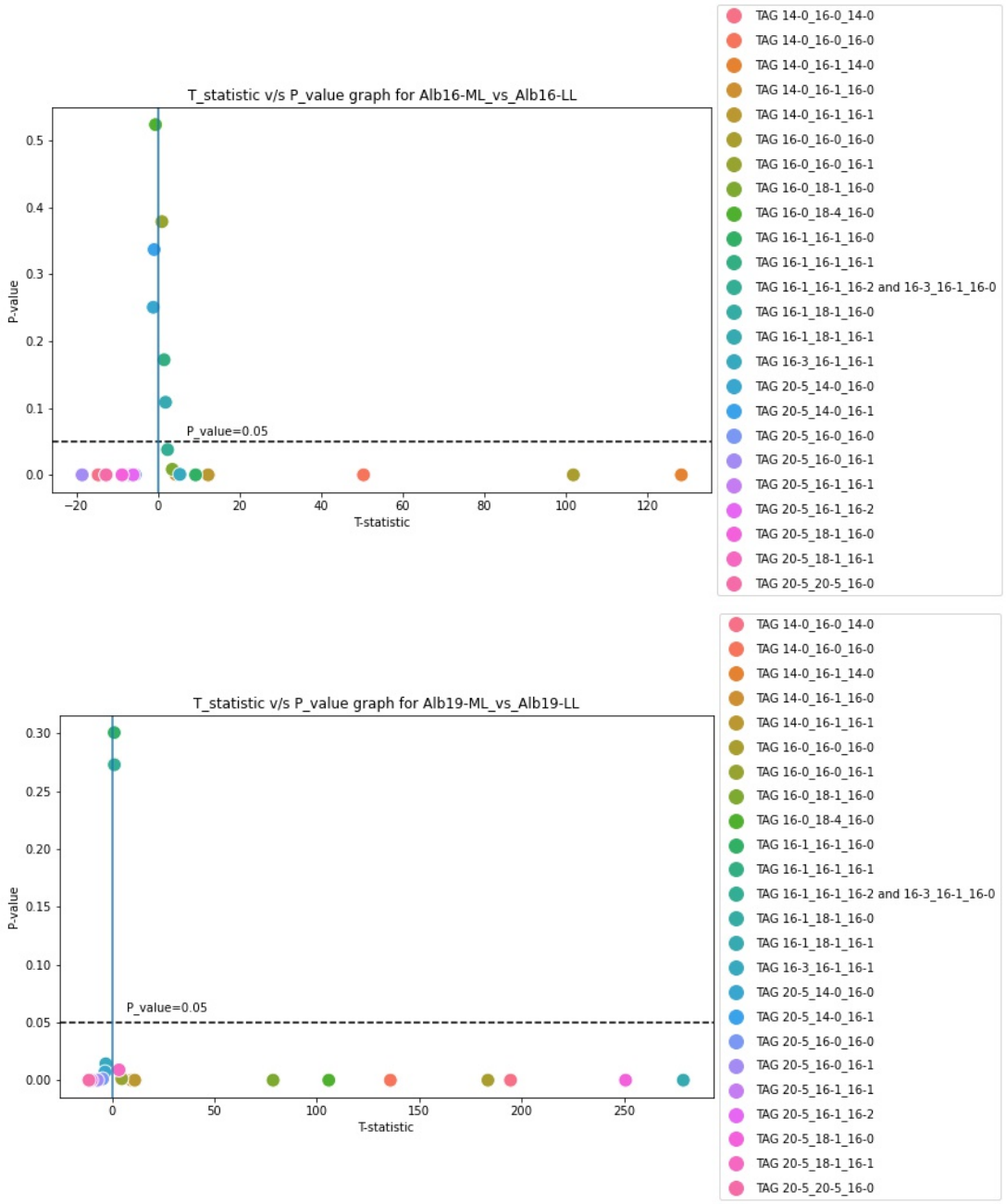


Figure 26: Results for T-test comparing concentrations of different fatty acid compositions of TAG in *Alb3b*-16(top) and 19(bottom) mutants acclimated to LL and ML conditions. The color coding of the compositions is the same as in Figure 25. Both the *Alb3b*-16 and 19 mutants show the opposite trend to that of the WT in terms of change in TAG composition between ML and LL. However, both of them differ from *Alb3b*-14 mutants, in terms of the range of increments in most of the FAs that are significantly increased, wherein this range is considerably broad in the *Alb3b*-16 and 19 compared to *Alb3b*-14 as observed from the T-statistic values in this figure and Figure 25. The samples compared included measurements from 3 technical replicates for every 3 biological replicates of one of the studied cell lines in one of the light conditions.

7.3 Results from light experiments

This results section involves the diverse observations or measurements from the experiments conducted on both the wild-type and mutant cell lines. As mentioned in the Material and Methods section, three experiments were conducted with three different light conditions. These are:

- Low-light(LL):35 µmol m^{-2} s^{-1}
- Medium-light(ML): 200 µmol m^{-2} s^{-1}
- High-light(HL): 700 µmol m^{-2} s^{-1}

The samples exposed to these three light conditions for two weeks were observed for various parameters using the various equipment mentioned. These include:

Table 3: Equipment and Parameters Measured

Technique	**Parameters Measured**
Flow cytometry	• Fluorescence intensity for BODIPY 505/515 staining • Fluorescence intensity for Chlorophyll-A • Forward scattering of incident light • Side scattering of incident light
Pulse amplitude modulation	• Non-photochemical quenching • Electron transport rate • Photosynthetic efficiency measured as Fv/Fm ratio • Light utilization efficiency
Plate reader	• Relative fluorescence units in time series
Confocal laser scanning Microscopy	• Structure of BODIPY 505/515 stained Lipid droplets • Autofluorescence
Quantitative PCR	• Relative expression levels of different genes involved in lipid metabolism

7.3.1 Results from autofluorescence growth curve measurements using plate reader

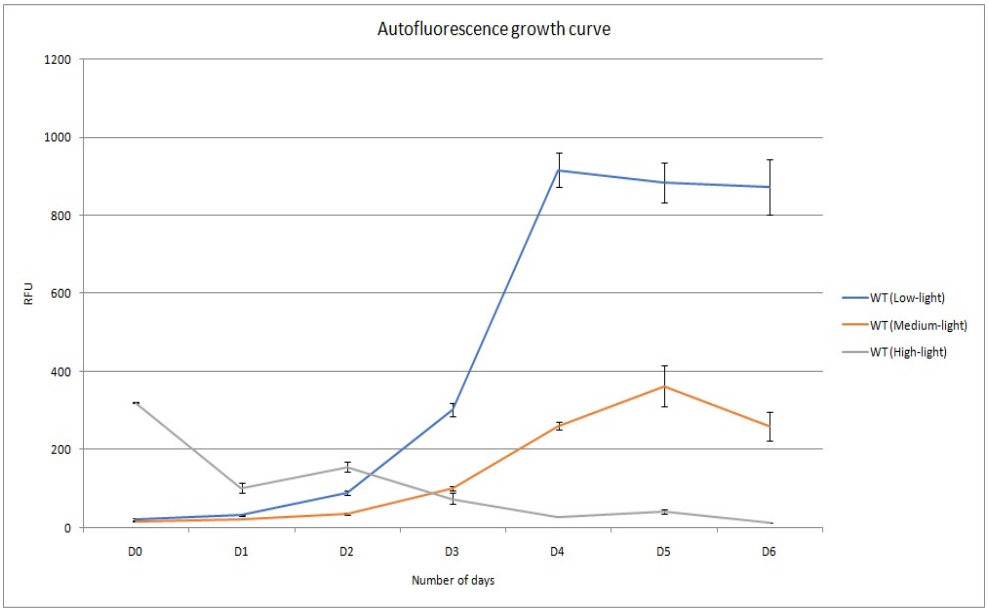
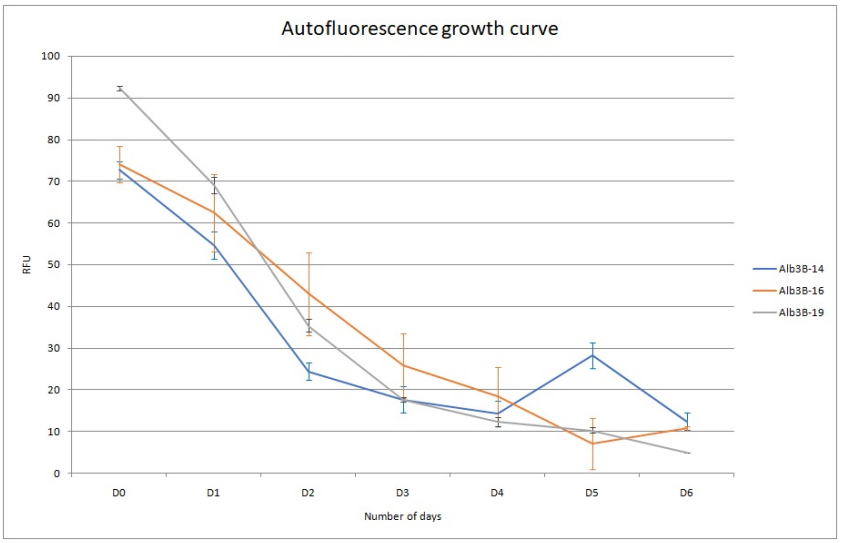


Figure 27: Autofluorescence growth curves for the WT cells obtained from plate reader by measuring the relative fluorescence units over 7 days. LL(blue) and ML(yellow) treatments were done with an initial cell count of 50,000 cells/ml. HL(grey) treatment was done with an initial cell count of 0.5 million cells/ml, The measurements are the mean of RFUs from 3 biological replicates of the WT in each light condition.

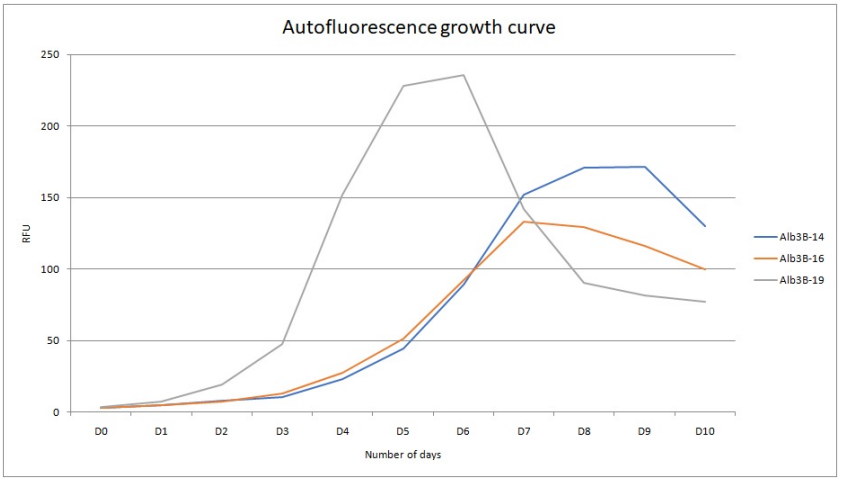
Growth curves using relative fluorescence units(RFU) from chlorophyll autofluorescence in the wild-type cells are depicted in figure 27. It can be observed that the LL and ML levels, where the initial cell count was set to around 50,000 cells/ml, resulted in normal growth curves with the lag, exponential, and stationary phases, followed by the beginning of the declining phase in 6 days in LL and ML conditions. In the HL levels, where the initial cell count was set to 0.5 million cells/ml, only a declination in the RFU levels was observed to approach near zero level by the 6th day.

Results for the Autofluorescence growth curves for all the *Alb3b* mutant lines are depicted in figure 28. It is observed that the same pattern of growth curves occurs for all the mutant lines as the wild-type. They form normal growth curves with lag, exponential, and stationary phases in the LL and ML levels, whereas they just decline in HL conditions. The *Alb3b* mutants took approximately 10 days to reach the stationary phase under LL compared to just 6 days of the WT. in the ML,n. In HL, all the mutants undergo declination, The RFU levels in the mutants can be seen as generally lower than the WT in all the light conditions.

(a) *Alb3b* mutants in HL



(b) *Alb3b* mutants in ML



(c) *Alb3b* mutants in LL

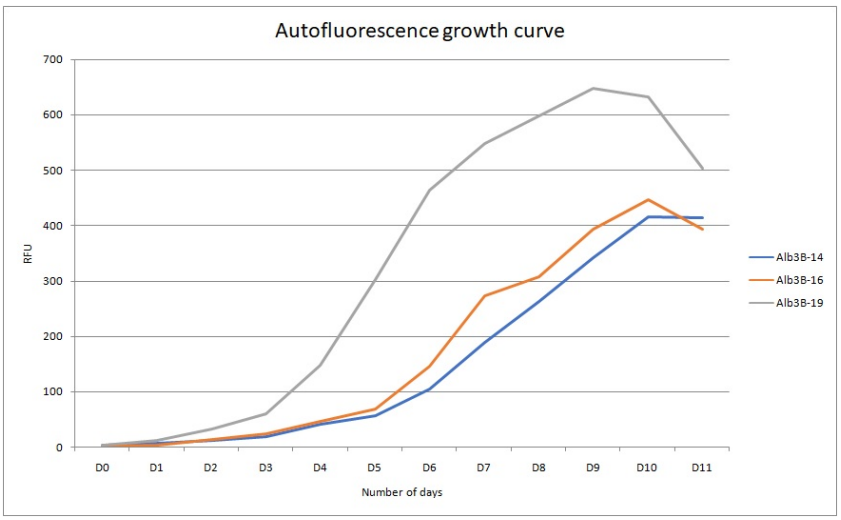


Figure 28: Growth curves made for *Alb3b* mutants under different light conditions by measuring the relative fluorescence units over 7 days for HL and 12 days for LL and ML treatments. The initial cell counts in each light treatment are the same as that followed for the WT(Figure 27). The measurements are the mean of RFUs from 3 biological replicates for each cell line in each light condition.

7.3.2 Results from flow cytometry using BODIPY 505/515 staining

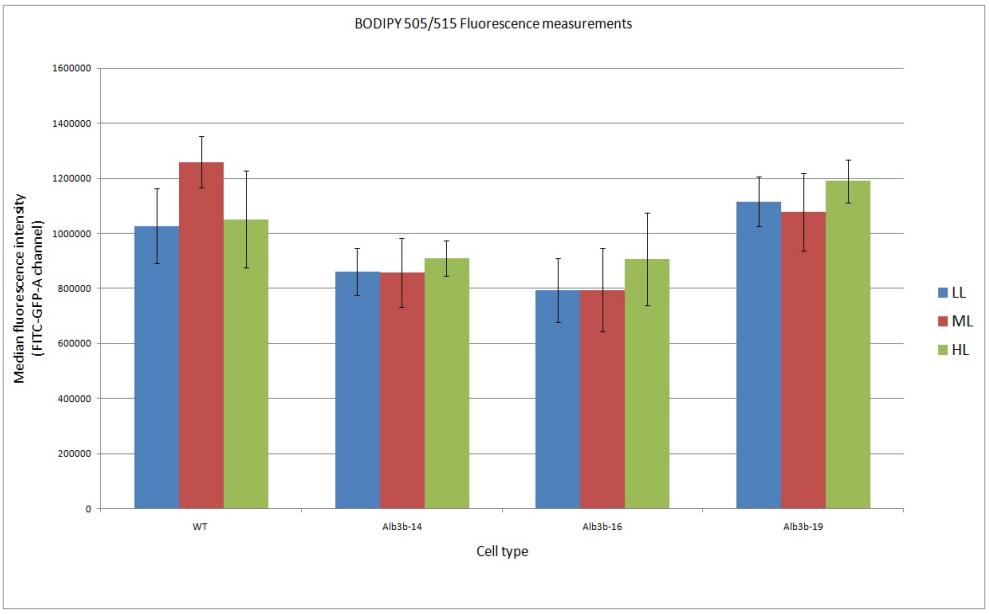


Figure 29: Median fluorescence intensity values from BODIPY 505/515 as observed in the FITC-GFP-A channel in flow cytometry. The values are the mean of median fluorescence intensities from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The quantity of lipid droplets as indicated by BODIPY 505/5015 fluorescence measurements in the FITC-GFP-A channel in the flow cytometer(29) does not indicate a clear pattern of increasing neutral lipid accumulation with increasing stress by irradiance as expected. In the wild-type cells, the fluorescence increases by 22% in ML conditions whereas it decreases by about 16% in HL conditions. However, the measurements in HL conditions have a high standard deviation indicating high variability between the replicates.In the mutant cell Lines, the fluorescence measurements appear to be more or less the same in LL and ML conditions. In HL conditions fluorescence measurements increase by 6, 14, and 10% for the *Alb3b*-14,16, and 19 mutants respectively.

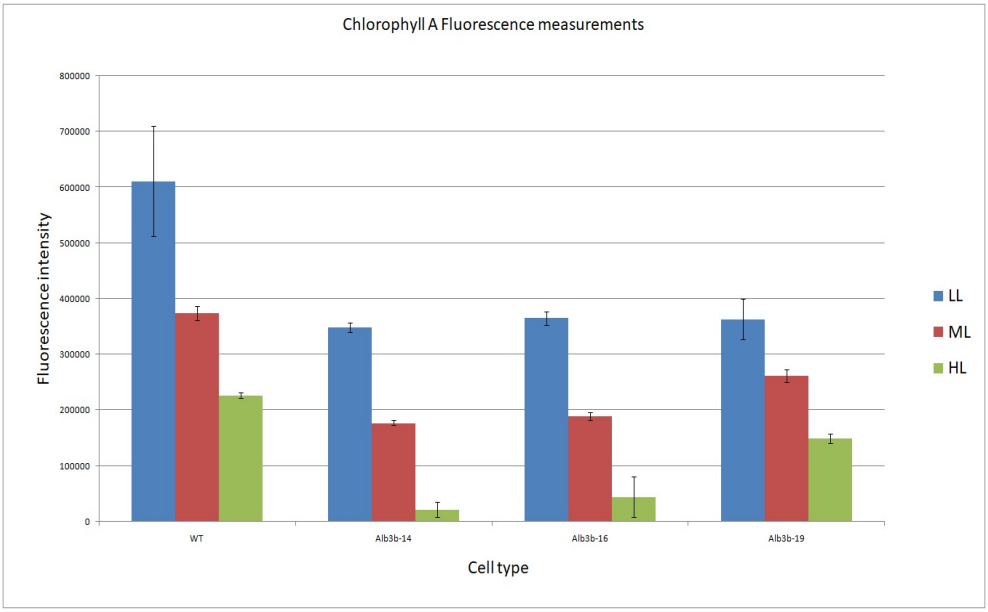


Figure 30: Median fluorescence intensity values from Chlorophyll as observed in the respective channel(Chlorophyll-A) in flow cytometry. The chlorophyll levels can be seen decreasing with increasing light intensity in all the cell lines and the values are lower in the mutants compared to WT in all conditions. The values are the mean of median fluorescence intensities from 3 technical replicates for every 3 biological replicates of the studied cell lines in each light condition.

The chlorophyll autofluorescence measurements from the flow cytometer indicate the expected results, wherein the values for mutants are significantly lower compared to those of the Wild type in all three conditions. Also, the chlorophyll levels are reduced for all the cell lines with increasing light levels. Under LL, all the mutant lines show a 40% decreased chlorophyll A autofluorescence compared to the wild type. Under, HL, this decrement becomes around 50% for the *Alb3b*14 and 16 lines and around 30% for the Alb3b 19 line. In the HL treatment, the mutant lines except Alb3b19 show extreme reductions in chlorophyll autofluorescence values, which are around 90% less for Alb3b 14 and 80% less for the Alb16 mutants compared to that of wild type in similar conditions. Whereas Alb 19 lines show just a 34% reduction in chlorophyll levels compared to the wild type in HL levels.

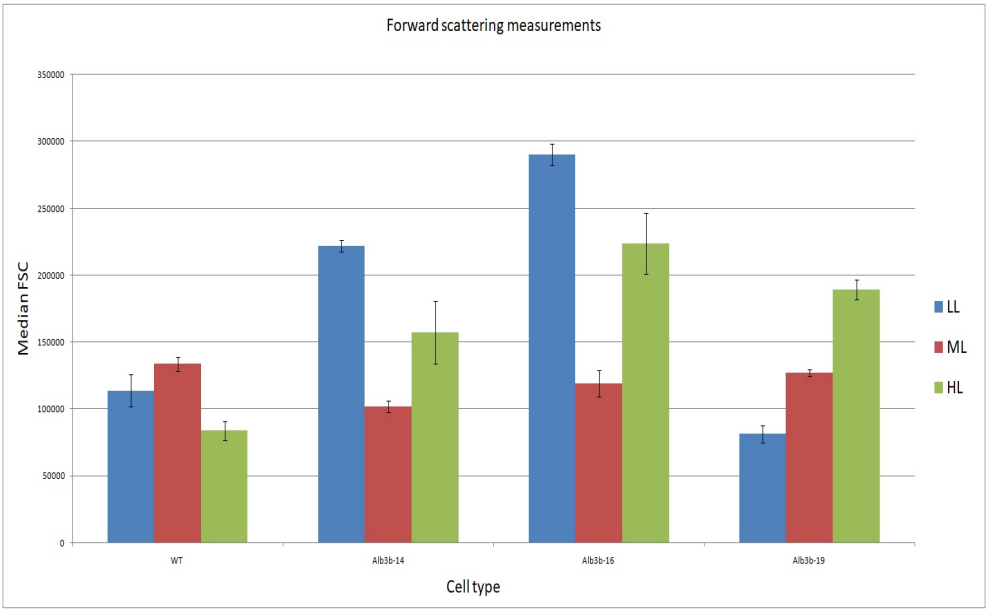


Figure 31: Median forward scattering values from flow cytometry for all the cell lines(WT, Alb3b-14,16, 19) in three different light conditions(HL, ML, and LL) measured as the average of medians of three technical replicates for each of the three biological replicates of all cell lines in each of the light treatments.

The results from forward and side scattering measurements in flow cytometry(figure 31 and figure 32) indicate that the mutant line, *Alb3b* 14 and 16, changes almost similarly with different treatments. The forward scattering for these mutant lines is the highest under LL and lowest in HL. Whereas in HL they are shown to have an intermediate forward scatter. This contrasts with what is observed in the wild type, where the highest forward scatter occurs in HL and the lowest in HL. The *Alb3b*-19 mutant line, shows a completely different change pattern, with the forward scattering increasing with increasing light intensity. Another notable observation is that the forward scattering is generally higher in the mutant lines compared to the wild type.

Regarding the side scattering measurements, although the values differ among all the cell lines, a general pattern can be observed. That is, the side scatter values decrease from low to HL levels and then increase from medium to HL levels. This pattern is seen as more pronounced and similar in the *Alb3b* 14 and 16 mutant lines. Also, the side scatter values are observed to be generally higher in the mutants compared to the wild type.

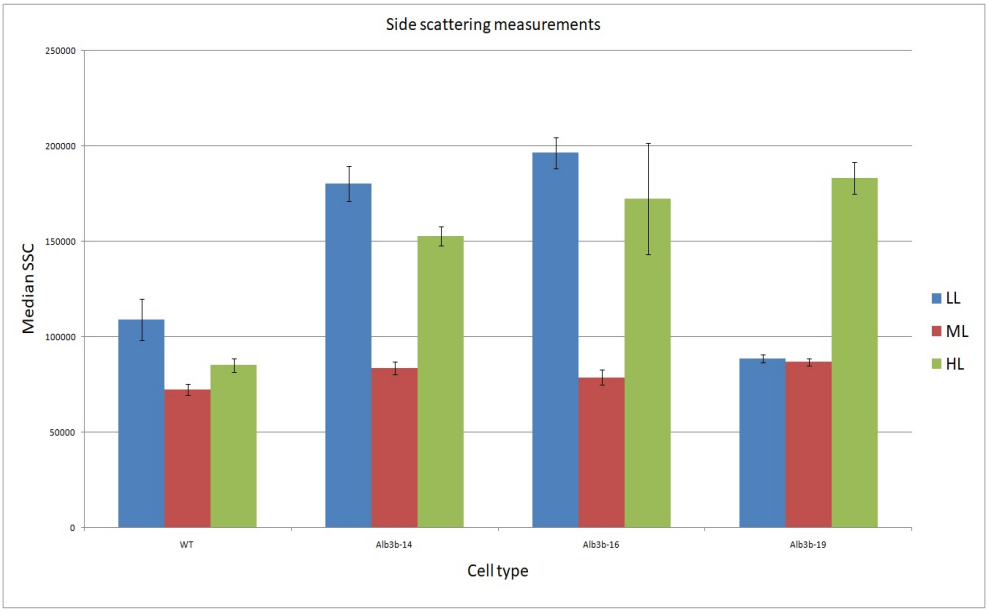


Figure 32: Median side scattering values from flow cytometry for all the cell lines(WT, *Alb3b*-14,16, and 19) in three different light conditions(HL, ML, and LL) measured as the average of medians of three technical replicates for each of the three biological replicates of all cell lines in each of the light treatments.

The results from two-way ANOVA for the measurements from flow cytometry are presented in figure 33. It can be seen that both the cell line and light condition have significant effects on all the measured parameters from flow cytometry with p-values less than 0.05. The interaction effect of the cell line and light condition can be seen as significant for all the parameters except the FITC-GFP-H. The validation of the assumptions for ANOVA was done using the 'residuals v/s fitted values' plot, q-q plot, Shapiro Wilk's test, and Levene's test. It was observed that, except for the FITC-GFP parameters, all the other parameters deviate from the assumption of the ANOVA model's homogeneously varying and normally distributed residuals. This is presented in the figure 12

	P-value(cell_type)	P-value(treatment)	P-value(interaction)
FITC_GFP_H	4.3257e-14	3.00995e-06	0.0621712
FITC_GFP_A	6.32242e-21	0.000307916	0.0279978
FSC_A	1.90516e-50	5.33477e-40	4.86774e-55
SSC_A	5.77121e-41	1.65765e-50	2.12674e-42
Chlorophyll_A	4.34201e-47	4.17084e-60	3.04911e-08

Figure 33: Table from ANOVA results from Jupyter Notebook indicating P-values for cell line, light condition, and the interaction effect of both for each of the measured variables from flow cytometry. Significant influences by both cell line and light treatment and interaction effect can be seen on all the parameters except for the interaction effect of the variables on the FITC-GFP-H parameter.

Since only the FITC-GFP parameters followed the underlying assumptions in ANOVA and only the FITC-GFP-A had a significant interaction effect for cell line and light condition, the results from residuals v/s fitted values plotting and q-q plotting, and Tukey's-HSD analysis is presented just for the

FITC-GFP-A. The results for the other parameters are added to the appendix.

The variance homogeneity and normality of the residuals, as proved to be significant in Levene's and Shapiro Wilk's test respectively, can be seen in figure 25. It can be noticed that the residual values across all points are almost homogeneous around the fitted values line. Similarly, the q-q plot, with the quantile values of the samples plotted against theoretical quantiles for normally distributed data, depicts how almost perfectly the residual points align with the red line representing normal distribution.

group1	group2	p-value
('WT', 'LL')	('WT', 'HL')	0.002750822
('WT', 'LL')	('WT', 'ML')	0.007050826
('WT', 'LL')	('Alb16', 'LL')	0.006456198
('WT', 'HL')	('Alb14', 'HL')	0.001
('WT', 'HL')	('Alb16', 'HL')	0.001
('WT', 'ML')	('Alb14', 'ML')	0.001
('WT', 'ML')	('Alb16', 'ML')	0.001
('Alb14', 'LL')	('Alb19', 'LL')	0.001773061
('Alb14', 'HL')	('Alb19', 'HL')	0.001
('Alb14', 'ML')	('Alb19', 'ML')	0.014640567
('Alb16', 'LL')	('Alb19', 'LL')	0.001
('Alb16', 'HL')	('Alb19', 'HL')	0.001
('Alb16', 'ML')	('Alb19', 'ML')	0.001

Figure 34: Post hoc (Tukey's HSD) test results for FITC-GFP measurements presenting the comparisons with significant changes(p-value<0.05) between individual samples.The color coding is as follows; Yellow: compares the wild type in different treatments, Blue: compares the wild type with the mutant lines in the same condition, Green: Compares different mutant lines under same conditions, Orange: compares the same mutant line under different conditions.

The subset of Tukey's HSD analysis results for FITC-GFP-A, in which only the comparisons with p value<0.05 are presented in figure 34. The comparative observations from these results are as follows:

- The wild type in LL differs significantly from the same cell line in medium and HL
- The wild type in HL and HL significantly differs from the *Alb3b* 14 and 16 mutant lines in the same conditions. Whereas the wild-type cell in LL differs significantly just with the *Alb3b* 16 mutant lines.
- The *Alb3b* 14 and 16 mutant lines under all three light treatments significantly differ from the *Alb3b* 19 mutant lines in the same conditions.
- There is no significant difference between the same mutant lines under any of the different conditions.

group1	group2	p-value
('WT', 'LL')	('WT', 'HL')	0.001
('WT', 'LL')	('WT', 'ML')	0.001
('WT', 'HL')	('WT', 'ML')	0.001
('WT', 'LL')	('Alb14', 'LL')	0.001
('WT', 'LL')	('Alb16', 'LL')	0.001
('WT', 'LL')	('Alb19', 'LL')	0.001
('WT', 'HL')	('Alb14', 'HL')	0.001
('WT', 'HL')	('Alb16', 'HL')	0.001
('WT', 'HL')	('Alb19', 'HL')	0.001
('WT', 'ML')	('Alb14', 'ML')	0.001
('WT', 'ML')	('Alb16', 'ML')	0.001
('WT', 'ML')	('Alb19', 'ML')	0.001
('Alb14', 'LL')	('Alb14', 'HL')	0.001
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'HL')	('Alb14', 'ML')	0.001
('Alb14', 'HL')	('Alb19', 'HL')	0.001
('Alb14', 'ML')	('Alb19', 'ML')	0.001
('Alb16', 'LL')	('Alb16', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001
('Alb16', 'HL')	('Alb16', 'ML')	0.001
('Alb16', 'HL')	('Alb19', 'HL')	0.001
('Alb16', 'ML')	('Alb19', 'ML')	0.001
('Alb19', 'LL')	('Alb19', 'ML')	0.001
('Alb19', 'HL')	('Alb19', 'ML')	0.001

Figure 35: Post hoc (Tukey's HSD) test results for chlorophyll measurements from flow cytometry, presenting the comparisons with significant changes(p-value<0.05) between individual samples. The color coding is the same as explained in Figure 34.

The post hoc analysis of the chlorophyll measurements from flow cytometry(Figure 35) shows that the chlorophyll autofluorescence significantly changes in all the cell lines from LL to ML and then from ML to HL treatments. Additionally, the chlorophyll levels are significantly different between WT and all the mutant lines in similar light conditions, It can also be observed that the *Alb3b*-19 mutant line has significantly different chlorophyll levels from the *Alb3b*-14 and 16 mutant lines under ML and HL treatments.

group1	group2	p-value
('WT', 'LL')	('WT', 'ML')	0.017605231
('WT', 'LL')	('Alb14', 'LL')	0.001
('WT', 'LL')	('Alb16', 'LL')	0.001
('WT', 'LL')	('Alb19', 'LL')	0.001
('WT', 'HL')	('Alb14', 'HL')	0.001
('WT', 'HL')	('Alb16', 'HL')	0.001
('WT', 'HL')	('Alb19', 'HL')	0.001
('WT', 'ML')	('Alb14', 'ML')	0.001
('Alb14', 'LL')	('Alb14', 'HL')	0.001
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'LL')	('Alb16', 'LL')	0.001
('Alb14', 'LL')	('Alb19', 'LL')	0.001
('Alb14', 'HL')	('Alb14', 'ML')	0.001
('Alb14', 'HL')	('Alb16', 'HL')	0.001
('Alb14', 'HL')	('Alb19', 'HL')	0.001
('Alb14', 'ML')	('Alb19', 'ML')	0.001
('Alb16', 'LL')	('Alb16', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001
('Alb16', 'LL')	('Alb19', 'LL')	0.001
('Alb16', 'HL')	('Alb16', 'ML')	0.001
('Alb16', 'HL')	('Alb19', 'HL')	0.001
('Alb19', 'LL')	('Alb19', 'HL')	0.001
('Alb19', 'LL')	('Alb19', 'ML')	0.001
('Alb19', 'HL')	('Alb19', 'ML')	0.001

(a)

group1	group2	p-value
('WT', 'LL')	('WT', 'HL')	0.001
('WT', 'LL')	('WT', 'ML')	0.001
('WT', 'LL')	('Alb14', 'LL')	0.001
('WT', 'LL')	('Alb16', 'LL')	0.001
('WT', 'LL')	('Alb19', 'LL')	0.002927751
('WT', 'HL')	('Alb14', 'HL')	0.001
('WT', 'HL')	('Alb16', 'HL')	0.001
('WT', 'HL')	('Alb19', 'HL')	0.001
('Alb14', 'LL')	('Alb14', 'HL')	0.001
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'LL')	('Alb16', 'LL')	0.048718769
('Alb14', 'LL')	('Alb19', 'LL')	0.001
('Alb14', 'HL')	('Alb16', 'HL')	0.005301745
('Alb14', 'HL')	('Alb19', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001
('Alb16', 'HL')	('Alb16', 'ML')	0.001
('Alb19', 'LL')	('Alb19', 'HL')	0.001
('Alb19', 'HL')	('Alb19', 'ML')	0.001

(b)

Figure 36: Post hoc (Tukey's HSD) test results for forward scatter(a) and side scatter(b) measurements from flow cytometry, presenting the comparisons with significant changes(p-value < 0.05) between individual samples. The color coding is the same as in Figure 34.

The post hoc analyses of FSC and SSC measurements from flow cytometry indicate several significant differences within cell lines under different light levels and also between the cell lines under similar conditions as seen in figure 36. The general observations from these analyses are as follows:

- The FSC and SSC values for WT in LL and HL significantly differ from those of all the mutant lines.
- The FSC and SSC values for all the mutant lines under LL significantly differ from those under ML and HL.
- There are significant differences in the FSC values between all the different mutant lines in LL and HL treatments.
- The FSC values for WT significantly change between LL and ML, but not between LL and HL or ML and HL. Additionally, the SSC values in WT significantly change between LL and both ML and HL treatments.

7.3.3 Results from pulse amplitude modulation

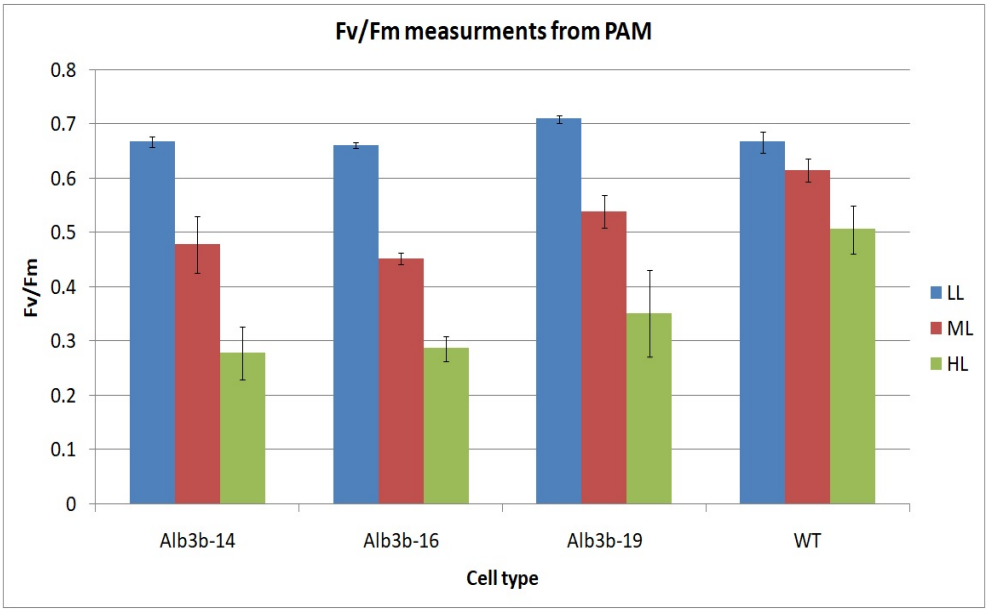


Figure 37: F_v/F_m measurements from PAM for the different cell lines(WT, *Alb3b*-14,16, 19) in three different light treatments (HL, ML, and LL) measured as mean of values from 3 biological replicates for each cell line in each light condition. The values F_v/F_m can be seen decreasing with increasing light intensity in all the cell lines, but at different extents between mutants and WT.

The F_v/F_m ratios of the different cell lines under different light treatments follow the same expected pattern. That is, the values are the highest in LL levels and decrease with increasing light intensity(Figure 37. However, the extent of this decrease varies between the wild type and the mutants. The decrease in the F_v/F_m values from LL to HL ranges between 24% to 32% in the mutants, whereas, for the wild type it is just approximately 7%. Similarly, the decrease in the F_v/F_m values from HL to HL ranges from 35% to 42% in the mutants, while that of the wild type is around 18%.

The results from posthoc analysis for the F_v/F_m values with only the significant changes are presented in figure 38. It can be observed that all the mutant lines significantly change in the F_v/F_m values from the same cell line under all different light treatments. In the wild type, the significant change is only between low and HL levels. Also, comparing the values between the wild type and the mutants shows that they have similar values under LL. The *Alb3b*14 and 16 line significantly differs from the wild type under the other two light conditions and Alb13b 19 differs from the wild type just under HL levels.

group1	group2	p-value
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'LL')	('Alb14', 'HL')	0.001
('Alb14', 'ML')	('Alb14', 'HL')	0.001
('Alb14', 'ML')	('WT', 'ML')	0.005301
('Alb14', 'HL')	('WT', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001
('Alb16', 'LL')	('Alb16', 'HL')	0.001
('Alb16', 'ML')	('Alb16', 'HL')	0.001
('Alb16', 'ML')	('WT', 'ML')	0.001
('Alb16', 'HL')	('WT', 'HL')	0.001
('Alb19', 'LL')	('Alb19', 'ML')	0.001
('Alb19', 'LL')	('Alb19', 'HL')	0.001
('Alb19', 'ML')	('Alb19', 'HL')	0.001
('Alb19', 'HL')	('WT', 'HL')	0.004624
('WT', 'LL')	('WT', 'HL')	0.002936

Figure 38: Post hoc (Tukey's HSD) test results for F_v/F_m values presenting the comparisons with significant changes(p-value<0.05) between individual samples.The color coding is as follows; Yellow: compares the wild type in different treatments, Blue: compares the wild type with the mutant lines in the same condition, Green: Compares different mutant lines under the same conditions, Orange: compares the same mutant line under different conditions.

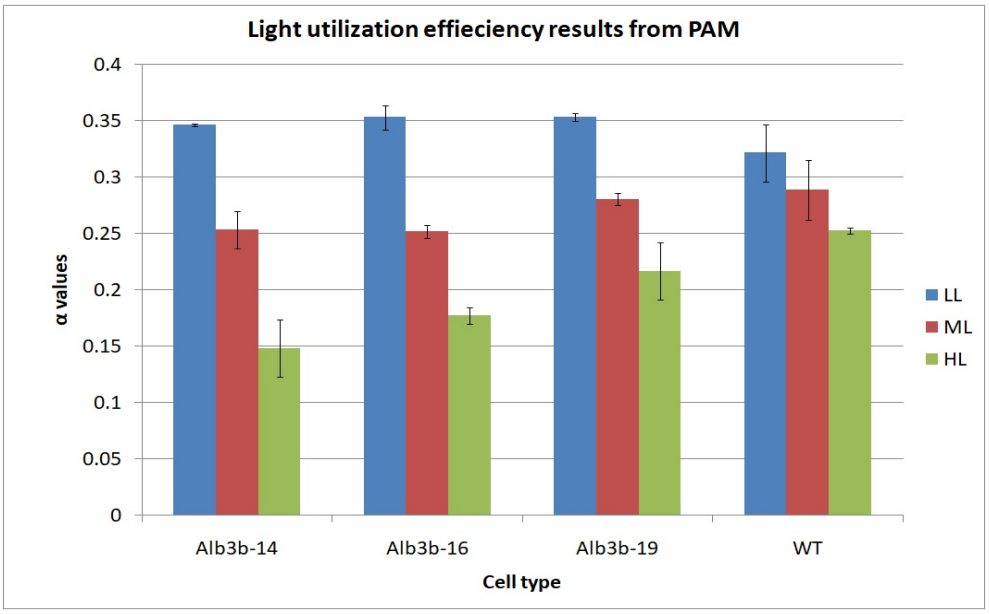


Figure 39: Light utilization efficiency(α) measurements from PAM for the different cell lines(WT, *Alb3b*-14,16, 19) in three different light treatments (HL, ML, and LL) measured as mean of values from 3 biological replicates for each cell line in each light condition. The α values can be seen decreasing with increasing light intensity in all the cell lines, but differently between mutants and WT similar to the F_v/F_m values.

The measurements of light utilization efficiency from PAM show a similar pattern to the F_v/F_m ratio values. It is observed in figure 39 that with the increasing light intensity, the light utilization efficiency decreases for all the cell lines. The extent of this decrease, Similar to F_v/F_m values, varies between the

mutants and the wild type. The decrease in light utilization values for the mutants when they are shifted from LL to HL ranges between 22% to 27%, And when they are shifted from HL to highlight it ranges between 22% to 41% with the *Alb3b*14 mutant showing the highest decrease. however, these values are just around 10% and 12%,respectively,in the wild type.

It is seen from the post hoc analysis results(Figure 40) that all the mutant lines' light utilization efficiency changes' are significant in all the different light treatments. Whereas, for the wild type the only significant difference in light utilization efficiency is between low and highlight levels. When the mutants are compared to the wild type it can be observed the values for all the mutants in low and HL levels are comparable to the same light levels for the wild type. However, the *Alb3b*-14 and 16 mutant lines show significant differences from the wild type under HL levels.

group1	group2	p-value
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'LL')	('Alb14', 'HL')	0.001
('Alb14', 'ML')	('Alb14', 'HL')	0.001
('Alb14', 'HL')	('Alb19', 'HL')	0.001373
('Alb14', 'HL')	('WT', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001
('Alb16', 'LL')	('Alb16', 'HL')	0.001
('Alb16', 'ML')	('Alb16', 'HL')	0.001
('Alb16', 'HL')	('WT', 'HL')	0.001
('Alb19', 'LL')	('Alb19', 'ML')	0.001
('Alb19', 'LL')	('Alb19', 'HL')	0.001
('Alb19', 'ML')	('Alb19', 'HL')	0.003013
('WT', 'LL')	('WT', 'HL')	0.001145

Figure 40: Post hoc (Tukey's HSD) test results for α values presenting the comparisons with significant changes(p-value<0.05) between individual samples. The color coding is the same as explained in Figure 38.

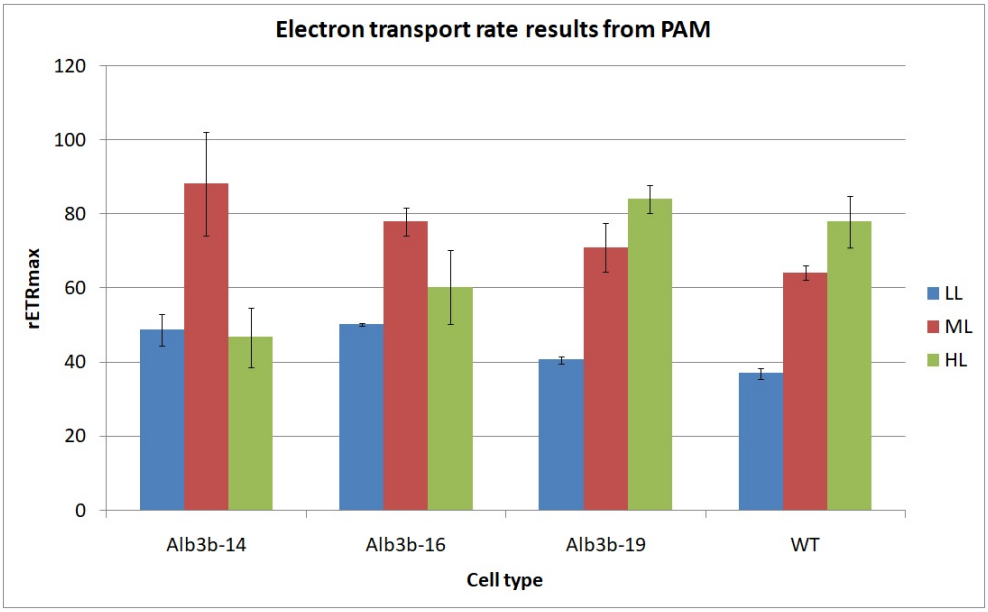


Figure 41: Relative maximum Electron transport rate($rETR_{max}$) measurements from PAM for the different cell lines(WT, *Alb3b*-14,16, 19) in three different light treatments (HL, ML, and LL) measured as mean of values from 3 biological replicates for each cell line in each light condition. The $rETR_{max}$ values can be seen increasing with increasing light intensity in WT and the *Alb3b*-19 mutants in a similar fashion.However, the *Alb3b*-14 and 16 mutants show a different behavior with the $rETR_{max}$ values increasing and then decreasing from LL to ML and then from ML to HL, respectively.

The maximum relative electron transport rate values($rETR_{max}$), as measured using PAM, indicated in figure 41 some unexpected results. Here, the *Alb3b*-19 mutants and wild type followed the same pattern, in which, as the light intensity increased, the $rETR_{max}$ values increased. This increase is around 74% from low to HL and around 20% from medium to HL. However, the *Alb3b*-14 and 16 lines show a different pattern in which the $rETR_{max}$ values increase from low to HL and then decrease under HL by around 47 and 22% respectively.

Post-hoc analysis of the $rETR_{max}$ measurements (Figure 42)indicate that, for all the mutant lines and the wild type, the increase in values from LL to HL is significant. It can also be observed that the decrease in values from HL to highlight is significant for the *Alb3b*-14. Whereas the increase in values from HL to HL is significant for both the *Alb3b*-19 and the wild type. As expected, there is a significant difference between the values for *Alb3b*-14 and 16 when compared to *Alb3b*-19 under highlight.

group1	group2	p-value
('Alb14', 'LL')	('Alb14', 'ML')	0.001
('Alb14', 'ML')	('Alb14', 'HL')	0.001
('Alb14', 'ML')	('WT', 'ML')	0.00603
('Alb14', 'HL')	('Alb19', 'HL')	0.001
('Alb14', 'HL')	('WT', 'HL')	0.001
('Alb16', 'LL')	('Alb16', 'ML')	0.001077
('Alb16', 'HL')	('Alb19', 'HL')	0.006996
('Alb19', 'LL')	('Alb19', 'ML')	0.001
('Alb19', 'LL')	('Alb19', 'HL')	0.001
('WT', 'LL')	('WT', 'ML')	0.001416
('WT', 'LL')	('WT', 'HL')	0.001

Figure 42: Post hoc (Tukey's HSD) test results for $rETR_{max}$ values presenting the comparisons with significant changes(p-value<0.05) between individual samples. The color coding is the same as explained in Figure 38.

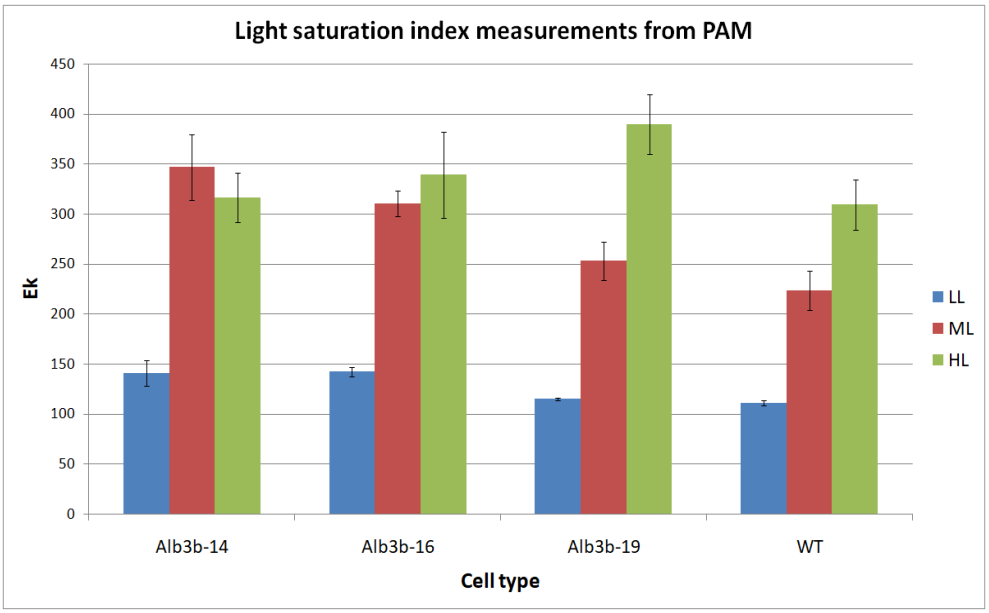


Figure 43: Light saturation index(E_k) from PAM for the different cell lines(WT, *Alb3b*-14,16, 19) in three different light treatments (HL, ML, and LL) measured as mean of values from 3 biological replicates for each cell line in each light condition. The E_k generally appears to be increasing with increasing light intensity in all the cell lines but to various extents.

The measurements of the light saturation indices of different cell lines under different light treatments are depicted in Figure 43. A general trend of increasing Ek values with increasing light intensity can be observed for all the cell lines, except for *Alb3b*-14, wherein the Ek values slightly decrease from ML to HL treatment. The common observation for all the cell lines is the steep increase in Ek from LL to ML conditions, the increments being 146,118,120, and 101 % for *Alb3b*-14,16,19 and WT, respectively. The *Alb3b*-16 mutant shows just a 9% increase in Ek from ML to HL. Whereas for the *Alb3b*-19 and WT,

this increase is around 54 and 38 %, respectively. Additionally, the mutants generally appear to have a greater Ek than the WT in all the light levels. The post hoc analysis to compare individual samples did not indicate a significant difference between any of the samples, which is not the case. This could be explained by the great deviation of the residuals in the ANOVA model from the homogeneous variance and Normality assumptions as indicated in Figure 29 in the Appendix.

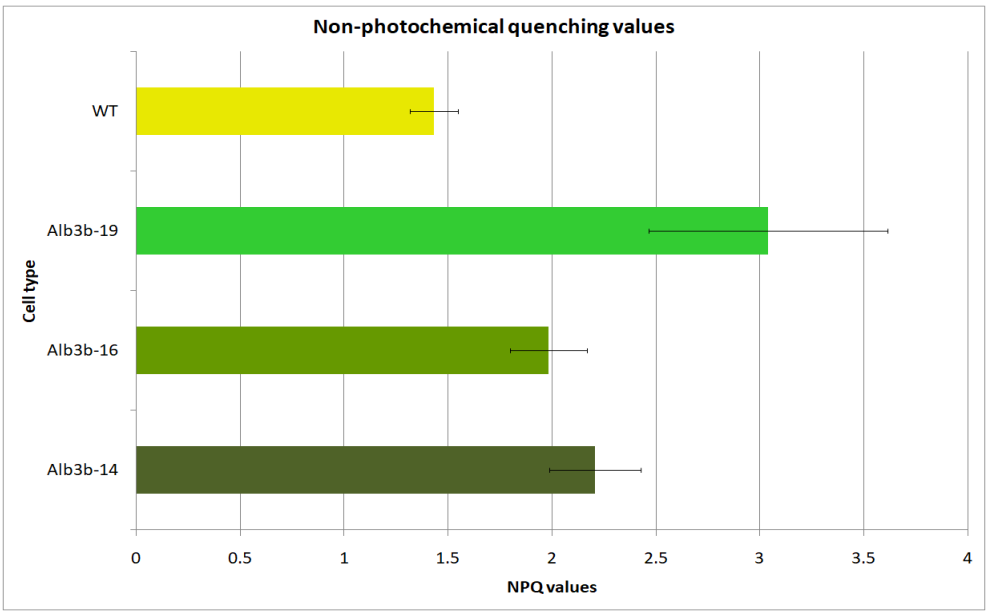


Figure 44: NPQ values values all the cell lines(WT, *Alb3b*-14,16, 19) measured from PAM after acclimation to LL conditions for 14 days. The values are averages from three biological replicates of each cell line.

The non-photochemical quenching measurements for the mutant lines and wild type using PAM fluorometry for the LL acclimated cells show that the mutants generally have higher NPQ values than the wild type, with the *Alb3b*-19 mutant having the highest NPQ values.

The results from ANOVA analysis conducted on the NPQ values (Figure 45) for different cell lines show that the p-value is about 0.002, meaning that the cell line significantly affects the NPQ values. Additionally, the results from post hoc analysis (Figure 45) indicate that the *Alb3b*-19 mutants have significantly different NPQ values from the wild-type.

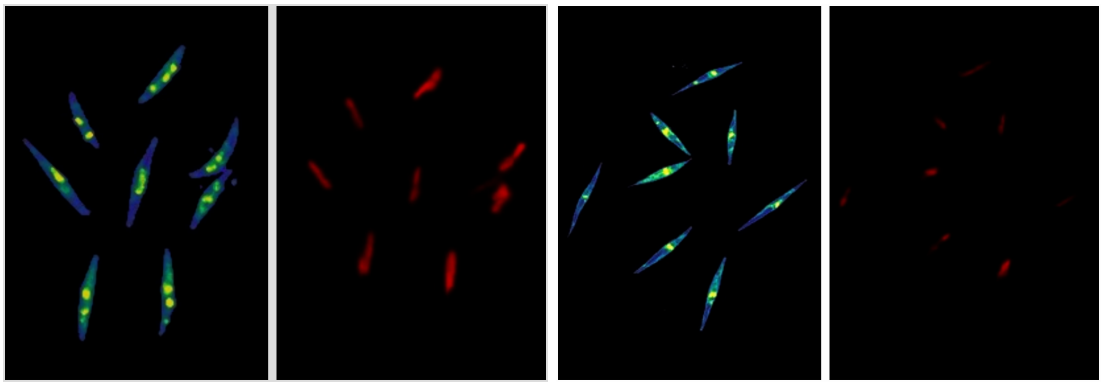
```
ANOVA Test Result:
F-statistic: 12.545033652270462
p-value: 0.002154731387449071

Tukey's HSD Post Hoc Test Results:
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
 Alb14  Alb16  -0.2223 0.8372 -1.0761  0.6314  False
 Alb14  Alb19   0.8343 0.0554 -0.0194  1.6881  False
 Alb14     WT  -0.7737 0.0764 -1.6274  0.0801  False
 Alb16  Alb19   1.0567 0.0175  0.2029  1.9104   True
 Alb16     WT  -0.5513 0.2417 -1.4051  0.3024  False
 Alb19     WT   -1.608 0.0014 -2.4618 -0.7542   True
-----------------------------------------------------
```

Figure 45: Results from One-way ANOVA and post-hoc(Tukey's HSD) for NPQ indicating a significant effect of cell line ($p<0.05$) on the NPQ levels and significant difference between the *Alb3b*-19 and WT and also one of the other mutant line(*Alb3b*-16).

7.3.4 Results from CLSM

(a) Wild type (b) *Alb3b*-14



(c) *Alb3b*-16 (d) *Alb3b*-19

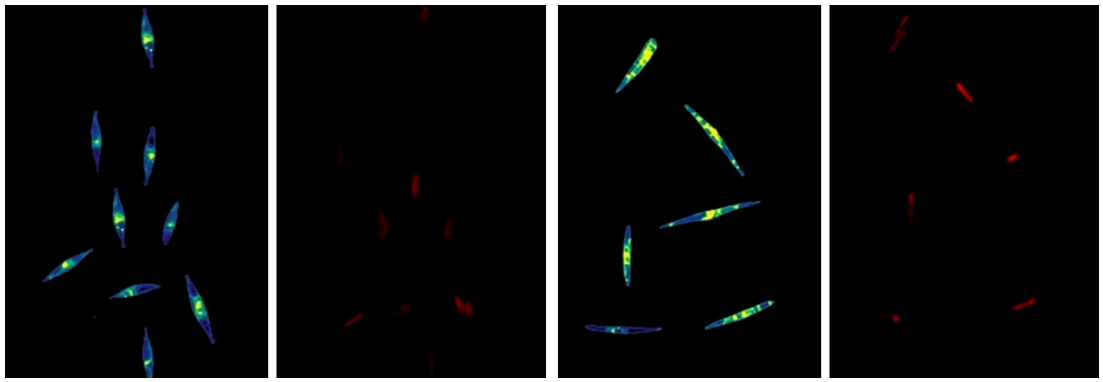


Figure 46: CLSM images of different cell lines exposed to LL levels of 35 µmol photons m^{-2} s^{-1} for 2 weeks and then stained with BODIPY 505/515. Images include those from channel 0, with BODIPY signals(Left), and from channel 1, with autofluorescence signals of the Leica SP8 microscopes. Images were processed using FIJI for applying appropriate look-up tables, removing background, and adjusting contrast levels.

The CLSM images from LL acclimated cells(Figure 46a) indicate variations in the lipid droplet structures among the different cell lines. The wild-type cells show multiple LDs(2-3) that appear compact and separated. Most of the cells in the *Alb3b*-14 and 16 mutant lines(Figure 46b and Figure 46c)seem to have a single compact LD having a similar size to that of LDs in the wild type. In contrast to all the other cell lines, the cells in the *Alb3b*-19(Figure 46d) line show much more dispersed LDs spread almost throughout the cell without conspicuous separation. In terms of autofluorescence images, all the cell lines emitted detectable levels of autofluorescence, with the wild-type cell having the strongest autofluorescent emissions compared to the *Alb3b* mutants. Also, the cell morphotypes for most of the cells are comparable between the cell lines.

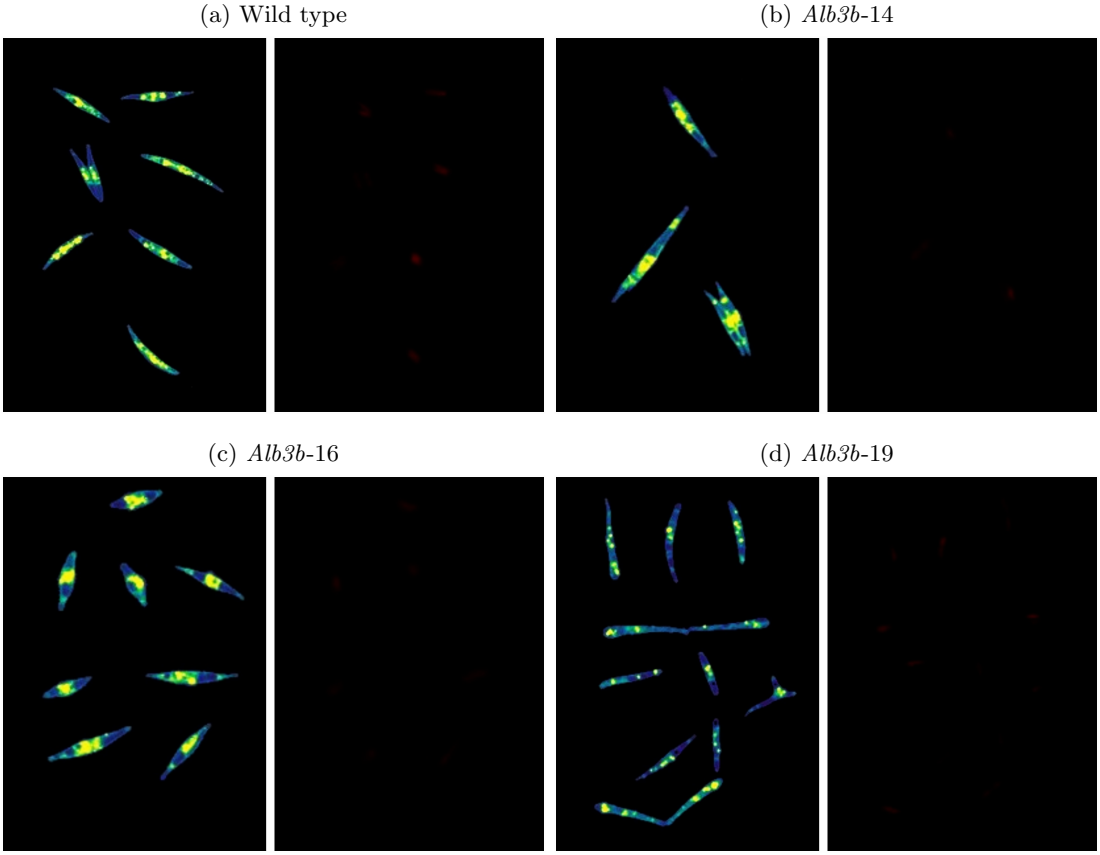


Figure 47: CLSM images of different cell lines exposed to ML levels of 200 µmol photons m^{-2} s^{-1} for 2 weeks and then stained with BODIPY 505/515. Images include those from channel 0, with BODIPY signals(Left), and from channel 1, with autofluorescence signals of the Leica SP8 microscopes. Images were processed similarly to those in Figure 46.

The Lipid droplet size and structure change considerably in the HL-treated cells for all the cell lines compared to their LL counterparts. In the wild-type, *Alb3b*-14, and *Alb3b*-16 cell lines, it can be observed that the LDs have increased in size due to stress and no longer appear compact and have become more dispersed. The morphology of these cells seems to be unaffected under HL exposure. However, the *Alb3b*-19 mutants show great variation in LD and cell morphology as compared to others. Here, it is seen that most cells have several compact LDs (3-8) spread across cells that appear elongated than the normal fusiform morphology. It was also seen that some of the tri-radiate cells in these cultures had at least one

of their arms extended beyond the others. Two common observations among all the cell lines include the tendency of some cells to cluster with others and significantly lower autofluorescence emissions compared to their LL counterparts, as observed in figure 46

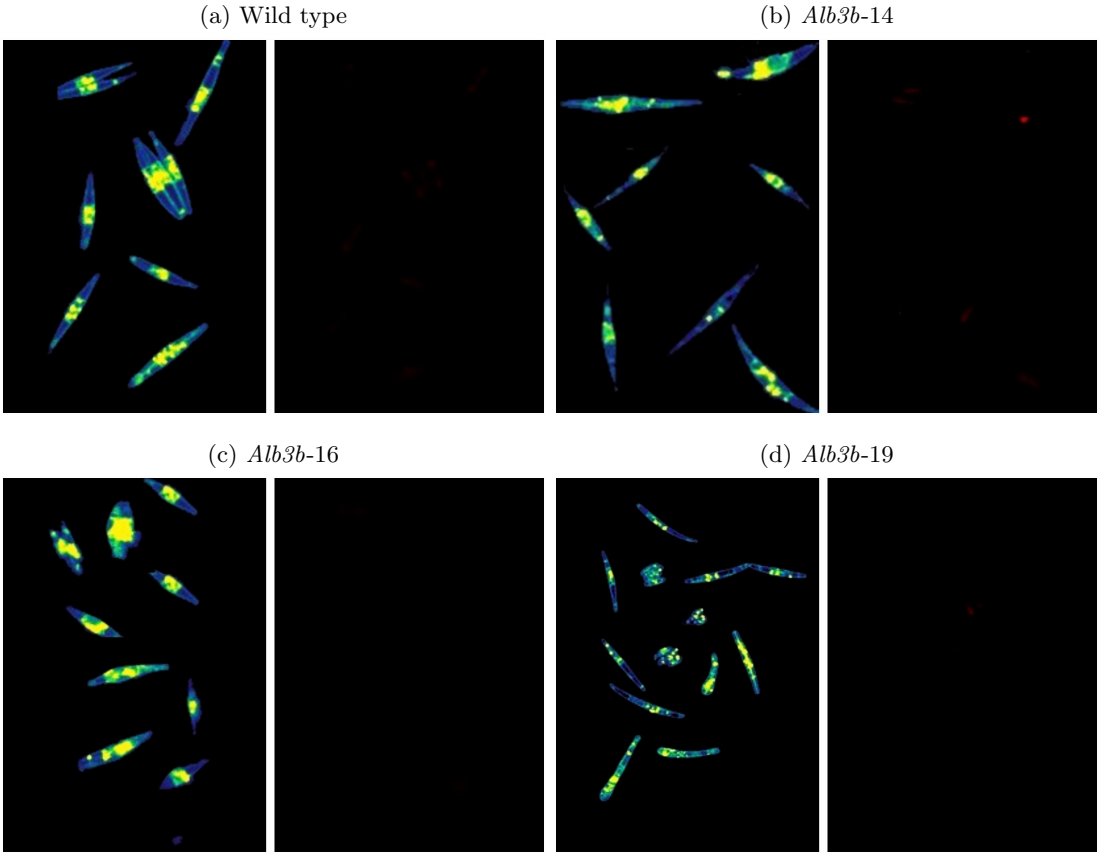


Figure 48: CLSM images of different cell lines exposed to HL levels of 680 µmol photons m^{-2} s^{-1} for 2 weeks and then stained with BODIPY 505/515. Images include those from channel 0, with BODIPY signals(Left), and from channel 1, with autofluorescence signals of the Leica SP8 microscopes. Images were processed similarly to those in Figure 46.

The imaging results for HL-treated cells(Figure 48) yielded almost the same results as the ML samples, except that the cells in HL were more stressed and consequently had bigger LDs. Even though The LDs in HL acclimated wild-type cells appear less compact and distinguished from each other than the ones in LL, they appear to congregate in the center of the cell(Figure 48a). LDs in *Alb3b*-14 and 16 cells(Figure 48b and Figure 48c) have almost the same structural and distribution features as their counterparts in medium-light but with an increased size. In the case of the *Alb3b*-mutant lines, no clear pattern was found as cells with different LD structures and dispersion were observed. Additionally, all the cell lines in HL seem to have the stress-responsive clustering of cells with *Alb3b*-19 mutants having more clusters and more fractions of oval-shaped cells(Figure 48d). Furthermore, the majority of the cells in all cell lines had significantly lower autofluorescence levels than both medium and LL treatments, with many of them even being unable to be detected.

7.3.5 Results from Real-time q-PCR

The results from the nanodrop assessment of the RNA are presented in the table 20. It was observed that several of the samples(yellow) had A260/A80 ratios above the recommended level of 2, therefore indicating reduced protein contamination, but had the A260/A230 ratios below 2, thus indicating possible phenolic contamination from the RNA isolation procedure. Just one sample(red), Alb3B-16LL3, had both the A260/A280 and A260/A230 ratios below 2, hence is of bad quality. In terms of the nucleic acid concentrations, the same sample had concentrations below the recommended level of 200 ng/µL, while the rest had more than this, with some of them even above 1000 ng/µL.

The results of calculated RIN values from the bioanalyzer are presented in Figure 21. All the samples had RIN values above 4, which is the recommended value for RNAseq. Since no recommended RIN levels are established for q-PCR, this value was used as a threshold for qualifying the sample as good quality with less RNA degradation.

Fluorescence and melting curves from the q-PCR reaction for amplification of the two selected reference genes indicated possible contamination with genomic DNA in the majority of the samples. As seen in the Figures, there are two batches of fluorescent curves, one having a higher range of Cq values and the other one with a lower range of Cq values. The fluorescence curves with a higher range of Cq values come from the RT-ve samples and those with a lower range of Cq values come from the cDNA samples, thus indicating the possible presence of genomic DNA at concentrations lower than that of the cDNA in the samples. This is further supported by the melting curves shown in Figures, in which all the melting curves including those from RT-ve samples and cDNA samples peak at the melting point of around 80°C

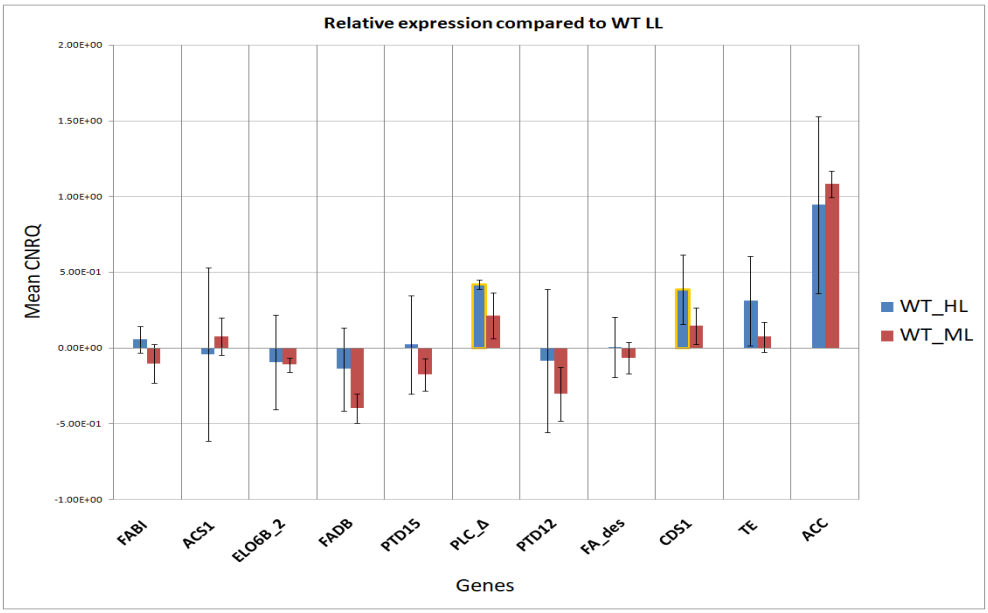


Figure 49: Results from q-PCR comparing the expressions of the different genes in WT in HL and ML compared to WT in LL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of WT in each light condition. The bars highlighted with yellow outline indicate significant upregulation or downregulation based on post hoc analysis. The PLC and CDS1 genes are seen as significantly up-regulated in HL samples of WT.

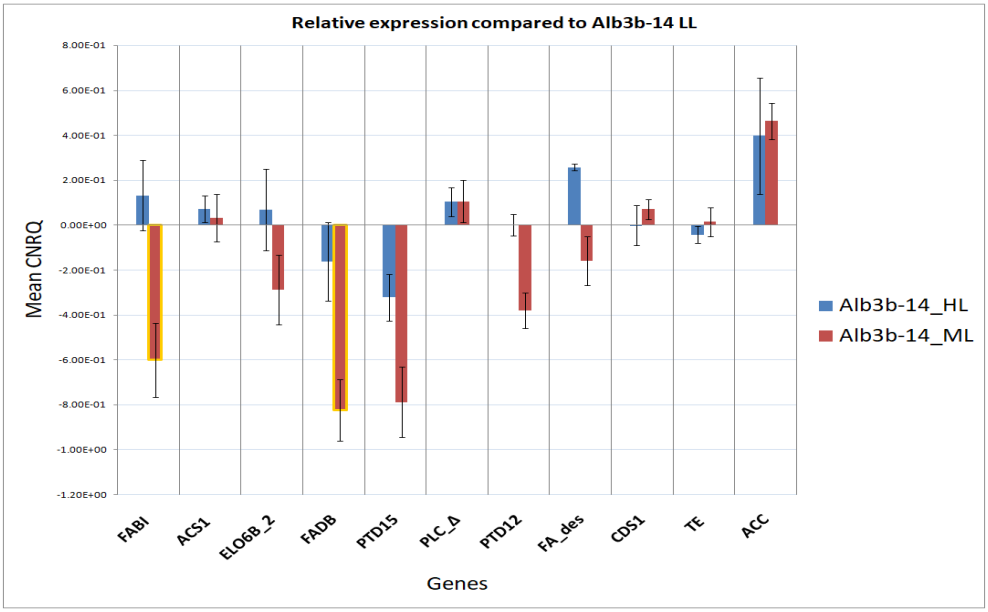


Figure 50: Results from q-PCR comparing the expressions of the different genes in *Alb3b*-14 mutants in HL and ML compared to the same mutant in LL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis. The FABI and FADB genes are seen as significantly down-regulated in ML samples of *Alb3b*-14.

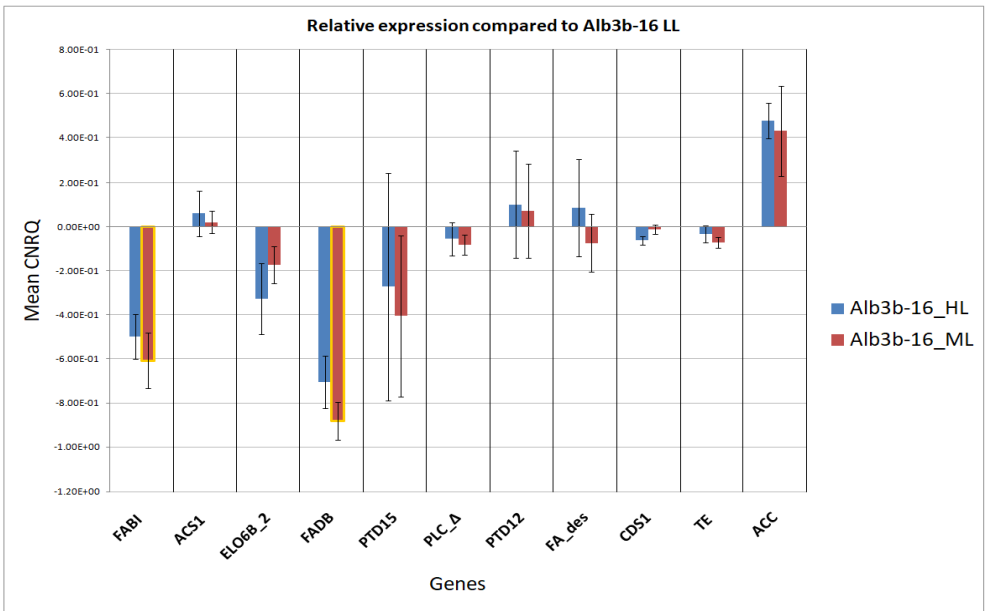


Figure 51: Results from q-PCR comparing the expressions of the different genes in *Alb3b*-16 mutants in HL and ML compared to the same mutant in LL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis. The observations are the same as that of the *Alb3b*-14 mutants(Figure 50) with significant downregulation in FABI and FADB enzymes in ML-treated mutants.

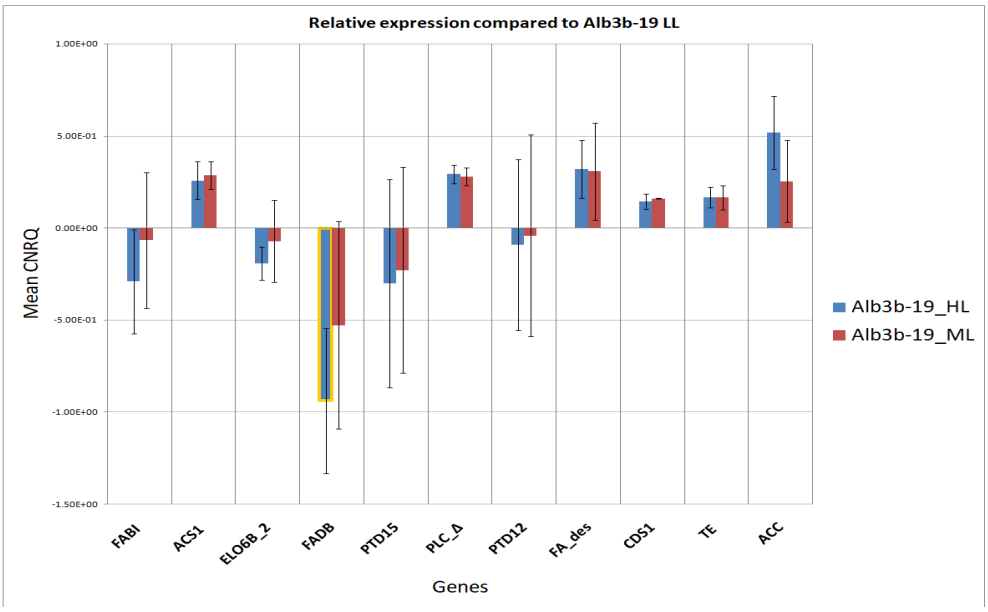


Figure 52: Results from q-PCR comparing the expressions of the different genes in *Alb3b*-19 mutants in HL and ML compared to the same mutant in LL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis.In contrast to the *Alb3b*-14 and 16 mutants(Figures 50, 51), there are no significant changes in FADB and FABI under ML. However, FADB is significantly down-regulated under HL.

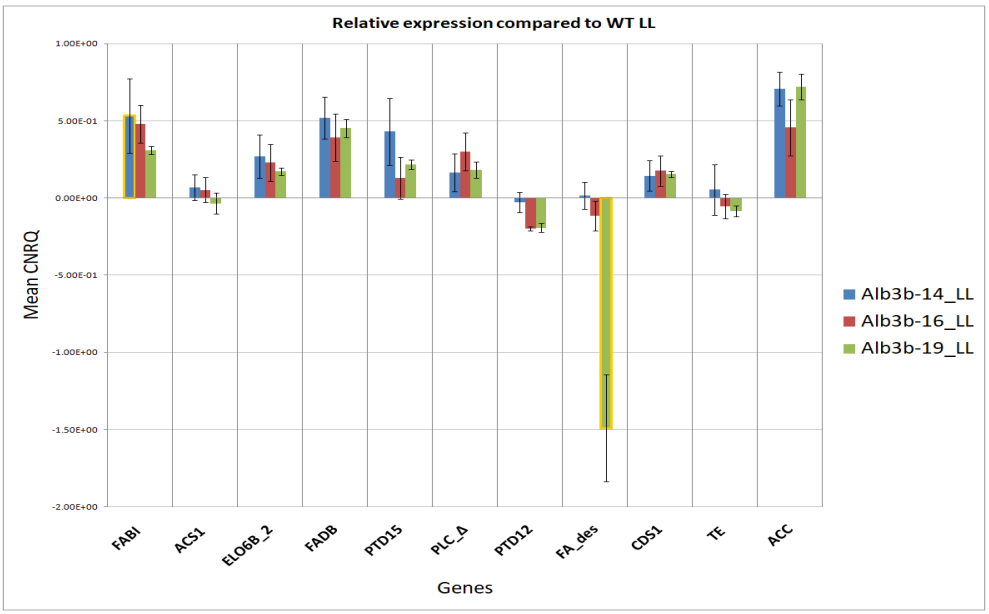


Figure 53: Results from q-PCR comparing the expressions of the different genes in all the *Alb3b* mutant lines LL compared to WT in LL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis. The FABI gene in *Alb3b*-14 is significantly upregulated and the FA-desaturase in *Alb3b*-19 is significantly down-regulated.

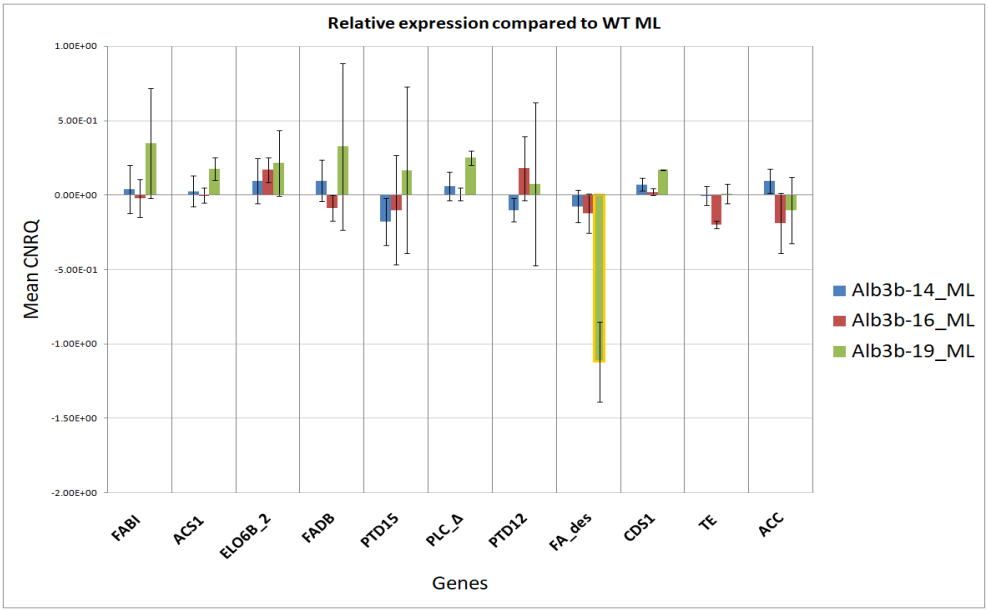


Figure 54: Results from q-PCR comparing the expressions of the different genes in all the *Alb3b* mutant lines ML compared to WT in ML. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis. The FA-desaturase in *Alb3b*-19 is significantly down-regulated.

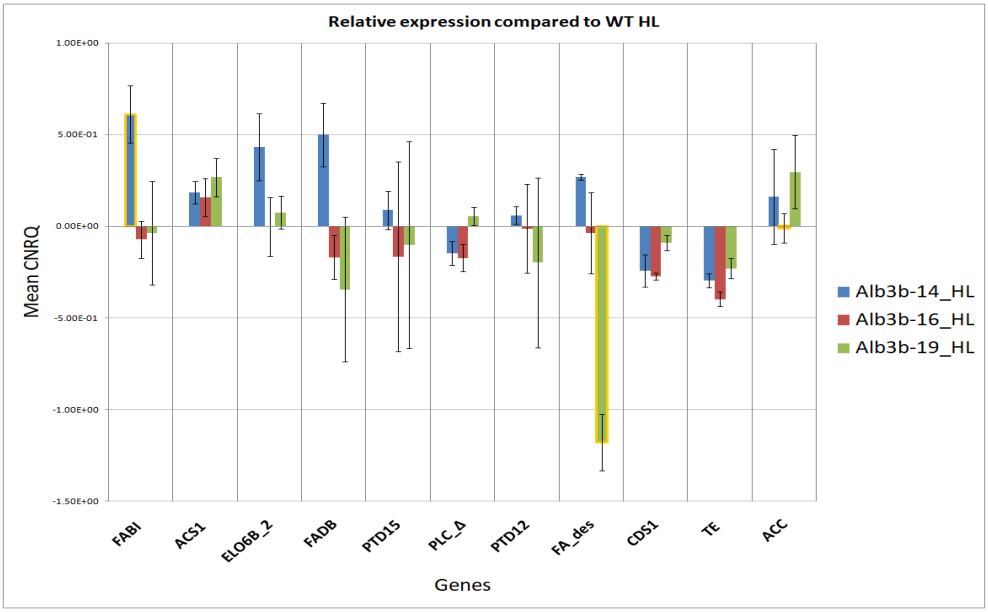
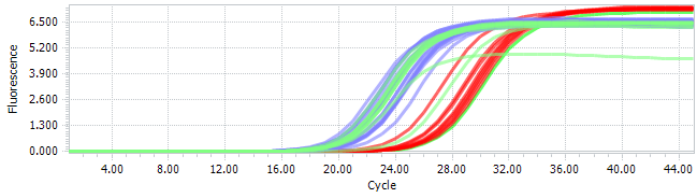


Figure 55: Results from q-PCR comparing the expressions of the different genes in all the *Alb3b* mutant lines HL compared to WT in HL. The values are calculated as the mean of CNRQ values obtained from the analysis of Cq values in qBASE+ from three biological replicates of the mutant in each light condition. The bars highlighted with a yellow outline indicate significant changes in expression post hoc analysis. The observations are the same as the LL samples(Figure 53with significant upregulation of the FABI gene in *Alb3b*-14 and downregulation of FA-desaturase in *Alb3b*-19.

The results from the q-PCR reaction indicate a high level of variability in many of the genes among the biological replicates of all the samples under all the different light conditions. This can be seen in the wide error bars in Figures 49, 50,51,52,53,54,55.

The results indicate differences in the regulation of the different genes studied between the WT and mutant cell lines. The WT cells in HL treatment show a significant upregulation of two of the phospholipid metabolism enzymes, namely CDS1 and PLC compared to WT in LL(Figure 49). However, this upregulation is not observed in any of the mutants as can be observed in Figures 50, 51, and 52. Another contrasting difference is observed in the expression levels of two of the enzymes involved in the acyl chain elongation cycle in the plastid, which are the FADB(Malony coA- ACP transacylase) and FABI(Enoyl ACP reductase). These two enzymes appear to be significantly down-regulated under ML treatment in the *Alb3b*-14 and 16 mutants compared to the same mutants in the LL treatment, whereas they do not significantly change in WT under ML or HL levels compared to WT in LL. In the *Alb3b*-19 cells, only the FADB enzyme is significantly down-regulated, but in the HL treatment, compared to the same mutant under LL condition. However, when the mutants under each of the light treatments are compared to WT in the same light treatment, no significant expression change is observed in the mentioned phospholipid metabolic and the acyl chain elongation enzymes(Figures 53,54,55), except for a significant upregulation of the FABI in *Alb3b*-14 cells in LL compared to WT in LL. Another observation in the later comparison is the significantly down-regulated levels of a predicted fatty acid desaturase enzyme localized in the chloroplast(Phatr2 50443) in the *Alb3b*-19 cells under all the light treatments compared to WT in each light level. Analysis of the melting curves of this particular gene shows that the *Alb3b*-19 mutants show a separate melting peak from the other mutants and the wild type(Figure 56).

(a) Amplification curves of the Phatr50443 gene in WT and mutants



(b) Melting curves of the Phatr50443 gene in WT and mutants

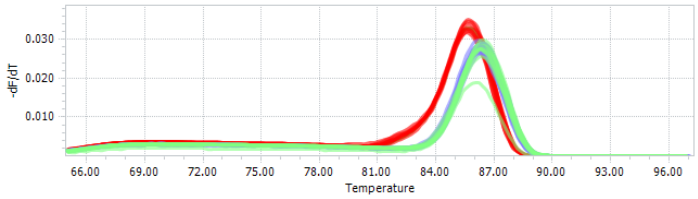


Figure 56: The results for amplification and melting curves of the Phatr50443 gene from the Light cycle 96 software. The WT cells are represented by the green lines, the *Alb3b*-14, and 16 cells by the blue lines, and the *Alb3b*-19 cells by the red lines. a) The amplification curves showing the red lines(*Alb3b*-19) generally expressed low compared to the other samples as the curves appear later than those of the other samples *Alb3b*-14, *Alb3b*-16, and WT). b) Melting peaks showing the red lines(*Alb3b*-19) forming a separate melting peak compared to the blue and green lines(*Alb3b*-14, *Alb3b*-16, and WT).

8 Discussion

Research on lipid metabolism in diatoms or microalgae is of great significance because of their potential to be used as feedstocks for the sustainable production of lipids or fatty acids to be used in various industries like bio-fuel, human nutraceuticals, and aquaculture feed production. Studying the lipidomics in knock-out mutants with a non-functional protein, which is otherwise important for the normal performance of a major metabolic pathway in the cell, can help us understand the influences of other pathways on lipid metabolism. This can lead to insights to be used for further research or development of strategies for commercial microalgal culturing for industrial lipid production. This project aimed at studying the lipidome of one such knock-out mutant, the Albino-3b mutant, compared to the wild-type cells.

As explained, the project involved developing a standardized workflow for processing and analyzing the lipidomics data from GC=MS, and conducting light treatment experiments to observe changes in various parameters (Table 3)associated with cell structure and function. The factors affecting the accuracy of the developed pipeline, the results obtained after passing the lipidomics data through the same, and how these results can be compared to the observations from previous research and those made from the experiments in this project are discussed sections below.

8.1 Principle component analysis and statistical modeling of MS-MS data: Challenges and Limitations

One main component of the pipeline developed for this project was to perform PCA after all the necessary pre-processing steps such as outlier imputation and scaling. This was performed on subsets of the GC=MS lipid class dataset containing measurements for wild-type samples and one of the mutant lines. That accounts for 3 subsets of data comparing each mutant line with the wild-type in two different light conditions.

The greatest advantage obtained by doing PCA in this project was generating visualizations that gave an overall overview of the entire dataset, The scatter plots in the space of principle components in not useful by themselves except for showing whether the samples get clustered in a discernible manner. The scatter plots were overlaid with corresponding loadings plots to obtain more information on which variables contribute most to the variance. These graphical representations gave a picture of how the various samples of different cell lines under different treatment conditions are differentiated based on the concentrations of different lipid classes in a lower dimensional space. This has facilitated circumventing the impossibility of plotting all the samples in 10 dimensions representing the 10 lipid classes in the dataset. Although the length and orientation of the arrows representing the loadings give valuable information as mentioned in Section 5.16.2, it was not possible to present the exact values of loading scores or angles between the arrows, as it can make the graphs redundant and hard to interpret. This limitation also applies to performing PCA on the Fatty acid composition datasets. Although perfectly separated clusters of different samples were observed for the fatty acids composition measurements for most of the lipid classes, the biplots generated were highly crowded because of the high number of variables, that is fatty acid compositions, making them hard to understand or interpret.

One main problem faced during the performing PCA was choosing between a standard PCA or a kernel PCA. Standard PCA assumes a linear correlation between the studied variables, whereas kernel PCA is adapted to non-linearly separate variables. For the MS-MS dataset, Linear relationships between

the measurements of concentrations of different lipid classes were observed for many of the lipid class pairs for WT and the mutant lines as presented in Figure 57. The heat maps presented here indicate that many of the lipid pairs have the absolute value of the person correlation coefficient(R) above 0.5 in all the cell lines, indicating strong linear correlations. However, non-linearity($R<0.5$) can also be observed in almost as many lipid class pairs in all the cell lines. However, standard PCA was chosen for the pipeline, even though the violation of the linearity assumption could potentially lead to wrong inferences. This was done mainly because in kernel PCA, the influence of individual variables cannot be represented using individual loading scores like in standard PCA, Additionally, selecting kernel PCA will lead to further difficult choices of the best kernel for the dataset, like Gaussian, tanH, or Neural net kernel, based on the kind of relationship between lipid pairs. Furthermore, the results obtained by executing a standard PCA on the dataset can be checked for correctness based on the results obtained from the second most important component in the pipeline, which is statistical modeling. Another issue that potentially could lead to misinterpretation is the fact that when comparisons are made between mutant and wild types, measurements of both these cell lines are considered for calculating the eigenvalues in PCA. Thus, the correlation interpreted from the angle between the loading score vectors in the loading plot, between lipids, will be based on all the measurements. This necessitates a cross-verification using the original dataset, to check if the correlation exists in the values within each cell line or between the cell lines. For example, in the 3D biplot comparing *Alb3b*-3b mutants and wild-type, a strong negative correlation is observed between PE and TAG as the loading score vectors are almost at a 180-degree angle. However, it is hard to interpret whether this negative correlation exists in both the cell lines or across the cell lines. In this case, we could cross-check it with the pair plot generated from the original dataset (Figure 2. It can be seen in the regression plot comparing PE to TAG this a negative correlation exists, just in the *Alb3b*-14 mutants, whereas in the wild type, they exhibit a positive correlation.

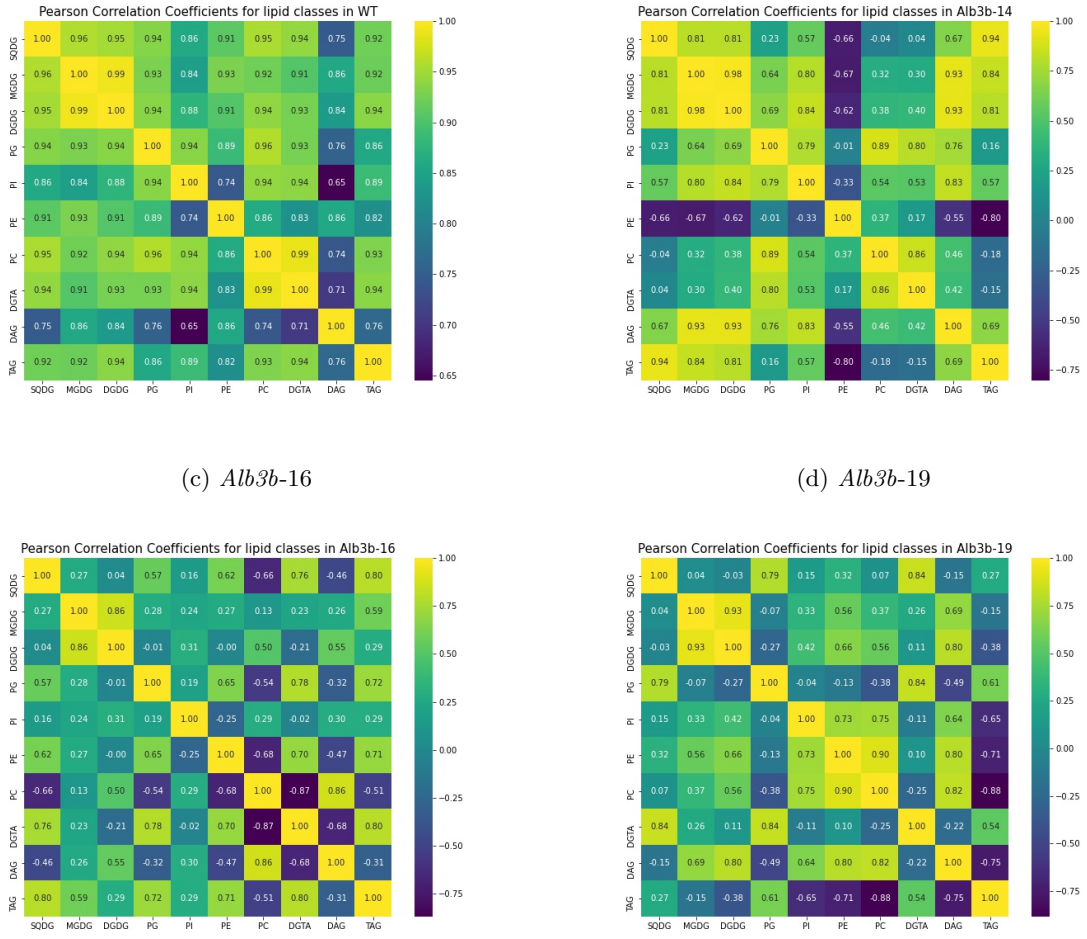
Misinterpretations could also arise during PCA, as the explained variance is not cent percent and the resultant biplots may not represent some hidden structures in the data. One reason for this is the inherent orthogonality of the principle components, that is the defined principle components are perpendicular to each other. There may be non-orthogonal components that possibly explain better variance and reveal unseen structures in distribution. One example of this is the correlations observed between TAG and the betaine lipid DGTA. It is seen in figure 16 that these two lipids are possibly positively correlated because of the acute angle between the loading score vectors. Additionally, in the figures 19 and 21, a negative correlation is observed for the same lipid classes. However, comparing the regression models for these two lipid classes in WT and the mutant lines in Figure 3 shows contrasting results, wherein the *Alb3b*-14 shows negative correlation and the wild-type, *Alb3b*-16, and *Alb3b*-19 shows positive correlation.

As mentioned above, the fatty acid composition data set was also passed through the same data processing and analysis pipeline. It was observed that for each of the lipid classes the data for the measurements of Concentrations of each of the different fatty acid compositions, yielded clear clustering of the different samples in the PCA scatter plot. that is, 4 different clusters representing mutants In two light conditions and the wild-type cells in two light conditions were observed during each of the comparisons. this indicates that the fatty acid composition varies between the cell lines and also between the different light conditions. Additionally, how this variation occurs between the different light conditions is different between the cell lines. However, making biplots for these comparisons was challenging because of the high number of unique fatty acid compositions that each lipid class has, which consequently resulted

in very crowded plotting plots that are hard to interpret. Regardless of this complexity, close observation of these graphs was done To check for any patterns in change of fatty acid compositions for each of the lipid classes. Nevertheless, no inclination for any of the cell lines or any of the Light treatments toward particular fatty acid compositions was observed in any of the lipid classes except for the neutral lipid TAG. This was cross-verified with the results obtained from statistical modeling of the fatty acid composition data for each of the lipid classes and similar results were obtained. That is a conspicuous difference in variation of concentrations of fatty acid compositions in TAG was observed Between the cell lines, but this was not the case for other lipid classes. This can also be considered a limitation of the PCA in this pipeline because even though a clustering was observed for the different samples in all the lipid classes, The reason for these differences between the samples could not be delineated.

(a) Wild-type

(b) *Alb3b*-14



(c) *Alb3b*-16

(d) *Alb3b*-19

Figure 57: Heat maps showing the Pearson coefficient values(R) between different lipid classes for WT and mutants. The R values were calculated to assess the linearity between the different lipid classes before performing the standard PCA. R values were estimated using the 'corr' function in the pandas library and the heat maps were generated using the 'seaborn' library in Python.

Statistical modeling for the MS-MS data was done based on results from unpaired t-tests comparing the means of the concentrations of each lipid class or fatty acid composition between the light treatments

for WT and the mutant lines. As Shapiro Wilk's tests yielded results showing normal distribution of concentrations of the majority of the lipids, and Levene's tests resulted in an indication of non-homogeneous variances in a considerable number of compared samples, Welch's T-test was chosen as the best choice of statistical test. A log transformation was added to the compared samples to scale the data and also to improve the normal distribution curve characteristic of the samples.

The main limitation here is the low number of observations in each sample compared, that is just 9 values in each treatment comprising of three biological replicates, with three technical replicates for each of them. This significantly reduces the statistical power and reduces confidence in the interpretation made from these tests. Another alternative was to use a two-way ANOVA model to fit this data with two defined categories, that is cell line, with 4 values(WT, *Alb3b*14, *Alb3b*16, and *Alb3b*19) and the treatment(LL and ML). Even though this test was performed, and significant interaction effects were observed between cell line and light treatment for each of the lipid classes, ANOVA was not chosen as a step in the pipeline as the residuals of concentrations of all the lipid classes significantly deviated from the normal distribution and were not homogeneously varying across the fitted values in ANOVA OLS model. However, many of the inferences made from the statistical tests were comparable to the results from PCA. For instance, in the PCA biplot comparing *Alb3b*-14 mutants with Wild type(Figure 16, two of the lipids causing the most variation between the cell lines are observed to be PE and TAG, and they also appear to have a strong negative correlation. In comparison, the T-test results show that, under high light, PE is the most significantly reduced lipid class and TAG is the most significantly increased one in ALb3b-14 mutants, indicating its negative correlation in the mutants. In contrast, both of these lipids significantly increase in the wild type in high light compared to LL, thus showing a difference in variation of these two lipids in the mutant and the wild type.

8.2 Interpretations of the results from MS-MS data analysis

It has been previously proven that diatoms have an intricate mechanism consisting of photo-receptive and other sensory systems, with associated metabolic pathways that are utilized for sensing and consequently modulating their photosynthetic Machinery to acclimatize to the changes in the ambient light conditions(Wilhelm et al., 2006). Since the metabolic pathways including those associated with photosynthesis and lipid metabolism are intertwined, the changes in one of the metabolic pathways can potentially influence the other metabolic pathways(Wilhelm et al., 2006). Changes in concentrations of total lipids and individual lipid classes have been discovered in diatom cells treated with high light in previous research. For example, in a study conducted by Ding et al., 2023, the light stress-induced cells, with exposure to a high light level of 300 µmol photons m^{-2} s^{-1} for 3 days showed an increased amount of the total lipid content in the cells compared to the control in LL level of 50 µmol photons m^{-2} s^{-1}(Ding et al., 2023). These high Light treated cells also indicated higher levels of the neutral lipid TAG, upregulation of certain phospholipid remodeling enzymes like the PDAT, and decreased levels of the plastidic membrane lipids like the MGDG, DGDG, and SQDGs(Ding et al., 2023). The wild-type cells continuously exposed to highlight treatment after three days in this study were found to have higher total lipid content and tag levels than cells treated in highlight for three days. Furthermore, the lipid levels of cells that were recovered from the highlight treatment for three days were also measured and it was seen that most of these stress responses reversed. In the cells recovered from high light the total lipid levels went back to

normal levels and the glycolipid MGDG increased back to the normal level. This confirms that changes in lipid metabolism are also a part of the stress response towards light treatment in the cells. It also points towards possible phospholipid and glycolipid remodeling of the membrane constituents of the cell to produce storage lipids (Ding et al., 2023). This master's project aimed to study the changes in the lipid profile of the wild-type and *Alb3b* mutant cell lines under continuous highlight and LL exposure for two weeks. Additionally, changes were observed in mutants compared to the wild type in several parameters, including photosynthetic performance, pigment concentrations, thylakoid membrane structures, and photo-protection in the research conducted by Nymark et al., 2019, so change in the lipid profile between the mutants and the wild type was also expected.

The inferences from the analysis of lipid classes is summarized in the following table:

	LL v/s ML treated samples			
Lipids	**Phospholipids**	**Glycolipids**	**Betaine Lipids**	**Neutral Lipids**
Classes	**PI, PC, PG, PE**	**MGDG, DGDG, SQDG**	**DGTA**	**TAG,DAG**
WT	SI	SI	SI	SI
***Alb3b*-14**	SD&NC	SI	NC	SI
***Alb3b*-16**	SI&SD&NC	SD&NC	SI	SI(TAG)&SD(DAG)
***Alb3b*-19**	SI&(SD)*	SD&NC	SI	SI(TAG)&SD(DAG)

Table 4: A summary of inferences from statistical analysis of lipid class data from MS. The abbreviations are as follows; SI: significant increase, SD: significant decrease, NC: No significant changes. * means a general trend towards the inferred change. Significance is defined by p-values<0.05.

As presented in the results section and summarized in table 4, clear differences were observed both in PCA and T-test results in the concentration of several lipid classes between LL and ML-treated samples in all cell lines. Also, obvious differences can be observed between the wild-type and the mutant cell lines. As seen in figure 22, in the wild-type samples treated at an ML level of 200 µmol photons m^{-2} s^{-1} for 2 weeks all the lipid classes have significantly increased compared to the LL treated samples at 35 µmol photons m^{-2} s^{-1}. This is consistent with results from a study by Ding et al., 2023, where the total lipid content in the cell in continuous HL treatment increased significantly compared to the control. However, the significantly increased amount of phospholipids and glycolipids in the HL-treated cell is not in accord with possible membrane lipid remodeling suggested by Ding et al., 2023. The possible explanation for this is that the long-term exposure has led to acclimation of the cells to the condition and reveal of the stress responses resulting in increases of the glyco- and phospholipids. Contrastingly, the mutants show significantly decreased levels of certain phospholipids, and increased levels of TAG (Figures 23 24. The group of significantly decreased, remained unchanged, or significantly increased phospholipids are not similar among the mutants. In terms of glycolipids, the *Alb3b*-14 have significantly increased amounts of MGDG AND DGDG under HL treatment, but the *Alb3b*-16 and 19 have significantly decreased DGDG and unchanged MGDG. These results could mean that the stress responses are not reversing in the mutants in the same way as in the wild-type and even if there is a reversal to some extent it is not occurring similarly among the different mutants. Another notable observation is the significantly increased amount of either the sulpholipid SQDG or the phospholipid PG in the ML-treated mutants. These are polar lipids that are known to replace the glycolipids in thylakoid membranes under light

stress and consequent remodeling events(Lepetit et al., 2011). These results lead to the inference that membrane lipid remodeling is occurring also in the mutants but, might be in a different fashion than in the wild-type.

Regarding the analysis of the fatty acid composition data, as presented in the Results section, a difference was observed in the TAG fatty acid composition change between mutants and the WT, when the LL and ML treatments were compared. It was observed that most of the long-chain polyunsaturated fatty acids were significantly higher and most of the medium-chain saturated fatty acids or those with a low degree of unsaturation were significantly lower in the ML treatment compared to the LL treatment for the WT cells. In contrast, the mutant cell lines had just the opposite scenario. This general trend was not observed in the fatty acid compositions of other lipid classes, but specific observation of the FA composition data analysis for the glycolipids MGDG and the phospholipid PC shows that the fraction of these lipids with most of the compositions containing the PUFAs like EPA and DHA changes differently between the mutants and WT when comparison is made between the different light levels. It can be seen in Figures 7a,8a, and 9a that the mutants have significantly increased amount of many of the PUFA containing molecular compositions, including those with both EPA and DHA, in their PC fraction under ML treatment, whereas, in the WT, most of the same are significantly lower in ML as seen in figure 10a. As opposed to this, In the MGDG fraction, the mutants have significantly lower levels of most of the PUFA containing compositions, including the ones with at least one EPA, under ML condition(Figures 7b,8b, and 9b), while the WT has significantly higher levels of the same in ML(Figure 10b)

Previous studies have shown that the fatty acid composition of the lipids in diatom changes in response to environmental factors including irradiance(Qiao et al., 2016,Guihéneuf et al., 2008). For instance, the study by Guihéneuf et al., 2008 has found that the diatom *Skeletonema costatum* had the highest level of the PUFA, EPA under saturating light levels of 340 µmol photons m^{-2} s^{-1} and the levels of the same lipid had shown a significant decrease under limiting light conditions. The research by Qiao et al., 2016 has observed a significant increase in another PUFA called DHA, with increasing irradiance. The changes in the fatty acid composition in the membrane lipids like the glyco- and phospholipids can be interpreted as an attempt to adapt to changed conditions by adjusting the membrane fluidity(Tanaka et al., 2022). The change in fluidity or stability of the membranes can have implications on photosynthetic performance as it will affect the rate of electron flow through these membranes(Mock and Kroon, 2002a). For instance, a previous study by Mock and Kroon, 2002b has shown an increased amount of EPA in the MGDG fraction under limited light conditions in the antarctic sea ice diatoms. This was also associated with an increased electron transport rate. It is also known that the presence of membrane pigment-protein complexes in the thylakoid membrane affects the lipid composition of these membranes(Mock and Kroon, 2002b). For example, a sufficient level of these pigment-protein complexes is required to obtain a bilayer membrane with a considerable amount of non-bilayer lipids like MGDG to ensure appropriate fluidity and electron transport(Mock and Kroon, 2002b). Thus the difference between the *Alb3b* mutant lines and WT observed in how this membrane lipids changes in adaptation to light conditions can be linked to the difference in the pigment-protein complexes in the mutants' thylakoid membrane as observed by Nymark et al., 2019. Additionally, the differences observed in the thylakoid membrane structures(Nymark et al., 2019) in the mutants compared to WT could also be due to this difference in the membrane lipid compositions.

8.3 Interpretations of results from lab work

8.3.1 BODIPY fluorescence measurements

No discernible pattern was observed in the BODIPY fluorescence for all the cell lines under the different light treatments as seen in figure 29. The observation that, in the wild-type cells, the quantity of lipid droplets, as indicated by BODIPY fluorescence, increases in medium light and then decreases in high light tends to be erroneous because of the considerably high standard deviation values in the HL samples. Additionally, the wild-type samples under HL appeared to be more stressed than the medium light samples in the autofluorescence-based growth curves(Figures 28 and 27), thus making the BODIPY measurements questionable. Similar observations do not occur for the *Alb3b* mutants, However, high standard deviation levels can also be observed for many of the *Alb3b* samples. This is because of the presence of outliers, which were consequently imputed with the median values of samples before the ANOVA and post hoc tests. The post hoc tests conducted on the BODIPY measurements also yielded unexpected results(Figure 34 like no significant differences in the BODIPY fluorescence levels between the different light conditions for any of the mutants. Nevertheless, as expected from the mass spectrometry results, a significant difference was observed between WT and two of the mutant lines (*Alb3b*-14 and 16) under HL and ML treatment. Also, a significant difference is seen between WT cells under LL and the same under ML and HL, showing increased accumulation of lipids in the WT due to stress from irradiance.

One possible explanation for the unexpected BODIPY measurements and the high variance between the biological replicates is connected to the staining protocol used during the measurements. The optimum staining concentration of 0.067% according to Govender et al., 2012 was tested for other microalgal species, namely, *Chaetoceros calcitrans, Dunaliella primolecta, and Chlorella vulgaris*. It could be possible that this concentration is not ideal for effectively staining the LDs in *P.tricornutum*. Additionally, the stock solutions of BODIPY in DMSO had been thawed and re-frozen again multiple times, before the measurement, which appeared to affect the stain as was indicated by the color change in the frozen form of the solutions. The freshly frozen solutions were bright red, whereas the red color decreased considerably in the refrozen samples. Furthermore, the sensitivity issues associated with such fluorometric measurements such as issues like signal loss, necessitate the development of standard protocols by testing staining methods for a wide range of microalgal species.(Govender et al., 2012,Rumin et al., 2015)

Another interesting observation from figure 29 is that the difference in lipid quantities, as measured by BODIPY fluorescence, between the wild-type and the mutants is not comparable to the TAG measurements from mass spectrometry as seen in figures 17 and 14. For example, the TAG level in *Alb3b*-14 in LL is around 85% lower than that of wild type in mass spectrometry. In contrast, this difference in lipid level is 16% in the BODIPY measurements. The same comparison is also applicable to the *Alb3b*-16 mutants in LL, for which the decrease in TAG levels from mass spectrometry is a strong 92%, whereas the decrement in lipids compared to wild type is 22% in the BODIPY experiment, Furthermore, the steep spike in TAG level in *Alb3b*-14 in medium light compared to LL as observed from mass spectrometry cannot be observed in the BODIPY experiment. *Alb3b*-19 mutants also show contrasting results in the BODIPY fluorescence as compared to mass spectrometry, These cells in both low and medium light have shown significantly low TAG levels(98% and 68% lower than the wild type in the same conditions respectively) in mass spectrometry data. However, the lipid levels be 8% higher and 14% lower in LL and

medium light respectively compared to wild type in similar treatments in the BODIPY measurements. It must be noted as mentioned above that the high standard deviations in the measurements for BODIPY fluorescence make the validity of these observations uncertain.

The reduced difference between WT and the mutants in their Lipid content in the BODIPY measurements as compared to the mass spectrometry was unexpected and led the way to some hypotheses. It could be possible that the BODIPY 505/515 is not just binding to the TAG as this is not explicitly mentioned in the previous studies like Cooper et al., 2010 and Govender et al., 2012. They just indicate that the BODIPY 505/515 targets the LDs or neutral lipids in the cells and not any other organelles. This led to the inference that there are possibly other neutral lipids in the LDs of the observed cells. It is known that the neutral lipids generated under stress conditions as energy reserves do not solely include TAG but also other compounds like cholesterol and wax esters(Turkish and Sturley, 2009). Also, BODIPY staining has been previously used for labeling other neutral lipid molecules like Cholesterol, stearyl esters, and free fatty acids(Elle et al., 2010), thus indicating that they could bind to similar molecules in the stained cells' LDs. Furthermore, a previous study by Lupette et al., 2019 detected the presence of brassicasterol in the LDs isolated from *P.tricornutum*, thus reassuring the possibility of accumulation of sterols in LDs.

8.3.2 Photophysiology and growth

The chlorophyll measurements from flow cytometry, the growth curves based on auto-fluorescence from the plate reader, and the measurements of photosynthetic parameters from PAM can be linked together.

As observed in Figure 30, the chlorophyll levels decrease with increasing light in density and the mutants have lower chlorophyll levels than the wild-type. Additionally, How the decrease happens is different between the mutants and the wild-type. From LL to ML treatments the decrease in chlorophyll is 38% in the wild type whereas for the *Alb3b*-14 and 16 mutants, it is around 50%. But the same value for the *Alb3b*-19 mutants is just around 27%. The decrease in chlorophyll levels from ML to HL treatments is also around 38% for the wild type whereas the *Alb3b*-14 and 16 mutants show a steep decrease of 87 and 76% respectively. the same value for *Alb3b*-19 mutants is 42%. These values hold sensible when compared with the measurements of photosynthetic parameters from PAM. For instance, the decrement observed in Photosynthetic performance, measured as Fv/Fm values, is on par with the decrement in chlorophyll levels when compared between the wild type and the mutant cells. As explained in the results section and depicted in Figure 37, the decrease in Fv/Fm values Is around 3 to 4 times in the mutants compared to the wild type from LL to ML treatments and about 2 times from ML to HL treatments, based on the increased decrement in chlorophyll levels of mutants compared to the wild type. Similarly, when the light utilization efficiencies are compared between wild type and mutants under different light treatments as depicted in figure 39, the decrement in the light of utilization efficiency is around 2 times for the mutants compared to WT from LL to ML treatment, and around 2-3 times from ML to HL. One possible confusion that might arise in this case is the comparable levels of the Fv/Fm or α values between the mutants and the wild type in the same light conditions, especially in the LL treatments as shown by the post hoc tests(Figures 38 and 40). This is counterintuitive because of the lower chlorophyll levels in the mutants compared to the wild type in the same light conditions Which should probably lead to lower photosynthetic performance in the mutants. this is accounted for in the previous study by Nymark et al., 2019, wherein the photosynthetic parameters were measured again using oxygen

evolution from photosynthesis, and subsequently normalized to chlorophyll a value from HPLC, which is a standard step in photosynthetic measurements(Consalvey et al., 2005). These repeated measurements led to the conclusion that the mutants had truncated antennae for light harvesting with higher maximum photosynthetic rate(P_{max}) and light saturation coefficient(E_s), and lower maximum utilization efficiency or α, similar to the observations in a previous study on cyanobacteria by Kirst et al., 2014.

The implications of these reduced chlorophyll levels and consequent decrease in photosynthetic performance with increasing light conditions are reflected in the growth curves measured by autofluorescence emissions using the plate reader. In the figure 27, it can be seen that the wild type reaches the stationary phase at around the fifth day of growth with the relative fluorescence units reaching a maximum of around 950 under LL treatment. However, the mutants take around 10 to 11 days to attain the stationary phase with comparatively lower RFU values ranging between 400 to 650(Figure 28). Furthermore, under ML treatment the WT cells reach the stationary phase at around 6 days, whereas the *Alb3b*-14 and 16 mutants do this at about the 9th to 10th day. The *Alb3b*-19 reached the stationary phase around the 6th day, the same as the wild type. This is possibly an outlier because of contamination in the *Alb3b*-19 cultures with the WT cells, as the post hoc analysis indicated significantly.

These growth rates are based on auto-fluorescence and do not indicate the actual cell division rate. Additionally, the varying chlorophyll levels in different light conditions further make these measurements unfit for comparing growth rates between mutants and WT, unlike using other methods like Bürker-Türk counting chamber or flow cytometry.

The decreased chlorophyll levels, as indicated by the reduced level of auto-fluorescence in both flow cytometry and the plate reader, can be explained by the results from Nymark et al., 2009. This study found downregulation of the Light-harvesting complex proteins during all the phases in which the WT cells acclimated to HL treatment of 500 µmol photons m^{-2} s^{-1}. This also appeared to have an impact on the concentrations Chl*a*, Chl*c*, and Fucoxanthin(F*x*)(Nymark et al., 2009), which are some of the major light-harvesting pigments(LHPs) in diatoms cells(Brown, 1988). The concentration of Fx decreased as an immediate response to HL exposure whereas the Chl*a* and Chl*c* levels decreased only in the late acclimation phases to HL(Nymark et al., 2009). The reduction in Fx was also evident during this lab work from the visual properties of the cultures. All the cultures, including the mutant lines and WT, showed changing coloration with increasing light intensity. The WT were more brown colored under LL but became more green in ML, and light green in HL. The *Alb3b* mutants were green colored in LL as expected from the results from Nymark et al., 2019, and became increasingly light green colored under ML and HL. The normal golden brown coloration observed for WT cells is because of the high amounts of Fx in the diatom fucoxanthin chlorophyll a/c-binding protein complexes(FCP complex)(Gundermann and Büchel, 2014). Additionally, it is known that Fx changes its spectral properties in terms of absorption and emission wavelengths in response to protein binding(Premvardhan et al., 2009). Therefore the fact that the *Alb3b* mutants are green-colored compared to WT and the decreasing pigmentation of all cell lines with increasing light can be explained by the changed FCP complex concentrations in the thylakoid membrane due to the mutation in the ALBINO3B insertase and the downregulation of LHPs, respectively.

The significant increases observed in the maximum electron transport rate(Figures 41,42) and saturation coefficients(Figure 43) from LL to ML in all the cell lines can be explained by the results from Nymark et al., 2009. According to this, the increased electron flow and the increased threshold to reach photosynthetic saturation are consequences of the photo-acclimation mechanisms adopted by the cells

in response to high light and the subsequent ability to capture and utilize the increased available light energy. A notable observation when comparing the mutants and the WT is that the mutants have comparatively higher ETR and Ek values than the WT. The $rETR_{max}$ values are 32, 35, and 9 % higher in the *Alb3b*-14,16,19 mutants respectively compared to the WT in LL, and 37, 21, and 10 % respectively under ML. As explained before this shouldn't be misinterpreted as increased photosynthetic performance of the mutants compared to WT as the repeated measurements using oxygen evolution and normalization to Chla proves otherwise according to Nymark et al., 2019. However, under HL the $rETR^{max}$ values reduce in the *Alb3b*-14 and 16 mutants with the decrease being significant for *Alb3b*-14(Figures 41,42). The same parameter for the *Alb3b*-19 mutant and the WT increases under HL but not significantly(Figures 41,42). This can be interpreted as a result of photo-inhibition in the *Alb3b*-14 and 16 mutants as the Chlorophyll levels decreased drastically for them in HL compared to *Alb3b*-19 and the WT (Figure 30), According to Adir et al., 2003 the electron transport associated with photosynthesis will be drastically affected during photo-inhibition as the rate of light-induced damage exceeds the rate of repair of the PS II reaction centers. The increased light saturation coefficient in the mutants ranged from 26 to 33% from that of the WT in LL(Figure 43). In ML, the increase was 55, 38, and 13% for the *Alb3b*-14,16 and 19 respectively compared to WT. In HL, this gap is diminished greatly to around 2-9% for the mutants compared to WT. Also, as explained in the results section and depicted in figure 43, there is a general trend of increasing Ek values with increasing irradiance in each cell line. This could again be linked to the reduced changing pigmentation or LHPs in the cells. The reduced LHPs with increasing light irradiance due to the downregulation of LHC proteins as described above or due to the truncated antennae in the mutants leads to the requirement of higher light intensities to reach the saturation points as was observed for similar mutants of cyanobacteria generated in previous studies by Kirst et al., 2014 and green algae by Polle et al., 2003.

8.3.3 Connection of Lipid profile to photophysiology

Linking these results of the variations in photosynthetic pigments, electron transport rate and photo-inhibition to the lipid profile of the cells is challenging. It is known that the lipids, especially the membrane lipids interact with the integral membrane proteins and affect the functions of each other(A. G. Lee, 2004). The LHPs like chlorophyll and carotenoids are folded along with the integral membrane proteins called the Light-harvesting complexes(Natali et al., 2014 and a previous study has shown that these LHCs interact with one of the major constituent lipids in the thylakoid membrane, that is MGDG, thereby affecting each other functioning(Simidjiev et al., 2000). For instance, MGDG is known to affect the formation and maintenance of PS II dimers in the thylakoid membrane(Kern and Guskov, 2011). Since significant changes were observed in the total amounts and fatty acid compositions of the membrane lipids including MGDG, between different light treatments, from the MS-MS data analysis, it might have played a role in changing the pigment concentrations or compositions in the thylakoid membranes as suggested by the Flow cytometry, Autofluorescence, and PAM experiments in this study and the previous one by Nymark et al., 2019 as discussed above. The changes in these lipids between different light treatments were also different between the mutant and WT, which might be the reason for the variation in how the pigmentation parameters change between them.

Another possible connection is between the membrane fluidity and the electron transport rate. The fatty acid compositions of the lipids will affect the membrane's fluidity (Katarzyna and Wydro, 2007).

Generally, saturated fatty acids are known to reduce the fluidity or increase the rigidity of the membrane and unsaturated fatty acids do the opposite(Katarzyna and Wydro, 2007). The maintenance of high fluidity in thylakoid membranes by the accumulation of PUFAs like the ω 3 FAs, EPA, and DHA, can increase the electron flow rate through these membranes(Mock and Kroon, 2002a,Guihéneuf et al., 2009). The change in the fatty acid compositions in the MGDG fraction is discussed in detail in section 8.2 based on results from the MS-MS data analysis. The increase in Electron transport rate in the WT under ML compared to LL can be explained partly by the significant increase in the fraction of many of the EPA-containing MGDG compositions(Figure 10b). However, the increase in electron transport rate in the mutants in ML than the LL conditions, at a level generally higher than that of WT, is not in accord with the fact that most of the EPA-containing MGDG compositions reduced significantly in all the mutants under Ml (Figures 7b, 8b,9b) and require further investigation into other mechanisms that could increase the electron flow rate through the membranes. In terms of photoacclimation and photoprotection, the mutants appeared to perform better than the WT as indicated by the generally higher levels of NPQ as detailed in the results section 7.3.3 and depicted in figure 44, additionally the NPQ levels were also shown to be significantly affected by Cell-type according to the ANOVA analysis. One explanation for this could be the truncated light-harvesting antennae in the mutants(Nymark et al., 2019), which leads to lower absorption of light energy and thus reduced photodamage.

Lipids also play a role in the photoacclimation of the cells. One way to link the photo-acclimation to the lipids is the role of MGDG in acting as a solvent for the xanthophyll cycle pigments, diatoxanthin, and diadinoxanthin, thus increasing their accessibility in the de-epoxidation reaction essential in photo-protection(Goss et al., 2005). It has also been shown that the MGDG forms a non-bilayer region that facilitates the conversion reaction of these photo-protective pigments(Latowski et al., 2002). Since the changes observed in MGDG levels in the mutants and WT under different light conditions are found to be different, there might be implications of this in the differences observed in the photo-protective properties of the mutants and the WT. Additionally, free fatty acids, particularly, long chain-saturated FAs, like palmitic and stearic acids improve the rates of PS II repair mechanism by accelerating the D1 protein synthesis(Jimbo et al., 2020). This supports the hypothesis of accumulation of neutral lipids other than TAG, like FFAs, in the Lipid droplet fraction of the cells as explained before in this section.

8.3.4 Changes in cell and LD morphology

Notable differences were also observed in the cell morphology between the WT and mutants in both flow cytometry parameters(FSC and SSC) and CLSM. However, the reason for this change is not understood.

As explained in the results section 7.3.2 and depicted in the figures 31 and 32 there are conspicuous differences in forward scattering and side scattering respectively, between the WT and mutants and also among the mutants. This means there are differences in the cell sizes and granularity and how they change under different light treatments. Generally, the mutants were observed to be significantly bigger and more granular(Figures 31,36a & Figures 32,36b, respectively) than the mutants in LL and HL except for the *Alb3b*-19 mutants in LL being smaller and less granular than the WT in LL. significant differences in cell size and granularity were also observed within each cell line with changing light conditions(Figure 36). These changes are also evident in the CLSM images(Figures 48, 47, 46). The cell sizes in these images cannot be compared as the processing of images resulted in varying levels of zoom in and out to adjust the noises and to remove the background. However, bigger or longer cell sizes were observed in the

mutants compared to the WT. This is consistent with the FSC values observed in figure 31. The same inference holds for the cell granularity as the SSC values are generally higher in the mutants than in WT and the CLSM images show more dispersed LDs in the mutant cell compared to WT. The high granularity could also be due to components other than LDs. The fact that the very high side scattering observed in LL treatments for *Alb3b*-14 and 16 is not observed as highly dispersed LDs in the corresponding CLSM images(Figures 46b, 46c) supports this notion. One interesting observation was the odd-shaped cells in the *Alb3b*-19 mutant line in ML and Hl treatments under CLSM, wherein these cells appeared considerably elongated with one end being round and the other tapered. Some of these cells were also found to touch tips. Additionally, the same mutants also had many cells with round-shaped morphotypes, the majority of which appeared in clusters(Supplementary figure 19). These observations were not made for the other two mutants and the WT, which expressed the normal fusiform shape. This could explain the difference in the behavior of the *Alb3b*-19 mutants from the other mutants and WT in FSC and SSC measurements. The round morphotype has been previously reported to have occurred in the WT cells of *P.tricornutum* under various abiotic stress conditions like salinity and temperature(De Martino et al., 2011), However, this was only a consequence of a long-term or chronic exposure to such conditions. Additionally, this particular morphotype was also observed in a previous light stress experiment by Herbstová et al., 2017, wherein chronic acclimation to ambient modified light conditions with an enhanced red region of the spectrum, resulted in the same. Although this behavior can thus be treated as a normal stress response by the cell to abiotic stressors including light(De Martino et al., 2011,Herbstová et al., 2017), the reason why *Alb3b*-19 cells do this in a considerably short acclimation time still need explanation. The reason for the change in the Lipid droplet morphology between mutants and WT could be linked to the differences observed in phospholipids and their fatty acid compositions. The LDs generally have a hydrophobic core containing mostly the accumulation of the neutral lipid, TAG and this is surrounded by a mono-layer of polar amphipathic lipids like phospholipids and glycolipids(Murphy, 2001). The study by Lupette et al., 2019 indicates that the composition of the LD mono-layer in *P.tricornutum* consists of PC,DGTA and SQDG. Although the change in DGTA and SQDG between light treatment is not different between the mutants and WT, that of PC is different as observed in figures 22, 23, 24. As explained in the results section 7.2 and depicted in these figures, the PC significantly increased in the WT, remained almost the same in *Alb3b*-14, and significantly decreased in *Alb3b*-16 and 19 in ML compared to LL treatment. Additionally, as already mentioned multiple times in earlier parts of this section, the changes in fatty acid composition between light treatments are different between WT and mutants (Figures 10a,7a,8a9a). Therefore, this could alter the LD mono-layer and thus the overall LD morphology differently for the WT and mutants.

8.3.5 Differential gene expression in lipid metabolism

The differential expression of selected genes involved in lipid metabolism was studied using q-PCR for this study. As explained in the results section 7.3.5, differences were found in the expression of certain enzymes involved in the phospholipid metabolism and acyl chain elongation cycle. It should be noted that these results are greatly uncertain owing to the high variability among the biological replicates and the contamination of cDNA samples with genomic DNA. Additionally, data from three biological replicates do not give enough statistical power to be conclusive about the observed up- or downregulations.

The upregulation of the enzymes, PLC delta(Phospholipase C isoform delta: PHATRDRAFT 42683)

and the CDS1(Phosphatidate cytidiltransferase: PHATRDRAFT 54756) under HL treatment compared to WT in LL, indicates the possible occurrence of phospholipid remodeling. PLC, also called PIPLC is a membrane-bound enzyme that is involved in hydrolyzing PI in the membrane bi-layers to form the neutral lipid, DAG(Lyon and Tesmer, 2013,Tanaka et al., 2022). CDS1 is involved in synthesizing Cytidyl diphosphate diacyl glycerol(CDP-DAG) from Phosphatidic acid(Mishra et al., 2017). The CDP-DAG will be in turn used to synthesize phospholipids like PS, PE, and PI(Tanaka et al., 2022). One possibility is that the cells are remodeling the PI from the membranes, and the consequent reduction in the PI fraction of the membranes is compensated by increased synthesis of other phospholipids from CDP-DAG. However, this was not observed in any of the mutants as expected from the results of MS-MS data analysis. This was unexpected because the MS-MS data analysis indicated that the WT in ML had significantly higher levels of all detected phospholipids compared to LL and the mutants had significantly lower levels of most phospholipids in ML compared to LL, as seen in figures 22 23, 24, there is a significant decrease in most of the phospholipids.. However, no significant differential expressions are observed for the studied phospholipid metabolism genes in the mutants acclimated to ML and HL conditions(Figures 50,51, and 52). Also, the upregulations observed for the same genes were not significant in the WT in ML. As mentioned in section 8.2, the study by Ding et al., 2023 gave indications of phospholipid remodeling in WT by the upregulation of enzymes like PDAT at a light level of 300 µmol photons m^{-2} s^{-1} for 3 days, which was also reversed when the light stress was removed. Therefore, it could be possible that there was a phospholipid remodeling response initially at a light level of 200 µmol photons m^{-2} s^{-1} and then a subsequent reversal of the same as the cell got acclimated. However, the same WT cells at HL couldn't reverse this response probably due to reduced ability to acclimate to a higher stress level. The observation that the mutants acclimated to HL and ML didn't show significant differential expression compared to the same cell lines in LL and WT in the same light treatments might indicate their higher photo-acclimation abilities as mentioned before in this section.

In addition to the enzymes mentioned above the mutants and WT showed a difference in the differential expression of two acyl chain elongation cycle enzymes, FADB(Malonyl-coA-ACP transacylase: PHATRDRAFT 37652) and FABI(Enoyl-ACP reductase: PHATRDRAFT 10068). FADB is crucial in fatty acid synthesis as it catalyzes the transfer of the 2 C atom donor, Malonyl-coA, to ACP, thus participating in the actual FA elongation process(Zhang et al., 2007,Tanaka et al., 2022). The FABI enzyme is also important as it catalyzes the final step in the elongation cycle by reducing the trans-enoyl-ACP into Acyl-ACP, the substrate for thioesterases to make FFAs(Massengo-Tiassé and Cronan, 2009,Tanaka et al., 2022). These enzymes were found to be significantly downregulated in the *Alb3b*-14 and 16 mutant lines in ML and for *Alb3b*-19 in HL compared to the LL conditions for the same cell lines(Figures 50,51, and 52). Considerable downregulation of the same enzymes can be also observed in HL conditions for the *Alb3b*-14 and 16, but these are not deemed significant in ANOVA. This strong deregulation was not found in both ML and HL conditions. These results are comparable with the results from analysis of the fatty acid composition data from MS-MS. In the figures 25 and 26, in section 7.2, it can be seen that mutants under ML generally had a significantly high amount of medium-chain saturated Fatty acids and low amount of long-chain saturated fatty acids. In contrast, the WT indicates an opposite scenario. This leads to the interpretation that the acyl-chain elongation might be affected in the mutants as supported by the observed downregulation of FABI and FADB.

The observed differences in stress responses in the lipid metabolism between the WT and mutants are

probably because of the compounded stress created by the mutation itself over the light stress. microalgae, among other autotrophs, are known for their physiological flexibility to adjust their metabolisms in response to diverse and extreme stress conditions to acclimate to their unpredictable natural surroundings(Gorelova et al., 2019). One major stress response among these is the alteration in the composition of pigments and functioning of the photosynthetic machinery(Gorelova et al., 2019). As these properties were already found to be different in the *Alb3b* mutants by Nymark et al., 2019, the change in the corresponding stress responses is not surprising. The numerous and complex interconnections and cross-talks between metabolic pathways within these cells, including those associated with photosynthesis, photoprotection, pigment, and lipid metabolism make the stress responses observed in these different cellular processes linked to each other. For instance, it is known that the molecular mechanism that triggers TAG accumulation upon stress is brought about by the reduced cell growth or division rate as a result of the reduced photosynthesis(Teh et al., 2021). Additionally, A link between pigment and Lipid metabolism reported as a coordinated synthesis of carotenoids and TAG has been found in previous studies by Whitelam and Codd, 1986 and Solovchenko et al., 2010. TAG synthesis and its accumulation in LDs can be directly linked to light stress by the fact that they act as a sink and reservoir of the excess energy produced through excessive photosynthesis. Additionally, the photo-protective role of the accumulated LDs under light stress has been previously reported where it acts as a quencher of excessive light to avoid photo-damage of the chloroplasts. These connections lead to the inference that the changes already observed in the photosynthetic or photo-protective properties of the mutants compared to WT, in the previous study, could also be the reason for the observed differences in the Lipid profile in this study. However, the differences observed among the different mutant lines are inconclusive. One plausible explanation is off-target mutations caused in other genes of these mutant lines while they were developed by knocking out the ALBINO3B genes. Proof for the occurrence of such off-target mutations was observed during q-PCR for the *Alb3b*-19 mutant's fatty acid desaturase gene, Phatr50443, which gave a significant down-regulated expression profile (Figures 53,54, and 55)separate melting peak from the WT and other mutant lines as seen in figure 56.

9 Conclusion

The results from the lipidomics data analysis pipeline developed for this project, and the lab work that followed it. and the discussed interpretations and predictions based on the literature search led to the conclusion that there are differences in the stress responses within lipid and fatty acid metabolism between the *Alb3b* mutants and the WT and among the different *Alb3b* mutant lines towards light stress.

The main conclusions made are as follows:

- There are differences in the way in which lipid profile changes between the WT and mutant lines, and also among the mutant lines when compared between LL and ML acclimated samples. The difference between WT and mutants is generally associated with the phospholipid and glycolipids profile changes between the light treatments as summarized in table 4. This suggests a difference in the membrane lipid remodeling response between the WT and mutants.
- It can be hypothesized that there is a possible accumulation of other neutral lipid molecules, like sterols and FFAs in the LDs of mutants unless the BODIPY staining protocol used for the LD

analyses is non-optimal.

- The mutants have altered photophysiological components compared to WT as previously demonstrated by Nymark et al., 2019, especially due to their truncated light-harvesting antennae. This mainly includes reduced chlorophyll levels and a difference in how the photophysiological parameters like Fv/Fm, Ek, and rETRmax change between the WT and mutant lines when compared between different light treatments. Also, the mutants generally have higher photoprotection and photoacclimation properties. It is hypothesized that the changes observed in the lipid profile can be contributing to these differences to some extent.

- There are visible changes in the cell and LD morphology between different light treatments and between the different cell lines. These differences observed between the cell lines are possibly due to the differences in their phospholipid profile changes between light treatments.

The inferences from the Differential gene expression analysis in section 8.3.5 cannot be used for drawing solid conclusions because of the high variability and low number of replicates, which yields insufficient statistical power for the results.

The analysis of the performance of the pipeline based on comparisons with regression plots and heat maps led to the conclusion that there are some challenges associated with dealing with complex metabolomics data like the MS-MS dataset used in this project and some potential pitfalls within the pipeline that could lead to uncertain results or misinterpretation. However, the pipeline succeeded in providing an overview of the entire data, with the majority of results being comparable to results from previous literature.

To conclude, the differences among cell lines in their lipid profile and associated aspects include changes within the actual concentrations of the various lipid classes, their different fatty acid compositions, and the structural characteristics of the LDs within the cells. The variations between the cell lines in their photosynthetic and photo-protective parameters like the light utilization efficiencies, electron transport rate, and non-photochemical quenching measured using PAM, along with the variations in their autofluorescence characteristics measured using flow cytometry and the plate reader, were able to be connected to these differences in lipid metabolic responses to some extent. This concludes the existence of connections or cross-talks between the Pigment composition, photosynthetic apparatus, photo-protective pathways, cell growth and division, TAG accumulation, lipid chemistry of the various membranes like the thylakoid membrane, and LD mono-layer, and the FA compositions within different lipid classes as reported previously in individual research projects.

10 Future research

The prospective future works to progress the findings from this project can include both attempts to improve the developed pipeline and further improve the knowledge about the *Alb3b*-mutant lines. This can include:

- **Finding a better way to impute the outliers, especially the null values in the dataset, rather than using 0.01% of the total lipid concentration:** One suggested approach is to use machine learning algorithms like the Random Forest or K-nearest neighbors as in the study

by Kokla et al., 2019. The performance of these imputation techniques can then be assessed and compared to the technique used in this study by measuring the Normalized Root Mean Squared Error (NRMSE) as mentioned in Kokla et al., 2019. It would also be interesting to see the exact nature of these missing values in the dataset and compare them with imputation techniques.

- **Studying the variations in the membrane structures in detail using transmission electron microscopy(TEM):** This can be done to build on the predictions of membrane lipid remodeling made in this study and the actual observation in structural changes in thylakoid membrane observed by Nymark et al., 2019. The changes in the membrane lipids involved in LD architecture are a reason to study the LD mono-layer in detail using TEM. Also, the possible differences in lipid class compositions in the membranes between the WT and mutants suggest using electron dispersive spectroscopy or electron energy loss spectroscopy in integration with TEM to detect differences in membrane elemental compositions.

- **DNA sequencing and RNAseq of the mutants to detect off-target mutations and detailed differential gene expression analysis:** the inference of possible mutations in genes other than the ALBINO3B in the different mutant lines can give more insights into the peculiar behavior of mutants like the formation of round morphotype and clustering, more rapidly than WT, in the *Alb3b*-19 mutants observed in this study and the increased NPQ values in the mutants. Additionally, performing RNAseq could yield results of differential regulation of genes in different metabolic pathways associated with light stress responses, and could explain changes in lipid metabolism in more detail than what is understood from q-PCR.

References

Abida, H., Dolch, L.-J., Meı, C., Villanova, V., Conte, M., Block, M. A., Finazzi, G., Bastien, O., Tirichine, L., Bowler, C., et al. (2015). Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in phaeodactylum tricornutum. *Plant physiology*, *167*(1), 118–136.

Adir, N., Zer, H., Shochat, S., & Ohad, I. (2003). Photoinhibition–a historical perspective. *Photosynthesis research*, *76*, 343–370.

Apt, K. E., Zaslavkaia, L., Lippmeier, J. C., Lang, M., Kilian, O., Wetherbee, R., Grossman, A. R., & Kroth, P. G. (2002). In vivo characterization of diatom multipartite plastid targeting signals. *Journal of Cell Science*, *115*(21), 4061–4069.

Arao, T., & Yamada, M. (1994). Biosynthesis of polyunsaturated fatty acids in the marine diatom, phaeodactylum tricornutum. *Phytochemistry*, *35*(5), 1177–1181.

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, *459*(7244), 185–192. https://doi.org/10.1038/nature08057

Bozarth, A., Maier, U.-G., & Zauner, S. (2009). Diatoms in biotechnology: modern tools and applications. *Applied Microbiology and Biotechnology*, *82*(2), 195–201. https://doi.org/10.1007/s00253-008-1804-8

Brown, J. S. (1988). Photosynthetic pigment organization in diatoms (bacillariophyceae) 1. *Journal of phycology*, *24*(1), 96–102.

Bustin, S. A., Beneš, V., Garson, J. A., Hellemans, J., Huggett, J. F., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE Guidelines: Minimum information for publication of Quantitative Real-Time PCR experiments. *Clinical chemistry (Baltimore, Md.)*, *55*(4), 611–622. https://doi.org/10.1373/clinchem.2008.112797

Canette, A., & Briandet, R. (2014, January). *MICROSCOPY — Confocal Laser Scanning Microscopy*. https://doi.org/10.1016/b978-0-12-384730-0.00214-7

Consalvey, M., Perkins, R., Paterson, D. M., & Underwood, G. J. C. (2005). PAM FLUORESCENCE: A BEGINNERS GUIDE FOR BENTHIC DIATOMISTS. *Diatom Research*, *20*(1), 1–22. https://doi.org/10.1080/0269249x.2005.9705619

Cooper, M. S., Hardin, W. R., Petersen, T. W., & Cattolico, R. A. (2010). Visualizing" green oil" in live algal cells. *Journal of bioscience and bioengineering*, *109*(2), 198–201.

Dan Margalit, J. R. (n.d.). Interactive Linear Algebra. https://textbooks.math.gatech.edu/ila/eigenvectors.html

David T. Harvey, B. A. H. (n.d.). Understanding Scores and Loadings. https://bryanhanson.github.io/LearnPCA/articles//Vig_04_Scores_Loadings.html

De Martino, A., Bartual, A., Willis, A., Meichenin, A., Villazán, B., Maheswari, U., & Bowler, C. (2011). Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom phaeodactylum tricornutum. *Protist*, *162*(3), 462–481.

Ding, W., Ye, Y., Yu, L., Liu, M., & Liu, J. (2023). Physiochemical and molecular responses of the diatom Phaeodactylum tricornutum to illumination transitions. *Biotechnology for biofuels and bioproducts*, *16*(1). https://doi.org/10.1186/s13068-023-02352-w

Elle, I. C., Olsen, L. C. B., Pultz, D., Rødkær, S. V., & Færgeman, N. J. (2010). Something worth dyeing for: Molecular tools for the dissection of lipid metabolism in caenorhabditis elegans. *FEBS letters*, *584*(11), 2183–2193.

Elliott, A. D. (2019). Confocal Microscopy: Principles and Modern Practices. *PubMed Central*, *92*(1). https://doi.org/10.1002/cpcy.68

García-Alegría, A. M., Anduro-Corona, I., Pérez-Martínez, C. J., Guadalupe Corella-Madueño, M. A., Rascón-Durán, M. L., Astiazaran-Garcia, H., et al. (2020). Quantification of dna through the nanodrop spectrophotometer: Methodological validation using standard reference material and sprague dawley rat and human dna. *International journal of analytical chemistry*, *2020*.

García-Plazaola, J. I., Fernández-Marín, B., Duke, S. O., Hernández, A., López-Arbeloa, F., & Becerril, J. M. (2015). Autofluorescence: Biological functions and technical applications. *Plant Science*, *236*, 136–145.

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, *24*(3). https://doi.org/10.1214/09-sts301

Ge, F., Huang, W., Chen, Z., Zhang, C., Xiong, Q., Bowler, C., Yang, J., Xu, J., & Hu, H. (2014). Methylcrotonyl-CoA Carboxylase Regulates Triacylglycerol Accumulation in the Model Diatom Phaeodactylum tricornutum. *The Plant Cell*, *26*(4), 1681–1697. https://doi.org/10.1105/tpc.114.124982

Gorelova, O., Baulina, O., Ismagulova, T., Kokabi, K., Lobakova, E., Selyakh, I., Semenova, L., Chivkunova, O., Karpova, O., Scherbakov, P., et al. (2019). Stress-induced changes in the ultrastructure of the photosynthetic apparatus of green microalgae. *Protoplasma*, *256*, 261–277.

Goss, R., Lohr, M., Latowski, D., Grzyb, J., Vieler, A., Wilhelm, C., & Strzalka, K. (2005). Role of hexagonal structure-forming lipids in diadinoxanthin and violaxanthin solubilization and de-epoxidation. *Biochemistry*, *44*(10), 4028–4036.

Govender, T., Ramanna, L., Rawat, I., & Bux, F. (2012). Bodipy staining, an alternative to the nile red fluorescence method for the evaluation of intracellular lipids in microalgae. *Bioresource technology*, *114*, 507–511.

Guihéneuf, F., Mimouni, V., Ulmann, L., & Tremblin, G. (2009). Combined effects of irradiance level and carbon source on fatty acid and lipid class composition in the microalga pavlova lutheri commonly used in mariculture. *Journal of Experimental Marine Biology and Ecology*, *369*(2), 136–143.

Guihéneuf, F., Mimouni, V., Ulmann, L., & Tremblin, G. (2008). Environmental factors affecting growth and omega 3 fatty acid composition in skeletonema costatum. the influences of irradiance and carbon source: Communication presented at the 25ème congrès annuel de l'association des diatomistes de langue francaise (adlaf), caen, 25-28 september 2006. *Diatom Research*, *23*(1), 93–103.

Gundermann, K., & Büchel, C. (2014). Structure and functional heterogeneity of fucoxanthin-chlorophyll proteins in diatoms. In *The structural basis of biological energy generation* (pp. 21–37). Springer.

Guo, R., Chen, Y., Borgard, H., Jijiwa, M., Nasu, M., He, M., & Deng, Y. (2020). The function and mechanism of lipid molecules and their roles in the diagnosis and prognosis of breast cancer. *Molecules*, *25*(20), 4864.

Hao, X., Luo, L., Jouhet, J., Rébeillé, F., Maréchal, E., Hu, H., Pan, Y., Tan, X., Chen, Z., You, L., et al. (2018). Enhanced triacylglycerol production in the diatom phaeodactylum tricornutum by inactivation of a hotdog-fold thioesterase gene using talen-based targeted mutagenesis. *Biotechnology for biofuels*, *11*, 1–18.

Haslam, R. P., Hamilton, M. L., Economou, C. K., Smith, R., Hassall, K. L., Napier, J. A., & Sayanova, O. (2020). Overexpression of an endogenous type 2 diacylglycerol acyltransferase in the marine diatom phaeodactylum tricornutum enhances lipid production and omega-3 long-chain polyunsaturated fatty acid content. *Biotechnology for biofuels*, *13*, 1–17.

Herbstová, M., Bína, D., Kaňa, R., Vácha, F., & Litvín, R. (2017). Red-light phenotype in a marine diatom involves a specialized oligomeric red-shifted antenna and altered cell morphology. *Scientific Reports*, *7*(1), 11976.

Jahns, P., & Holzwarth, A. R. (2012). The role of the xanthophyll cycle and of lutein in photoprotection of photosystem II. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1817*(1), 182–193. https://doi.org/10.1016/j.bbabio.2011.04.012

Jimbo, H., Takagi, K., Hirashima, T., Nishiyama, Y., & Wada, H. (2020). Long-chain saturated fatty acids, palmitic and stearic acids, enhance the repair of photosystem ii. *International Journal of Molecular Sciences*, *21*(20), 7509.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Katarzyna & Wydro. (2007). The influence of fatty acids on model cholesterol/phospholipid membranes. *Chemistry and physics of lipids*, *150*(1), 66–81.

Kern, J., & Guskov, A. (2011). Lipids in photosystem ii: Multifunctional cofactors. *Journal of Photochemistry and Photobiology B: Biology*, *104*(1-2), 19–34.

Kirst, H., Formighieri, C., & Melis, A. (2014). Maximizing photosynthetic efficiency and culture productivity in cyanobacteria upon minimizing the phycobilisome light-harvesting antenna size. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, *1837*(10), 1653–1664.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: A comparative study. *BMC bioinformatics*, *20*, 1–11.

Králík, P., & Ricchi, M. (2017). A basic guide to real time PCR in microbial diagnostics: definitions, parameters, and everything. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00108

Kuang, J., Yan, X., Genders, A. J., Granata, C., & Bishop, D. J. (2018). An overview of technical considerations when using quantitative real-time pcr analysis of gene expression in human exercise research. *PloS one*, *13*(5), e0196438.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., et al. (2006). The real-time polymerase chain reaction. *Molecular aspects of medicine*, *27*(2-3), 95–125.

Latowski, D., Kruk, J., Burda, K., Skrzynecka-Jaskier, M., Kostecka-Gugała, A., & Strzałka, K. (2002). Kinetics of violaxanthin de-epoxidation by violaxanthin de-epoxidase, a xanthophyll cycle en-

zyme, is regulated by membrane fluidity in model lipid bilayers. *European Journal of Biochemistry*, *269*(18), 4656–4665.

Lebeau, T., & Robert, J.-M. (2003). Diatom cultivation and biotechnologically relevant products. Part II: Current and putative products. *Applied Microbiology and Biotechnology*, *60*(6), 624–632. https://doi.org/10.1007/s00253-002-1177-3

Lee, A. G. (2004). How lipids affect the activities of integral membrane proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1666*(1-2), 62–87.

Lee, J. M., Lee, H., Kang, S., & Park, W. J. (2016). Fatty acid desaturases, polyunsaturated fatty acid regulation, and biotechnological advances. *Nutrients*, *8*(1), 23.

Lepetit, B., Goss, R., Jakob, T., & Wilhelm, C. (2011). Molecular dynamics of the diatom thylakoid membrane under different light conditions. *Photosynthesis research*, *111*(1-2), 245–257. https://doi.org/10.1007/s11120-011-9633-5

Levitan, O., Dinamarca, J., Zelzion, E., Lun, D. S., Guerra, L. T., Kim, M. K., Kim, J., Van Mooy, B. A., Bhattacharya, D., & Falkowski, P. G. (2015). Remodeling of intermediate metabolism in the diatom phaeodactylum tricornutum under nitrogen stress. *Proceedings of the National Academy of Sciences*, *112*(2), 412–417.

Libretexts. (2023, March). 7.1: Eigenvalues and Eigenvectors of a Matrix. https://math.libretexts.org/Bookshelves/Linear_Algebra/A_First_Course_in_Linear_Algebra_(Kuttler)/07%3A_Spectral_Theory/7.01%3A_Eigenvalues_and_Eigenvectors_of_a_Matrix

Lu, Z. L., & Yuan, K.-H. (2010). Welch's t test. *ResearchGate*. https://doi.org/10.13140/RG.2.1.3057.9607

Lupette, J., Jaussaud, A., Seddiki, K., Morabito, C., Brugière, S., Schaller, H., Kuntz, M., Putaux, J.-L., Jouneau, P.-H., Rébeillé, F., et al. (2019). The architecture of lipid droplets in the diatom phaeodactylum tricornutum. *Algal Research*, *38*, 101415.

Lyon, A. M., & Tesmer, J. J. (2013). Structural insights into phospholipase c-β function. *Molecular pharmacology*, *84*(4), 488–500.

Maeda, Y., Nojima, D., Yoshino, T., & Tanaka, T. (2017). Structure and properties of oil bodies in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160408.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., De Vargas, C., Bittner, L., Zingone, A., & Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(11). https://doi.org/10.1073/pnas.1509523113

Manning, S. R. (2022). Microalgal lipids: Biochemistry and biotechnology. *Current Opinion in Biotechnology*, *74*, 1–7.

Maréchal, E., & Lupette, J. (2020). Relationship between acyl-lipid and sterol metabolisms in diatoms. *Biochimie*, *169*, 3–11.

Massengo-Tiassé, R. P., & Cronan, J. E. (2009). Diversity in enoyl-acyl carrier protein reductases. *Cellular and molecular life sciences*, *66*, 1507–1517.

McKinnon, K. (2018). Flow Cytometry: An Overview. *Current protocols in immunology*, *120*(1). https://doi.org/10.1002/cpim.40

Medium, B. (2018). How to read PCA biplots and scree Plots - BioTuring Team - Medium. https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063

Mishra, N. N., Tran, T. T., Seepersaud, R., García-De-La-Mària, C., Faull, K., Yoon, A., Proctor, R., Miro, J. M., Rybak, M. J., Bayer, A. S., et al. (2017). Perturbations of phosphatidate cytidylyltransferase (cdsa) mediate daptomycin resistance in streptococcus mitis/oralis by a novel mechanism. *Antimicrobial agents and chemotherapy*, *61*(4), 10–1128.

Mock, T., & Kroon, B. M. (2002a). Photosynthetic energy conversion under extreme conditions—i: Important role of lipids as structural modulators and energy sink under n-limited growth in antarctic sea ice diatoms. *Phytochemistry*, *61*(1), 41–51.

Mock, T., & Kroon, B. M. (2002b). Photosynthetic energy conversion under extreme conditions—ii: The significance of lipids under light limited growth in antarctic sea ice diatoms. *Phytochemistry*, *61*(1), 53–60.

Montecillo-Aguado, M., Tirado-Rodriguez, B., & Huerta-Yepez, S. (2023). The involvement of polyunsaturated fatty acids in apoptosis mechanisms and their implications in cancer. *International Journal of Molecular Sciences*, *24*(14), 11691.

Murphy, D. J. (2001). The biogenesis and functions of lipid bodies in animals, plants and microorganisms. *Progress in lipid research*, *40*(5), 325–438.

Natali, A., Roy, L. M., & Croce, R. (2014). In vitro reconstitution of light-harvesting complexes of plants and green algae. *JoVE (Journal of Visualized Experiments)*, (92), e51852.

Nymark, M., Valle, K. C., Brembu, T., Hancke, K., Winge, P., Andresen, K., Johnsen, G., & Bones, A. M. (2009). An Integrated Analysis of Molecular Acclimation to High Light in the Marine Diatom Phaeodactylum tricornutum. *PloS one*, *4*(11), e7743. https://doi.org/10.1371/journal.pone.0007743

Nymark, M., Volpe, C., Hafskjold, M. C. G., Kirst, H., Serif, M., Vadstein, Ô., Bones, A. M., Melis, A., & Winge, P. (2019). Loss of ALBINO3b insertase results in truncated Light-Harvesting antenna in diatoms. *Plant Physiology*, *181*(3), 1257–1276. https://doi.org/10.1104/pp.19.00868

Polle, J. E., Kanakagiri, S.-D., & Melis, A. (2003). Tla1, a dna insertional transformant of the green alga chlamydomonas reinhardtii with a truncated light-harvesting chlorophyll antenna size. *Planta*, *217*, 49–59.

Premvardhan, L., Bordes, L., Beer, A., Büchel, C., & Robert, B. (2009). Carotenoid structures and environments in trimeric and oligomeric fucoxanthin chlorophyll a/c2 proteins from resonance raman spectroscopy. *The Journal of Physical Chemistry B*, *113*(37), 12565–12574.

Puchta, M., Boczkowska, M., & Groszyk, J. (2020). Low rin value for rna-seq library construction from long-term stored seeds: A case study of barley seeds. *Genes*, *11*, 1190. https://doi.org/10.3390/genes11101190

Qiagen. (2021, March). Qiazol handbook for efficient lysis of fatty tissues and all other types of tissue before RNA purification. https://www.qiagen.com/cn/resources/download.aspx?id=61c3ddbd-69c1-4b68-ab89-a428f14a9245&lang=en

Qiagen. (2023, June). RNeasy Mini Handbook from Qiagen. https://www.qiagen.com/de/resources/download.aspx?id=f646813a-efbb-4672-9ae3-e665b3045b2b&lang=en

Qiao, H., Cong, C., Sun, C., Li, B., Wang, J., & Zhang, L. (2016). Effect of culture conditions on growth, fatty acid composition and dha/epa ratio of phaeodactylum tricornutum. *Aquaculture*, *452*, 311–317.

Raschka, S. (2015, January). Principal component analysis. https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#:~:text=The%20eigenvectors%20and%20eigenvalues%20of,the%20eigenvalues%20determine%20their%20magnitude.

Rost, F. (1999). Fluorescence microscopy, applications. *Encyclopedia of spectroscopy and spectrometry*, 565–570.

Rumin, J., Bonnefond, H., Saint-Jean, B., Rouxel, C., Sciandra, A., Bernard, O., Cadoret, J.-P., & Bougaran, G. (2015). The use of fluorescent Nile red and BODIPY for lipid measurement in microalgae. *Biotechnology for Biofuels*, *8*(1). https://doi.org/10.1186/s13068-015-0220-4

Sayanova, O., Mimouni, V., Ulmann, L., Morant-Manceau, A., Pasquet, V., Schoefs, B., & Napier, J. A. (2017). Modulation of lipid biosynthesis by stress in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160407.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, *63*(324), 1343–1372. https://doi.org/10.1080/01621459.1968.10480932

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv (Cornell University)*. http://cs.gmu.edu/~hrangwal/files/pca.pdf

Simidjiev, I., Stoylova, S., Amenitsch, H., Jávorfi, T., Mustárdy, L., Laggner, P., Holzenburg, A., & Garab, G. (2000). Self-assembly of large, ordered lamellae from non-bilayer lipids and integral membrane proteins in vitro. *Proceedings of the National Academy of Sciences*, *97*(4), 1473–1476.

Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry–based proteomics. *The biochemist*, *42*(5), 64–69. https://doi.org/10.1042/bio20200057

Solovchenko, A., Merzlyak, M. N., Khozin-Goldberg, I., Cohen, Z., & Boussiba, S. (2010). Coordinated carotenoid and lipid syntheses induced in parietochloris incisa (chlorophyta, trebouxiophyceae) mutant deficient in δ5 desaturase by nitrogen starvation and high light 1. *Journal of phycology*, *46*(4), 763–772.

Tanaka, T., Maeda, Y., Veluchamy, A., Tanaka, M., Abida, H., Maréchal, E., Bowler, C., Muto, M., Sunaga, Y., Tanaka, M., et al. (2015). Oil accumulation by the oleaginous diatom fistulifera solaris as revealed by the genome and transcriptome. *The Plant Cell*, *27*(1), 162–176.

Tanaka, T., Yoneda, K., & Maeda, Y. (2022). Lipid metabolism in diatoms. In *The molecular life of diatoms* (pp. 493–527). Springer.

Teh, K. Y., Loh, S. H., Aziz, A., Takahashi, K., Effendy, A. W. M., & Cha, T. S. (2021). Lipid accumulation patterns and role of different fatty acid types towards mitigating salinity fluctuations in chlorella vulgaris. *Scientific reports*, *11*(1), 438.

Turkish, A. R., & Sturley, S. L. (2009). The genetics of neutral lipid biosynthesis: An evolutionary perspective. *American Journal of Physiology-Endocrinology and Metabolism*, *297*(1), E19–E27.

Valle, K. C., Nymark, M., Aamot, I., Hancke, K., Winge, P., Andresen, K., Johnsen, G., Brembu, T., & Bones, A. M. (2014). System Responses to Equal Doses of Photosynthetically Usable Radiation of Blue, Green, and Red Light in the Marine Diatom Phaeodactylum tricornutum. *PLOS ONE*, *9*(12), e114211. https://doi.org/10.1371/journal.pone.0114211

Wang, X., Yu, H., Gao, R., Liu, M., & Xie, W. (2023). A comprehensive review of the family of very-long-chain fatty acid elongases: Structure, function, and implications in physiology and pathology. *European Journal of Medical Research*, *28*(1), 532.

Whitelam, G., & Codd, G. (1986). Damaging effects of light on microorganisms.

Wilhelm, C., Büchel, C., Fisahn, J., Goss, R., Jakob, T., LaRoche, J., Lavaud, J., Lohr, M., Riebesell, U., Stehfest, K., Valentin, K., & Kroth, P. G. (2006). The Regulation of Carbon and Nutrient Assimilation in Diatoms is Significantly Different from Green Algae. *Protist*, *157*(2), 91–124. https://doi.org/10.1016/j.protis.2006.02.003

Williams, D. M. (2007). aaDiatom phylogeny: Fossils, molecules and the extinction of evidence. *Comptes Rendus Palevol*, *6*(6-7), 505–514. https://doi.org/10.1016/j.crpv.2007.09.016

Wu, Z., Bagarolo, G. I., Thoröe-Boveleth, S., & Jankowski, J. (2020). "Lipidomics": Mass spectrometric and chemometric analyses of lipids. *Advanced Drug Delivery Reviews*, *159*, 294–307. https://doi.org/10.1016/j.addr.2020.06.009

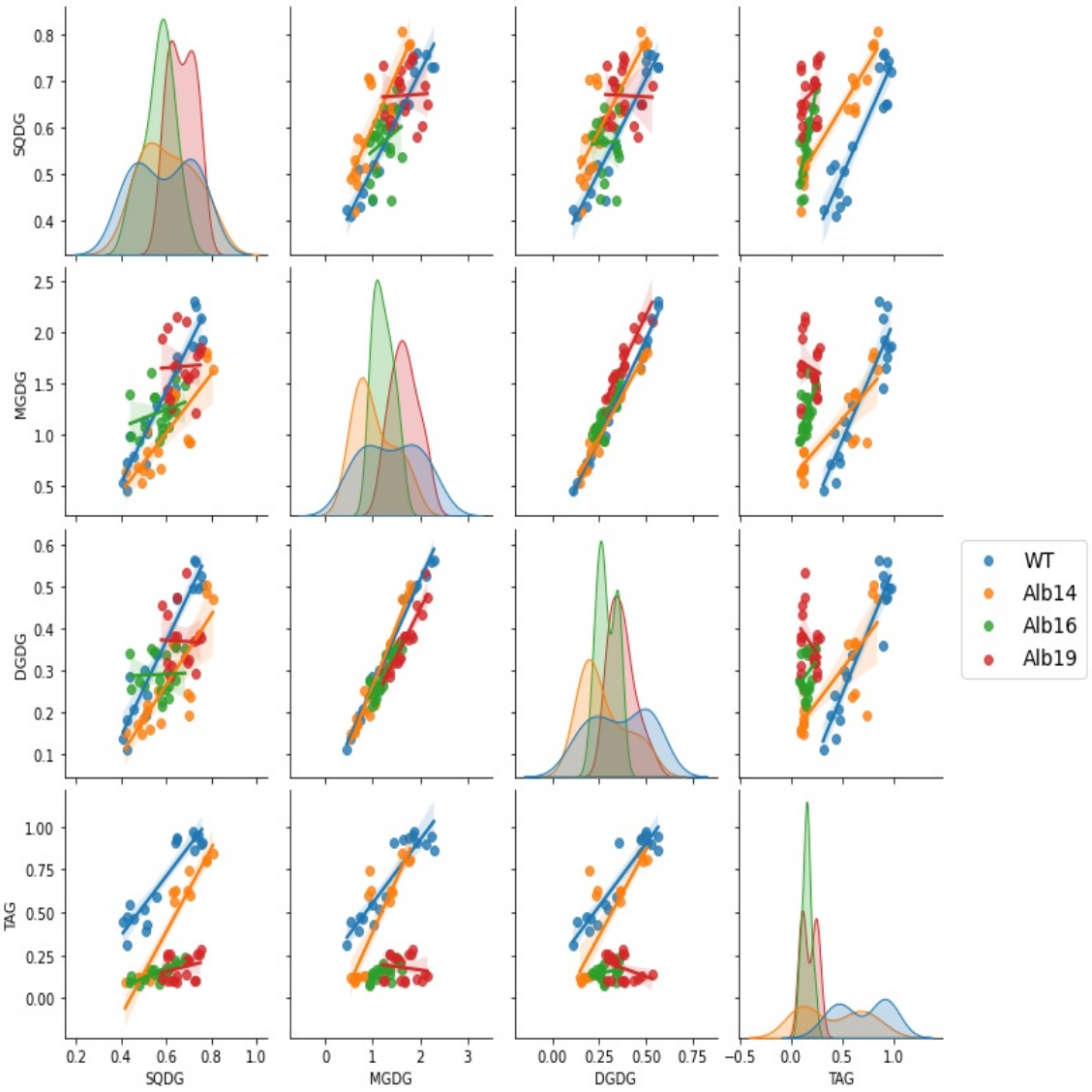
Yang, S., & Rothman, R. E. (2004). Pcr-based diagnostics for infectious diseases: Uses, limitations, and future applications in acute-care settings. *The Lancet infectious diseases*, *4*(6), 337–348.

Yi, Z., Xu, M., Di, X., Brynjólfsson, S., & Fu, W. (2017). Exploring valuable lipids in diatoms. *Frontiers in Marine Science*, *4*. https://doi.org/10.3389/fmars.2017.00017

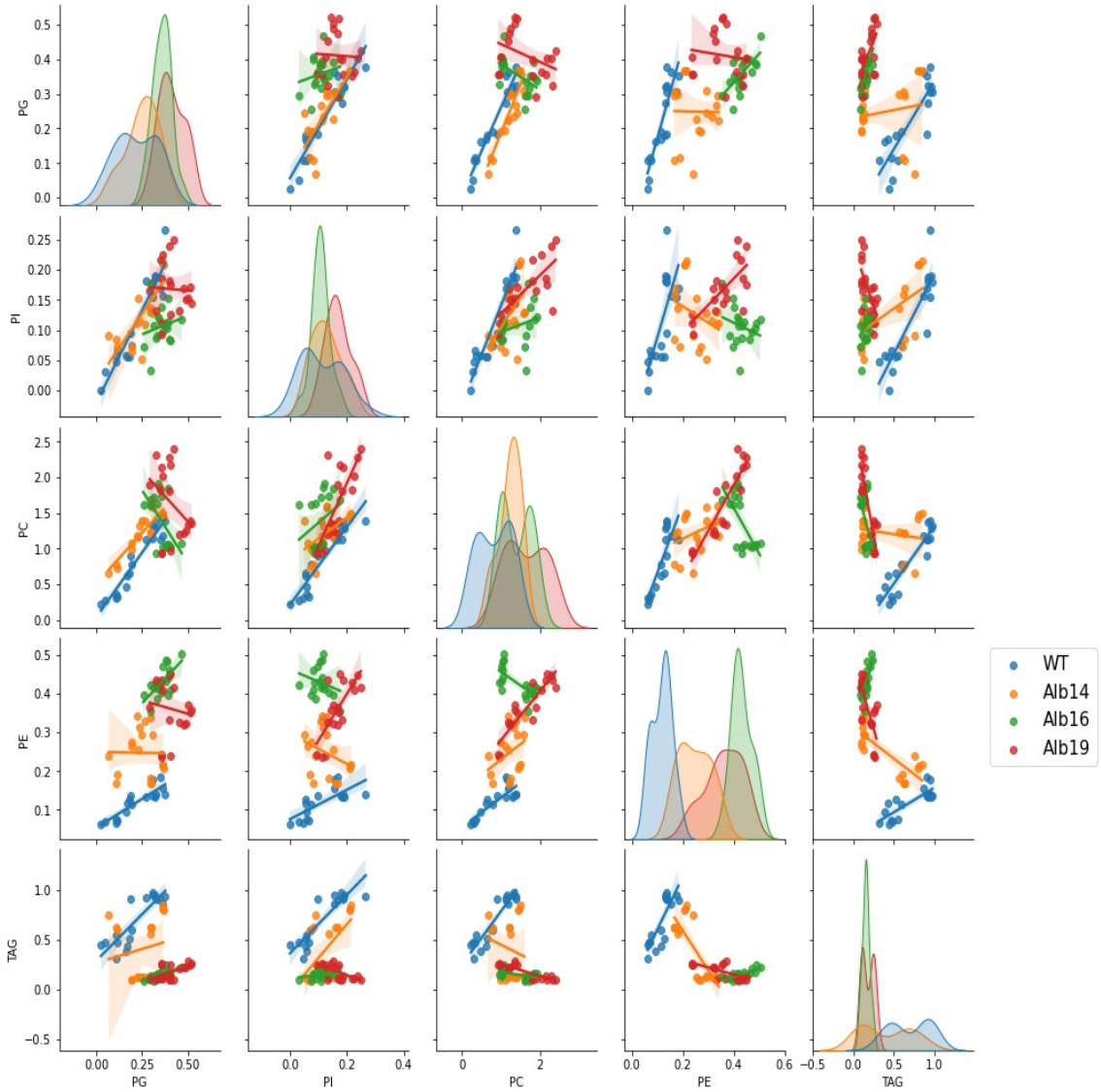
Zhang, L., Liu, W., Xiao, J., Hu, T., Chen, J., Chen, K., Jiang, H., & Shen, X. (2007). Malonyl-coa: Acyl carrier protein transacylase from helicobacter pylori: Crystal structure and its interaction with acyl carrier protein. *Protein Science*, *16*(6), 1184–1192.

11 Appendix

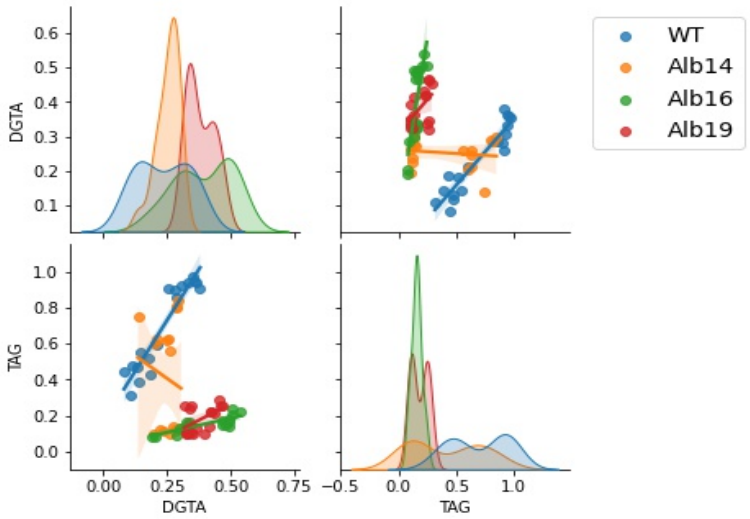
11.1 Supplementary analyses for assessing pipeline efficiency



Supplementary Figure 1: A pair plot presenting regression plots between different Glycolipids classes and the neutral lipid TAG. This was created to cross-verify the observed correlation observed in the Biplots generated after PCA. Heat maps in Figure 57could also have been used for the same purpose, but these regression plots present all the cell lines together in a single Cartesian plane along with corresponding regression lines, thus allowing multiple comparisons. The pair plot was generated using the 'pairplot' function in the seaborn library of Python. The histograms presented in the diagonal of the pair plot can be used to determine the distribution of the different lipid classes.

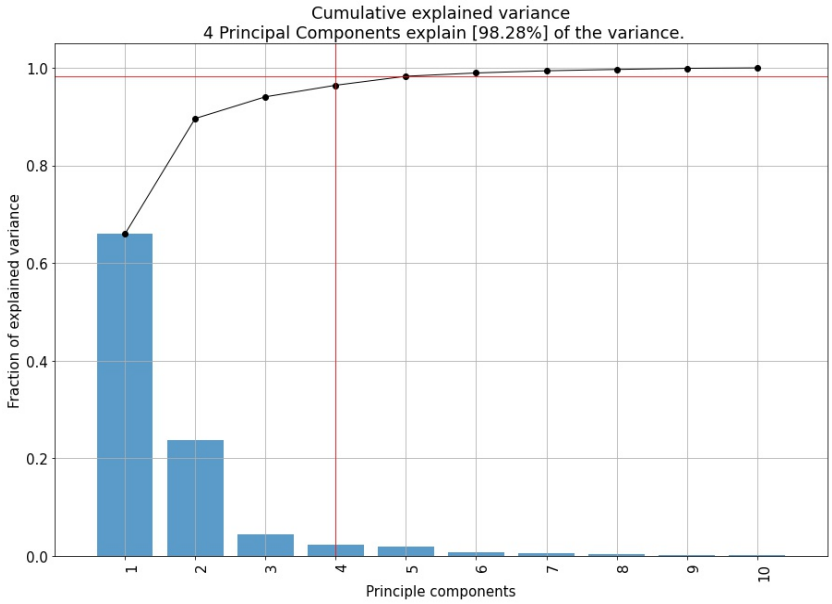


Supplementary Figure 2: A pair plot presenting regression plots between different Phospholipid classes and the neutral lipid TAG. The plot was created using 'seaborn' package from Python.

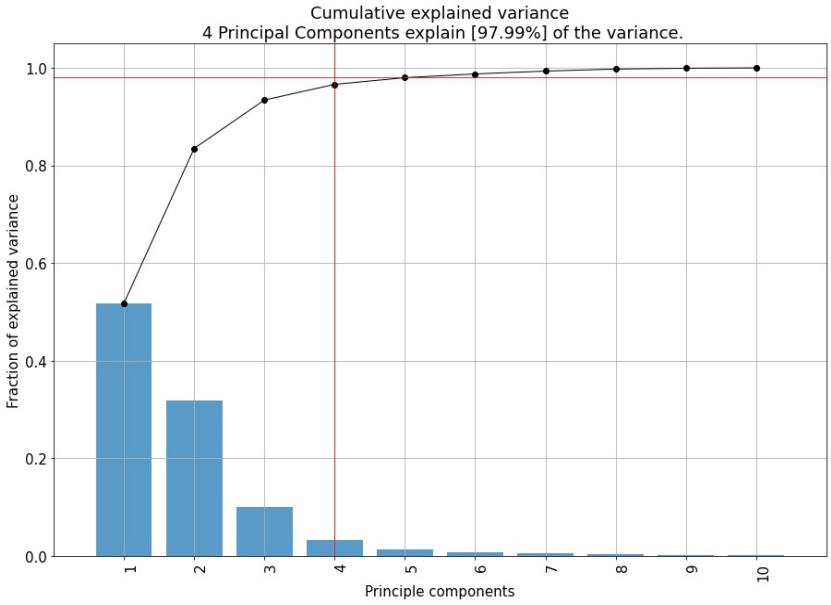


Supplementary Figure 3: A pair plot presenting regression plots between Betaine lipid DGTA and the neutral lipid TAG.The plot was created using 'seaborn' package from Python.

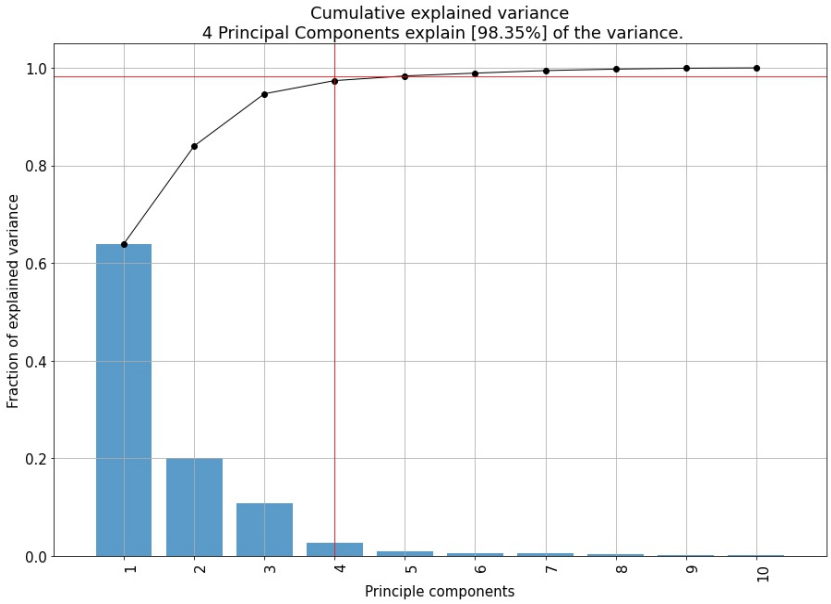
11.2 Supplementary results from PCA



Supplementary Figure 4: Scree plot for the principle components of the subset of the original dataset, which includes all the Alb14 and wild-type samples in both the light conditions.

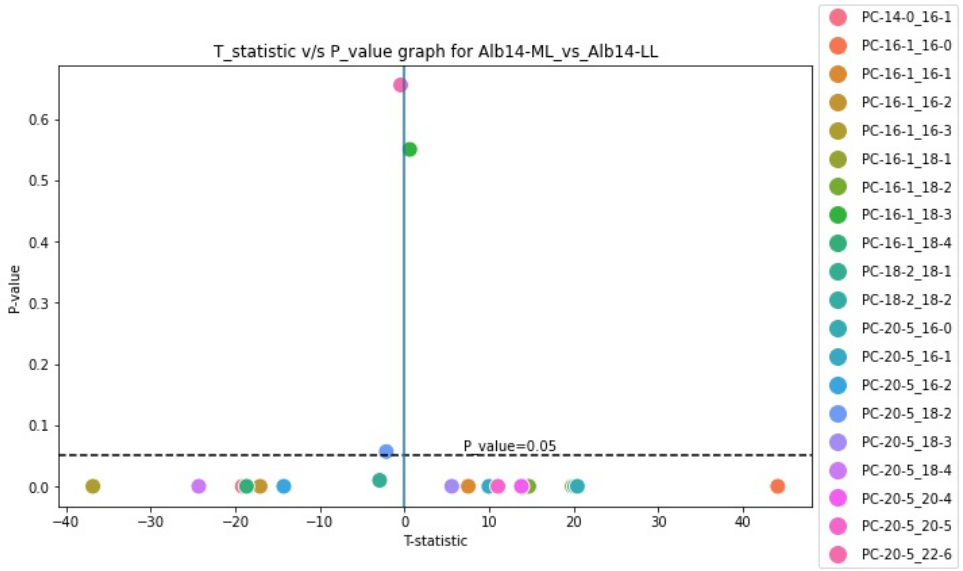


Supplementary Figure 5: Scree plot for the principle components of the subset of the original dataset, which includes all the Alb16and wild-type samples in both the light conditions.

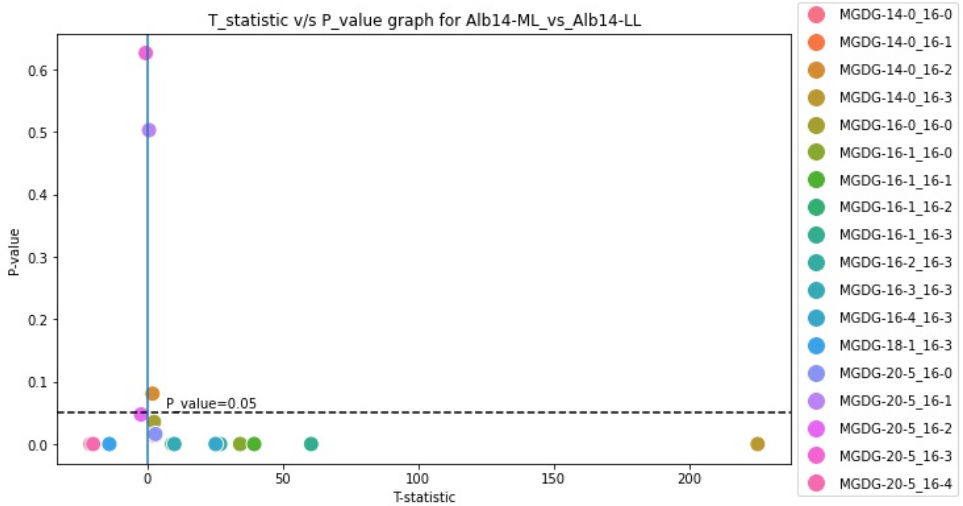


Supplementary Figure 6: Scree plot for the principle components of the subset of the original dataset, which includes all the Alb19 and wild-type samples in both the light conditions.

11.3 Supplementary results from T-tests

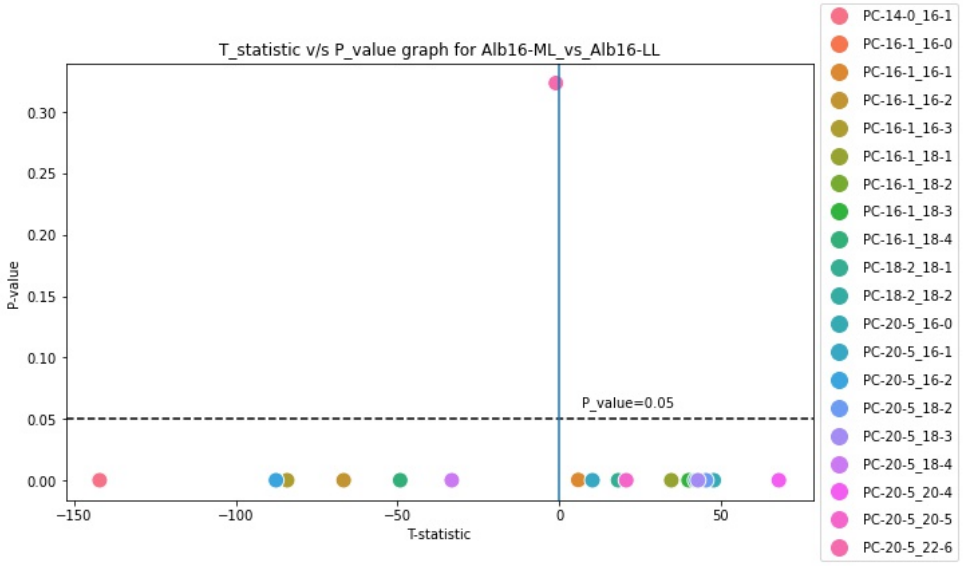


(a)

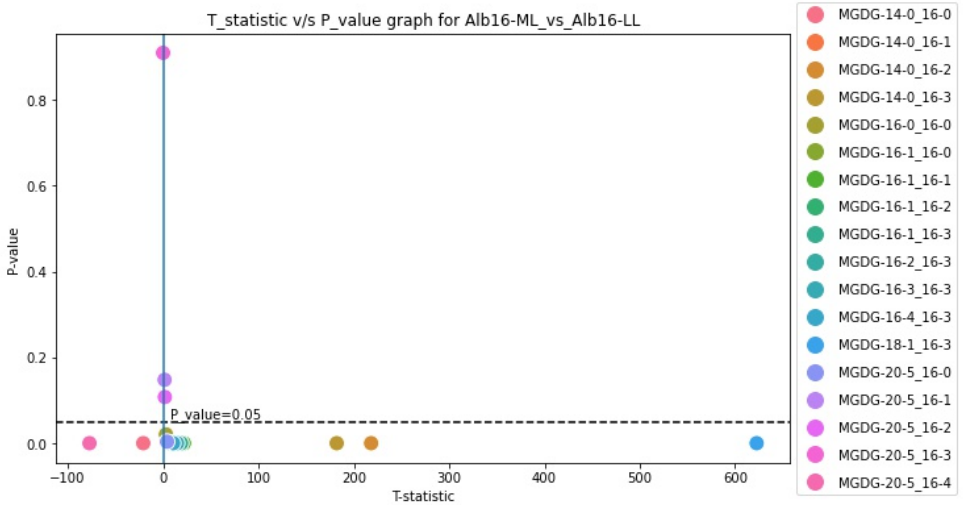


(b)

Supplementary Figure 7: Results for T-test comparing concentrations of different fatty acid compositions of PC and MGDG in *Alb3b*-14 in ML and LL conditions.

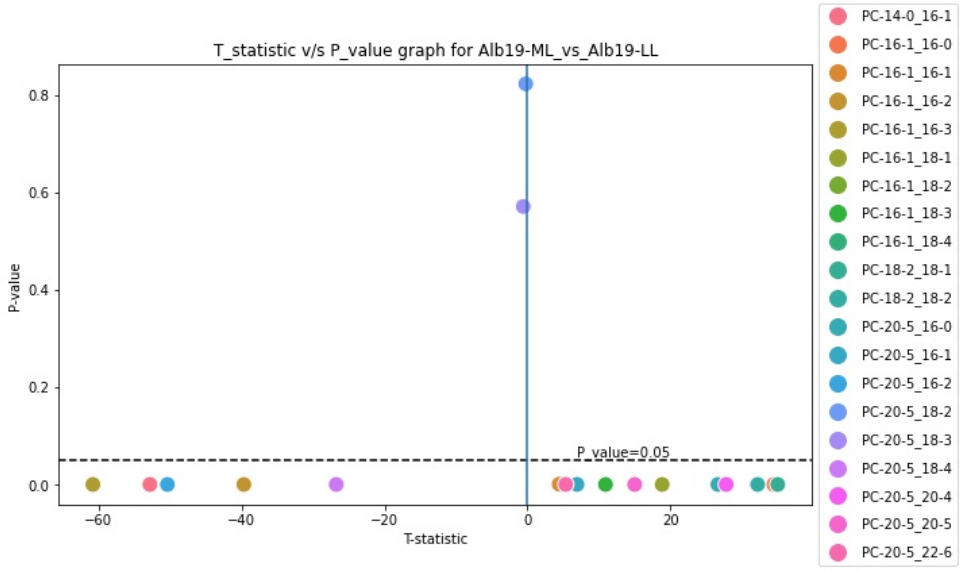


(a)

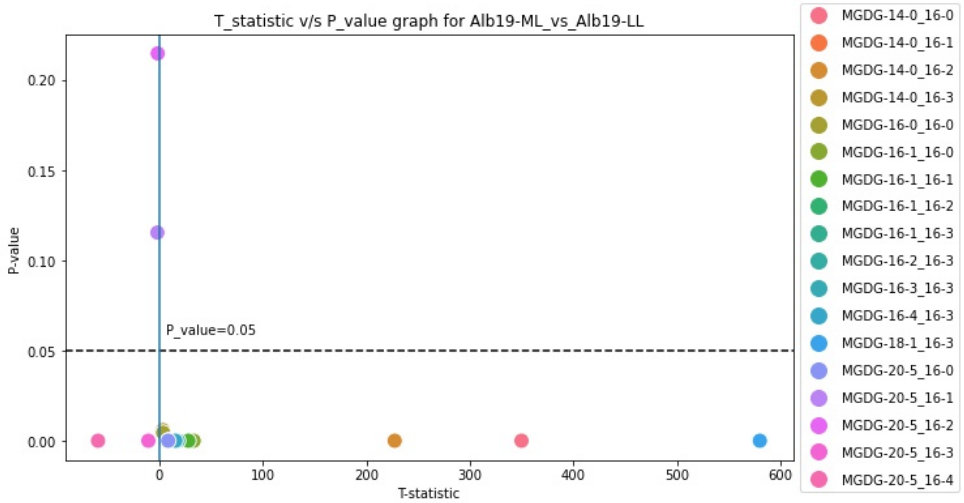


(b)

Supplementary Figure 8: Results for T-test comparing concentrations of different fatty acid compositions of PC and MGDG in *Alb3b*-16 in ML and LL conditions.

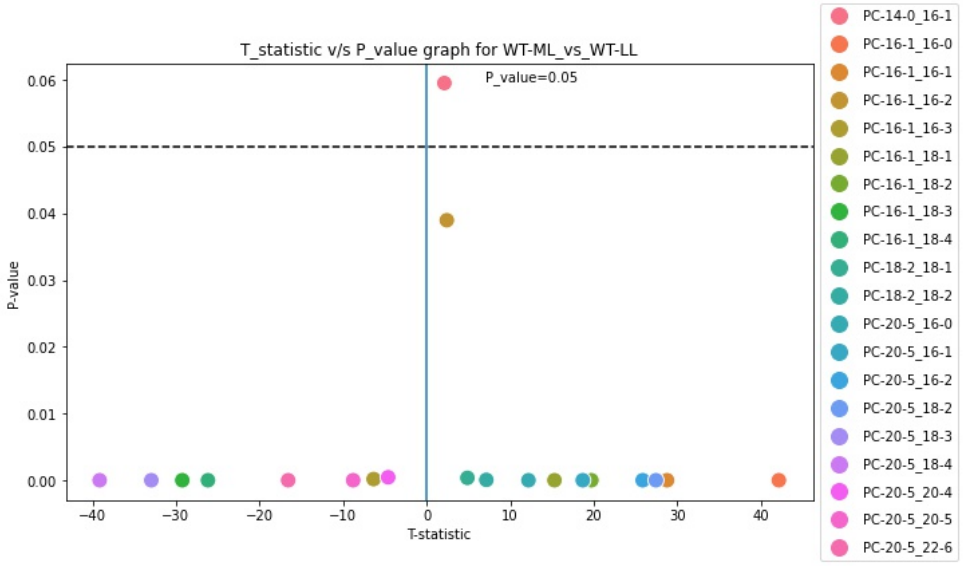


(a)

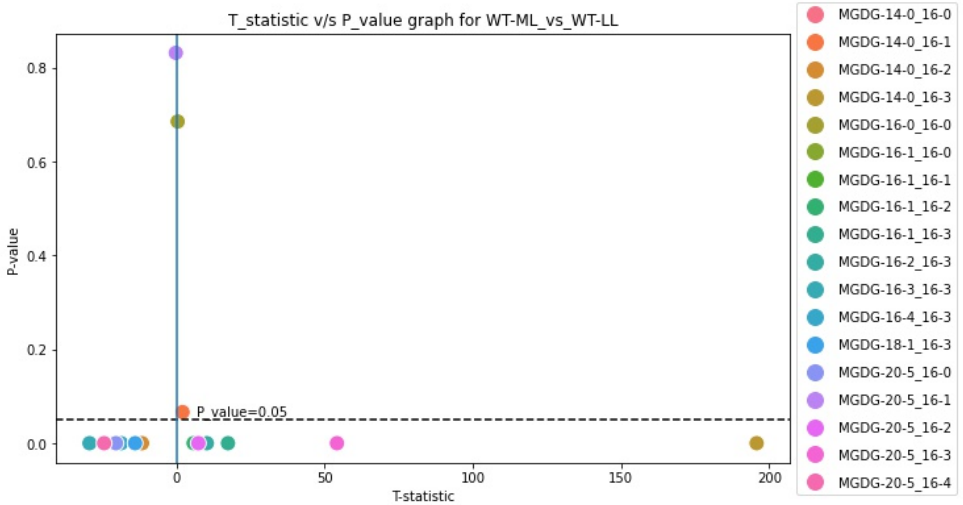


(b)

Supplementary Figure 9: Results for T-test comparing concentrations of different fatty acid compositions of PC and MGDG in Alb19 in ML and LL conditions.



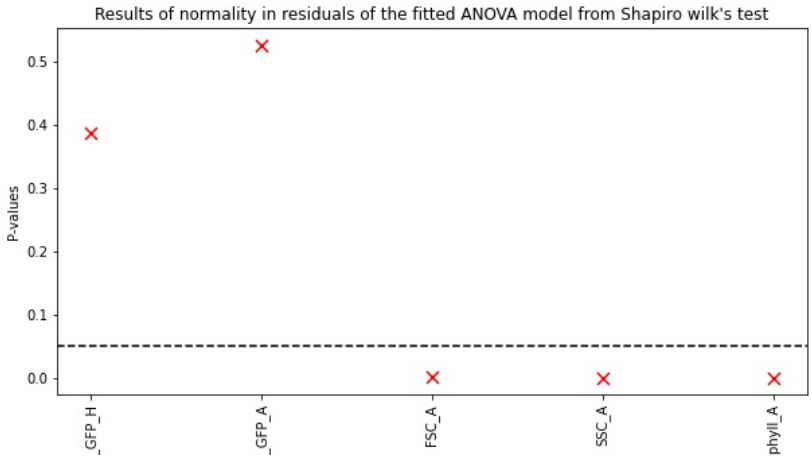
(a)



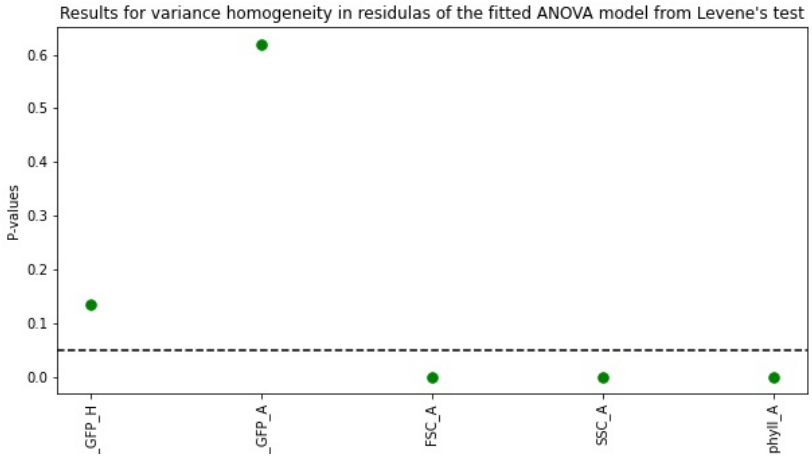
(b)

Supplementary Figure 10: Results for T-test comparing concentrations of different fatty acid compositions of PC and MGDG in Wild-type in ML and LL conditions.

11.4 ANOVA assumptions tests for flow cytometry parameters

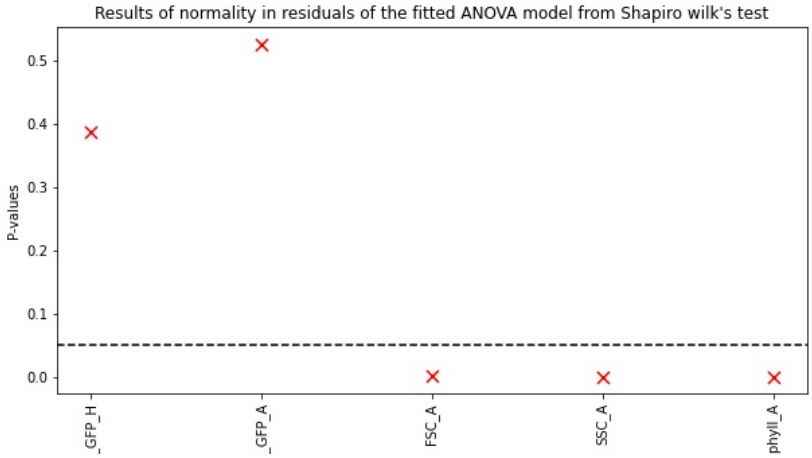


(a) Shapiro Wilk's test results for residuals from the fitted ANOVA model.

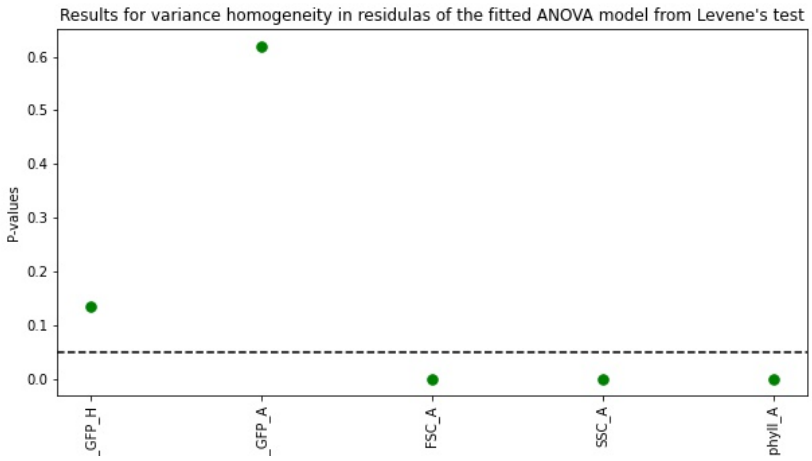


(b) Levene's test results for residuals from the fitted ANOVA model.

Supplementary Figure 11: Results from statistical tests conducted for checking whether the assumptions of ANOVA are being followed by the parameters measured using flow cytometry. Shapiro Wilk's tests(a) and Levene's tests(b) indicate that all parameters except those from the FITC-GFP channel, violate both normal distribution and homogeneous variation assumption for residuals in the ANOVA model.



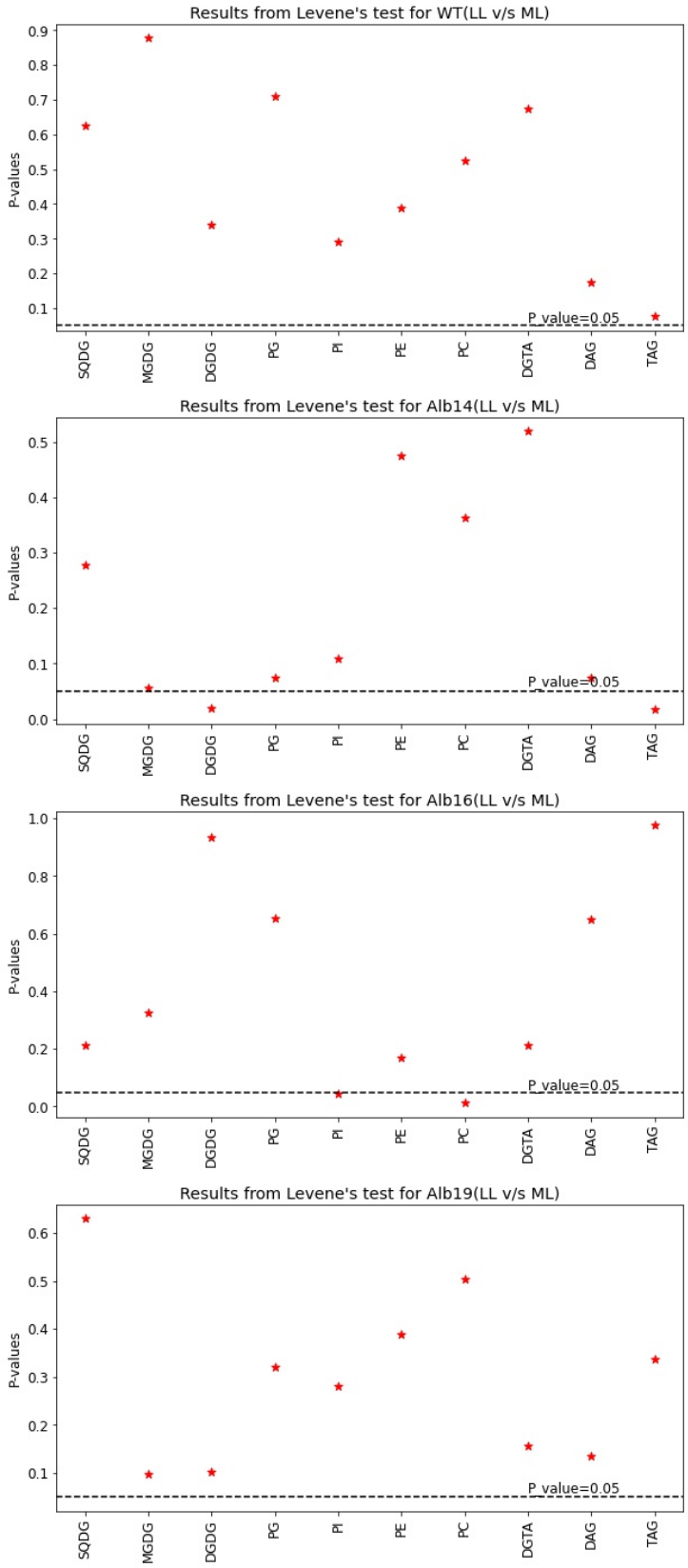
(a) Shapiro Wilk's test results for residuals from the fitted ANOVA model



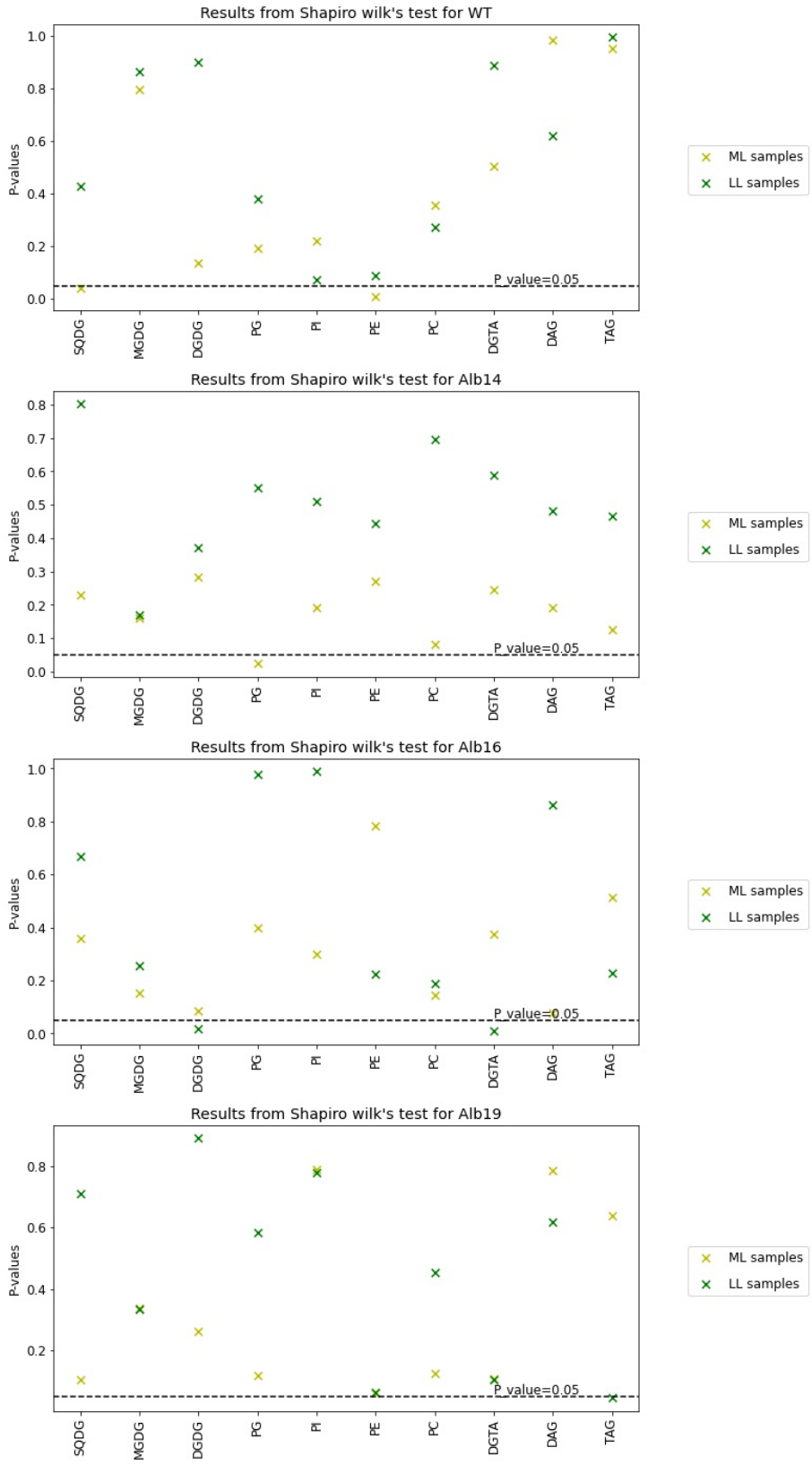
(b) Levene's test results for residuals from the fitted ANOVA model

Supplementary Figure 12: Results from statistical tests conducted for checking whether the assumptions of ANOVA are being followed by the parameters measured using flow cytometry. Shapiro Wilk's tests(a) and Levene's tests(b) indicate that all parameters except those from the FITC-GFP channel, violate both normal distribution and homogeneous variation assumption for residuals in the ANOVA model.

11.5 Results from testing assumptions of T-tests on the MS-MS data



Supplementary Figure 13: Results from Levene's test for the different cell lines comparing the equality of variances in data from all lipid classes between two different light conditions(LL and ML) in WT, *Alb3b*-14,16,19 (in order from top to bottom). The horizontal dotted line indicates p=0.05.

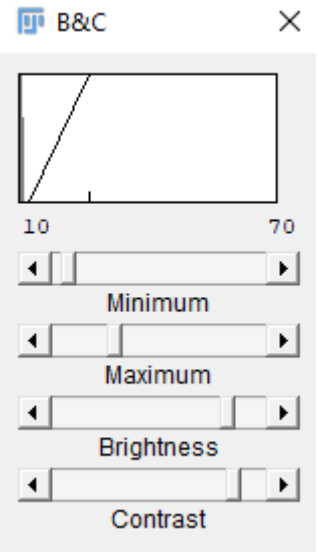


Supplementary Figure 14: Results from Shapiro Wilk's test for the different cell lines for assessing the normal distributions of data from all lipid classes in two different light conditions(LL and ML) in WT, *Alb3b*-14,16,19 (in order from top to bottom). The horizontal dotted line indicates p=0.05. Thus, all the points below indicate deviation from normal distribution.

11.6 Processing steps and supplementary results from CLSM

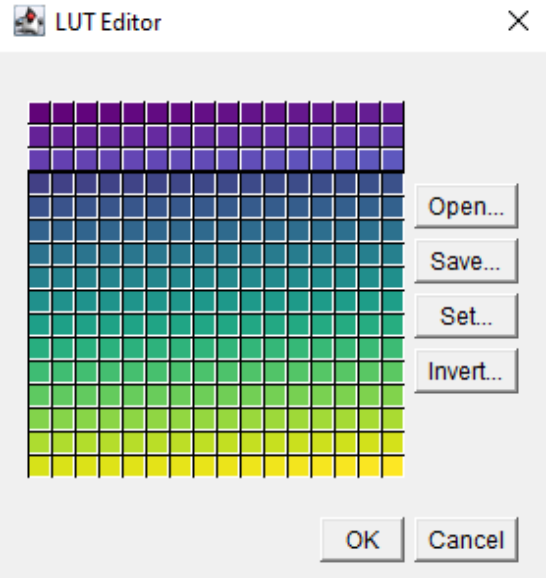
```
selectImage("z_stack_image.lif - C=0");
run("mpl-viridis");
run("Z Project...", "projection=[Max Intensity]");
run("Edit LUT...");
run("Despeckle");
selectImage("z_stack_image.lif - wtll1az - C=1");
run("Red");
run("Z Project...", "projection=[Max Intensity]");
//run("Brightness/Contrast...");
run("Apply LUT");
run("Close");
run("Despeckle");
```

Supplementary Figure 15: The macro applied in FIJI/ImageJ for editing the z-stacks obtained from CLSM to obtain the final version of the images shown in section 5.3.4(Figures 46,47,48). C=0 represents images from channel zero representing the BODIPY signals and C=1 represents channel 1 images for auto fluorescence. The LUT editing part in step 4 is explained in detail in figure 17. The brightness and contrast adjustment part for images from channel 1 for auto fluorescence in step 9 is presented in figure 16.

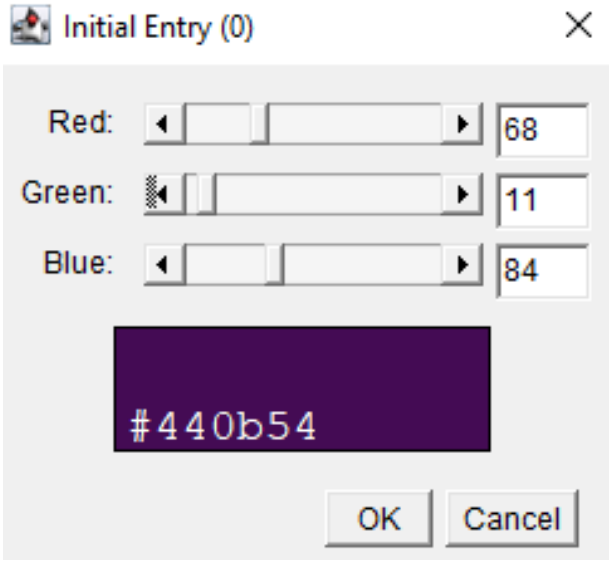


Supplementary Figure 16: Brightness and contrast setting applied for all the images in the auto fluorescence channel(c=1) during the 9^{th} step in the macro given in Figure 15.

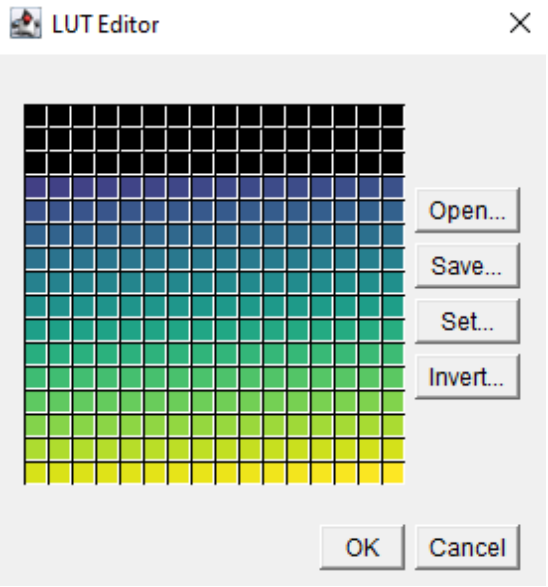
(a) Original color distribution of 'mpl-viridis'



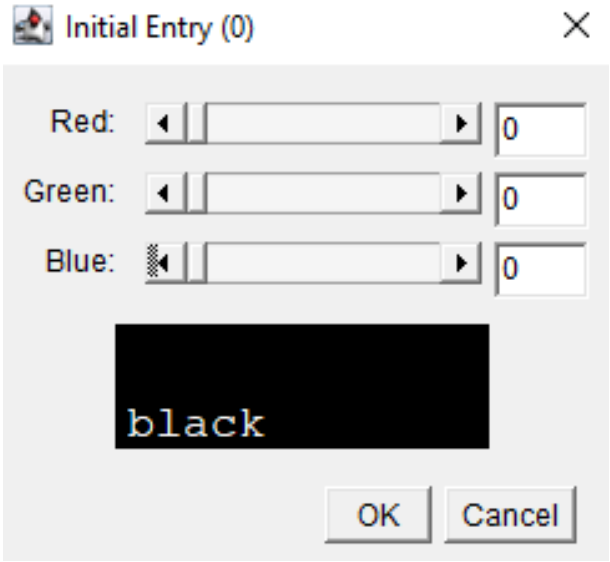
(b) Original RGB values of 'mpl-viridis'



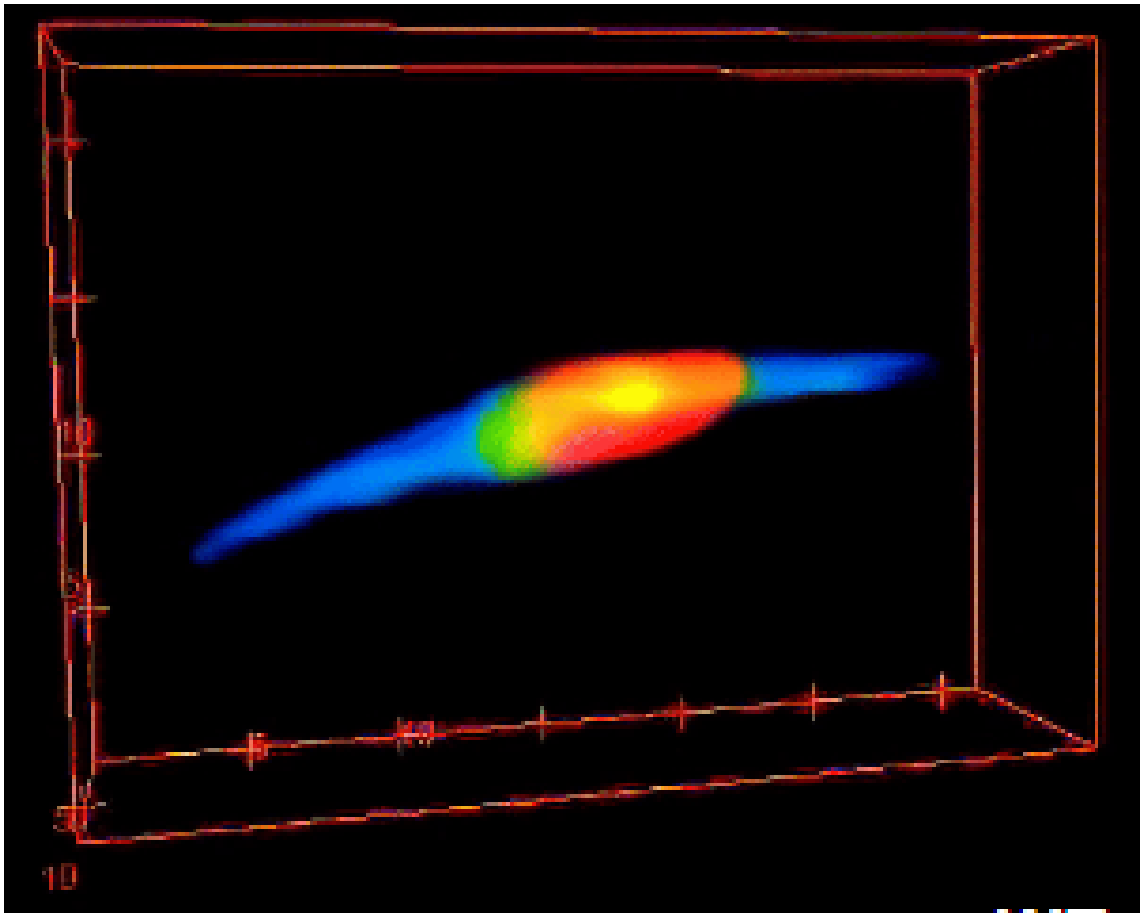
(c) Edited color distribution of 'mpl-viridis'



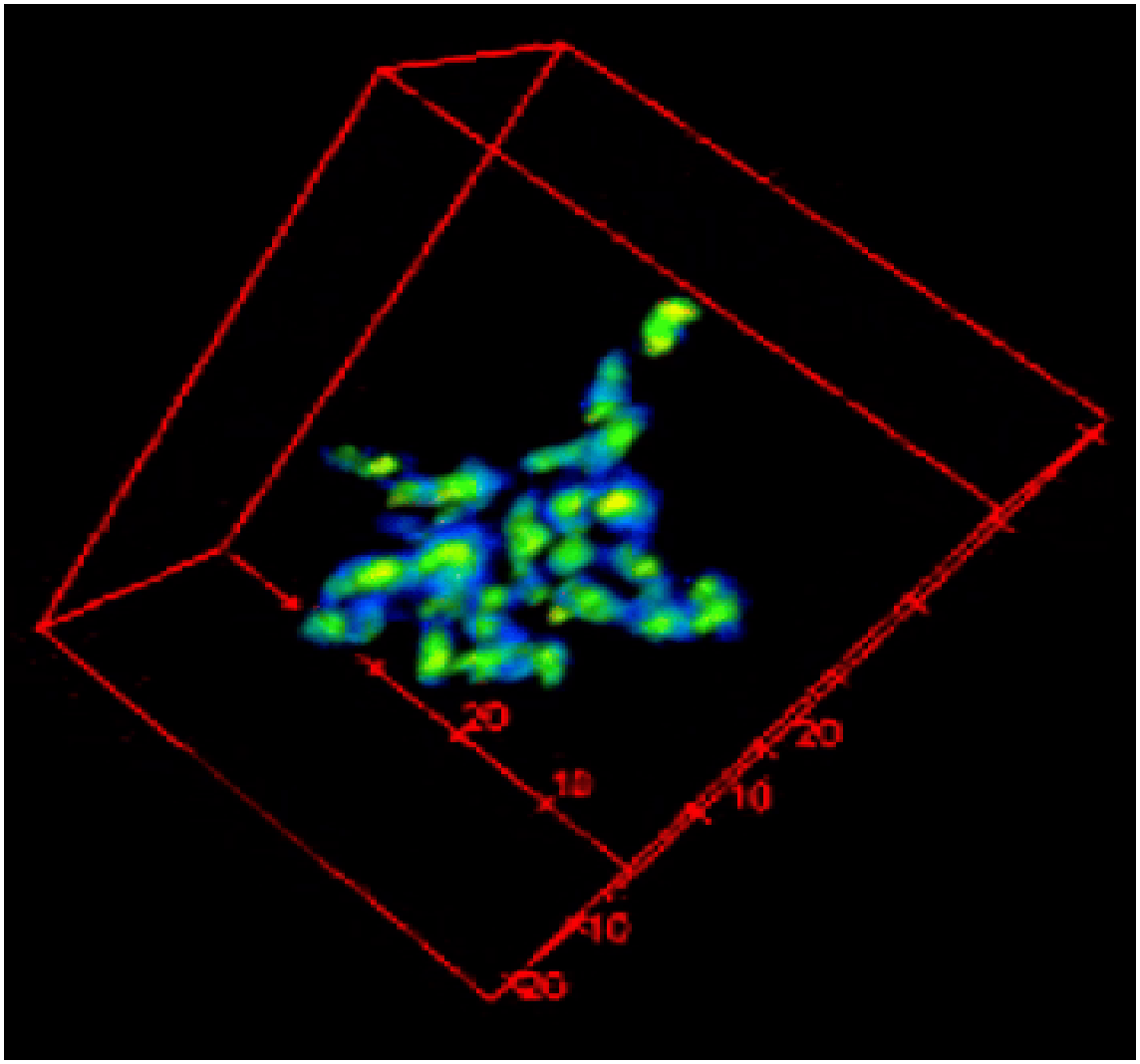
(d) Edited RGB values of 'mpl-viridis'



Supplementary Figure 17: Figures showing the process of editing the 'mpl-viridis' LUT in FIJI/ImageJ. The purple background was removed by changing the original RGB values for the first three rows in the LUT into zero resulting in a completely black background with blue cells with yellow colored LDs.a) Color distribution of the original 'mpl-viridis' LUT with the first three rows representing the various shades of purple. b) Original RGB values of the purple region in 'mpl-viridis' LUT. c) and d) Edited 'mpl-viridis' color distribution and RGB values for the purple shade region respectively, indicating the change of all shades into complete black.



Supplementary Figure 18: A 3D model of a WT cell under LL conditions showing a compact LD in the center (Yellow). The red color indicates the auto fluorescence emissions from the cell. The model was formed using FIJI using Z stacks captured from CLSM.



Supplementary Figure 19: A 3D model of a cell cluster formed by the *Alb3b*-19 mutants under HL acclimation for 14 days indicating the formation of round morphotypes and aggregation of the same as a stress response to HL. The model was generated using FIJI using Z stacks captured from CLSM.

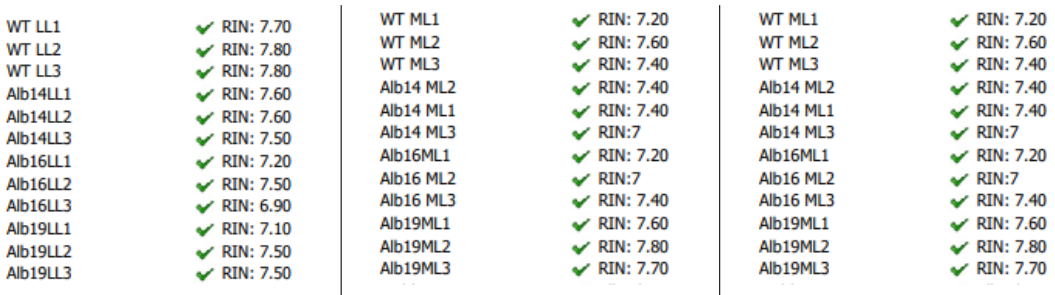
11.7 supplementary results from q-PCR and associated steps

Gene ID	Primer Designs
PHATRDRAFT_3765	Fw: TCCCACAAGACGGGTGACGTG Rw: CCGGCTGTTGCCTCCTTGGAA
PHATRDRAFT_48423	Fw: ACAGATACCGATGTTCCGCA Rw: GCATCGGTAGCAATCTGAGC
PHATRDRAFT_20508	Fw: TCCTCACCATCGTCTTGAACGGT Rw: TCGTGATGAACTGCACCAGC
PHATRDRAFT_50443	Fw: CGGAAGCAGATGCGGTCCCT Rw: ACCGTCTGGTCGATCTTCGCC
PHATRDRAFT_10068	Fw: CCGGTAACGGTTCGGGCCAT Rw: TGGCAAAGAGCCCGGTCAGA
PHATRDRAFT_20143	Fw: GACTCAAGGGTGACTGGGAG Rw: GTCCACGGTGAGTCCAAAAG
PHATRDRAFT_bd765	Fw: TGGTTGGAAAGGCGAGCCGA Rw: TCGGGGACTACTACGCCTCT
PHATRDRAFT_54756	Fw: GATCGAGAATGCCGCCGTGC Rw: GCGGTACACTGTCGGAGCGT
PHATRDRAFT_42683	Fw: CTCGCCGCAGGCAGGATACA Rw: AGAAAACTGGGCAGTCGCTATGTT
PHATRDRAFT_bd976	Fw: CCAATGGTCAGTGCTCAACGC Rw: TGGTCTTTACAGGGGAAGATTG
PHATRDRAFT_41570	Fw: CGGGGCGCCTTTCAAACTGT Rw: TCGTCGCTTCTTCGAGCCTGT

Table 5: Primer designs for the genes used in real-time PCR. Fw stands for forward primer and Rw stands for reverse primer.

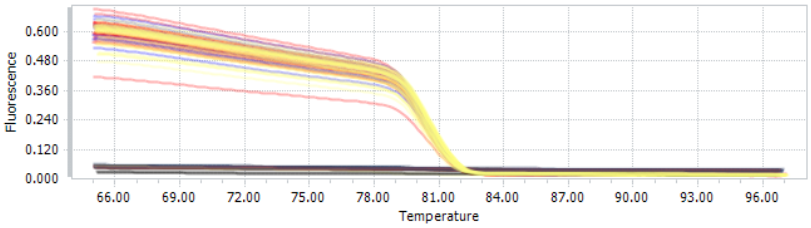
Sample Name	A260/A280	A260/A230	Nucleic Acid(ng/uL)
WT HL 1	2.21	2.488	1190.205
WT HL 2	2.207	2.506	1088.807
WT HL 3	2.2	2.44	784.4
WT ML 1	2.163	2.535	899.298
WT ML 2	2.208	2.55	968.493
WT ML 3	2.21	2.6	1451.9
WT LL 1	2.183	2.611	1500.36
WT LL 2	2.205	2.554	1125.748
WT LL 3	2.23	2.51	698.6
Alb14 HL 1	2.178	2.578	1381.934
Alb14 HL 2	2.175	2.315	477.543
Alb14 HL 3	2.21	1.88	924.6
Alb14 ML 1	2.18	2.586	1310.302
Alb14 ML 2	2.203	2.274	934.158
Alb14 ML 3	2.24	1.76	707.6
Alb14LL1	2.168	2.519	352.453
Alb14LL2	2.227	1.921	718.034
Alb14LL3	2.164	2.486	312.094
Alb16 HL 1	2.177	2.5	1395.425
Alb16 HL 2	2.211	1.929	1314.174
Alb16 HL 3	2.2	1.74	554.7
Alb16 ML 1	2.172	2.561	1027.603
Alb16 ML 2B	2.132	0.552	210.644
Alb16 ML 3	2.24	2.23	581
Alb16LL1	2.176	1.881	442.699
Alb16LL2	2.183	2.283	472.143
Alb16LL3	1.98	1.429	126.802
Alb19 HL 1	2.185	2.495	1058.18
Alb19 HL 2	2.194	1.603	725.109
Alb19 HL 3	2.19	2.04	845.4
Alb19 ML 1	2.182	2.606	1319.609
Alb19 ML 2	2.195	2.476	1122.338
Alb19 ML 3	2.03	1.43	354.4
Alb19LL1	2.183	2.478	404.044
Alb19LL2	2.181	2.32	399.637
Alb19LL3	2.227	2.184	658.853

Supplementary Figure 20: Results from nanodrop assessment showing extent of contamination based on A260/A80 and A260/A230 ratios. The samples highlighted in yellow indicate A260/A230 values less than 2, but A260/A280 values greater than 2 and those in red indicate both ratios below 2. Only one sample, i.e *Alb3b*-16 LL3, is considered to be of bad quality as it has both values below 2, pointing towards the possible presence of both protein and phenolic contamination.

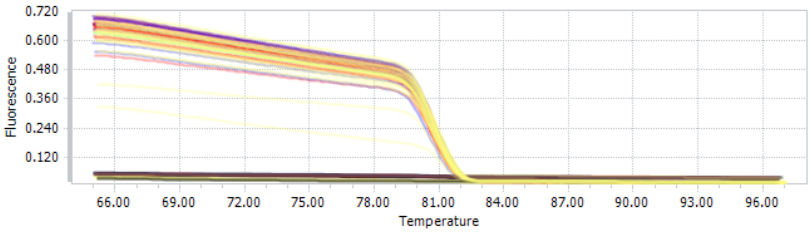


(a) LL treated samples (b) ML treated samples (c) HL treated samples

Supplementary Figure 21: RNA integrity numbers calculated from samples of the four different cell lines treated under the three different light conditions. All samples had RIN numbers above the value of 4 recommended for RNAseq(Although RNAseq was not done, this value was taken as a reference for good RNA quality for q-PCR because of the unavailability of any standards.

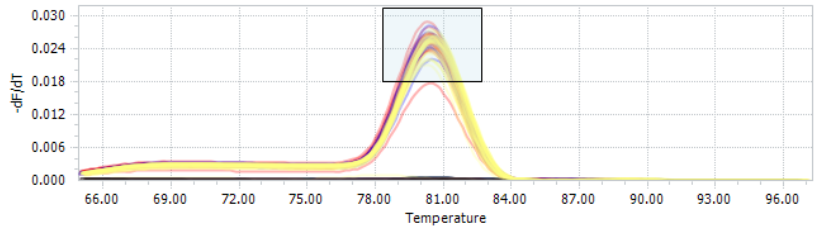


(a) Fluorescence curves from q-PCR reaction for the hiv binding rev protein



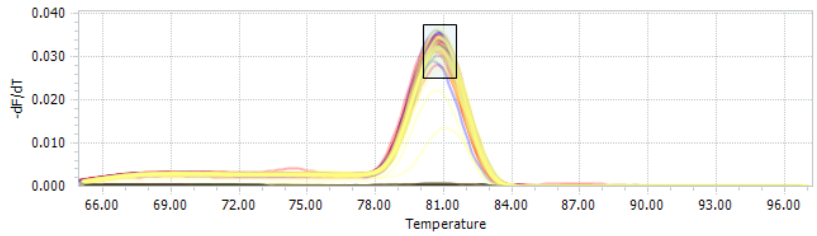
(b) Fluorescence curves from q-PCR reaction for the RPS5 protein

Supplementary Figure 22: Fluorescence curves from the q-PCR reaction for the two reference genes in both the RT-ve and cDNA samples. The yellow lines represent all the RT-ve samples including wild-type and mutant lines in all the different light treatments. The red lines represent cDNA samples from all the mutant lines in all different light conditions, the blue lines represent the wild-type cDNA samples in all the light treatments, and the black lines indicate blanks with just the master mix containing primers, RT enzyme and dNTPs in buffer solution.



(a) Melting curves from q-PCR reaction for the hiv binding rev protein

(b)

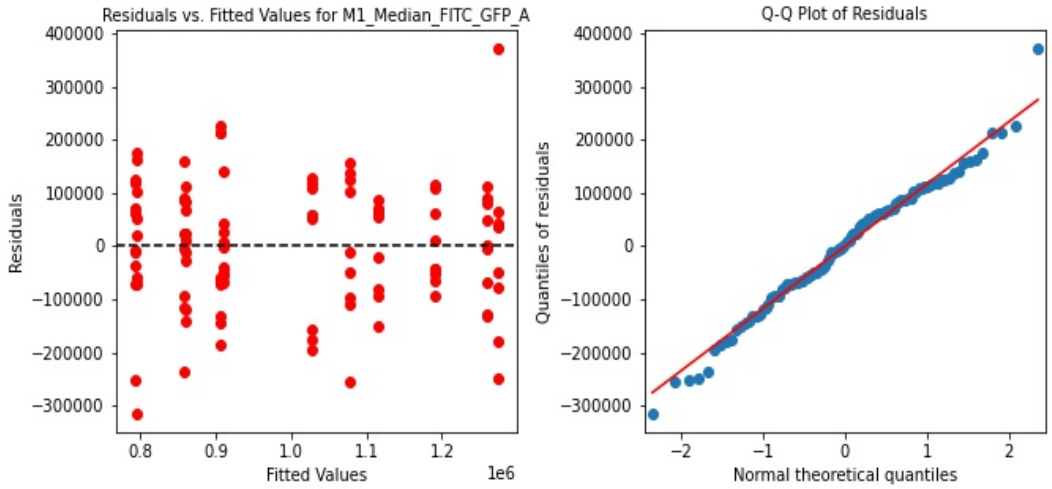


(c) Melting curves from q-PCR reaction for the RPS5 protein

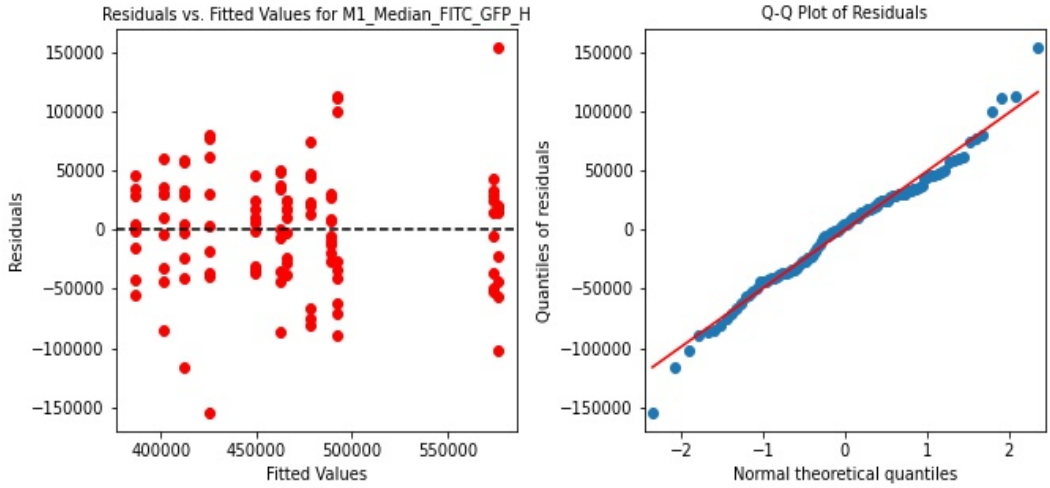
Supplementary Figure 23: Melting curves from the last stage of q-PCR reaction for the two reference genes in both the RT-ve and cDNA samples. the color coding of lines is the same as Figure 22. The majority of the curves, including the RT-ve samples and the cDNA samples appear to have a melting peak around 80°C

11.8 Results from testing assumptions for ANOVA on various data collected during lab work

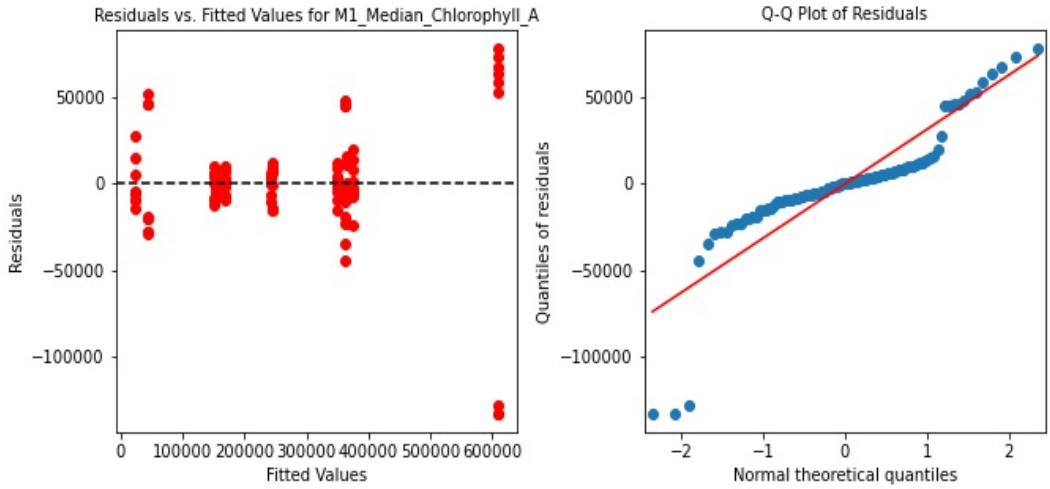
11.9 ANOVA assumptions tests for Flow cytometry measurements



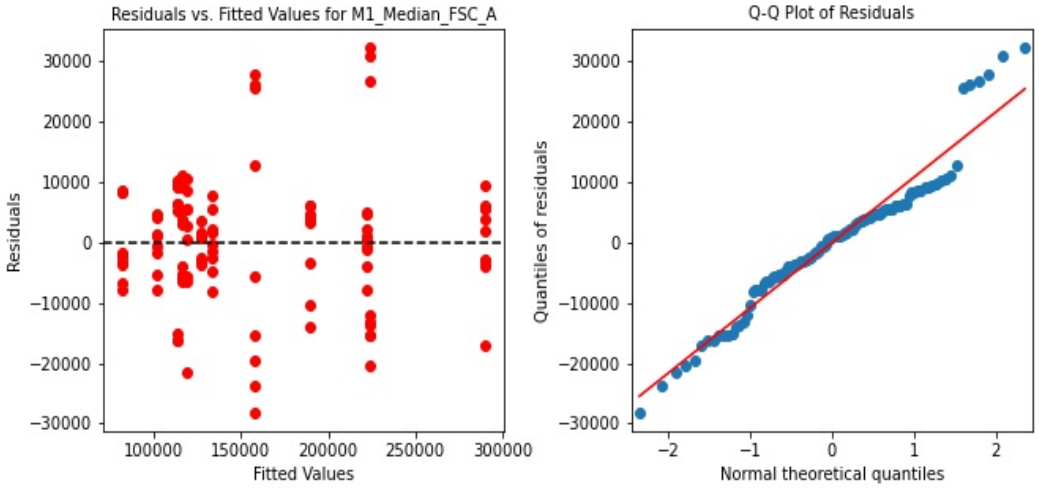
Supplementary Figure 24: Plots showing characteristics of Residuals in the fitted ANOVA model for median FITC-GFP-A parameter. The total number of sample points and thus the number of residuals will be 108, including 3 technical replicates for each of the 3 biological replicates of 4 cell lines in 3 light treatments. The number of sample categories thus fitted values is 12 combinations including 4 cell lines in 3 light conditions. The 'Residuals v/s fitted values' plot(Left) indicates almost homogeneously varying residuals of different sample points(red circles) across the fitted values along the fitted model(dashed black line) as assumed in ANOVA. The q-q plot(quantile-quantile plot)(right) plotting the theoretical quantile values for a standard normal distribution along the x-axis plotted against the observed quantiles for residuals indicates the residuals almost aligned to the line representing normally distributed residuals as assumed in ANOVA.



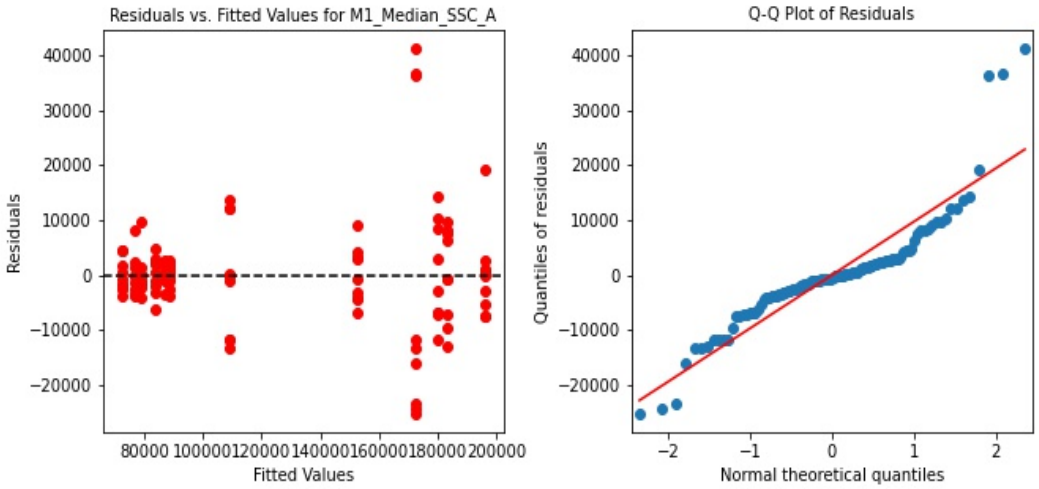
Supplementary Figure 25: Plots showing characteristics of Residuals in the fitted ANOVA model for the measured FITC-GFP-H parameters. The description of the graph is the same as Figure 24. Although imperfect, homogeneity in distribution and normality can be observed for the residuals to some extent.



Supplementary Figure 26: Plots showing characteristics of Residuals in the fitted ANOVA model for median Chlorophyll-A parameter measurements. The description of the graph is the same as Figure 24. Residuals for one of the sample categories around the fitted value are not homogeneous with that of the rest and notable deviations from normal distribution are observed.

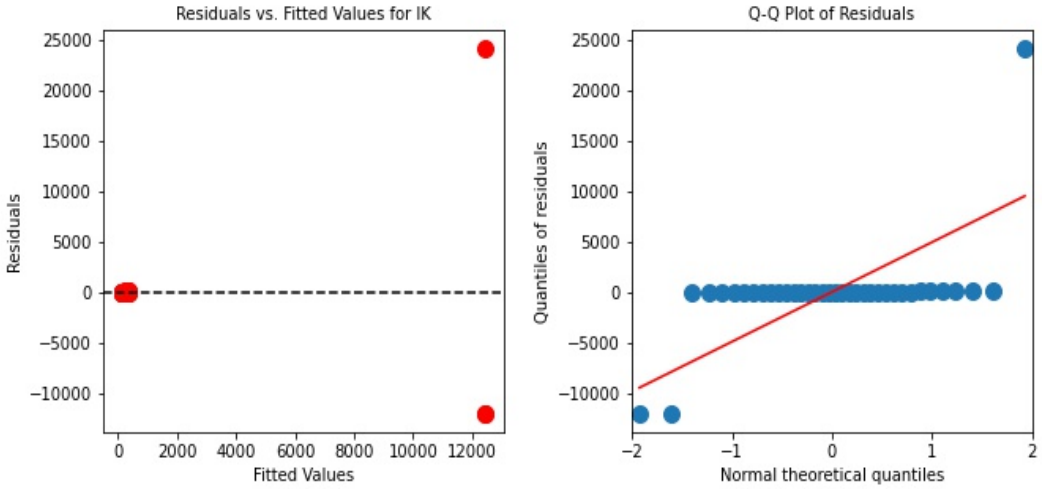


Supplementary Figure 27: Plots showing characteristics of Residuals in the fitted ANOVA model for median FSC-A parameter measurements. The description of the graph is the same as Figure 24. Although not perfect, some level of homogeneity in distribution and normality can be observed for the residuals.

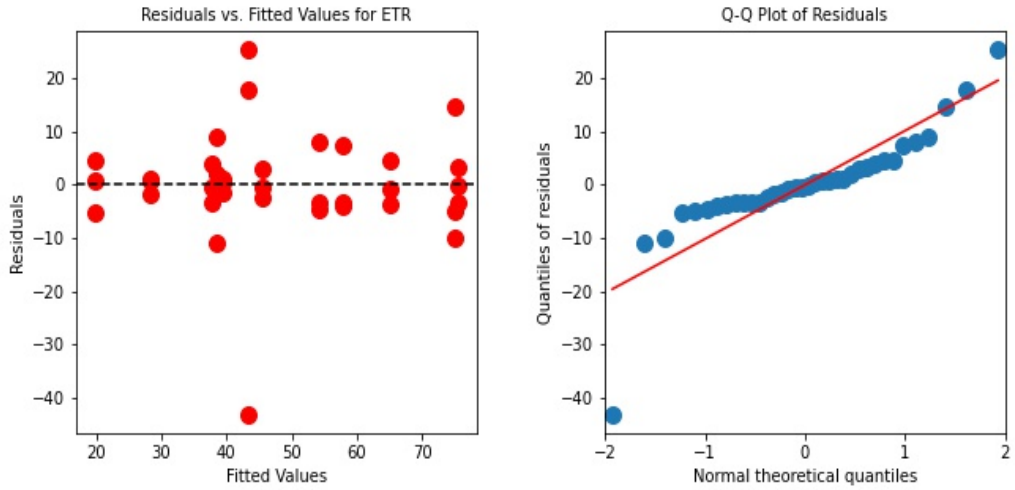


Supplementary Figure 28: Plots showing characteristics of Residuals in the fitted ANOVA model for median SSC-A parameter measurements. The description of the graph is the same as Figure 24. Residuals for almost half of the sample categories appear to be distributed in a short range, and the rest in a long range, thus deviating from the homogeneity assumption. The normal distribution also does not appear perfect.

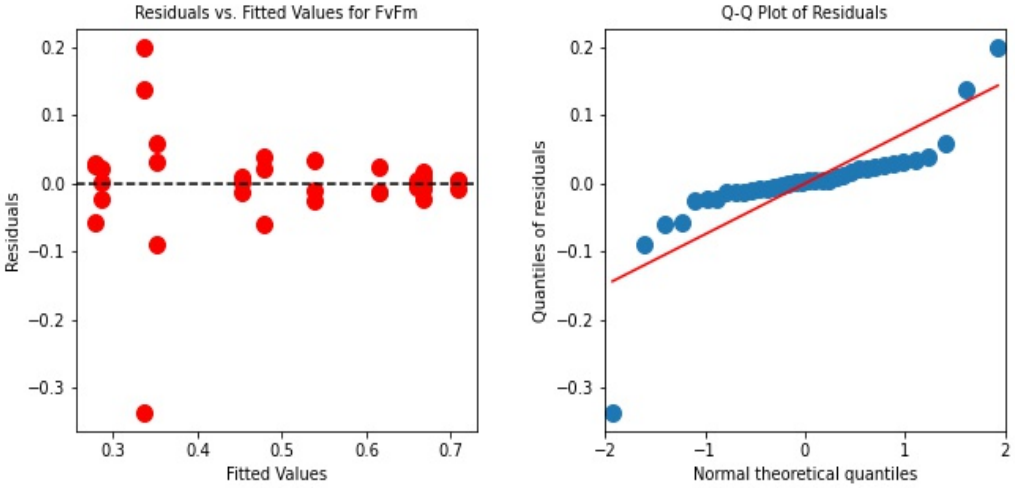
11.10 ANOVA assumptions tests for PAM measurements



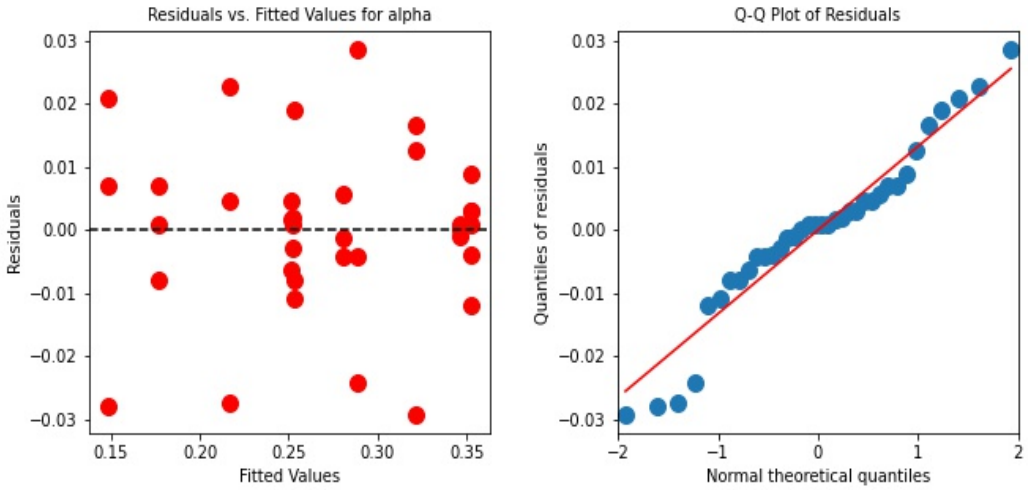
Supplementary Figure 29: Plots showing characteristics of Residuals in the fitted ANOVA model for the measured Ek parameters. The total number of sample points and thus the number of residuals will be 36, including 3 biological replicates of 4 cell lines in 3 light treatments. The number of sample categories thus fitted values is 12 combinations including 4 cell lines in 3 light conditions The 'Residuals v/s fitted values' plot(Left) indicates a strong violation of the homogeneous variance assumption for the residuals of different sample points(red circles) across the fitted values along the fitted model(dashed black line)). The q-q plot(quantile-quantile plot)(right) plotting the theoretical quantile values for a standard normal distribution along the x-axis plotted against the observed quantiles for residuals indicates a clear violation of the normal distribution assumption for the residuals in ANOVA.



Supplementary Figure 30: Plots showing characteristics of Residuals in the fitted ANOVA model for measured rETRmax parameters The description of the graph is the same as Figure 29. Although imperfect, homogeneity in distribution and normality can be observed for the residuals to some extent. One sample category shows a great difference in residual distribution.

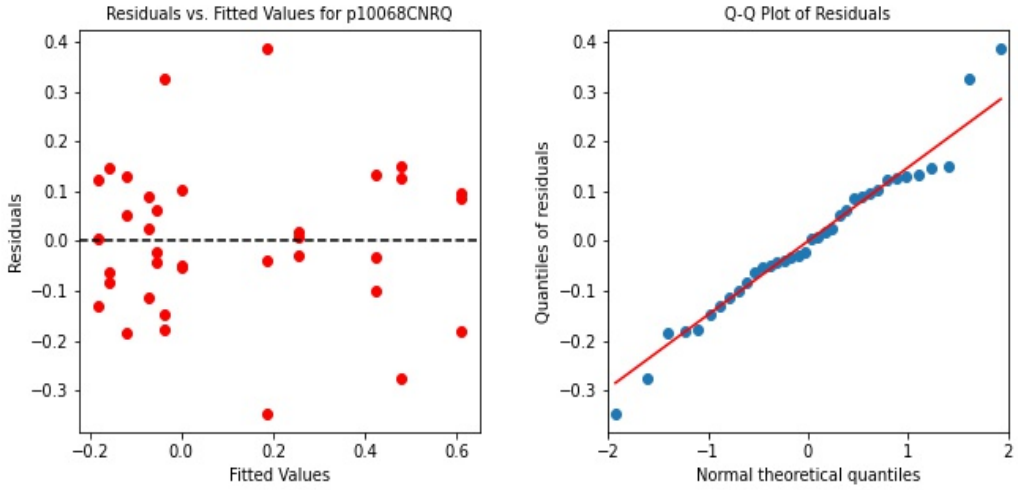


Supplementary Figure 31: Plots showing characteristics of Residuals in the fitted ANOVA model for measured Fv/Fm ratios. The description of the graph is the same as Figure 29. One sample category appears to be greatly deviating in residual distribution compared to the rest and there is a notable deviation from normal distribution.

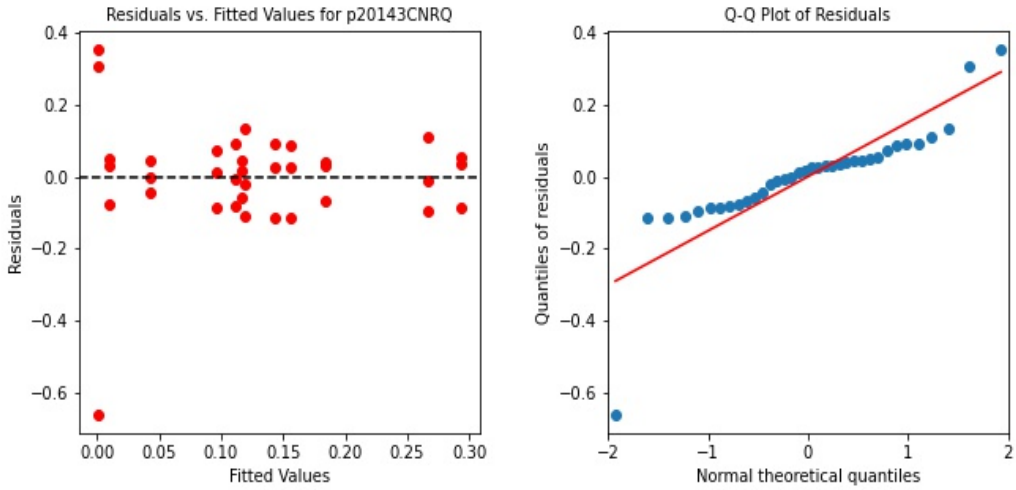


Supplementary Figure 32: Plots showing characteristics of Residuals in the fitted ANOVA model for measured alpha values. The description of the graph is the same as Figure 29. An almost homogeneous residual distribution and normal distribution can be observed.

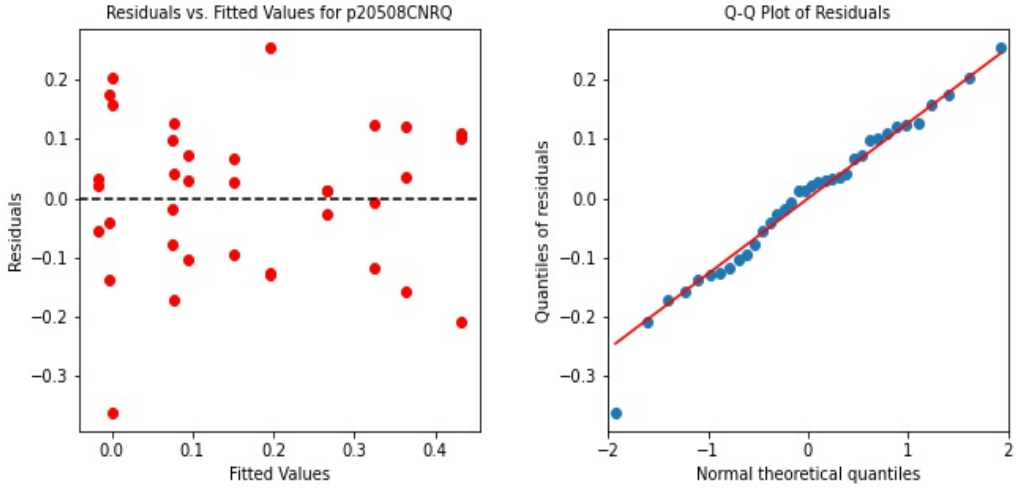
11.11 ANOVA assumptions tests for q-PCR measurements



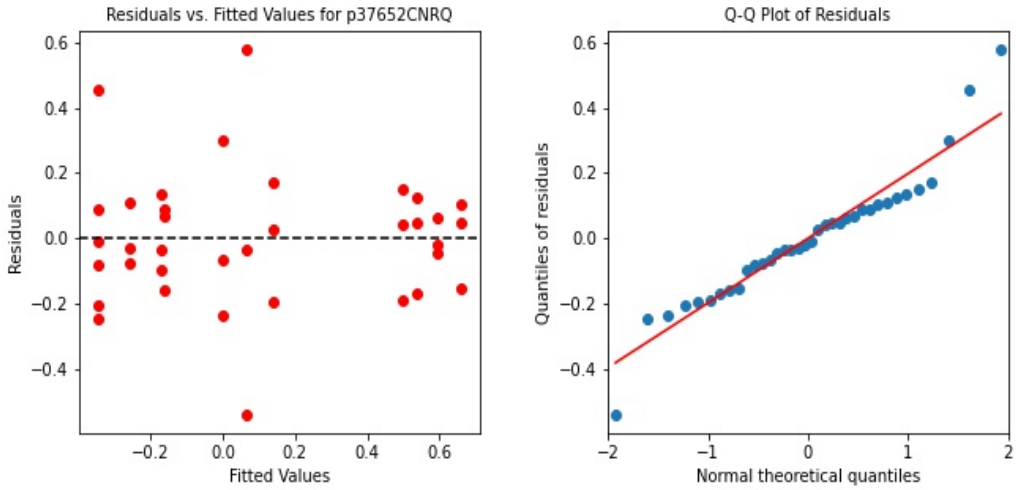
Supplementary Figure 33: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-10068 Gene. The total number of sample points and thus the number of residuals will be 36, including 3 biological replicates of 4 cell lines in 3 light treatments. The number of sample categories and thus fitted values is 12 combinations including 4 cell lines in 3 light conditions. The 'Residuals v/s fitted values' plot(Left) indicates a homogeneous variance between residuals of the different sample points(red circles) across the fitted values along the fitted model(dashed black line)). The q-q plot(quantile-quantile plot)(right) plotting the theoretical quantile values for a standard normal distribution along the x-axis plotted against the observed quantiles for residuals indicates almost normally distributed residuals in ANOVA.



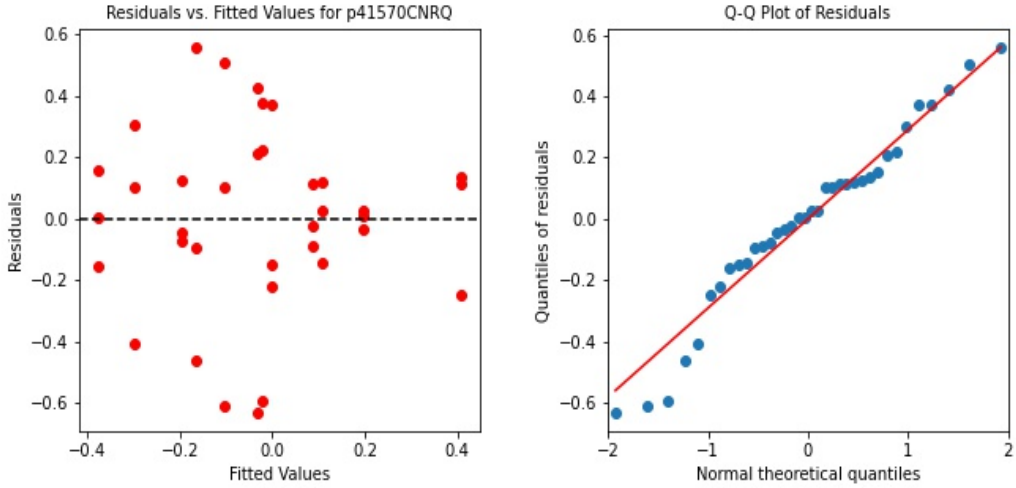
Supplementary Figure 34: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-20143 Gene. The description of the graph is the same as Figure 33. Just one of the sample categories has the residuals considerably deviating in distribution compared to the rest and there is a notable deviation from the normal distribution.



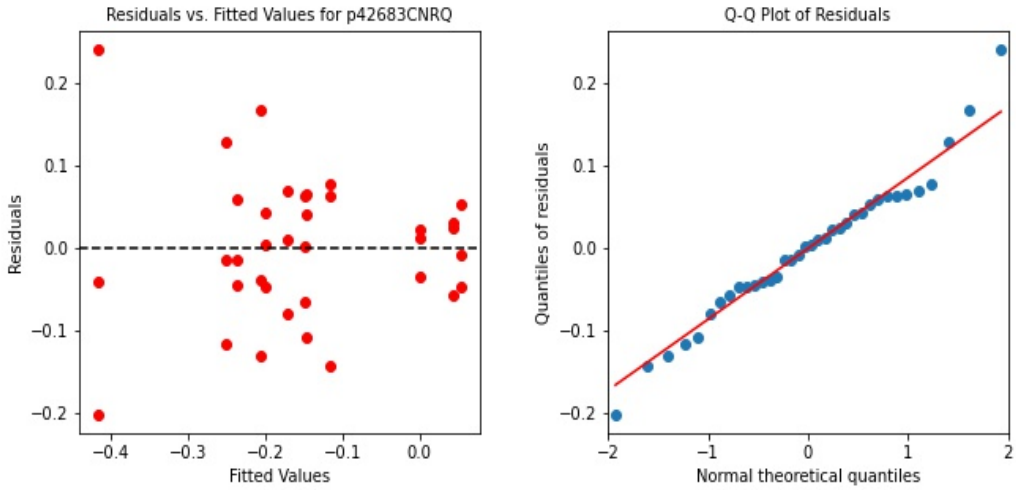
Supplementary Figure 35: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-20508 Gene. The description of the graph is the same as Figure 33. All sample categories appear to have a homogeneous distribution of residuals that are also almost normally distributed.



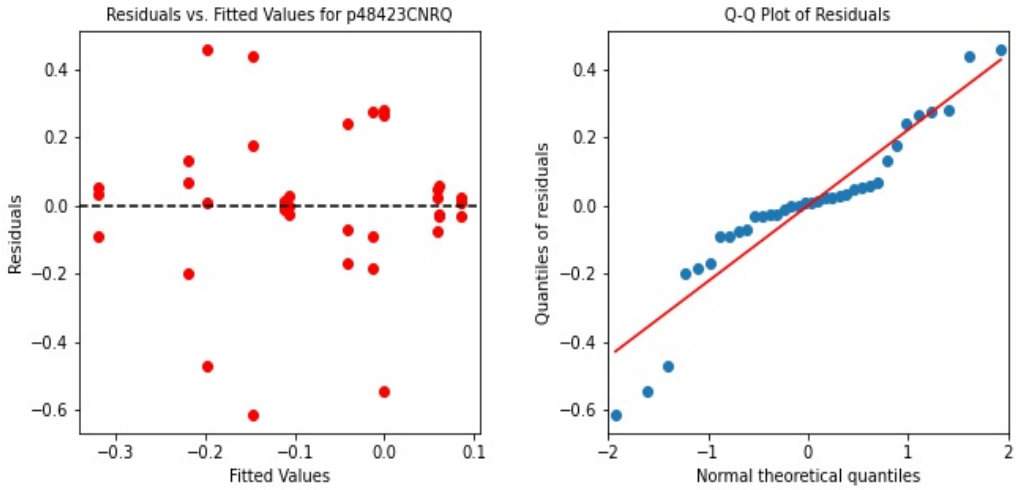
Supplementary Figure 36: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-37652 Gene. The description of the graph is the same as Figure 33. Homogeneous distribution and normal distribution of residuals observed to some extent, though not perfect,



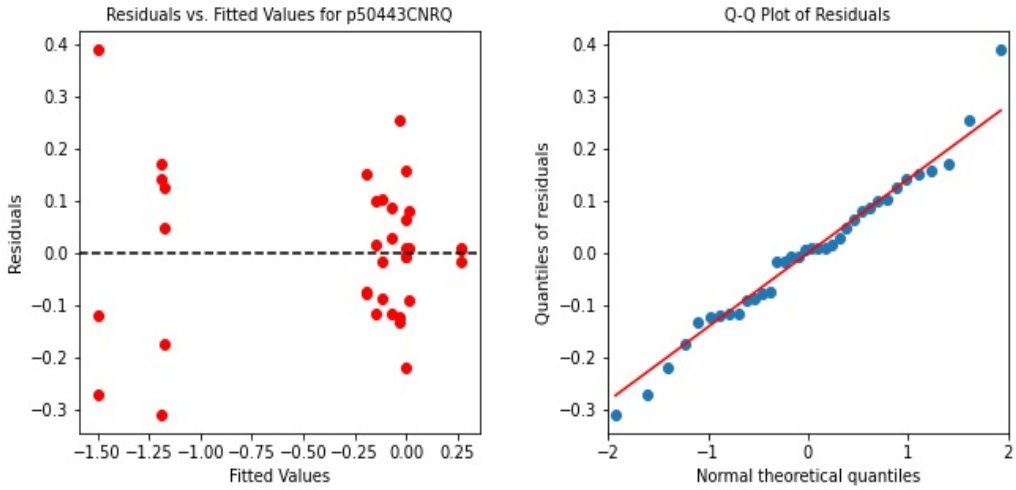
Supplementary Figure 37: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-41570 Gene. The description of the graph is the same as Figure 33. Residuals for almost half of the sample categories appear to be distributed in a short range, and the rest in a long range, thus deviating from the homogeneity assumption. An almost perfect normal distribution can also be observed.



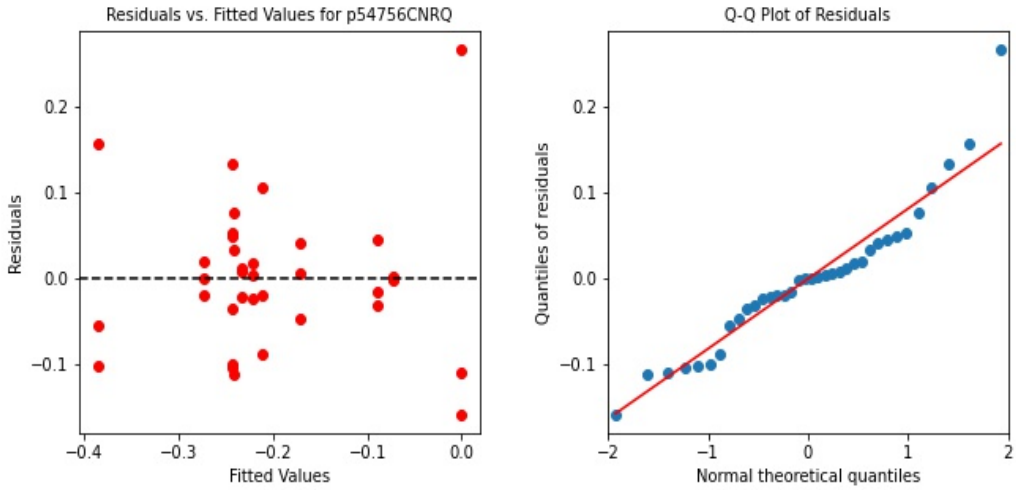
Supplementary Figure 38: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-42683 Gene. The description of the graph is the same as Figure 33. There is a deviation from homogeneous residual distribution, with short, intermediate, and long ranges among different sample categories. The residuals also appear to be almost normally distributed.



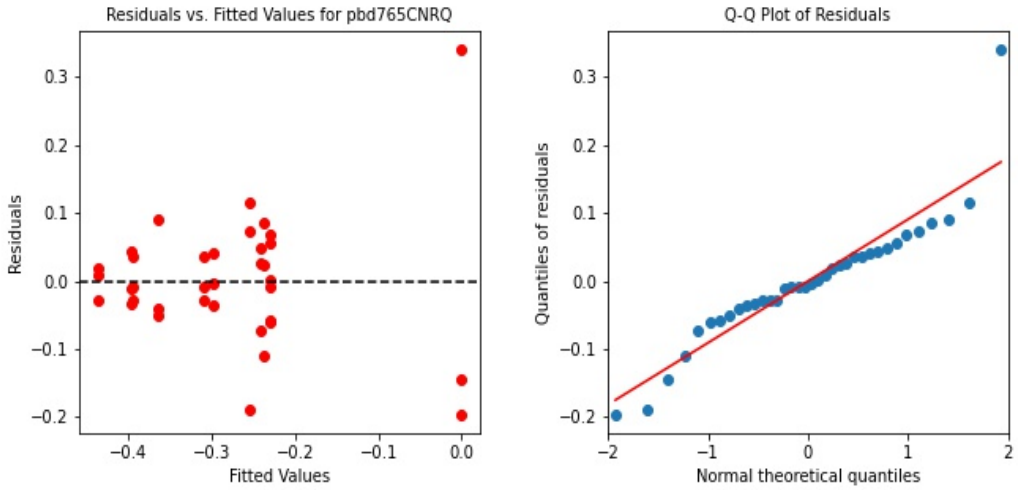
Supplementary Figure 39: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-48423 Gene. The description of the graph is the same as Figure 33. Residuals for almost half of the sample categories appear to be distributed in a short range, and the rest in a long range, thus deviating from the homogeneity assumption. Also, there is a notable deviation from the normal distribution.



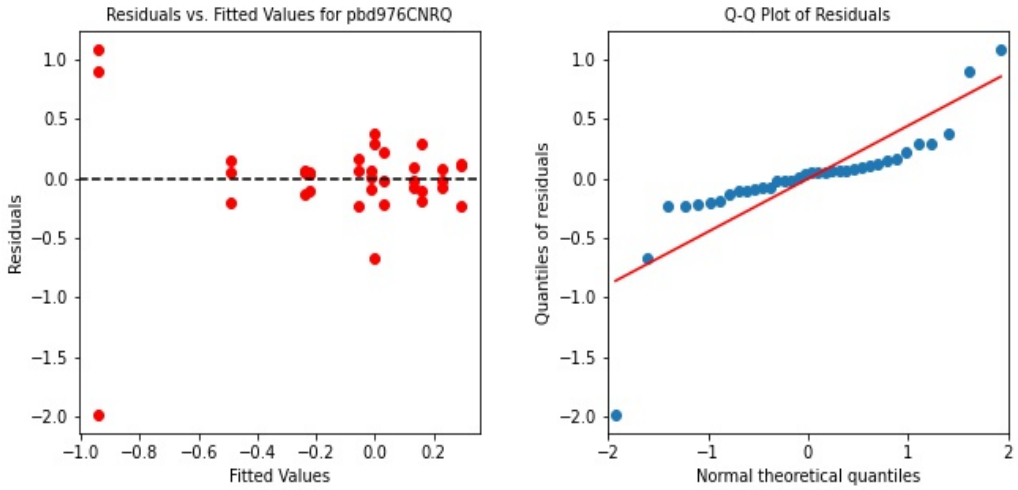
Supplementary Figure 40: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-50443 Gene. The description of the graph is the same as Figure 33. Residuals for one of the samples are distributed in a considerably short range compared to the rest. A normal distribution of residuals was observed to some extent, though not perfect,



Supplementary Figure 41: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-54756 Gene. The description of the graph is the same as Figure 33. There is a deviation from homogeneous residual distribution, with short, intermediate, and long ranges among different sample categories. The residuals also appear to be almost normally distributed.



Supplementary Figure 42: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-pbd765 Gene. The description of the graph is the same as Figure 33. There is a deviation from homogeneous residual distribution, with short, intermediate, and long ranges among different sample categories. The residuals also appear to be almost normally distributed.



Supplementary Figure 43: Plots showing characteristics of Residuals in the fitted ANOVA model for CNRQ values of the PHATRDRAFT-pbd976 Gene. Residuals for one of the samples are distributed in a considerably long range compared to the rest. A notable deviation from the normal distribution can also be observed.

11.12 Python script for the data analysis pipeline development

11.12.1 Importing packages

```
# Importing necessary packages
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import scale
import os
from scipy.stats import ttest_ind
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
os.chdir('C:\\Users\\arshad\\Downloads')
df = pd.read_excel('Lipidclassnew.xlsx','Sheet2',index_col=0)
df.drop('Total',axis=1,inplace=True)
from scipy.stats import shapiro, levene
```

11.12.2 Custom-made data preprocessing functions for the MS-MS dataset

```
# Function for mapping the dataset for extracting the needed sections of the
    data
def mapper(name):
    l=[]
    for i in df.index:
        if str(name) in i:
            l.append(i)
    return l
# Defining sample names
samples=['Alb14-LL',
      'Alb14-HL',
      'Alb16-LL',
      'Alb16-HL',
      'Alb19-LL',
      'Alb19-HL',
      'WT-LL',
      'WT-HL']
# Function for imputing outliers with the median values
def outlier_imputer(dataframe):
    for sample in samples:
        for lipid in lipids:
            dft=dataframe.loc[mapper(sample)][lipid]
            Q1 = dft.quantile(0.25)
```

```
            Q3 = dft.quantile(0.75)
            IQR = Q3 - Q1
            lower = Q1 - 1.5*IQR
            upper = Q3 + 1.5*IQR
            upper_array = np.where(dft>=upper)[0]
            lower_array = np.where(dft<=lower)[0]
            if len(dataframe.index[list(upper_array)+list(lower_array)])!=0:
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][lipid]='Nan'
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    a=mapper(dft.index[list(upper_array)+list(lower_array)][i
                        ][0:11])
                    median=np.median(dataframe.loc[a][[lipid]])
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][lipid]=median
    return dataframe
# Function for imputing zero/null values with 0.01 percent of total value of the
     lipid class measurements
def zero_imputer(sample,lipid):
    df_z=df.loc[mapper(sample)][lipid]
    indices=df_z[df_z==0].index
    df.loc[indices,lipid]=[random.uniform((0.01/100)*sum(df_z),(0.01/100)*sum(
        df_z)) for i in range(len(indices))]
```

11.12.3 Script for principle component analysis and related plots

```
# extracting a subsection of the dataset for PCA
l=[]
for i in df.index:
    if 'WT' in i or 'Alb19' in i:   #This is an example script to compare Alb3b
        -19 with the WT using PCA(This line can be changed to the desired mutant
         line)
        l.append(i)
df=df.loc[l]
print('dataframe:','\n',df)
# Performing PCA after scaling on the subsection of the dataset
scaled_df=scale(df)
pca=PCA()
pca.fit(scaled_df)
pca_data=pca.transform(scaled_df)
pca_data
per_var=np.round(pca.explained_variance_ratio_*100,decimals=1)
labels=['PC'+str(i) for i in range(1,len(per_var)+1)]
```

```
pca_df=pd.DataFrame(pca_data,index=df.index,columns=labels)
# Plotting a 2D PCA scatterplot
plt.rcParams.update({'font.size': 12})
fig = plt.suppfigure(figsize=(10,5))
albhl=plt.scatter(pca_df.loc[mapper('Alb19-HL')].PC1,pca_df.loc[mapper('Alb19-HL
    ')].PC2,marker='*',color='yellowgreen',s=150,alpha=1)
albll=plt.scatter(pca_df.loc[mapper('Alb19-LL')].PC1,pca_df.loc[mapper('Alb19-LL
    ')].PC2,marker='*',color='lime',s=150,alpha=1)
wthl=plt.scatter(pca_df.loc[mapper('WT-HL')].PC1,pca_df.loc[mapper('WT-HL')].PC2
    ,marker='o',color='peru',s=150,alpha=1)
wtll=plt.scatter(pca_df.loc[mapper('WT-LL')].PC1,pca_df.loc[mapper('WT-LL')].PC2
    ,marker='o',color='saddlebrown',s=150,alpha=1)
plt.xlabel('PC1 - {0}%'.format(per_var[0]))
plt.ylabel('PC2 - {0}%'.format(per_var[1]))
plt.legend((albhl,albll,wthl,wtll),('Albino3b-19 mutant line in high light','
    Albino3b-19 mutant line in low light',
                                    'Wild-type in high light','Wild-type in low
                                       light' ),loc='center left',
           bbox_to_anchor=(1, 0.5), ncol=1,labelspacing=1)
plt.title('Scattering at_'+ str(per_var[0]+per_var[1]) + '%_explained variance')
plt.savefig('Alb192Dscatter.jpg',bbox_inches='tight')
# Plotting a 3D PCA scatterplot
fig = plt.figure(figsize=(10,10))
plt.rcParams.update({'font.size': 13})
ax = fig.add_subplot(projection='3d')
albhl=ax.scatter(xs=pca_df.loc[mapper('Alb19-HL')].PC1,ys=pca_df.loc[mapper('
    Alb19-HL')].PC2, zs=pca_df.loc[mapper('Alb19-HL')].PC3, marker='*',
           color='yellowgreen',s=150,alpha=1)
albll=ax.scatter(xs=pca_df.loc[mapper('Alb19-LL')].PC1,ys=pca_df.loc[mapper('
    Alb19-LL')].PC2, zs=pca_df.loc[mapper('Alb19-LL')].PC3, marker='*',
           color='lime',s=150,alpha=1)
wthl=ax.scatter(xs=pca_df.loc[mapper('WT-HL')].PC1,ys=pca_df.loc[mapper('WT-HL')
    ].PC2, zs=pca_df.loc[mapper('WT-HL')].PC3, marker='o',
           color='peru',s=150,alpha=1)
wtll=ax.scatter(xs=pca_df.loc[mapper('WT-LL')].PC1,ys=pca_df.loc[mapper('WT-LL')
    ].PC2, zs=pca_df.loc[mapper('WT-LL')].PC3, marker='o',
           color='saddlebrown',s=150,alpha=1)
ax.legend((albhl,albll,wthl,wtll),('Albino3b-19 mutant line in high light','
    Albino3b-19 mutant line in low light',
                                    'Wild-type in high light','Wild-type in low
                                       light' ),loc='center left',
           bbox_to_anchor=(1.1, 0.5), ncol=1,labelspacing=1)
ax.set_xlabel('PC1 - {0}%'.format(per_var[0]),labelpad=10)
ax.set_ylabel('PC2 - {0}%'.format(per_var[1]),labelpad=10)
ax.set_zlabel('PC3 - {0}%'.format(per_var[2]),labelpad=10)
```

```
plt.title('Scattering at_'+str(np.round(per_var[0]+per_var[1]+per_var[3]))+'%
    _explained variance')
plt.savefig('Alb193Dscatter.jpg',bbox_inches='tight')
# Plotting a 2D biplot
load_df=pd.DataFrame(pca.components_,columns=df.columns,index=pca_df.columns)
PC1 = pca.fit_transform(scaled_df)[:,0]
PC2 = pca.fit_transform(scaled_df)[:,1]
ldngs = pca.components_
scalePC1 = 1.0/(PC1.max() - PC1.min())
scalePC2 = 1.0/(PC2.max() - PC2.min())
features = df.columns
plt.rcParams.update({'font.size': 12})
fig, ax = plt.subplots(figsize=(19, 9))

for i, feature in enumerate(features):
    ax.arrow(0, 0, ldngs[0, i],
             ldngs[1, i],
             head_width=0.01,
             head_length=0.01,
             color="black")
    ax.text(ldngs[0, i]*1.1,
            ldngs[1, i]*1.1 ,
            feature,color="red", fontsize=13)

    albhl=ax.scatter(pca_df.loc[mapper('Alb19-HL')].PC1*scalePC1,pca_df.loc[
        mapper('Alb19-HL')].PC2*scalePC2,marker='*',color='yellowgreen',s=150,
        alpha=1)
    albll=ax.scatter(pca_df.loc[mapper('Alb19-LL')].PC1*scalePC1,pca_df.loc[
        mapper('Alb19-LL')].PC2*scalePC2,marker='*',color='lime',s=150,alpha=1)
    wthl=ax.scatter(pca_df.loc[mapper('WT-HL')].PC1*scalePC1,pca_df.loc[mapper('
        WT-HL')].PC2*scalePC2,marker='o',color='peru',s=150,alpha=1)
    wtll=ax.scatter(pca_df.loc[mapper('WT-LL')].PC1*scalePC1,pca_df.loc[mapper('
        WT-LL')].PC2*scalePC2,marker='o',color='saddlebrown',s=150,alpha=1)
    ax.legend((albhl,albll,wthl,wtll),('Albino3b-19 mutant line in high light','
        Albino3b-19 mutant line in low light',
                                       'Wild-type in high light','Wild-type in
                                           low light' ),loc='center left',
             bbox_to_anchor=(1, 0.5), ncol=1,labelspacing=1)

    ax.set_xlabel('PC1', fontsize=20)
    ax.set_ylabel('PC2', fontsize=20)
    ax.set_title('Biplot with loading scores by each lipid at_'+ str(per_var[0]+
        per_var[1]) + '%_explained variance', fontsize=15)
plt.savefig('Alb19scatterbiplot.jpg',bbox_inches='tight')
# Finding the best loading scores
```

```
from pca import pca as new_pca
model=new_pca(normalize=True,n_components=0.95)
results=model.fit_transform(df)
results_df=results['topfeat']
print(results_df)
# Plotting a screeplot for the PCA
from bioinfokit.visuz import cluster
plt.rcParams.update({'font.size': 3})
plt.ylabel('Fraction of explained variance')
plt.xlabel('Principle components')
plt.savefig('Alb19scree.jpg',bbox_inches='tight')
# Plotting a 3D PCA scatterplot
plt.rcParams.update({'font.size': 10})
PC1 = pca.fit_transform(scaled_df)[:,0]
PC2 = pca.fit_transform(scaled_df)[:,1]
PC3= pca.fit_transform(scaled_df)[:,3]
ldngs = pca.components_

scalePC1 = 1.0/(PC1.max() - PC1.min())
scalePC2 = 1.0/(PC2.max() - PC2.min())
scalePC3= 1.0/(PC3.max() - PC3.min())
features = df.columns
plt.rcParams.update({'font.size': 13})
fig = plt.figure(figsize=(10,10),)
ax = fig.add_subplot(projection='3d')

for i, feature in enumerate(features):
    ax.quiver(0,0,0, ldngs[0, i],
              ldngs[1, i],
              ldngs[2, i],
              color="black")
    ax.text(ldngs[0, i]*1.3,
            ldngs[1, i]*1.3,
            ldngs[2, i]*1.3,
            feature,color="black", fontsize=13)
    albhl=ax.scatter(xs=pca_df.loc[mapper('Alb19-HL')].PC1*scalePC1,ys=pca_df.
        loc[mapper('Alb19-HL')].PC2*scalePC2, zs=pca_df.loc[mapper('Alb19-HL')].
        PC3*scalePC3, marker='*',
            color='yellowgreen',s=150,alpha=1)
    albll=ax.scatter(xs=pca_df.loc[mapper('Alb19-LL')].PC1*scalePC1,ys=pca_df.
        loc[mapper('Alb19-LL')].PC2*scalePC2, zs=pca_df.loc[mapper('Alb19-LL')].
        PC3*scalePC3, marker='*',
                color='lime',s=150,alpha=1)
    wthl=ax.scatter(xs=pca_df.loc[mapper('WT-HL')].PC1*scalePC1,ys=pca_df.loc[
        mapper('WT-HL')].PC2*scalePC2, zs=pca_df.loc[mapper('WT-HL')].PC3*
```

```
        scalePC3, marker='o',
                color='peru',s=150,alpha=1)
    wtll=ax.scatter(xs=pca_df.loc[mapper('WT-LL')].PC1*scalePC1,ys=pca_df.loc[
        mapper('WT-LL')].PC2*scalePC2, zs=pca_df.loc[mapper('WT-LL')].PC3*
        scalePC3, marker='o',
                color='saddlebrown',s=150,alpha=1)
    ax.legend((albhl,albll,wthl,wtll),('Albino3b-19 mutant line in high light','
        Albino3b-19 mutant line in low light',
                                      'Wild-type in high light','Wild-type in
                                          low light' ),loc='center left',
              bbox_to_anchor=(1.1, 0.5), ncol=1,labelspacing=1)

    ax.set_xlabel('Loading scores for PC1 - {0}%'.format(per_var[0]),labelpad
        =10)
    ax.set_ylabel('Loading scores for PC2 - {0}%'.format(per_var[1]),labelpad
        =10)
    ax.set_zlabel('Loading scores for PC3 - {0}%'.format(per_var[2]),labelpad
        =10)
    plt.title('3D Biplot with Scattering at_'+str(np.round(per_var[0]+per_var
        [1]+per_var[3]))+'%_explained variance and loading scores')
    plt.savefig('Alb193Dbiplot.jpg',bbox_inches='tight')
```

11.12.4 T-tests

```
# Custom-made function for  performing Welch's T-test on the MS-MS dataset
def Ttester(sample1,sample2,lipid):
    df1=np.log2(outlier_remover(sample1,lipid))
    df2=np.log2(outlier_remover(sample2,lipid))
    plt.rcParams.update({'font.size': 10})
    plt.figure(figsize=(5,5))
    fig,ax=plt.subplots(1,2)
    sns.violinplot(df1,ax=ax[0])
    sns.violinplot(df2,ax=ax[1])
    print(df1)
    print(df2)
    t_stat, p_value = ttest_ind(df1, df2,equal_var=False)
    print('Results for',sample1,'vs',sample2,':')
    print('T-statistic value:', t_stat)
    print("P-Value:", p_value)
    measures=[t_stat,p_value]
    return measures
# Defining the samples and variables
samples=[['Alb14-HL','Alb14-LL'],['Alb16-HL','Alb16-LL'],['Alb19-HL','Alb19-LL'
    ],['WT-HL','WT-LL']]
lipids=list(df.columns)
#Performing the t-test and storing the results
```

```
df_list=[]
for s in samples:
    res_list=[]
    for l in lipids:
        t_test_values=Ttester(s[0],s[1],l)
        res_list.append(t_test_values)
    res_df=pd.DataFrame(np.array(res_list),index=lipids,columns=['T-statistic','
        P-value'])
    df_list.append(res_df)
df_p=pd.DataFrame()
a=0
for i in samples:
    df_p[str(i[0])+'_v/s_'+str(i[1])]=df_list[a]['P-value']
    a=a+1
df_T=pd.DataFrame()
a=0
for i in samples:
    df_T[str(i[0])+'_v/s_'+str(i[1])]=df_list[a]['T-statistic']
    a=a+1
# Plotting the stored Welch's T-test results
a=0
for i in samples:
    mpl.rcParams['lines.markersize'] = 12
    plt.rcParams.update({'font.size': 13})
    plt.figure(figsize=(10,6))
    plt.title('T_statistic v/s P_value graph for '+ i[0]+'_vs_'+i[1])
    g=sns.scatterplot(data=df_list[a],x=df_list[a]['T-statistic'],y=df_list[a]['
        P-value'],hue=df_list[a].index,marker='o')
    plt.axhline(0.05,linestyle="--",color='k')
    plt.text(7,0.048,'P_value=0.05')
    plt.axvline(0.047)
    g.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1,labelspacing=1)
    a=a+1
    plt.savefig('scatterinf'+str(i)+'.jpeg',bbox_inches='tight')
```

11.13 Shapiro Wilk's tests

```
# Creating an empty data frame to store p-values
shapiro_p_values_df = pd.DataFrame(columns=['cell_type', 'condition', 'lipid', '
    Shapiro-Wilk P-Value'])

# Iterating over each lipid class for running the test
for lipid in lipids:
    for cell_type in df['cell_type'].unique():
        for condition in df['condition'].unique():
```

```
            # Extract the data for the current combination
            data = df[(df['cell_type'] == cell_type) & (df['condition'] ==
                condition)][lipid]
            # Shapiro-Wilk test
            shapiro_stat, shapiro_p_value = shapiro(data)
            # Add the results to the p-values data frame
            shapiro_p_values_df = shapiro_p_values_df.append({
                'cell_type': cell_type,
                'condition': condition,
                'lipid': lipid,
                'Shapiro-Wilk P-Value': shapiro_p_value
            }, ignore_index=True)

# Display or use shapiro_p_values_df as needed
print(shapiro_p_values_df)
# Plotting the Shapiro results
import matplotlib as mpl
for ct in df.cell_type.unique():
    shapiro_df=shapiro_p_values_df[shapiro_p_values_df['cell_type']==ct]
    mpl.rcParams['lines.markersize'] =7.5
    plt.rcParams.update({'font.size': 12})
    plt.figure(figsize=(10,5))
    plt.title('Results from Shapiro wilk\'s test for '+str(ct))
    plt.ylabel('P-values')
    hl=plt.scatter(x=lipids,y=shapiro_df[shapiro_df['condition']=='HL']['Shapiro
        -Wilk P-Value'],marker='x',color='y')
    ll=plt.scatter(x=lipids,y=shapiro_df[shapiro_df['condition']=='LL']['Shapiro
        -Wilk P-Value'],marker='x',color='g')
    plt.legend((hl,ll),('High light samples','Low light samples'),bbox_to_anchor
        =(1.30, 0.6), ncol=1,labelspacing=1)
    plt.axhline(0.05,linestyle="--",color='k')
    plt.text(7,0.06,'P_value=0.05')
    plt.xticks(rotation=90)
    plt.savefig('shapiro'+str(ct)+'.jpeg',bbox_inches='tight')

# Creating an empty DataFrame to store the p-values
levene_p_values_df = pd.DataFrame(columns=['lipid', 'cell_type', 'Levene_P_Value
    '])

# Iterating through each lipid class for running the Levene's test
for lipid in lipids:
    for cell_type in df['cell_type'].unique():
        # Splitting the data into the two light conditions
        condition_ll = df[(df['cell_type'] == cell_type) & (df['condition'] == '
            LL')][lipid]
        condition_hl = df[(df['cell_type'] == cell_type) & (df['condition'] == '
```

```
            HL')][lipid]

        # Perform Levene's test
        stat, p_value = levene(condition_ll, condition_hl)

        # Appending the results to the p-values DataFrame
        levene_p_values_df = levene_p_values_df.append({'lipid': lipid, '
            cell_type': cell_type, 'Levene_P_Value': p_value}, ignore_index=True
            )

# Print or use levene_p_values_df as needed
print(levene_p_values_df)
# Plotting the Levene's test results

for ct in df.cell_type.unique():
    levene_df=levene_p_values_df[levene_p_values_df['cell_type']==ct]
    mpl.rcParams['lines.markersize'] =7.5
    plt.rcParams.update({'font.size': 12})
    plt.figure(figsize=(10,5))
    plt.title('Results from Levene\'s test for '+str(ct)+ '(low light v/s high
        light)')
    plt.ylabel('P-values')
    plt.scatter(x=lipids,y=levene_df['Levene_P_Value'],marker='*',color='red')
    plt.axhline(0.05,linestyle="--",color='k')
    plt.text(7,0.06,'P_value=0.05')
    plt.xticks(rotation=90)
    plt.savefig('levene'+str(ct)+'.jpeg',bbox_inches='tight')
```

11.14 Python scripts for the ANOVA and post hoc analyses on data from labwork

11.14.1 ANOVA and post hoc analyses of flow cytometry data

```
#Importing packages
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import scipy.stats as stats
import seaborn as sns
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
os.chdir("C:\\Users\\arshad\\Downloads")
import warnings
warnings.filterwarnings('ignore')
#loading the dataset
df=pd.read_csv('BODIPY COMPILED - Sheet1.csv',index_col=0)
df.drop(['All_Abs._Count','M1_Abs._Count'],axis=1,inplace=True)
# Defining the samples
samples=['Alb14ll','Alb14hl','Alb14ml','Alb16ll','Alb16hl','Alb16ml','Alb19ll','
    Alb19hl','Alb19ml','WTLL','WTHL','WTML']
# Making a custom mapping fucntion
def mapper(name):
    l=[]
    for i in df.index:
        if str(name) in i:
            l.append(i)
    return l
#Defining variables
Variables=df.columns[0:5]
# Imputing outliers with mean values using custom-made IQR rule-based function
def outlier_imputer(dataframe):
    for sample in samples:
        for variable in Variables:
            dft=dataframe.loc[mapper(sample)][variable]
            Q1 = dft.quantile(0.25)
            Q3 = dft.quantile(0.75)
            IQR = Q3 - Q1
            lower = Q1 - 1.5*IQR
            upper = Q3 + 1.5*IQR
            upper_array = np.where(dft>=upper)[0]
            lower_array = np.where(dft<=lower)[0]
            if len(dataframe.index[list(upper_array)+list(lower_array)])!=0:
```

```
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][variable]='Nan'
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    a=mapper(dft.index[list(upper_array)+list(lower_array)][i
                        ][0:11])
                    mean=np.mean(dataframe.loc[a][[variable]])
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][variable]=mean
    return dataframe
df=outlier_imputer(df)
#Performing ANOVA on dataset, Shapiro Wilk's test and Levenes's test on the
    residuals, and storing the results
anova_df=pd.DataFrame(index=Variables,columns=['P-value(cell_type)','P-value(
    treatment)',
                                               'P-value(interaction)','P-value(
                                                   shapiro)',
                                              'P-value(levene)'])
for i in Variables:
    test_data=df[[str(i),'cell_type','treatment']]
    #test_data[str(i)]=np.log2(test_data[str(i)])
    model = ols(str(i) +'~ cell_type*treatment', data=test_data).fit()
    anova_table = anova_lm(model, typ=2)
    cell_type_sig=anova_table['PR(>F)'][0]
    light_sig=anova_table['PR(>F)'][1]
    interaction=anova_table['PR(>F)'][2]
    anova_df.loc[i]['P-value(cell_type)']=cell_type_sig
    anova_df.loc[i]['P-value(treatment)']=light_sig
    anova_df.loc[i]['P-value(interaction)']=interaction
    residuals = model.resid
    shapiro_test_statistic, shapiro_p_value = stats.shapiro(residuals)
    anova_df.loc[i]['P-value(shapiro)']=shapiro_p_value
    fitted_values = model.fittedvalues
    plt.figure(figsize=(10, 15))
    fig,(ax1,ax2)=plt.subplots(1,2,figsize=(10,5))
    fig.tight_layout(pad=5.0)
    ax1.scatter(fitted_values, residuals, c='r', marker='o')
    ax1.axhline(y=0, color='k', linestyle='--')
    ax1.set_title('Residuals vs. Fitted Values for '+str(i),fontsize=10)
    ax1.set_xlabel('Fitted Values')
    ax1.set_ylabel('Residuals')
    sm.qqplot(residuals, line='s',ax=ax2)
    ax2.set_title('Q-Q Plot of Residuals',fontsize=10)
```

```
    ax2.set_ylabel('Quantiles of residuals')
    ax2.set_xlabel('Normal theoretical quantiles')
    plt.savefig(str(i)+'qq.jpeg')
    #sns.displot(residuals,ax=ax2)

    from bioinfokit.analys import stat
    res=stat()
    res.levene(df=test_data,res_var=str(i),xfac_var=['cell_type','treatment'])
    anova_df.loc[i]['P-value(levene)']=res.levene_summary['Value'][2]

print(anova_table)

print(anova_df)
# Performing Tukey-HSD analyses and storing the data
for i in Variables:
    test_data=df[[str(i),'cell_type','treatment']]
    from bioinfokit.analys import stat
    res = stat()
    res.tukey_hsd(df=test_data, res_var=str(i), xfac_var=['cell_type','treatment
        '], anova_model=str(i)+'~C(cell_type)+C(treatment)+C(cell_type):C(
        treatment)')
    print('post hoc results for_'+ str(i))
    tukey_df=pd.DataFrame(res.tukey_summary[['group1','group2','p-value']])
    print(tukey_df[tukey_df['p-value']<0.05])
    sig_df=tukey_df[tukey_df['p-value']<0.05]
    sig_df.to_excel(str(i)+'tukeyhsd.xlsx')
```

11.14.2 ANOVA and post hoc analyses of PAM data

The same packages used in section 11.14.1 are imported as the first step.

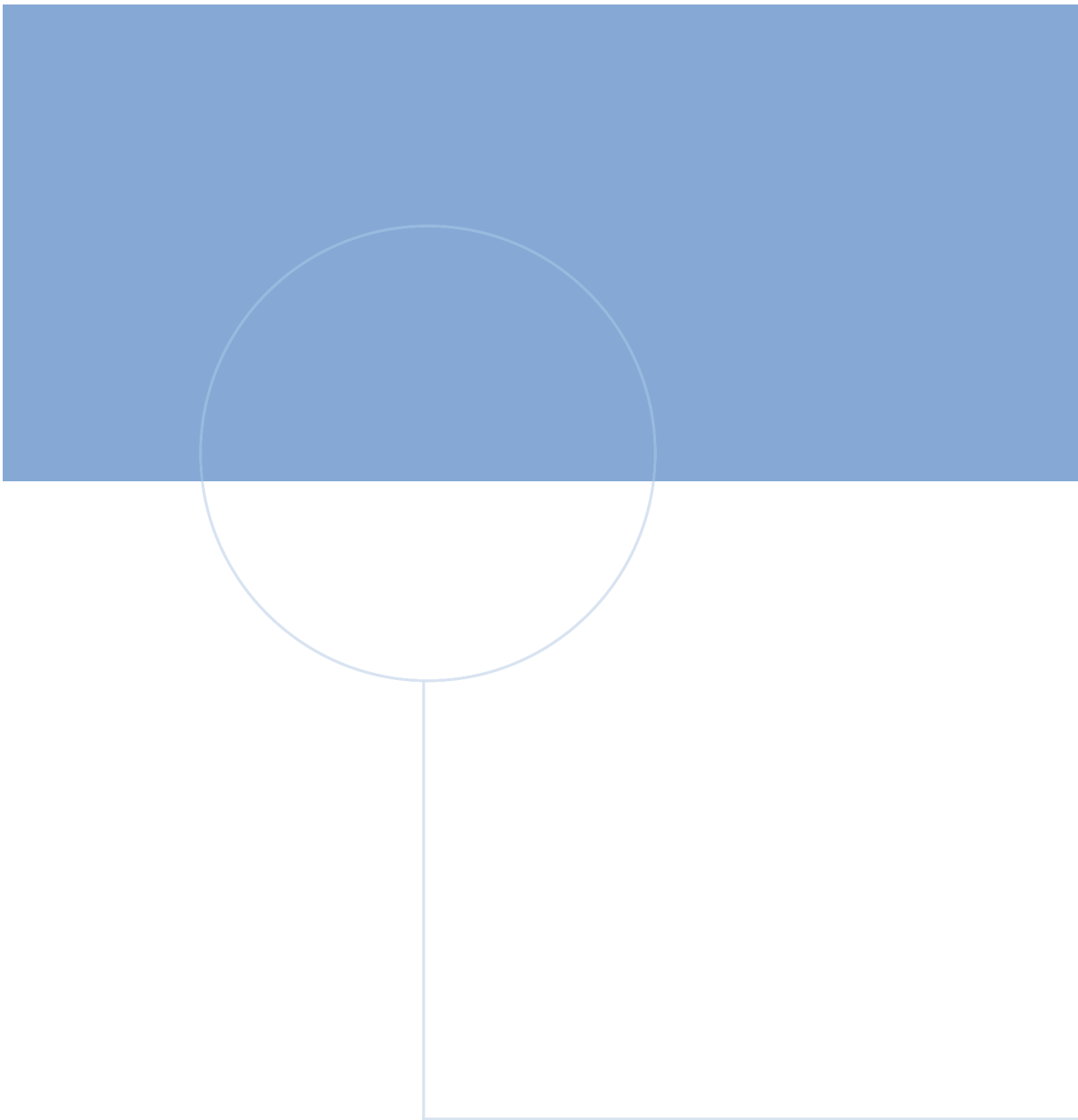
```
#loading the dataset
df=pd.read_csv('PAM - Sheet4.csv',index_col=0)
# Removing unnecessary variables(including NPQ in the first part as the analyses
    is separate for NPQ)
df.drop(['Fo','Fm','Fv','factor','NPQ'],axis=1,inplace=True)
#Defining the samples
samples=['Alb14LL','Alb14HL','Alb14ML','Alb16LL','Alb16HL','Alb16ML','Alb19LL','
    Alb19HL','Alb19ML','WTLL','WTHL','WTML']
# Making a custom mapping function
def mapper(name):
    l=[]
    for i in df.index:
        if str(name) in i:
            l.append(i)
    return l
```

```
#Defining variables
Variables=df.columns[0:5]
# Imputing outliers with mean values using custom-made IQR rule-based function
def outlier_imputer(dataframe):
    for sample in samples:
        for variable in Variables:
            dft=dataframe.loc[mapper(sample)][variable]
            Q1 = dft.quantile(0.25)
            Q3 = dft.quantile(0.75)
            IQR = Q3 - Q1
            lower = Q1 - 1.5*IQR
            upper = Q3 + 1.5*IQR
            upper_array = np.where(dft>=upper)[0]
            lower_array = np.where(dft<=lower)[0]
            if len(dataframe.index[list(upper_array)+list(lower_array)])!=0:
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][variable]='Nan'
                for i in range(len(dataframe.index[list(upper_array)+list(
                    lower_array)])):
                    a=mapper(dft.index[list(upper_array)+list(lower_array)][i
                        ][0:11])
                    mean=np.mean(dataframe.loc[a][[variable]])
                    dataframe.loc[dft.index[list(upper_array)+list(lower_array)
                        ][i]][variable]=mean
    return dataframe
#Performing ANOVA on the dataset, Shapiro Wilk's test and Levenes's test on the
    residuals, and storing the results
(# The same script as in section 10.11.1 for flow cytometry data is used for
    ANOVA of PAM data)
# Performing Tukey-HSD analyses and storing the data
(# The same script as in section 10.11.1 for flow cytometry data is used for
    Tukey-HSD analyses of PAM data)

# One-way ANOVA for NPQ
# Realoading the data
df=pd.read_csv('PAM - Sheet4.csv',index_col=0)
#Perform outlier imputation with the previous function
df=outlier_imputer(df)
# Making a new dataset with just NPQ values for LL-treated samples
df_npq=df[['NPQ','cell_type','treatment']]
df=df_npq[df_npq['treatment']=='LL']
#Performing ANOVA and Tukey HSD
from scipy.stats import f_oneway
```

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
anova_result = f_oneway(df['NPQ'][df['cell_type'] == 'Alb14'],
                        df['NPQ'][df['cell_type'] == 'Alb16'],
                        df['NPQ'][df['cell_type'] == 'Alb19'],
                        df['NPQ'][df['cell_type'] == 'WT'])

# Print ANOVA test result
print("ANOVA Test Result:")
print("F-statistic:", anova_result.statistic)
print("p-value:", anova_result.pvalue)
# Perform Tukey's HSD post hoc test
tukey_results = pairwise_tukeyhsd(df['NPQ'], df['cell_type'])
# Print Tukey's HSD post hoc test results
print("\nTukey's HSD Post Hoc Test Results:")
print(tukey_results)
```



NTNU

Norwegian University of
Science and Technology