Metabolic Network Reconstruction of Algal Strains of *Chlorella species*

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Arif Saeed

aus

Lahore, Pakistan

Bonn 2025

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer: Prof. Dr. rer. nat. Heiko Schoof

Gutachter: Prof. Dr. rer. nat. Jan Hasenauer

Tag der Promotion: 12.07.2024

Erscheinungsjahr: 2025

ABSTRACT OF THE THESIS

Chlorella is a photosynthetic, eukaryotic microalgae that has received profound interest as a prospective feed source for the production of biofuels. For efficient biotechnological application, the high light tolerance and thereby accelerated growth of some species is of special interest. To uncover the genetic basis of high light tolerance, genome metabolic network reconstruction of high light sensitive and tolerant strains was performed.

For this thesis research work, four *Chlorella* strains are used for a comparative analysis. Three strains were high light tolerant while one was high light sensitive. The genomes of these algal strains were annotated with gene ontology terms and EC numbers by using InterProScan, InterPro2GO and EC2GO mapping.

With *Arabidopsis thaliana* as reference genome, ortholog prediction was conducted by using OrthoMCL. BLASTP was used for sequence similarity analysis, whereas for pathway mapping KEGG Pathways mapping tool was used.

Pathway coverage of shared and strain-specific genes was analysed. The results show that in the Oxidative Phosphorylation pathway there is a missing gene in the sensitive strain (Cv11b). This gene belongs to NADH dehydrogenase, Complex-I and functions in the transfer of electrons from NADH to the respiratory chain, while it is present in tolerant strains. The glycerophospholipid metabolism pathway also shows a missing enzyme, PSD3, in the sensitive strain (Cv11b). This gene is present in all tolerant strains. Subsequently, synteny analysis was conducted for the PSD genomic region. The alignment of Cv264 and Cv11b algal strains showed

that PSD3-neighbouring genes were located on contig node-207 of the sensitive algal strain Cv11b, however flanking genes to one side of the expected PSD3 location were found in reverse order and orientation in Cv11b. However, the PSD3 gene itself could not be mapped to the Cv11b genome sequence. This indicates that the genomic region of PSD3 is sequenced and assembled in Cv11b, however the PSD3 gene has presumably been deleted.

These differences in metabolic networks could be candidates for further studies regarding their potential to enhance lipid production by using metabolic engineering. Especially PSD3 appears to be an interesting target for high light tolerance as it may be linked to lipid production and the capacity of membranes to maintain function under high light stress.

Acknowledgement

"Praise be to Allah (God)"

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Heiko Schoof, for the continuous support of my Ph.D. study, for his patience, encouragement, and immense knowledge, without which I would not have reached here. His valuable guidance helped me at all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D. study.

Besides my supervisor, I would also like to thank Prof. Dr. Jan Hasenauer. It is an honour to have him as the second supervisor for my thesis. My sincere gratitude also goes to Prof. Dr. Peter Dörmann and Prof. Dr. Gabriel Schaaf for agreeing to be part of my doctoral committee.

My special thanks go to Lena Altrogge for her valuable support during my entire PhD research work. I am also grateful to Dr. Florian Boecker for his valuable support in reviewing my dissertation. My sincere thanks also go to all my colleagues from my lab for contributing to good scientific research, thanks a lot for giving me a new experience in a wonderful working environment.

I am indebted to my parents for believing in me.

Last but not the least, I would like to thank my family and friends for always being supportive and encouraging throughout my studies. Thanks for all your encouragement!

Contents

Chapter 1

Introduction and Literature Review

1. Microalgae:

1.1. A Good Source of Biofuel

Microalgae have been known as a great source of lipids for health metabolites, protein and biofuels including vitamins, antioxidants and polyphenols [1]. *Chlorella* species has gathered substantial attention, as for its comparatively high nutritional value, its capability to change its metabolites with variations in its growth medium, reasonably rapid rates of reproduction and having a thick cell wall that guards its nutrients [2].

A wide range of metabolites can be produced by microalgae (like, *Chlorella*), even with growth under stress environments [3]. It is also established knowledge that various types of metabolites can be yielded by different types of microalgae [4]. In an optimal growth environment, the relative profiles and concentrations of different metabolites in microalgae remain the same. Though, in a sub-optimal growth environment, the metabolite profile changes considerably. Algal species use various approaches to manage these environmental changes. According to their capability to cope with different types of stresses, the microalgae produce a diverse range of secondary metabolites to enhance their probability of survival [4].

Several studies regarding the consequences of stress upon the microalgal metabolome have been established in literature [5,6]. However, our knowledge about microalgal molecular level response to physiological stress is mainly limited to model organisms, and the pertinent pathways have also not been fully established.

Likewise, microalgal appearance and size can be altered significantly according to the associated types and levels of stress and their environmental conditions. Foregoing studies proposed that, depending on different growth and environmental conditions, the colour of *Chlorella* can be altered from green to yellow or red, as a result of changes in pigment production [7,8,9,10].

Comparative transcriptome profiling can help to identify phenotypic details, like the effect of stress on the making of metabolites and pigments that play a role to microalgae survival [11,12]. Transcriptomic study is an appropriate approach, in a microalgal stress response, which offers a preliminary and comprehensive assessment of the derived metabolite pathways regulation.

Up to now, studies have concentrated on metabolite content screening and growth experiments, but regarding gene expression have generated partial information, in microalgae under stress and normal conditions [13,14,15,16].

With the identification of multiple new categories of RNA molecules like gene regulatory, along with transcriptional regulatory as well as protein based [17], transcriptome sequencing can offer a significant approach to find microalgae

functional genomics knowledge. Hence, it is essential to study the transcriptomic profile along with *Chlorella* metabolite composition.

It is significant and well-timed to identify the right prospective of these species and to establish the potential for microalgal genetic engineering, because they are getting a significant focus as substitute source of biofuel, health supplements and food items.

Global warming and the rising need for energy are a couple of big challenges confronting modern-day society. Depending only on fossil fuels to fulfil increasing demand of energy is unsustainable, because of growing levels of utilisation and a lack of novel sources for these non-renewables. Therefore, this issue has inspired researchers to identify unconventional energy sources like biomass, wind, solar, and water.

Biofuels are formed from sugar, cellulosic, starch or lipid-rich substrates, and they are good alternatives to liquid fossil fuels. Thus, biofuels are derived from feedstocks like cereal crops, involving wheat and corn [18]; sugar crops, like sugarcane [19]; energy crops, like switchgrass [20,21]; agricultural wastes, like straws [22–25]; and numerous aquatic species. Presently, ethanol is produced from sugarcane and corn in significant volumes as a supplemental fuel. Ethanol use as a transportation energy results in a decline of emission of greenhouse gas.

1.2. Algal Production:

1.2.1. Effects of Environmental Factors

Whether in closed photobioreactors or in open ponds, culturing algae demands consideration of several environmental conditions. Environmental factors such as *light, temperature, nutrients and pH level,* not only affect photosynthesis and rate of growth of the algae but also affect the endeavour of cellular composition and metabolism. During photosynthesis, algae produce carbohydrates, proteins and lipids, by only using light and nutrients. The relative volumes of these metabolic products are closely linked to environmental as well as nutrient conditions containing: the intensity and amount of CO_2 concentrations, sunlight, temperature, and the presence of other organisms.

1.2.2. Effects of Light on Algal Growth

Light is used as the energy source for the photoautotrophic development stage and plants use the energy of light to produce organic compounds (like, sugars) from carbon dioxide. The spectrum of light intensity changes with sturdy local and seasonal reliance [26]. Intensity of light stimulates algal progression through its influence on photosynthesis [27].

The algal growth rate is highest at saturation of light intensity and declines in both conditions either increase or decrease in intensity of light [28]. According to the accessibility of light and an escalation in photosynthetic proficiency, the process of photoadaptation in algae leads to alterations in cell properties [29].

These alterations may occur across various mechanisms like modifications in quantities and types of rates of growth, rate of dark respiration, pigments or the accessibility of necessary fatty acids [30]. Algae overwhelm light constraint by chloroplast membrane desaturation [31]. Intensity of light upsurge above drenching edges instigates photoinhibition [32,33]. This is because the interruption of the chloroplast lamellae triggered by high light strength [34] besides deactivation of enzymes concerned in C_2O complex [35]. It is like, growth rate of *Dunaliella viridis* (green algae) declined to sixty percent with rise in light concentration from seven hundred to fifteen hundred micromole per square meter for each second [32].

Intensity of light influences the cellular arrangement of algae. Green alga *Dunaliella* reveals a decline in protein substance and a rise in content of the lipid as light intensities rise up to the drenching level [36]. Alike outcomes were stated in a study that low-light led to a rise in the protein synthesis ratio [37]. High light intensity has been detected to outcome in enhanced polysaccharide content in extracellular region, while low light concentration results in more protein material [35]. Lack of light was detected to upsurge the entire lipid substance of the *Dunaliella virdis* but decrease sterols, triglycerides, and abandoned fatty acids [38].

Generally, high light indications impairment in polyunsaturated fatty acids oxidation. Several studies have proposed that the lipid substance of a cell and polyunsaturated fatty acids is reduced under more intense light [39-41]. Contrarily, cells of *Nannochloropsis* species were shown to have increased lipid content but reduced percentage of eicosapentaenoic acid under high light [42]. Validating this observed trend, another study described an upsurge in unsaturated fatty acids under low light primarily because of a rise in eicosapentaenoic acid and reduction in protein matter [30]. Under light-restricted growth circumstances, upsurge in polyunsaturated fatty acids is linked to a growth of cellular thylakoid membrane [43].

Additionally, the composition range of light cycles influence algae production. Moreover, research assessed the effect of dark and light rotations on the algal growth. It is detected that with intensifying the density of photon-flux, growth rate rises up to a certain threshold value of photon flux density, after that growth rate was declined [44]. It is also described that persistent high-light concentrations could be a reason of photoinhibition and minimise efficient use of light. Under high light circumstances, efficient use of light may be elevated by extending the dark period. This permits the photosynthesis equipment of the cell to completely operate and amend the photons into chemical energy after capturing them [45].

1.3. Systems Biology

Systems biology is a field of study that links biological systems information with computing and mathematical approaches. Systems Biology emphasises on the collaboration of the discrete components and it aims to comprehend the whole system. One module could be the group of biochemical reactions and other module maybe accountable for the metabolic processes that regulate the functionality of the cell, like the metabolic network.

By considering metabolic as well as genetic organisation, we can estimate phenotypical characters initiated due to variations of the metabolic network or genome. It may facilitate to find essential transporters or enzymes that can be prospective targets for new drugs. It can assist the pathways optimisation which can be accountable to produce specific compounds for biotechnological purposes.

This knowledge can be attained by the metabolic network modelling, a practise called as metabolic reconstruction. Specifically, by using signals from the genome sequence, metabolic reconstruction is the progression of composing a metabolic network map.

A promising methodology of network reconstruction is established by manual curation. Researcher for an explicit organism practises various accessible resources altogether with literature and experimental outcomes and afterwards review manually all the metabolic network annotations. This comprises steps like performing laboratory experiments and literature search, to discover proof that

accepts or rejects every annotation. It is a very time-consuming procedure. The growing pace to sequence a genome of any organism, helps bioinformatics to create significant source of information.

Presently, automatic role allocations are still mostly accomplished using similarity sequence approaches. The similarity sequence searches work satisfactory, if an annotation of a relevant organism is available already, while for many species, this may not be sensitive enough to identify all the enzymes, like as *Chlorella sp*ecies.

Enzymes can have some more tasks that cannot be signified in sequence databases, some practical analogues to other irrelevant proteins catalysing the similar reaction, or merely have deviated too far to be detectable. Concisely, further remotely linked proteins, where merely specific sequence characters or fundamental motifs are preserved, the resemblance between a pair of proteins cannot clearly be recognisable by pairwise alignment approaches, not even by other more perceptive profile-based techniques [46].

Due to this difficulty for assignment of certain enzymatic functions, preliminary metabolic reconstruction generally harvests networks with many gaps of significant reactions for a comprehensive biochemical pathway, though for that no enzyme has annotated in the genome. The presence of a gap in a pathway can be because of various reasons. It may be initiated by a fault or the gene is not yet or incorrectly annotated.

In the genomes of microbes, around three hundred pathway gaps are anticipated, whereas a large number of the gaps are supposed to be the consequence of a failure to find the accurate gene [47].

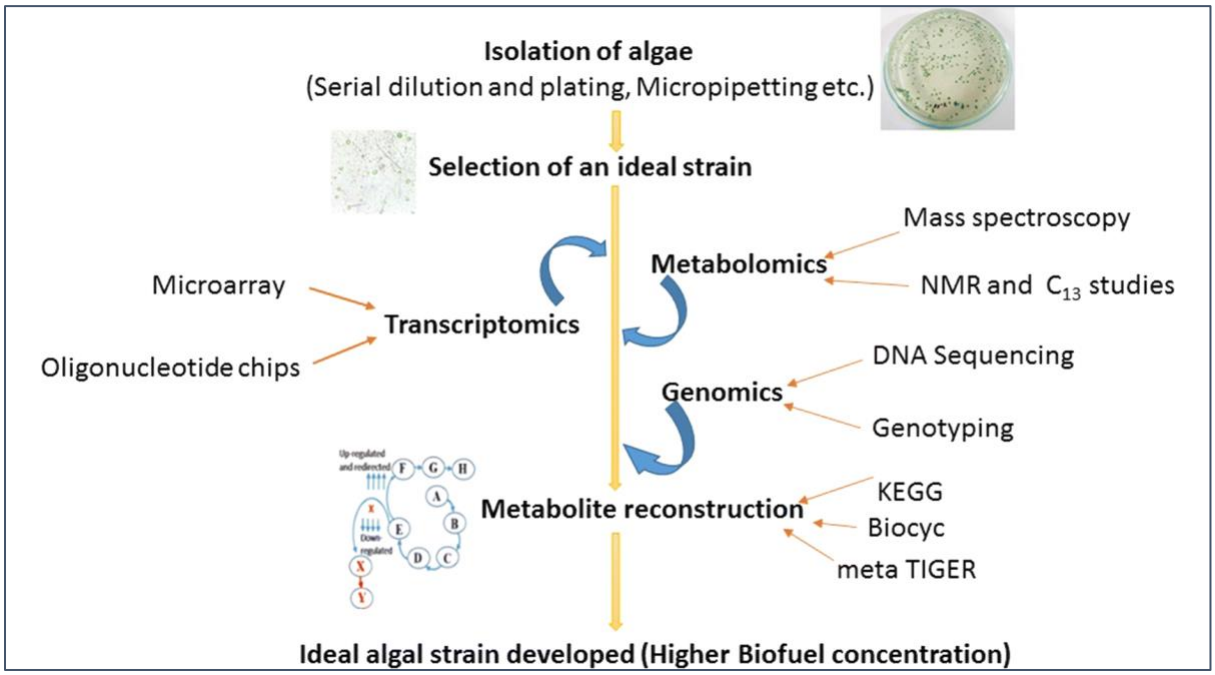


Figure 1.1: Ideal Algal strain development for Higher Biofuel concentration: *A workflow representation to develop an ideal algal strain; firstly, an algal strain is selected and isolated from the culture, by using serial dilution and plating and Micro-pipetting methodologies. Afterwards algal strain's metabolomics, transcriptomics and genomics analysis data is used for the metabolite reconstruction* [48].

To find the lost enzymes that catalyse reactions believed to be present, by using a comparative genomic approach [49] where information from intently linked genomes is used. While within comparative genomics, some studies are tried to discover functionally analogous genes [50]. Other methodologies use Machine

Learning practices in order to assess the nominee gene using genomic perspective, homology and pathway-based confirmation [51].

Novel and more accurate approaches for the metabolic pathways are required if we have to make complete use of systems biology in finding transporters or enzymes [Figure 1.1].

1.4. Background Gene and Enzymes

1.4.1. Enzymes

The genetic knowledge transferred in cell division is comprised of the deoxyribonucleic acid. Some specific fragments of the deoxyribonucleic acid sequence, named *genes*, be able to transcribe to a ribonucleic acid and translated to *proteins* made of amino acids connected by peptide bonds. The remaining part of the deoxyribonucleic acid sequence is yet not fully comprehended [52].

Proteins can have various different tasks that makes them accountable effectively for all functions of the cell.

A protein class is termed enzymes. Enzyme proteins have functions to catalyse chemical reactions. Virtually all reactions require an enzyme for catalysation. The reactions are entitled spontaneous which do not require an enzyme. The enzyme works by reducing the minimum energy essential to

initiate the reaction, subsequently increasing the rate of the reaction. Enzymes, their function play a significant role for metabolic networks.

1.4.2. Enzymatic Function Labelling

Associating an enzyme with a specific task can occasionally be a very challenging task. For that it is critical to have an organised enzymatic classification system. Presently the two most used classification systems are GO terms and EC Numbers.

1.4.2.1.EC Numbers

The EC number (Enzyme Commission number) is a numerical as well as hierarchical classification pattern for enzymes, established according to the chemical reactions they catalyse.

An enzyme code contains the EC letters followed by four numbers connected with dot. Those numbers indicate an increasingly progressive classification of the enzyme. The first number defines the following six groups of enzymatic functions: 1) Oxidoreductases: comprise all the oxidation and reduction reactions, which are symbolised by the handover of a hydrogen or oxygen atom and electrons in the molecules. 2) Transferases: dependable for the allocation of a functional group, like methyl, phosphate, etc. 3) Hydrolases: comprise all the hydrolysis reactions, accountable for the cleavage of a compound by addition of water. 4) Lyases: accountable for cleaving non-hydrolytic chemical bonds. 5)

Isomerases: convert one molecule into another, which has precisely the similar set of atoms. 6) Ligases: responsible for the chemical bond synthesis by breaking down ATP.

The next number in the EC number notation defines the finer groups of reactions until the 4th number, that classifies the substrate level reaction.

1.4.3. Gene Ontology

GO terms are a segment of a scheme with the objective of regulating the gene and protein annotations within species and diverse data sources [53]. This classification scheme also offers a number of tools to access its contents and reduces the time needed to search.

The GO terms divide gene products into three distinct ontologies: 1) cellular component, 2) biological process and 3) molecular function. Each individual ontology is a directed acyclic graph, a gene product can be allocated to one or more GO terms. Due to these pairs of fundamental characters a gene associated with a given node is automatically linked to all its inherited nodes.

1.5. Structural Domain Classification

The primary structure or sequence of amino acids of a protein defines its three-dimensional (3D) conformation. There are two types of patterns within the 3D conformation, which compose the secondary structure. Two types are the α-helices and the β-sheets. Manifold α-helices and β-sheets can join into more compact and complex units entitled domains.

These structures can be represented in diverse proteins and can be connected with each other in distinct groups in a single protein called multi-domain proteins or alone called single-domain protein, follow-on in different enzymatic tasks. Domains are also perceived as evolutionary units. The domains, within a multidomain protein, are often functionally and structurally autonomous.

There are three stages of structural domain classification: 1) fold, 2) superfamily and 3) family [54]. Fold is the uppermost stage. It clusters together domains which have the identical subordinate structure elements and the identical chain topology.

Subsequent to the fold, next is the superfamily. This stage clusters together domains which have functional and structure proof to share a common predecessor. So, in these groups are considered to be the most distant homologous genes. The lowermost final stage is called family. This stage groups collectively domains with well-defined sequence similarity. Same family domains manage to have similar tasks.

1.6. Metabolic Network

A Metabolic Network is a set of biochemical reactions. The reaction's interaction is accountable for the metabolic processes which regulate the cell functions. Usually, the reactions denote the nodes of this Metabolic network. For virtually all reactions have an enzyme behind catalysis reactions.

The sets of associated chemical reactions, that convert a preliminary molecule into another molecule called product, are termed as metabolic pathways. They generally represent the conversion of a main molecule into product. Furthermore, the pathways are dependent on each other, taking common molecules and reactions.

As different species have diverse biochemical properties, a particular pathway may differ between species or not exist in others species. Some databases gather all these different types of biochemical properties and have constructed pathway templates that demonstrate the complexity level of the metabolic networks and the variations between species. Such pathway templates can be located in KEGG (Kyoto Encyclopedia of Genes and Genomes) [55,56] and in RAST (Rapid Annotation using Subsystem Technology) [57].

1.6.1. Metabolic Reconstruction

Metabolic reconstruction is the methodology of constructing an organism's metabolic network map using confirmation from its genome sequence [58]. A motivation to make these models better is that more precise metabolic networks of parasites and pathogens will allow the discovery of key enzymes or transporters that can be prospective targets for new drugs.

The reconstruction process may be explained as a series of simple steps. Subsequently assembling the whole genome sequence, the first step is the discovery of the coding sequences of potential genes.

The approach used to interpret gene sequences can run from the discovery of the starting and ending codons, to the consumption of family profiles or sequence similarity. After identification of the gene's coding sequence, the predicted protein sequences are compared with the sequences from known and closely related genomes for enzymatic annotation where genes appear to be functionally relevant. The most common approach, for this step, relies on sequence comparison approaches such as BLAST.

Like this, acknowledged metabolic networks are assembled. The resulting steps are time consuming. They are associated to the manual curation of the networks and also to the nodes that were created for the missing nodes. Here, researchers try to reconcile the conclusive information with the known biology, especially with species-specific information [58].

1.6.2. Data Sources for Metabolic Networks

There are numerous databases that assemble and link essential biological information altogether, which can be used to curate metabolic networks. The KEGG (Kyoto Encyclopedia of Genes and Genome) knowledge base [59,60,61] provides an abundance of information of several species about biological systems, starting from genes and proteins to molecular wiring diagrams of network interactions and reactions.

Whereas, BRENDA (BRaunschweig ENzyme DAtabase) is an example of a more enzyme-specific database [62], which offers several levels of information including enzymes' nomenclature, relation between reactions and species specificity, etc for mapping and linking with other metabolic network databases.

There are a number of other databases, which provide details of the identification and classification of domains. Moreover, there are also several other databases that make use of these above-described databases and through Hidden Markov Models (HMM) construct profiles for different structural levels. Some of these types of databases are SUPERFAMILY [63], Pfam [64], and Gene3D [65].

1.6.3. Available Approaches for Metabolic Network Reconstruction

A manual curation-based approach is a reasonably promising approach for network reconstruction. However, this procedure just by itself is very time consuming. The improving speed of an organism's genome sequencing offers bioinformatics a progressively significant source of information.

There are various different approaches that can be used to assist with metabolic reconstruction. Some approaches cover most of the steps needed to construct a metabolic model for example Pathway Tools [66], ERGO [67] and RAST [68]. This metabolic reconstruction software combines multiple bioinformatic tools that cover the genome and gene sequence, protein function annotation, and visualization tools all together with integrated databases that assist to curate the metabolic network model.

However, the availability of a comprehensive annotated genome is indispensable, because most of the techniques of network reconstruction start with the functional annotation to the known and potential enzymes.

Some other software use different techniques to annotate sequences, like metabolic SearcH And Reconstruction Kit (metaSHARK) [69,70], ERGO and RAST. It uses the PRIAM library of its SHARKhunt tool [71], and profile models are used as the basis of a search tool to find the DNA sequence regions to known enzymes, based on significant similarity.

Instead, a very diverse methodology is used by GLIMMER2, it is integrated in RAST [72]. Here, the annotation is built by the use of incorporated Markov models trained by using curated structured gene data. Yet the annotation is presently typically established on sequence similarities, like BLAST [73].

Chapter 2

Material and Methods

2. Genome Analysis

2.1. Annotation and Mapping of Genomes of *Chlorella* strains

2.1.1. InterPro Annotation

Protein annotation for the analysed *Chlorella* strains (*Chlorella sorokiniana 211-8k, Chlorella vulgaris 264, Chlorella vulgaris C-1, Chlorella vulgaris 211-11b* [Table 2.1] [111] was assessed by using InterProScan [74]. InterProScan mapped genome sequences of algal strains with InterPro IDs, protein domains, families and motifs, including Pfam [75], PANTHER [76], PRINTS [77], SMART [78], SUPERFAMILY [79] and TIGRFAMs [80].

Strain	Full name	Origin	Strain
Cs-8k	*Chlorella sorokiniana 211-8k*	Austin, USA	High Light Tolerant
Cv264	*Chlorella vulgaris 264*	Köthen, Germany	High Light Tolerant
Cv-C1	*Chlorella vulgaris C-1*	Arkhangelsk, Russia	High Light Tolerant
Cv11b	*Chlorella vulgaris 211-11b*	Delft, The Netherlands	High Light Sensitive

Table 2.1: List of 4 *Chlorella* Strains: Cs-8k, Cv264, and Cv-C1 *are high light tolerant algal strains originally belong to the USA, Germany and Russia respectively; while Cv11b is high light sensitive algal strain belonging to the*

Netherlands. High light tolerant strains showed good growth rate under low light as well as high light intensity, while High light sensitive strain showed good growth rate only under low light. These algal cultures were grown under the high light intensity of 1,000 µmol /m2 /s1.

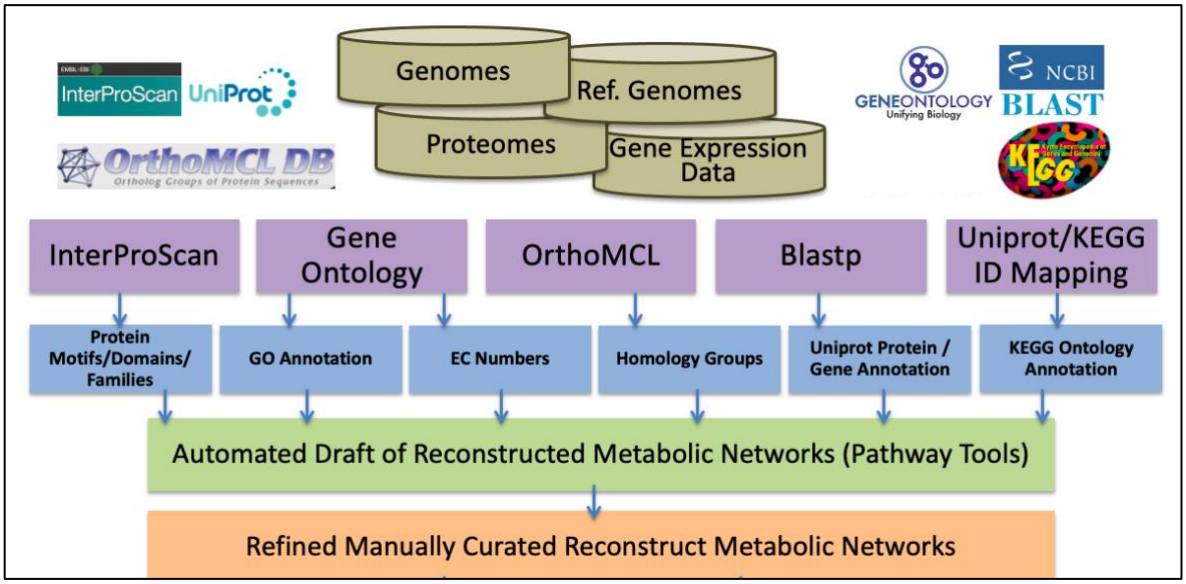


Figure 2.1: Framework designed for this Metabolic Network Reconstruction research work: *This pictorial illustration explains the workflow methodology, for this research work. All the databases and mapping tools, like InterProScan, OrthoMCL, Blastp and UniProt, were used to annotate the genome and protein sequence data. Then KEGG Pathway Mapping tool was used to generate automated drafts of reconstructed metabolic networks, then list of missing genes from metabolic pathway was extracted and tried to annotate them with tBlastn tool, finally automated drafts of reconstructed metabolic networks were generated again.*

2.1.2. Annotation of Gene Ontology

Genomic sequencing has made it evident that a large number of genes linked to the core biological functions are shared by all eucaryotes [81, 82]. Thus, the biological functions of these shared proteins are generally predicted by mapping between newly identified protein sequences and known functional genes with well-established knowledge of their biological roles in a model organism. Therefore, these annotated algal protein sequences are mapped with Gene Ontology (GO) terms [83] using an updated InterPro2GO mapping file [Figure 2.1] [84].

2.1.3. Annotation of Enzyme Commission (EC) Number

Mapping of multiple functional annotations can significantly improve metabolic network size, particularly for non-model organisms [86]. Correspondingly the count of EC numbers mapped with protein sequences suggests the size of the metabolic network [85]. Consequently, GO terms mapped with algal protein sequences are further linked with EC numbers by using the EC2GO mapping file [86].

2.2. Similarity Comparison Between Protein Sequences of Algal Strains

For similarity comparison between algal protein sequences, the OrthoMCL software [87] was used, which allows simultaneous classification of global relationships in a similarity space. Thus, Orthologous Groups between all

Chlorella strains i.e., Cs-8k, Cv264, Cv11b and Cv-C1, were identified with OrthoMCL-v2.0.9 [88] by using the protein sequences of annotated genes as input. Then the output result files were curated by removing protein sequences labelled with 'No-Group'.

For visualization of this similarity comparison, a four-way Venn diagram of orthologous groups was drawn between the analysed *Chlorella* strains by using the R package VennDiagram v1.6.20 [89].

2.3. OrthoMCL Clustering with Reference Genome

2.3.1. Reference Genomes (*Arabidopsis thaliana* and *Chlorella variabilis*)

In addition, OrthoMCL was run once again including the proteomes of *Chlorella variabilis*, and *Arabidopsis thaliana* with the *Chlorella* strains Cs-8k, Cv264, Cv11b and Cv-C1.

2.3.2. Comparison Between Algal Strains and Reference Genomes

The output result files of OrthoMCL were used for comparison between the *Chlorella* strains Cs-8k, Cv264, Cv11b and Cv-C1; and the reference genomes of *Chlorella variabilis*, and *Arabidopsis thaliana*.

2.4. Sequence Analysis by Using Blastp of Algal Strains with Reference Genome (*Arabidopsis thaliana*)

Sequence similarity searches between predicted genes of four algal strains and the database of *Arabidopsis thaliana* as a reference genome from

UniProtKB/Swiss-Prot were carried out with BLASTP v2.3.0+ [90], by setting the e-value (expect value) to 1^{e-6}.

2.4.1. Mapping Of UniProt and KEGG Ids by Using KEGG Mapper

The output result files of BLASTP were used to extract UniProt gene identifiers mapped with the algal strains' predicted gene ids. Then these UniProt identifiers of the four algal strains were mapped with KEGG [91] gene ids of *Arabidopsis thaliana* by using the reference organism-based Convert ID [92] tool of KEGG Mapper [93].

2.4.2. KEGG Pathway Mapping with *Arabidopsis thaliana*

For pathway mapping, the KEGG Pathways mapping tool [94] was used. The lists of KEGG gene ids of *Arabidopsis thaliana* were mapped with metabolic pathways including "Carbohydrate metabolism", "Energy metabolism", "Lipid metabolism", "Nucleotide metabolism", "Amino acid metabolism", "Metabolism of other amino acids", "Glycan biosynthesis and metabolism" and "Metabolism of cofactors and vitamins"; while *Arabidopsis thaliana* was used as a reference organism.

2.5. Finding Missing Functions with TBLASTN

2.5.1. Missing Functions in Algal Strains

Subsequently, lists of mapping and missing genes were extracted from the KEGG Mapping tool. Then missing KEGG gene ids of *Arabidopsis thaliana*

were mapped with KEGG Orthology (KO) ids to identify the missing functions.

2.5.2. Algal Proteins (Absent in *Arabidopsis*)

Similarly, lists of those predicted genes were also extracted, which were present in four algal strains but missing in the *Arabidopsis thaliana* fasta file to identify the different functions.

2.5.3. Sequence Analysis by Using TBLASTN

Afterwards, the missing gene list is used to extract genome sequences of relevant predicted genes from the genome fasta files for the four algal strains. Then, sequence similarity analysis was carried out by using tBLASTn and setting the e-value to 1^{e-6}, between these extracted genome fasta files and the protein database of *Arabidopsis thaliana.* This was a reference genome from UniProtKB/Swiss-Prot, to compare a protein query sequence against the six-frame translations of nucleotide sequences for finding homologous protein coding regions in unannotated nucleotide sequences.

2.5.4. Count of Newly Found Functions in Algal Strains

After sequence similarity analysis with tBLASTn, newly discovered genes were extracted from the result files, counted and added into previously existing lists of mapping genes.

Newly extracted UniProt gene ids were again mapped with KEGG gene ids of *Arabidopsis thaliana* and these new updated lists of KEGG gene ids of *Arabidopsis thaliana* were again mapped with metabolic pathways by using the KEGG Pathways mapping tool, while *Arabidopsis thaliana* was used as a reference organism.

2.6. Calculating Pathway Coverage for *Arabidopsis* and Four Algal Strains

The Percentage of pathway coverage was calculated by comparing function counts between *Arabidopsis* and all algal strains, for all pathways, while for pathway's percentage coverage, count of functions found in an algal strain was divided by all functions count in a pathway (for all functions of *Arabidopsis*).

2.6.1. Pathways and Algal Strains Comparison

After metabolic pathway mapping, pathway coverage was calculated for *Arabidopsis thaliana* and four algal strains. Missing genes were identified and counted for all four algal strains against all metabolic pathways with the reference of mapped list of genes of *Arabidopsis thaliana*.

2.6.2. Comparison Between Shared and Strain Specific Gene and Function Count

Afterwards, pathway coverage was again calculated for shared and strain specific genes and functions (by using KO ids) for all algal strains with pair-

wise analysis. These lists of shared and strain specific gene and KO ids counts were used for comparison between four algal strains against all metabolic pathways.

2.7. Fischer Test Calculations Between Tolerant and Sensitive Algal Strains

2.7.1. Fisher Test to Verify Difference Between Shared and Strain-Specific Gene and KO Ids Count for Cs8k-Cv11b

The intersection between each shared and strain-specific gene and KO ids' count was built and a Fisher's exact test was performed in R v3.2.3 [95] to test the hypothesis that difference between tolerant and sensitive algal strains are significantly enriched (p=0.05).

Fischer's exact test was calculated for all pathways and for all algal strains for:

- the number of shared (overlapped) proteins between tolerant algal strain (Cs-8k) and sensitive algal strain (Cv11b) in a pathway
- the number of tolerant algal strain (Cs-8k) specific proteins in a pathway
- the number of shared (overlapped) functions (KO) between tolerant algal strain (Cs-8k) and sensitive algal strain (Cv11b) in a pathway
- the number of functions present only in tolerant algal strain (Cs-8k) in a pathway

2.7.2. Bonferroni and Benjamini-Hochberg (BH (alias FDR)) Correction

Subsequently, Bonferroni and Benjamini-Hochberg (BH (alias FDR)) corrections were implemented to verify the results of Fischer exact tests for significant differences.

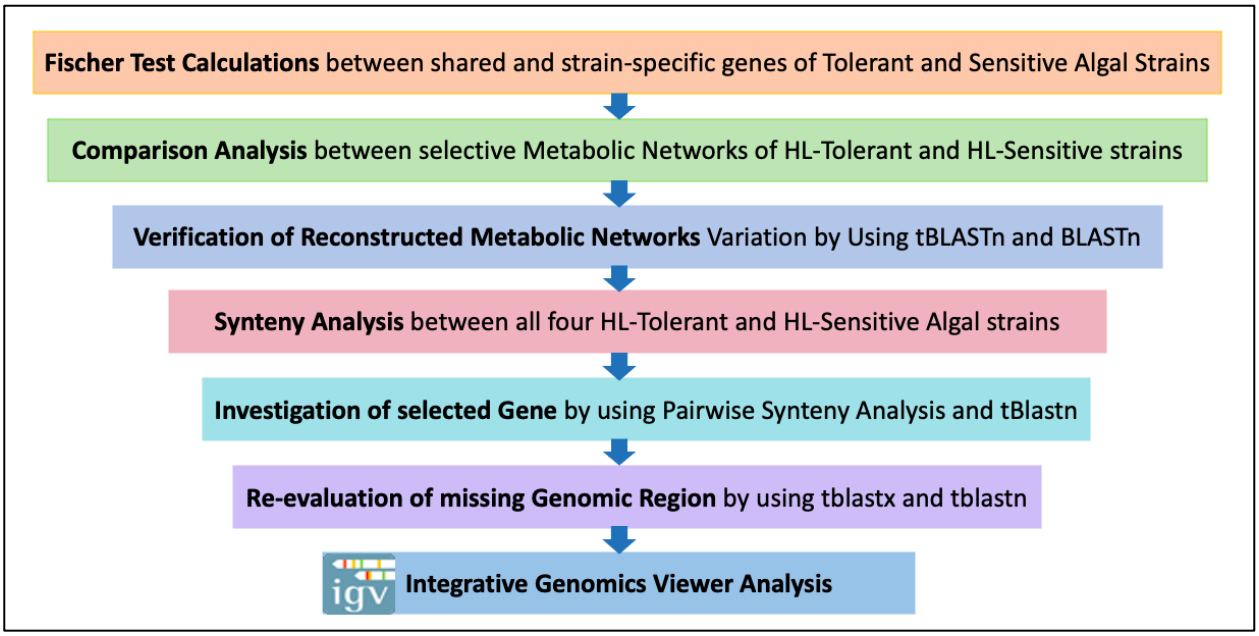


Figure 2.2: Workflow for Comparison & Verification: *This workflow shows the steps of metabolic networks comparison, their verification and re-evaluation for this research work. Fischer Exact test was used to calculate the statistical significance of the variation between HL-tolerant and HL-sensitive strains. Then highly significant metabolic networks were selected for further comparison, missing genes were again verified by using tBlastn and Blastn tools. Then synteny analysis was applied to identify the variant genomic regions, then selected gene was further investigated and its missing genomic region was re-evaluated and analysed by using Integrative Genomic Viewer.*

2.8. Identifying Pathways Variation Between Tolerant vs Sensitive Strains

After comparison between tolerant and sensitive strains, pathways were ranked according to results of Fischer's exact test. Then top ranked pathways were selected to identify variations between tolerant and sensitive strains, and those genes ids were discovered which were different in tolerant and sensitive algal strains [Figure 2.2].

2.9. Verification of Pathways Variation Between Tolerant vs Sensitive Strains

After identification of differences between tolerant and sensitive strains, these variant genes were verified with a sequence similarity check by using tBLASTn [96], BLASTn [97].

2.9.1. Sequence Analysis by Using tBLASTn (Tolerant vs Sensitive Strains)

To verify the variation between genes of tolerant and sensitive strains' sequence a similarity check was conducted by using tBLASTn against protein sequences of Cs8k and genome sequences of Cv11b for selected genes sequences.

2.9.2. Sequence Analysis by Using BLASTn (Tolerant vs Sensitive Strains)

Afterwards, a sequence similarity check was conducted by using BLASTn against the genome sequences of Cs8k (with upstream and downstream genome) and Cv11b

2.9.3. Sequence Analysis by Using tBLASTn (*Arabidopsis* and Sensitive Strain)

Subsequently, a sequence similarity check was conducted by using tBLASTn against protein sequences of *Arabidopsis* and the genome sequence of Cv11b.

2.10. Synteny Analysis

Synteny analysis refers to synteny as the conservation of blocks of order within two sets of chromosomes that are being compared with each other. This analysis was conducted to visualize the chromosome regions, which can help to identify chromosomal rearrangement processes. That is why, genome sequences of selected genes of high ranked pathways were used for synteny analysis and for that SimpleSynteny v1.5 [98] was used.

2.10.1. Pairwise Synteny Analysis for Selected Nodes of Algal Strains to Explore the PSD3 Gene

Pairwise synteny analysis was done systematically to check the genome orientation and positions of the genes. Firstly, without flipping and reordering to see their original alignment and later pairwise synteny analysis was redone after flipping and reordering the nodes to check the best fit orientation between the sequences of these nodes.

i. Node 117 and 218 of Cs8k and Node 22 and 72 of Cv264
ii. Node 117 and 218 of Cs8k and Node 5 and Node 68 of CvC1
iii. Node 117 and 218 of Cs8k and Node 207, 341, 377 and 383 of Cv11b
iv. Node 22 and 72 of Cv264 and Node 5 and Node 68 of CvC1
v. Node 22 and 72 of Cv264 and Node 207, 341, 377 and 383 of Cv11b
vi. Node 5 and 68 of CvC1and Node 207, 341, 377 and 383 of Cv11b

Subsequently, the neighbouring nodes of Node-341 and Node-383 of Cv11b (i.e., Node-342 and Node-382) were extracted and mapped firstly with the CvC1 Node 68 and then with the Cv264 Node 72.

2.11. Re-evaluation of Missing Gene by Synteny Analysis

2.11.1. Sequence Alignment by Using Contiguous Megablast

For reassessment and detailed review of the result of sequence alignment between all four algal strains and to retry the discovery of PSD3 gene; sequence alignment was done again between selected nodes of all algal strains by using Blast online available tool, sequence fasta files of genome and proteome were

uploaded and contiguous megablast was used to identify highly similar sequences.

Thus, sequence alignment was done for sequences of nodes of CvC1 (i.e., NODE_68), Cv264 (i.e., NODE_72) and Cs8k (i.e., NODE_117) with 3 nodes of the Cv11b strain (i.e., NODE_207, NODE_383 and NODE_341); sequence fasta files of the genome were uploaded and the contiguous megablast tool was used to identify highly similar sequences.

Secondly, if the contiguous megablast tool could not identify highly similar sequences, the alignment was done by using the discontinuous megablast tool.

2.11.2. Sequence Analysis by Using TBLASTN and TBLASTX

TBLASTN compares a protein query sequence to a database of nucleotide sequences dynamically translated in all six possible reading frames and is used to identify proteins in new, undescribed genomes or to ask whether a DNA database encodes a protein that matches your protein query of interest.

TBLASTX compares the dynamically translated six-frames of a nucleotide query sequence against the dynamically translated six-frames of a nucleotide sequence database. The TBLASTX program is more sensitive and computationally intensive than TBLASTN. It is used for a DNA sequence with no obvious database matches to identify if it encodes a protein with distant, statistically significant database matches in a database of expressed

sequence tags and it is therefore useful to reveal genes that encode proteins homologous to the query [Figure 2.3].

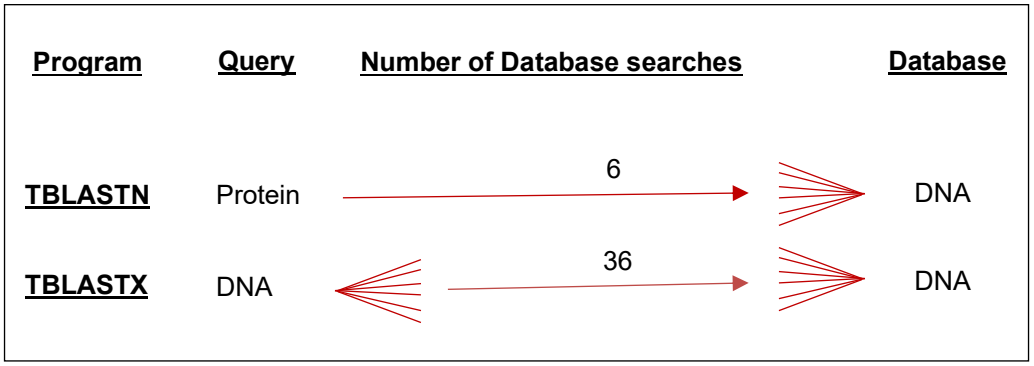


Figure 2.3: Finding Algal Genes with TBLASTN sequence Alignment: *To verify gene annotation, another sequence similarity analysis was carried out by using TBLASTN between algal genome fasta files and the protein database of Arabidopsis thaliana, and TBLASTX between algal and Arabidopsis thaliana genome fasta files. TBLASTN compares a protein query sequence to a database of nucleotide sequences dynamically translated in all six possible reading frames, while TBLASTX compares the dynamically translated six-frames of a nucleotide query sequence against the dynamically translated six-frames of a nucleotide sequence database, and is used to identify proteins in new, undescribed genomes or to ask whether a DNA database encodes a protein that matches your protein query.*

In order to investigate the patterns and degree of DNA sequence divergence between the **Cv11b** and **CvC1** algal genomes, the neighbouring nodes of Node-341 and Node-383 of Cv11b (i.e., Node-342 as a neighbouring node of Node-341 and Node-382 as a neighbouring node of Node-383 of Cv11b) were extracted and mapped firstly with the CvC1 Node 68 and then with the Cv264 Node 72 by using tblastn and tblastx. TBLASTN operates by translating database nucleotide sequences to hypothetical amino acid sequences in all six reading frames and then aligning the hypothetical amino acid sequences to the query.

For re-evaluation, the protein sequence file of the PSD3 gene of *Arabidopsis thaliana* was extracted from uniport KB. Moreover, nucleotide sequences of 3 relevant Nodes of Cv11b (i.e., Node_207, Node_341 and Node_383), Node-117 of Cs8k, Node-72 of Cv264, Node-68 of CvC1 and the complete genome of Cv11b were used to create a database. Subsequently tblastn was used to map the protein sequence with all the above-mentioned nucleotide databases.

Furthermore, nucleotide sequences of Node-117 of Cs8k, Node-72 of Cv264, Node-68 of CvC1 were mapped by using tblastx, with the databases of 3 relevant Nodes of Cv11b (i.e., Node_207, Node_341 and Node_383) and complete genome sequence of Cv11b individually.

2.11.3. IGV (Integrative Genomics Viewer) Analysis

IGV (Integrative Genomics Viewer) (version 2.3) [99,100] was used to figure out the micro-collinearity of genomic sequences between all four algal strains in the region of the PSD3 gene, and to confirm assembly and synteny, and to find possible traces of deletion of the PSD3 gene.

Thus, the sequence of Node117 of the Cs8k genome was explored to identify NNN's series (i.e., the sequencer could not resolve with enough confidence about which base was sequenced and assigns letter 'N' instead of A, T, C, G.) in the region between base pairs 156912 and 167183. Moreover, Cs8k NODE_117 was prepared for IGV analysis to check the annotation for PSD3 between base pairs 162000 and 173000.

Chapter 3

Results

3. Genome Analysis

3.1. Annotation and Mapping of Genomes of *Chlorella* strains

3.1.1. Genomes and Proteomes of Analysed *Chlorella* Strains

The initial purpose of this research work was the genetic analysis by evaluating similarity between the analysed Chlorella strains, *Chlorella sorokiniana 211-8k* (Cs8k), *Chlorella vulgaris 264* (Cv264), *Chlorella vulgaris C-1* (CvC1), and *Chlorella vulgaris 211-11b* (Cv11b).

The genome assemblies of the High Light (HL) tolerant strains *Chlorella sorokiniana 211-8k* (Cs8k), *Chlorella vulgaris 264 (*Cv264), and *Chlorella vulgaris C-1* (CvC1), and the HL-sensitive strain *Chlorella vulgaris 211-11b* (Cv11b) were estimated regarding their completeness in comparison to the reference genomes of *Chlorella variabilis NC64A*, and *Arabidopsis thaliana* (Thale cress) [Figure 3.1]. Furthermore, whole genome alignments between the strains were performed and the similarity between genomic sequences was assessed.

Thereafter, the genomes of the analysed *Chlorella* strains were annotated using two different genome annotation pipelines. The completeness of

genome annotations was assessed for both genome annotation pipelines. In addition, the completeness of genome annotations was further determined in comparison to the reference genome annotations of *Chlorella variabilis NC64A*, and *Arabidopsis thaliana*. The predicted proteins were grouped into Clusters of Orthologous Groups (COGs) to establish the similarity of proteomes between the analysed algae species and the selected reference genomes.

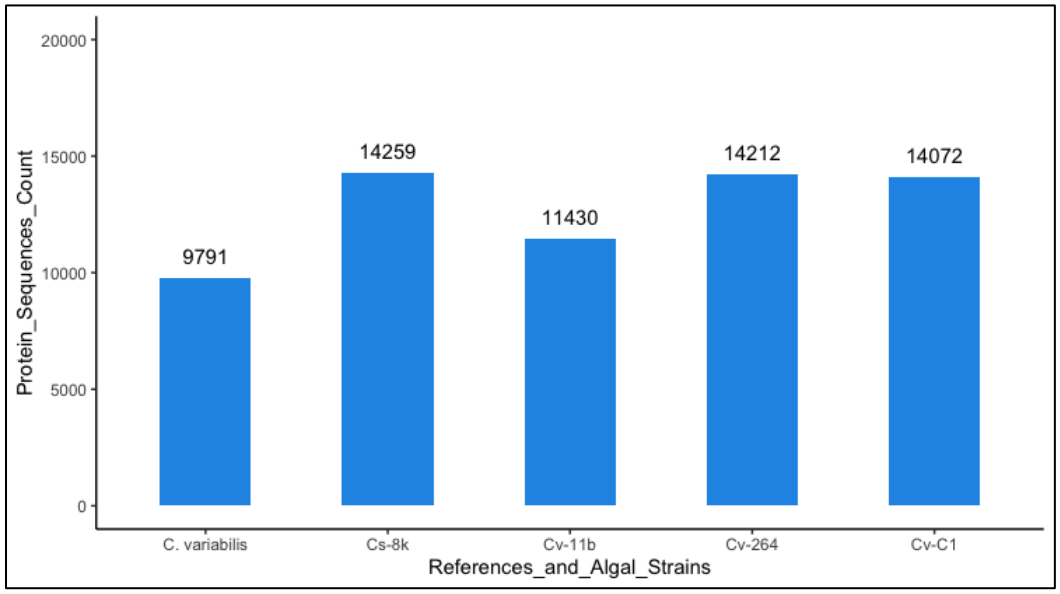


Figure 3.1: Comparison of different protein sequence count: *The Chart of comparison shows that HL-tolerant algal strains have more protein sequence count (i.e., Cs-8k:14259, Cv-264:14212, and Cv-C1:14072) than HL-sensitive algal strain (i.e., Cv-11b:11430), while protein sequence count for Chlorella variabilis is even less than HL-sensitive algal strain (i.e., C. variabilis:9791).*

Protein annotation for the analysed *Chlorella* strains (*Chlorella sorokiniana 211-8k, Chlorella vulgaris 264, Chlorella vulgaris C-1, Chlorella vulgaris 211-11b*) [111] was assessed by using InterProScan. InterProScan mapped genome sequences of algal strains with InterPro IDs, protein domains, families and motifs, including Pfam, PANTHER, PRINTS, SMART, SUPERFAMILY and TIGRFAMs.

3.1.2. Algal Genome Mapping and Annotation

Not all the protein sequences were able to find their mapping InterPro IDs and GO terms. Approximately half of them were mapped with GO terms, while before that 80% protein sequences were annotated perfectly with InterPro identification numbers [Figure 3.2].

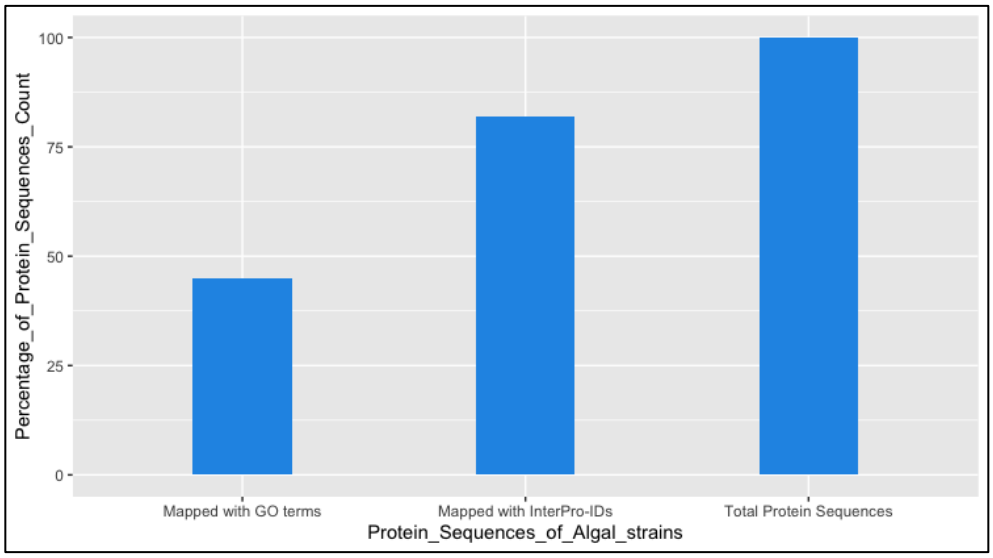


Figure 3.2: Ratio of InterPro and GO annotations: *The chart represents that all protein sequences of algal strains could not be annotated, with protein IDs. Only 80 percent protein sequences are mapped with InterPro IDs while half of these annotated protein sequences are linked with GO terms.*

3.2. Similarity Comparison Between Protein Sequences of Algal Strains

According to the output result files that are illustrated in the chart below; most of the proteins were grouped with the homologs, and only a few were left unassigned [Figure 3.3].

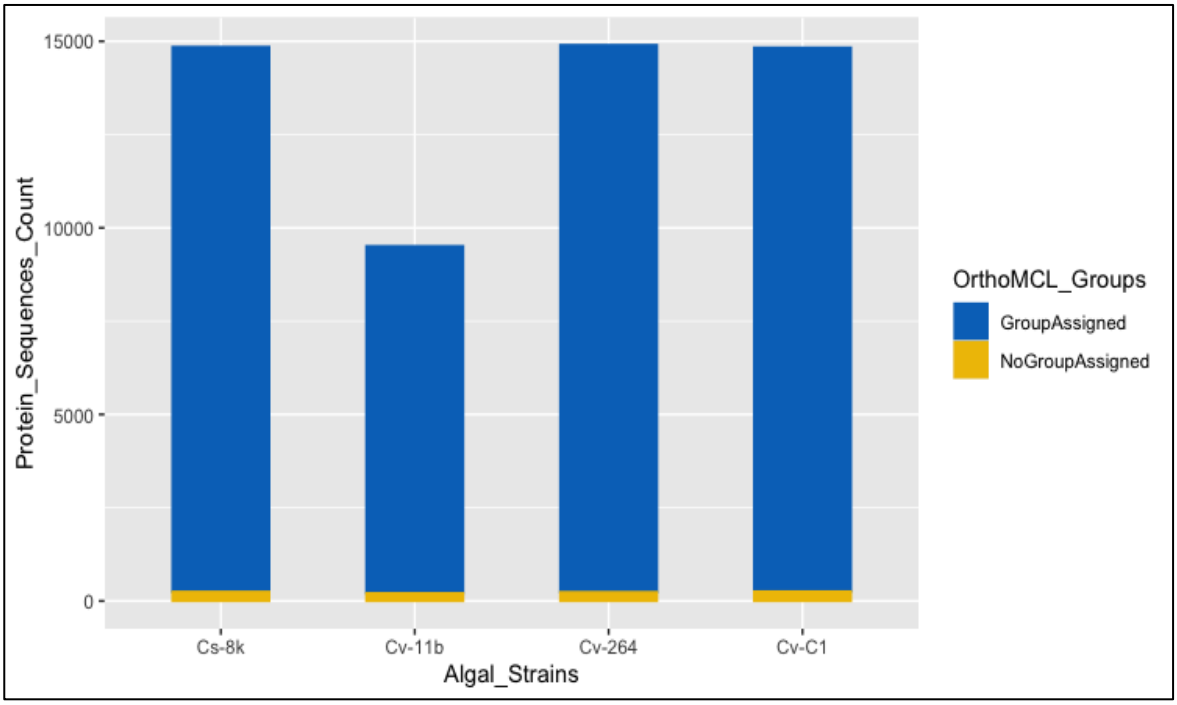


Figure 3.3: OrthoMCL: Assign-Group VS No-Group: *This chart represents a comparison for the protein sequences which are assigned to protein groups by OrthoMCL tool. It shows that a very small number of protein sequences could not be assigned to any group while majority of them are assigned to a group.*

3.3. OrthoMCL Clustering with Reference Genome

3.3.1. Comparison Between Algal Strains

According to OrthoMCL comparison, the predicted genes of these algal strains were sorted into a total number of 13,783 Orthologous Groups (OGs).

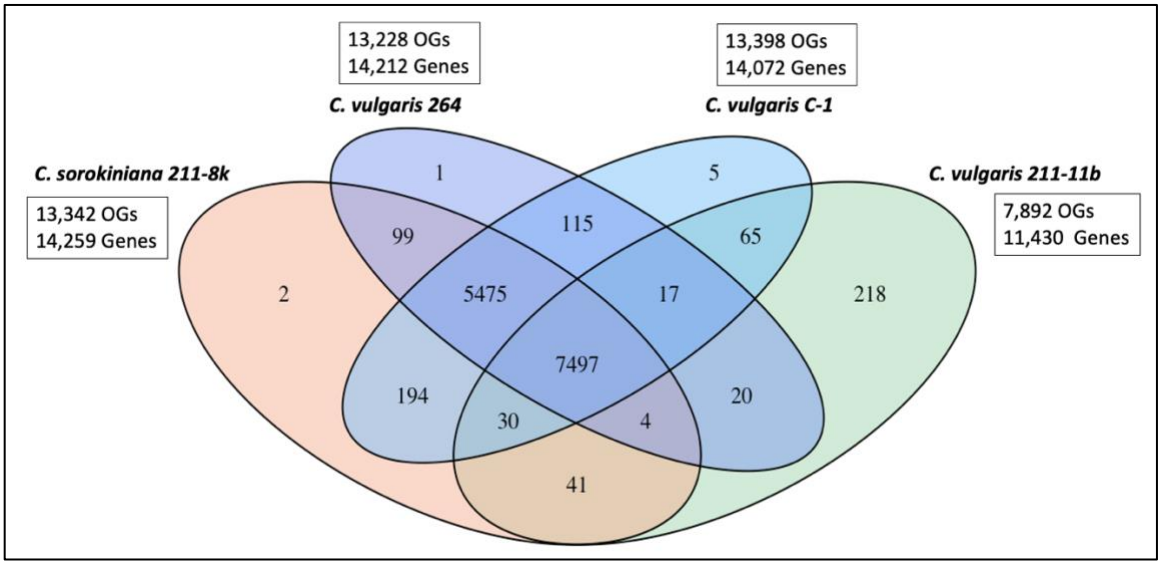


Figure 3.4: Venn Diagram showing the shared genome between four algal species: *This Venn Diagram represents that 7497 orthologous groups are shared between all four algal strains, while 12972 orthologous groups are shared only between three HL-tolerant algal strains. Additionally, genomes of HL-tolerant algal strains are associated with more than 13000 orthologous groups while genome of HL-sensitive algal strain is only linked with 7892 orthologous groups.*

The results indicate close phylogenetic relationships between the analysed *Chlorella* strains, especially between the HL-tolerant strains Cs8k, Cv264

and CvC1. While 7,497 OGs contained orthologs from all analysed strains, the HL-tolerant strains shared a large number of OGs, including 5,475 OGs specific to these strains. Cv11b showed the highest number of strain-specific OGs among these strains (i.e., 218) [Figure 3.4].

3.3.2. Comparison Between Algal Strains and Reference Genomes

The analysed *Chlorella* strains were compared to the reference proteomes of *Chlorella variabilis NC64A*, and *Arabidopsis thaliana*. The proteomes of the reference strains were included in the analysis of COGs. The OrthoMCL results including all strains revealed a total of 18,003 OGs.

3.4. Sequence analysis by Using BLASTP of Algal Strains with Reference Genome (*Arabidopsis thaliana*)

After sequence similarity searches, the lists of KEGG gene ids of *Arabidopsis thaliana* were mapped with metabolic pathways including “Carbohydrate metabolism”, “Energy metabolism”, “Lipid metabolism”, “Nucleotide metabolism”, “Amino acid metabolism”, “Metabolism of other amino acids”, “Glycan biosynthesis and metabolism” and “Metabolism of cofactors and vitamins”; while *Arabidopsis thaliana* was used as a reference organism.

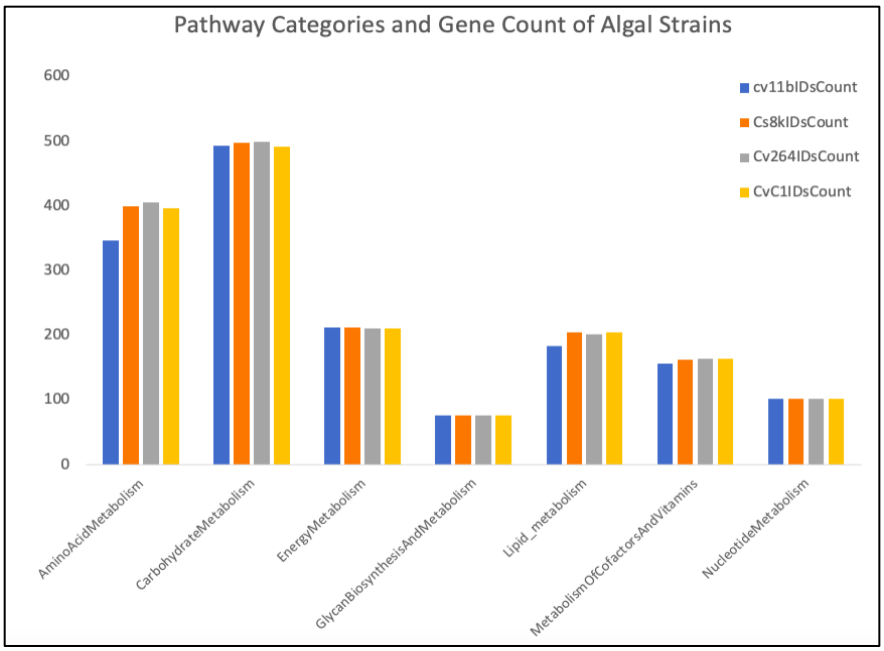


Figure 3.5: Pathway Categories and Gene Count of Algal Strains: *Pathway analysis of all different algal strains displays that the "Amino acid metabolism" pathway has maximum variation, whereas the "Carbohydrate metabolism", "Lipid metabolism" and "Metabolism of cofactors and vitamins" pathways have moderate variation. However, the "Energy metabolism", "Glycan biosynthesis and metabolism" and "Nucleotide metabolism" pathways have approximately similar gene counts.*

According to Pathway categories and the gene count of different algal strain analysis, it was revealed that the highest level of variation in gene count was observed for the "Amino acid metabolism" pathway, where Cv11b strain has the lowest number of gene count. While the "Carbohydrate metabolism", "Lipid metabolism" and "Metabolism of cofactors and vitamins" pathways showed moderate variation in gene count for all algal strains. However, the "Energy

metabolism", "Glycan biosynthesis and metabolism" and "Nucleotide metabolism" pathways showed approximately similar gene counts for all algal strains [Figure 3.5].

3.5. Finding Missing Functions with TBLASTN

3.5.1. Missing Functions in Algal Strains

The result of mapping between KEGG gene ids and KEGG Orthology (KO) ids showed that the maximum missing functions i.e., 267 functions, belong to the HL-sensitive strain Cv11b. While the HL- tolerant strains have an almost similar number of missing functions i.e., 240 functions.

3.5.2. Algal Proteins (Absent in *Arabidopsis*)

The list of functions which were present in algal strains but missing in the *Arabidopsis thaliana* fasta file, represented here that minimum number of functions belong to HL-sensitive strain Cv11b. While HL- tolerant strains have almost similar number of functions [Figure 3.6].

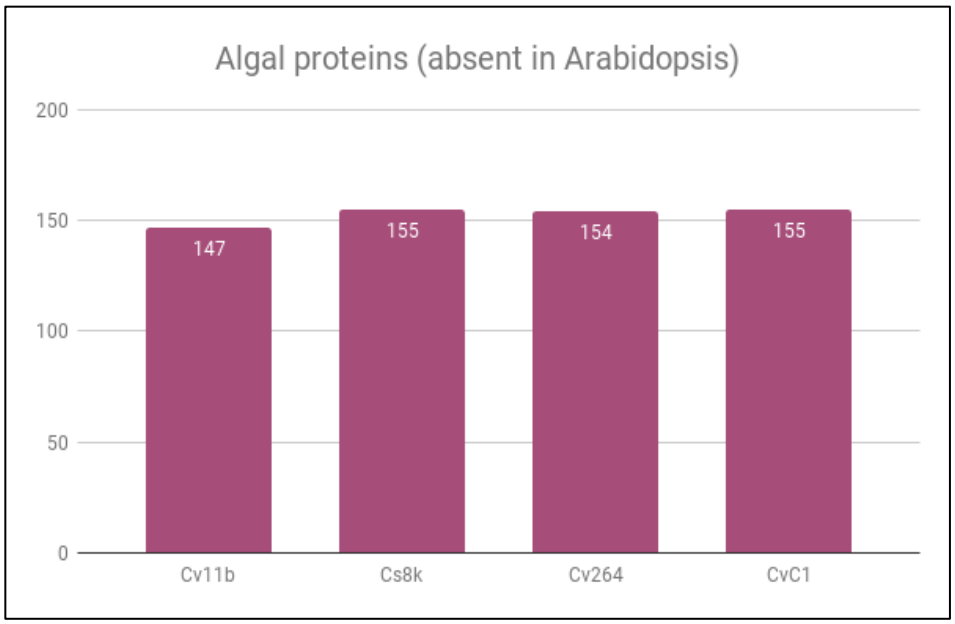


Figure 3.6: Gene Count of Algal Strains missing in *Arabidopsis* species: *This chart shows that all of 4 algal strains HL-tolerant as well as HL-sensitive have around 150 genes which are specific to algal genome and could not be mapped with the model genome of Arabidopsis species.*

3.5.3. Sequence Analysis by Using TBLASTN

As TBLASTN compares a protein query sequence against the six-frame translations of genome sequences for finding similar protein coding regions, therefore that tool helped to annotate several significant genes in unannotated genome sequences, which were missed in annotation earlier in similarity analysis while using BLASTP.

3.5.4. Count of Newly Found Functions in Algal Strains

Sequence similarity analysis with TBLASTN, assisted to identify 159 functions for Cv11b, while 133, 134 and 135 identified functions belonged to CvC1, Cv264 and Cs8k respectively [Figure 3.7].

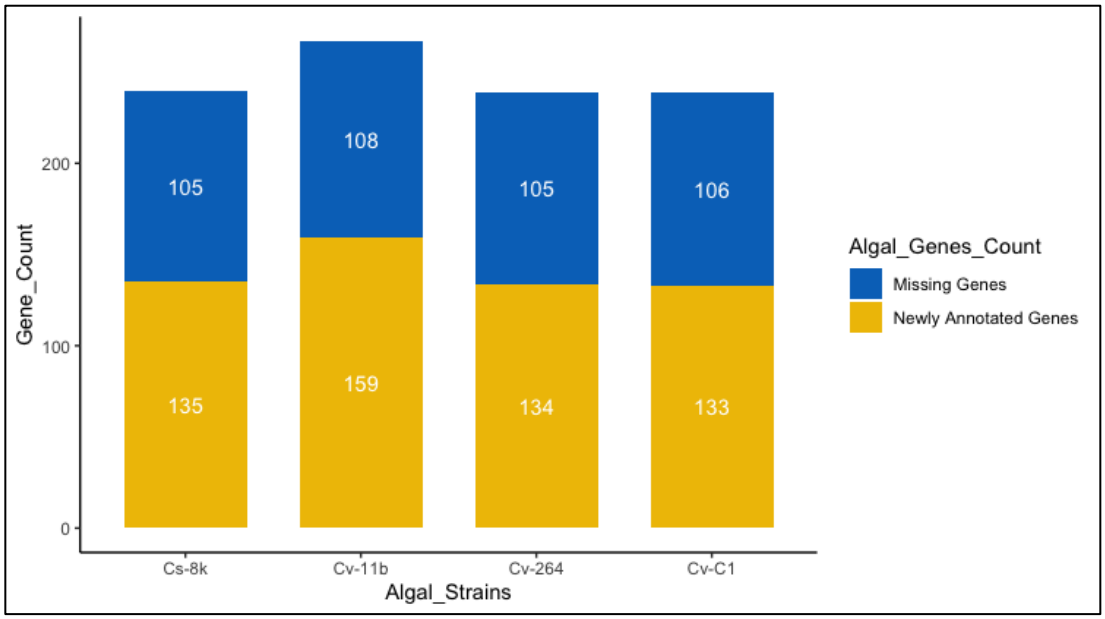


Figure 3.7: Count of Missing Genes in Algal Strains: *Sequence similarity analysis with TBLASTN tool supported to annotate around 134 and 159 new genes for HL-tolerant and HL-sensitive algal strains respectively.*

3.6. Calculating Pathway Coverage for *Arabidopsis* and Four Algal Strains

Pathway coverage for shared and strain specific genes and functions (by using KO ids) for all algal strains, demonstrated that highest number of shared and algal specific genes were associated with Carbohydrate Metabolism Pathways.

Whereas the algal strain Cs8k has more strain specific genes than the algal strain Cv11b [Figure 3.8].

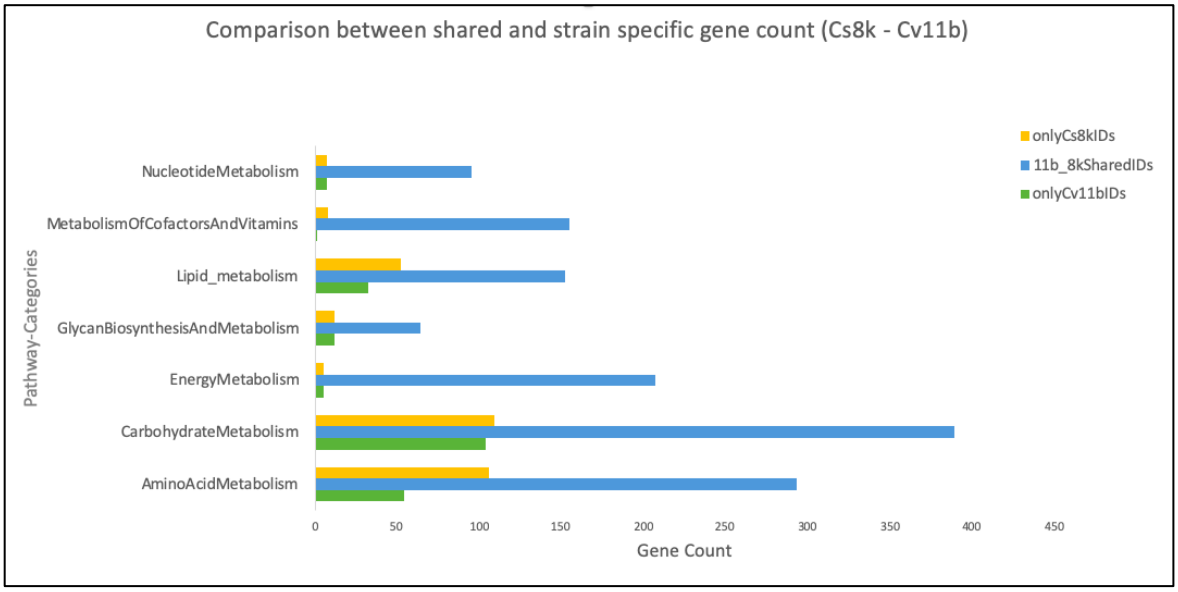


Figure 3.8: Comparison between shared and strain specific gene count: *This chart represents pathway analysis by comparing shared and strain specific gene count for HL-tolerant (Cs-8k) and HL-sensitive (Cv-11b) algal strains, where blue bars are presenting shared gene IDs, while yellow and green bars are showing strain specific gene count for HL-tolerant (Cs-8k) and HL-sensitive (Cv-11b) algal strains respectively.*

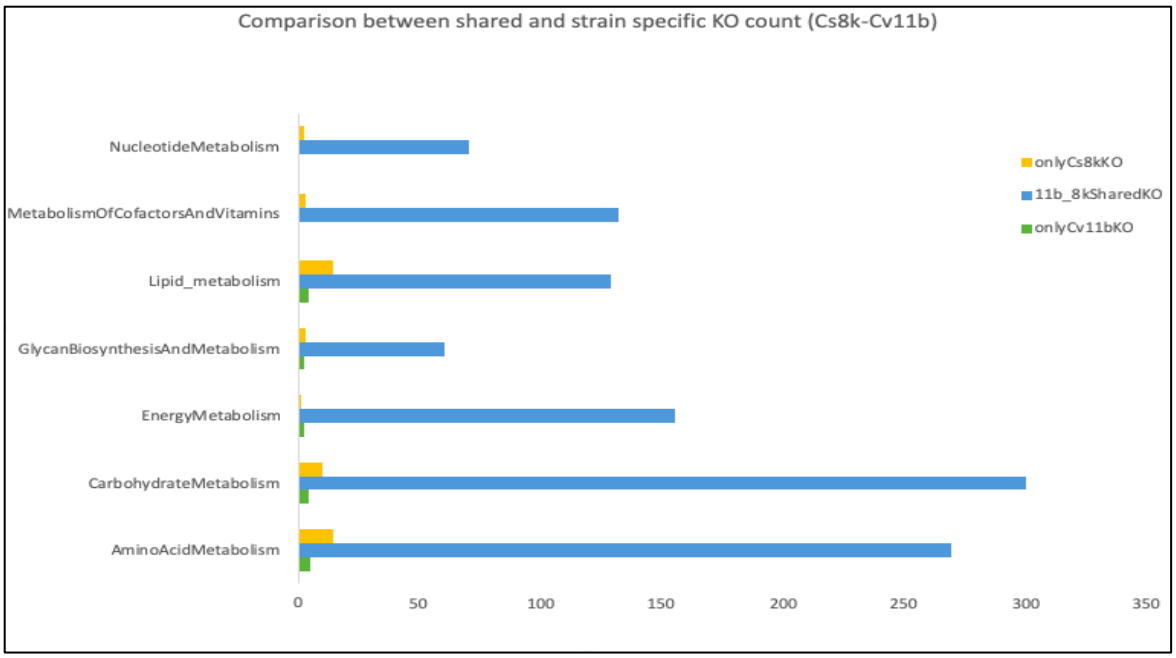


Figure 3.9: Comparison between shared and strain specific KO terms count: *This chart represents pathway analysis by comparing shared and strain specific* KEGG Orthology (KO) *IDs for HL-tolerant (Cs-8k) and HL-sensitive (Cv-11b) algal strains, where blue bars are presenting shared KO IDs, while yellow and green bars are showing strain specific KO IDs for HL-tolerant (Cs-8k) and HL-sensitive (Cv-11b) algal strains respectively.*

The results are showing that along with shared genes, there are a large number of algal strain specific genes but when these strain specific genes were mapped with the KO functions then the number of algal strain specific functions were reduced remarkably. Moreover, some notable algal specific functions belonged to HL-tolerant strain Cs8k, while the algal specific functions linked to HL-sensitive strain Cv11b are missing. Thus, even though both strains have different gene set but still these genes belong to the same KO functions [Figure 3.9].

3.7. Fischer Test Calculations Between Tolerant and Sensitive Algal Strains

3.7.1. Fisher Test to Verify Difference Between Shared and Strain-Specific Gene IDs Count for Cs8k-Cv11b

According to Fischer's exact test, three top ranked pathways were Cysteine and methionine metabolism, Glycerophospholipid metabolism and Galactose metabolism, with significant p-value i.e., < 0.05. Significant p-value result for these pathways is evidence of the hypothesis approval [Figure 3.10]. Thus, according to the hypothesis these pathways have shown the significant difference between genes of HL-tolerant and HL-sensitive algal strains [Figure 3.11].

```
> data
# A tibble: 68 × 8
   PathwayCategory                 PathwayName                                SharedIDs264_11b onlyCv264IDs onlyCv11bIDs `p-value`
   <chr>                           <chr>                                                 <dbl>        <dbl>        <dbl>     <dbl>
 1 AminoAcidMetabolism             Cysteine and methionine metabolism                       47           26            9    0.0165
 2 Lipid_metabolism                Glycerophospholipid metabolism                           23           14            3    0.0239
 3 CarbohydrateMetabolism          Galactose metabolism                                     17            2           12    0.0264
 4 MetabolismOfCofactorsAndVitamins Ubiquinone and other terpenoid-quinone biosyn…          18            6            0    0.0292
 5 AminoAcidMetabolism             Phenylalanine, tyrosine and tryptophan biosyn…           20           10            2    0.0511
 6 Lipid_metabolism                Glycerolipid metabolism                                  16            7            1    0.107
 7 Lipid_metabolism                alpha-Linolenic acid metabolism                           1            0           13    0.133
 8 CarbohydrateMetabolism          Pentose phosphate pathway                                21            2            7    0.159
 9 GlycanBiosynthesisAndMetabolism N-Glycan biosynthesis                                    31            2            7    0.161
10 CarbohydrateMetabolism          Citrate cycle (TCA cycle)                                22           11            5    0.248
# ℹ 58 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figure 3.10: A Ranked List of Metabolic Pathways: *A table showing the ranked list of Metabolic Pathways with p-values to measure the significant difference between genes of HL-tolerant and HL-sensitive algal strains, where Cysteine and methionine metabolism pathway of Amino Acid Metabolism is top of the list followed by Glycerophospholipid metabolism of Lipid metabolism.*

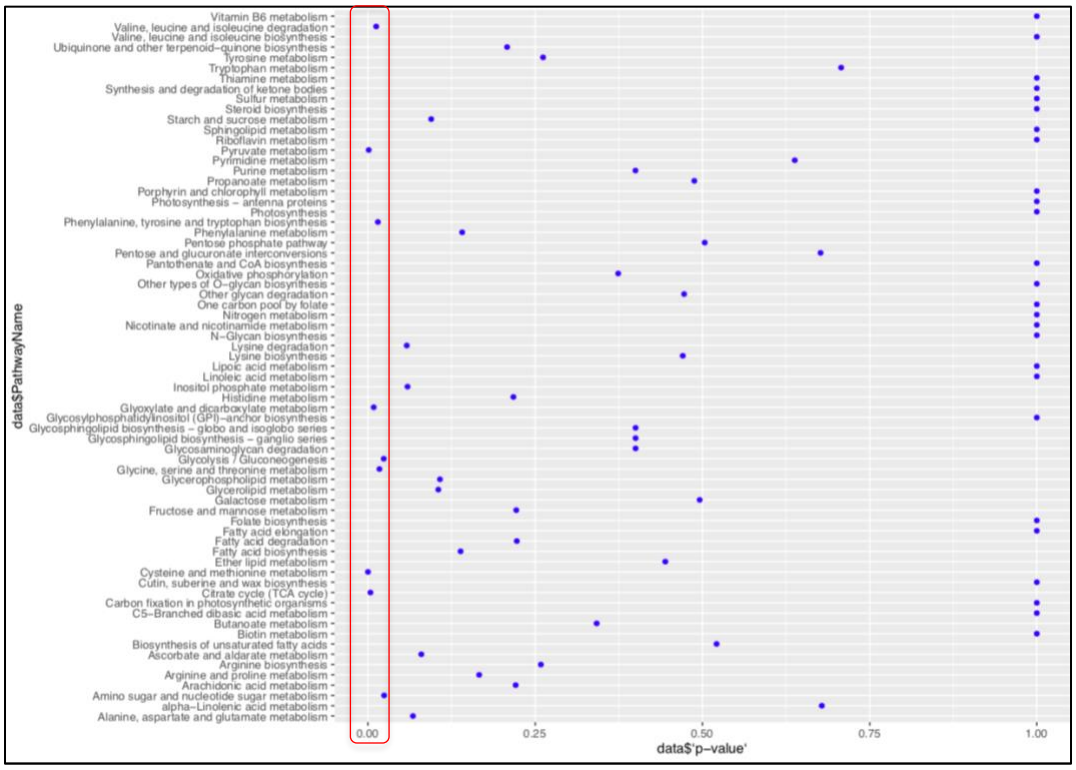


Figure 3.11: Graph of Metabolic Pathways: *Graph showing p-values for a list of Metabolic Pathways to measure the significant difference between genes of HL-tolerant and HL-sensitive algal strains.*

3.7.2. Bonferroni and Benjamini-Hochberg (BH (alias FDR)) Correction

The following plot of the Bonferroni and the Benjamini-Hochberg corrections, between Fischer exact test p-values and adjusted p-values, demonstrates that Bonferroni method is more conservative than the Benjamini-Hochberg method. The Bonferroni's adjusted p-values approach 1.0 very abruptly when the Fischer exact test p-values increase [Figure 3.12].

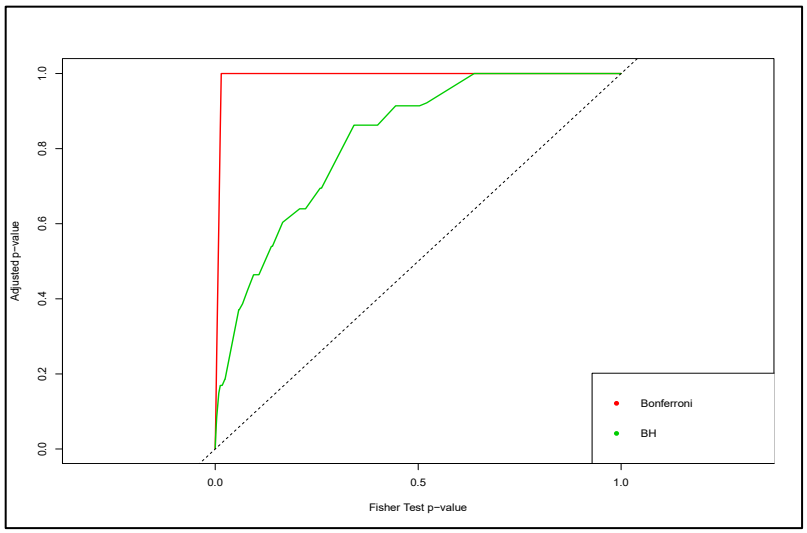


Figure 3.12: Graph showing the Bonferroni and Benjamini-Hochberg *corrections of Fischer exact test:* *This plot shows variance between two distinct multi-test correction methods, i.e. The Bonferroni and the Benjamini-Hochberg corrections, between Fischer exact test p-values and adjusted p-values.*

3.8. Identifying Pathways Variation Between Tolerant and Sensitive Strains

From the ranked pathway list, four pathways were selected based on their relevancy with the lipid production.

3.8.1. Oxidative Phosphorylation Pathway

In Energy metabolism, Oxidative Phosphorylation pathway is showing a missing gene in sensitive strain (Cv11b) this gene belongs to NADH dehydrogenase (NDUA1). This gene is present in tolerant strains [Figure 3.13].

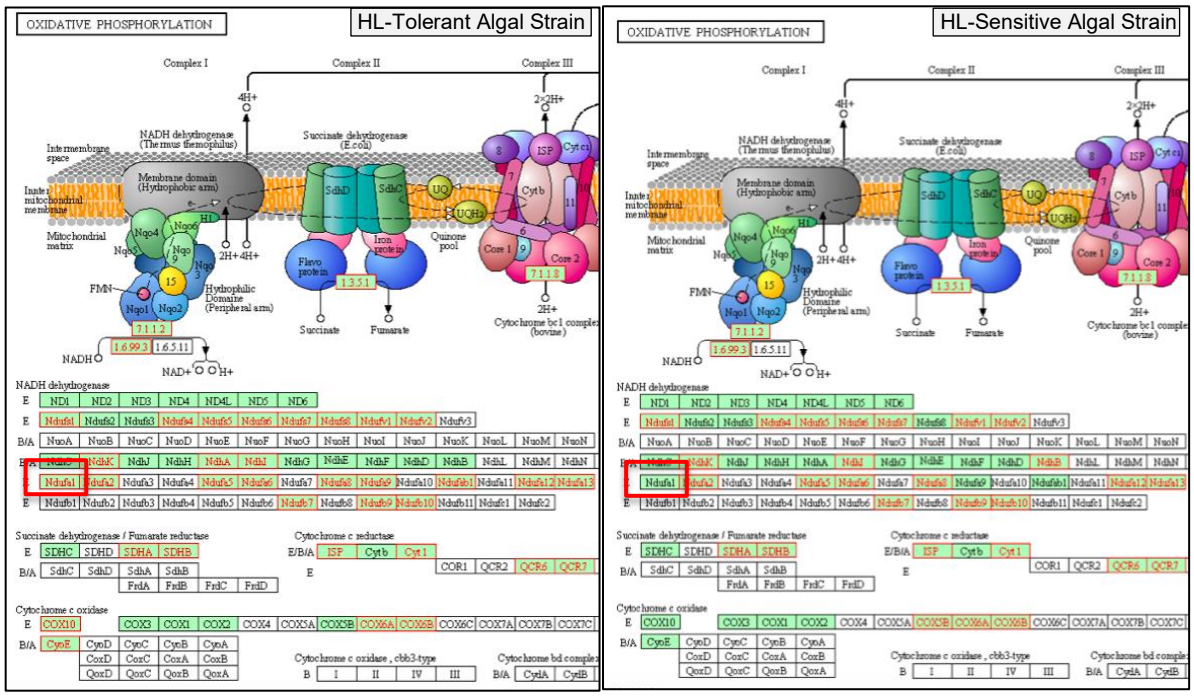


Figure 3.13: Oxidative Phosphorylation Pathway: *This demonstration shows Comparison of Oxidative Phosphorylation pathway mapping with genomes of HL-Tolerant as well as with HL-Sensitive Algal Strains. Where green shaded genes are*

present in the reference genome while red font represents genes' presence in algal genome.

3.8.2. Sulfur Metabolism Pathway

In Energy metabolism, Sulfur metabolism pathway is showing a missing gene in sensitive strain (Cv11b) This missing enzyme is linked to a gene, SOX (Sulfite oxidase). This gene is present in tolerant strains [Figure 3.14].

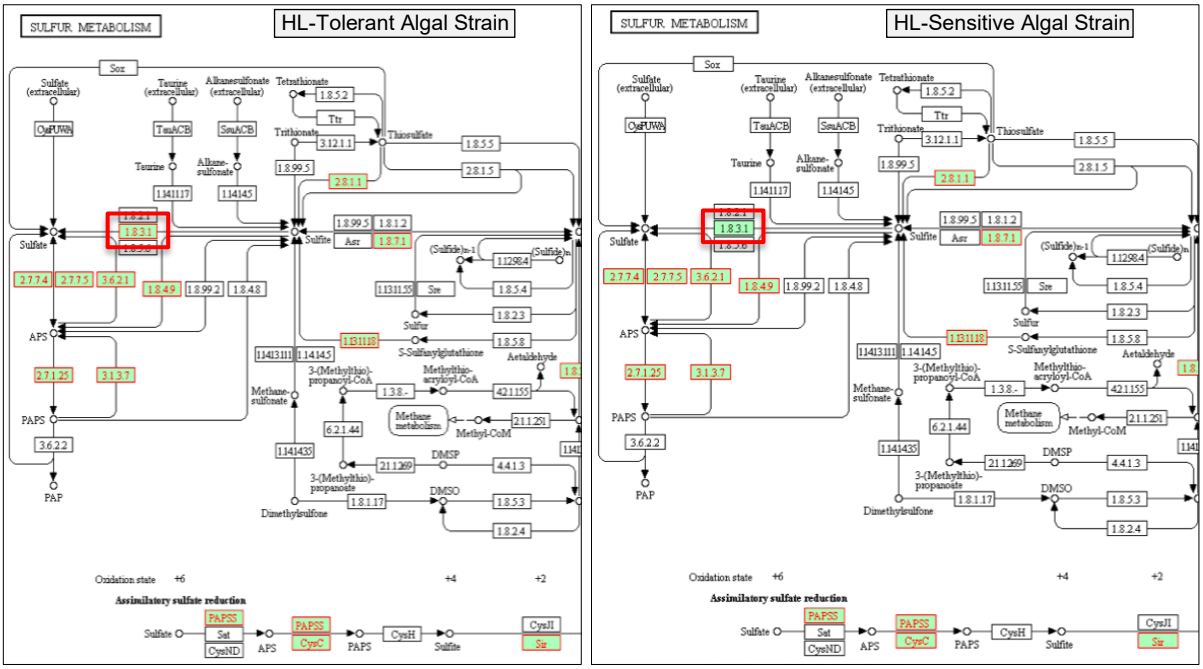


Figure 3.14: Sulfur metabolism Pathway: *This demonstration shows Comparison of Sulfur metabolism pathway mapping with genomes of HL-Tolerant as well as with HL-Sensitive Algal Strains. Where green shaded genes are present in the reference genome while red font represents genes' presence in algal genome.*

3.8.3. Glycerophospholipid Metabolism

In Lipid metabolism, Glycerophospholipid metabolism pathway is also showing a missing gene in sensitive strain (Cv11b). This missing enzyme is linked to three genes, PSD1, PSD2 and PSD3. PSD3 (Phosphatidyl Serine Decarboxylase) gene is present in tolerant strains [Figure 3.15].

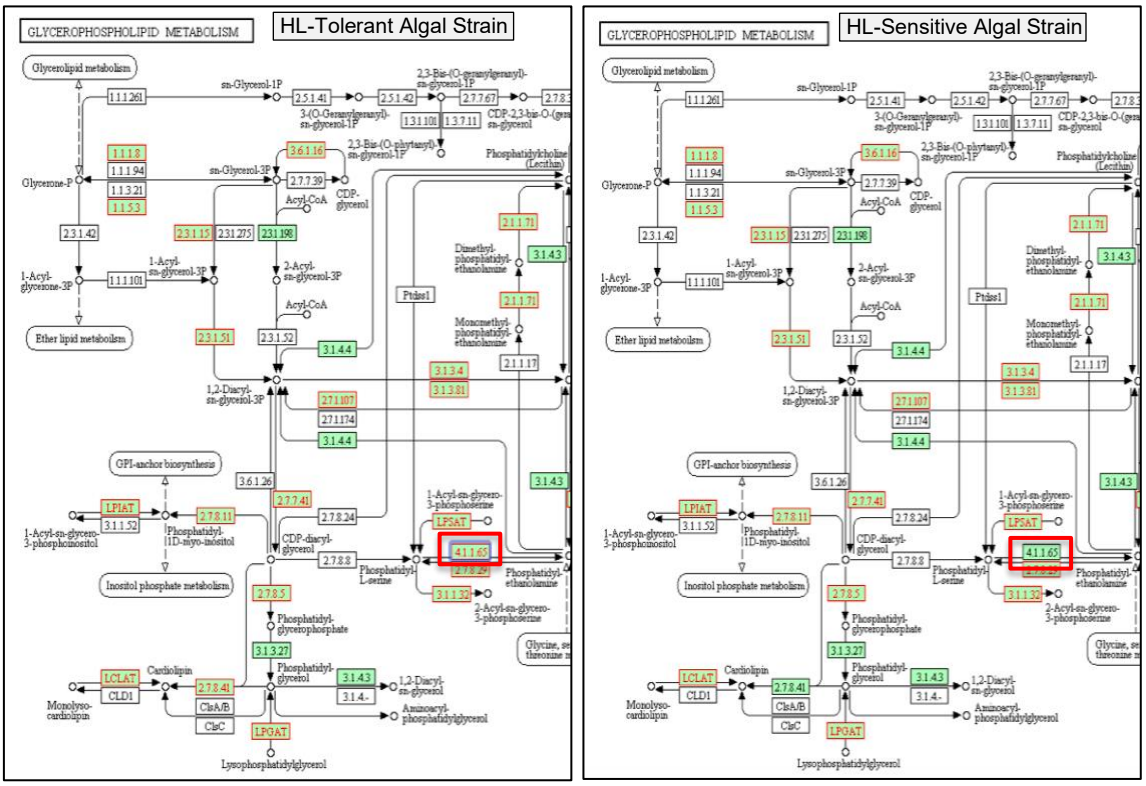


Figure 3.15: Glycerophospholipid metabolism Pathway: *This demonstration shows Comparison of Glycerophospholipid metabolism pathway mapping with genomes of HL-Tolerant as well as with HL-Sensitive Algal Strains. Where green shaded genes are present in the reference genome while red font represents genes' presence in algal genome.*

3.8.4. Sphingolipid Metabolism

In Lipid metabolism, Sphingolipid metabolism pathway is also showing a missing gene in sensitive strain (Cv11b). This missing enzyme is linked to ACD5 (Accelerated Cell Death 5) gene, gene is present in tolerant strains [Figure 3.16].

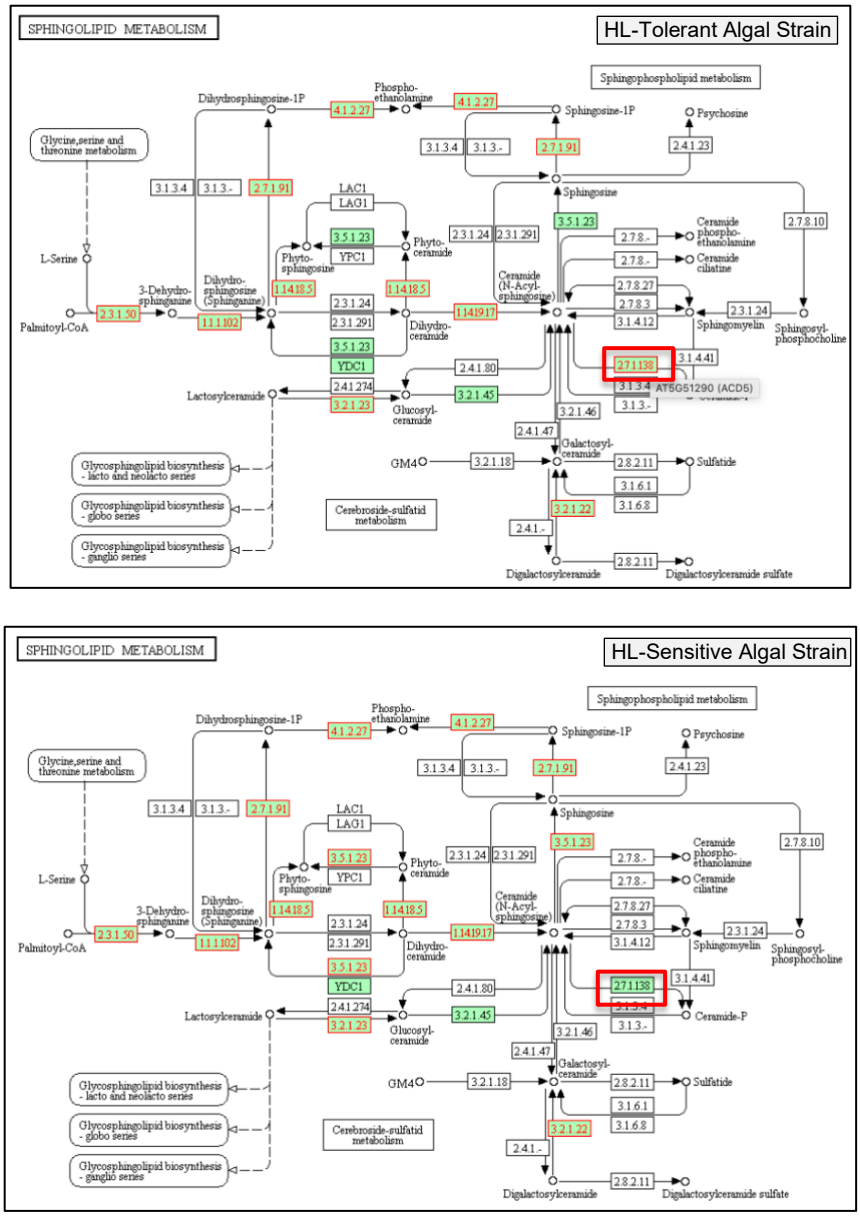


Figure 3.16: Sphingolipid metabolism Pathway: *This demonstration shows Comparison of Sphingolipid metabolism pathway mapping with*

genomes of HL-Tolerant as well as with HL-Sensitive Algal Strains. Where green shaded genes are present in the reference genome while red font represents genes' presence in algal genome.

3.9. Verification Pathways Variation Between Tolerant vs Sensitive Strains

This verification process between HL tolerant and HL sensitive genome, by using TBLASTN, revealed presence of some missing genes like SUOX (Sulfite oxidase) [Figure 3.17] and ACD5 gene (Accelerated Cell Death 5) [Figure 3.18] in sensitive strain (Cv11b). As PSD1 and PSD2 gene could not be mapped in tolerant genomes, PSD3 gene was selected for further investigation.

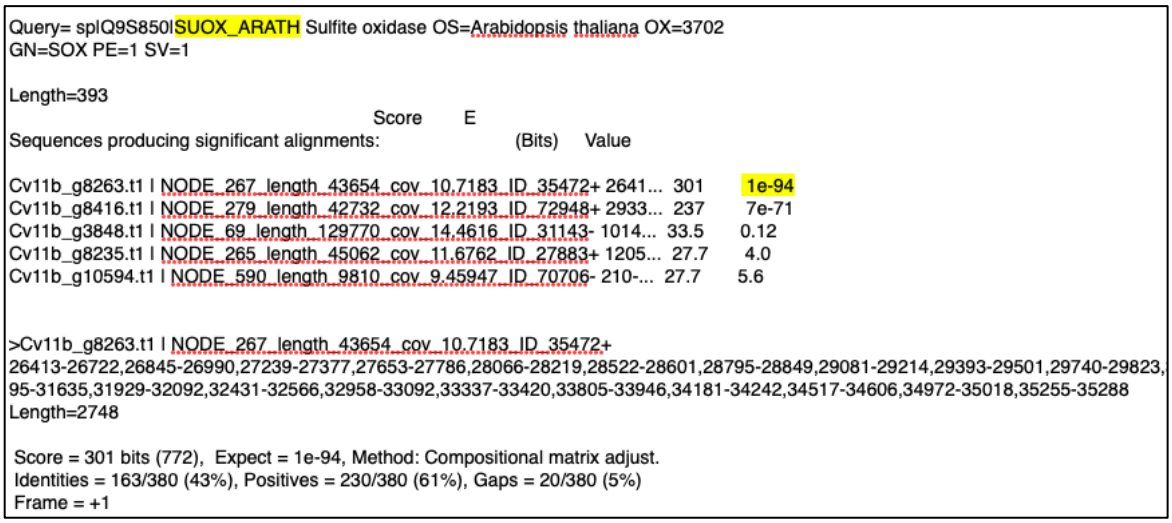


Figure 3.17: Blast query result of SOX (Sulfite oxidase) gene and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, revealed the presence of earlier missing gene SUOX (Sulfite oxidase).*

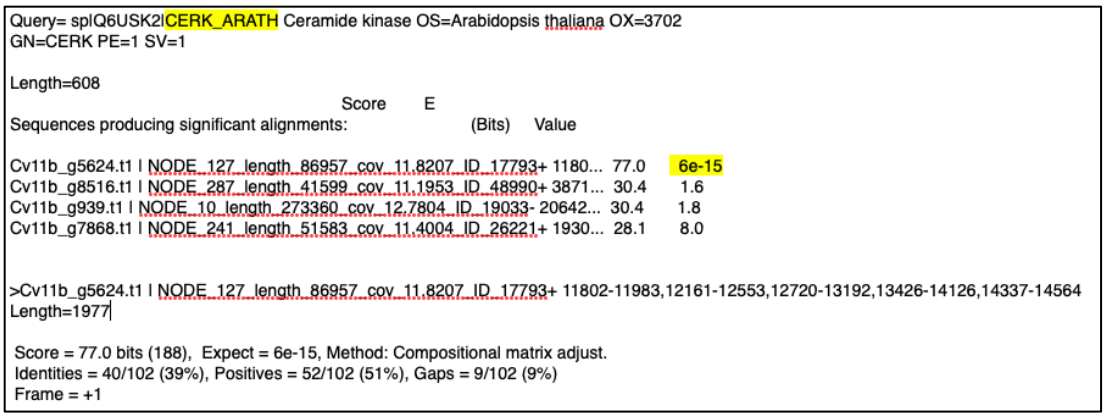


Figure 3.18: Blast query result of CERK (ACD5 gene (Diacylglycerol kinase family protein)) gene and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, revealed the presence of earlier missing gene ACD5 (CERK) gene (Accelerated Cell Death 5)*

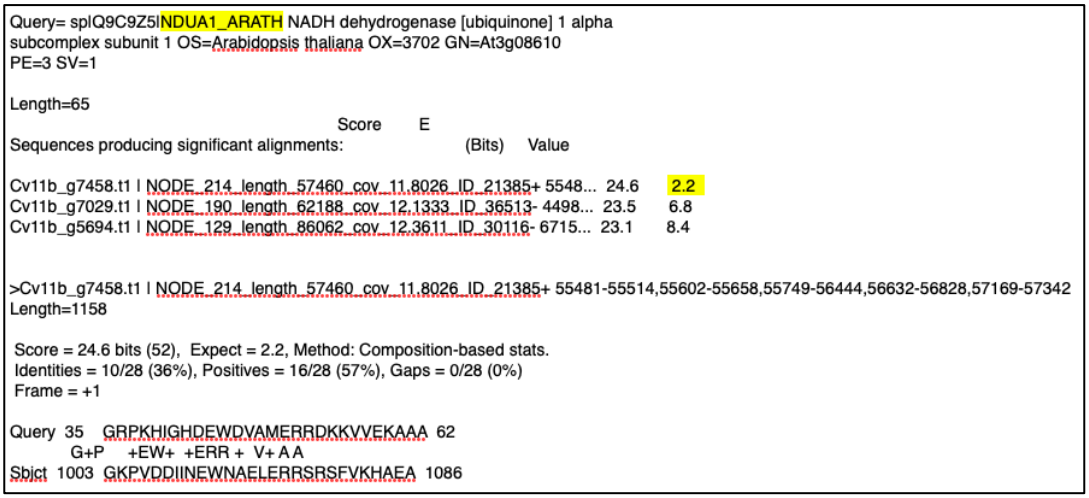
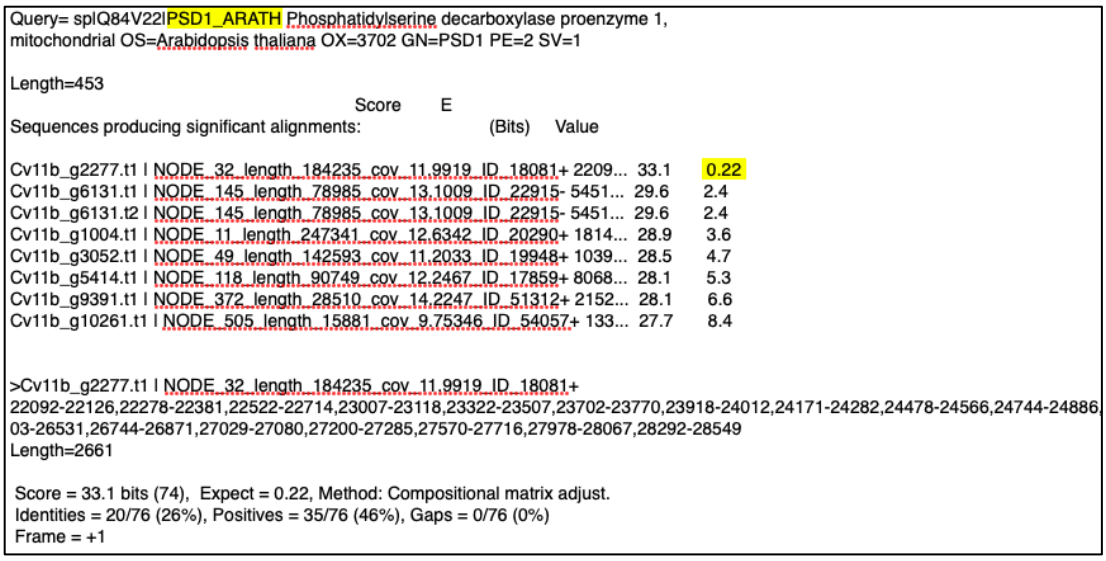


Figure 3.19: Blast query result of NADH dehydrogenase 1 alpha subcomplex subunit 1 and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, confirmed the absence of NDUA1 gene.*



Query= sp|Q84V22|PSD1_ARATH Phosphatidylserine decarboxylase proenzyme 1, mitochondrial OS=Arabidopsis thaliana OX=3702 GN=PSD1 PE=2 SV=1

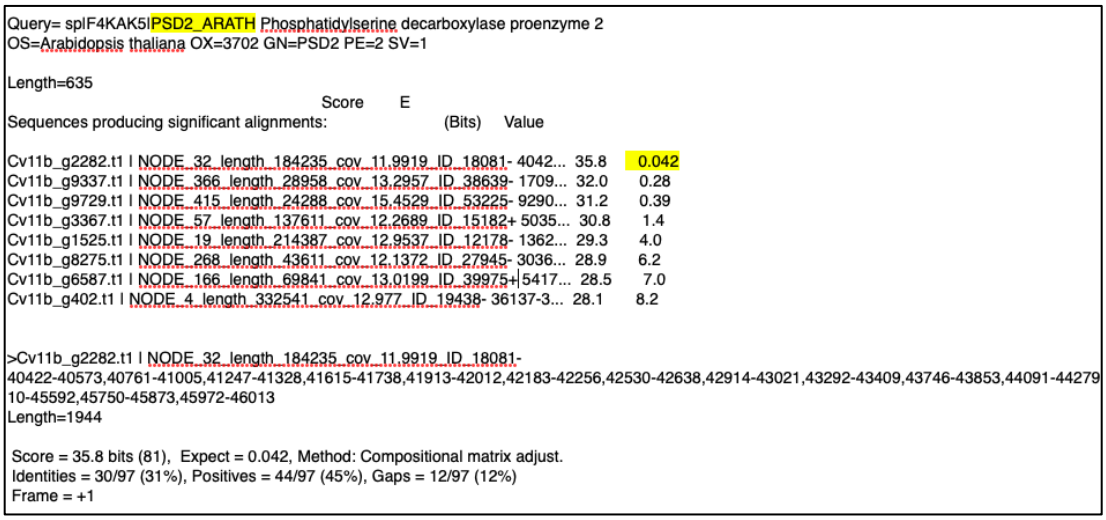
Length=453

Sequences producing significant alignments:	Score (Bits)	E Value
Cv11b_g2277.t1 | NODE_32_length_184235_cov_11.9919_ID_18081+ 2209...	33.1	0.22
Cv11b_g6131.t1 | NODE_145_length_78985_cov_13.1009_ID_22915- 5451...	29.6	2.4
Cv11b_g6131.t2 | NODE_145_length_78985_cov_13.1009_ID_22915- 5451...	29.6	2.4
Cv11b_g1004.t1 | NODE_11_length_247341_cov_12.6342_ID_20290+ 1814...	28.9	3.6
Cv11b_g3052.t1 | NODE_49_length_142593_cov_11.2033_ID_19948+ 1039...	28.5	4.7
Cv11b_g5414.t1 | NODE_118_length_90749_cov_12.2467_ID_17859+ 8068...	28.1	5.3
Cv11b_g9391.t1 | NODE_372_length_28510_cov_14.2247_ID_51312+ 2152...	28.1	6.6
Cv11b_g10261.t1 | NODE_505_length_15881_cov_9.75346_ID_54057+ 133...	27.7	8.4

>Cv11b_g2277.t1 | NODE_32_length_184235_cov_11.9919_ID_18081+
22092-22126,22278-22381,22522-22714,23007-23118,23322-23507,23702-23770,23918-24012,24171-24282,24478-24566,24744-24886,
03-26531,26744-26871,27029-27080,27200-27285,27570-27716,27978-28067,28292-28549
Length=2661

Score = 33.1 bits (74), Expect = 0.22, Method: Compositional matrix adjust.
Identities = 20/76 (26%), Positives = 35/76 (46%), Gaps = 0/76 (0%)
Frame = +1

Figure 3.20: Blast query result of PSD1_ARATH and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, confirmed the absence of PSD1 gene.*



Query= sp|F4KAK5|PSD2_ARATH Phosphatidylserine decarboxylase proenzyme 2 OS=Arabidopsis thaliana OX=3702 GN=PSD2 PE=2 SV=1

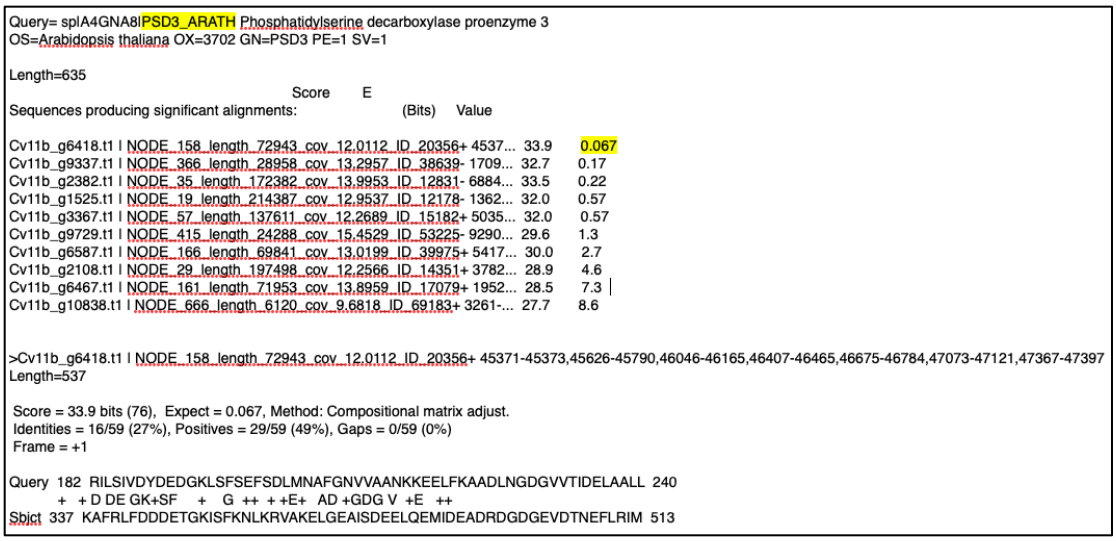
Length=635

Sequences producing significant alignments:	Score (Bits)	E Value
Cv11b_g2282.t1 | NODE_32_length_184235_cov_11.9919_ID_18081- 4042...	35.8	0.042
Cv11b_g9337.t1 | NODE_366_length_28958_cov_13.2957_ID_38639- 1709...	32.0	0.28
Cv11b_g9729.t1 | NODE_415_length_24288_cov_15.4529_ID_53225- 9290...	31.2	0.39
Cv11b_g3367.t1 | NODE_57_length_137611_cov_12.2689_ID_15182+ 5035...	30.8	1.4
Cv11b_g1525.t1 | NODE_19_length_214387_cov_12.9537_ID_12178- 1362...	29.3	4.0
Cv11b_g8275.t1 | NODE_268_length_43611_cov_12.1372_ID_27945- 3036...	28.9	6.2
Cv11b_g6587.t1 | NODE_166_length_69841_cov_13.0199_ID_39975+ 5417...	28.5	7.0
Cv11b_g402.t1 | NODE_4_length_332541_cov_12.977_ID_19438- 36137-3...	28.1	8.2

>Cv11b_g2282.t1 | NODE_32_length_184235_cov_11.9919_ID_18081-
40422-40573,40761-41005,41247-41328,41615-41738,41913-42012,42183-42256,42530-42638,42914-43021,43292-43409,43746-43853,44091-44279
10-45592,45750-45873,45972-46013
Length=1944

Score = 35.8 bits (81), Expect = 0.042, Method: Compositional matrix adjust.
Identities = 30/97 (31%), Positives = 44/97 (45%), Gaps = 12/97 (12%)
Frame = +1

Figure 3.21: Blast query result of PSD2_ARATH and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, confirmed the absence of PSD2 gene.*



```
Query= splA4GNA8|PSD3_ARATH Phosphatidylserine decarboxylase proenzyme 3
OS=Arabidopsis thaliana OX=3702 GN=PSD3 PE=1 SV=1

Length=635
                                                                    Score     E
Sequences producing significant alignments:                         (Bits)    Value

Cv11b_g6418.t1 | NODE_158_length_72943_cov_12.0112_ID_20356+ 4537...  33.9      0.067
Cv11b_g9337.t1 | NODE_366_length_28958_cov_13.2957_ID_38639- 1709...  32.7      0.17
Cv11b_g2382.t1 | NODE_35_length_172382_cov_13.9953_ID_12831- 6884...  33.5      0.22
Cv11b_g1525.t1 | NODE_19_length_214387_cov_12.9537_ID_12178- 1362...  32.0      0.57
Cv11b_g3367.t1 | NODE_57_length_137611_cov_12.2689_ID_15182+ 5035...  32.0      0.57
Cv11b_g9729.t1 | NODE_415_length_24288_cov_15.4529_ID_53225- 9290...  29.6      1.3
Cv11b_g6587.t1 | NODE_166_length_69841_cov_13.0199_ID_39975+ 5417...  30.0      2.7
Cv11b_g2108.t1 | NODE_29_length_197498_cov_12.2566_ID_14351+ 3782...  28.9      4.6
Cv11b_g6467.t1 | NODE_161_length_71953_cov_13.8959_ID_17079+ 1952...  28.5      7.3
Cv11b_g10838.t1 | NODE_666_length_6120_cov_9.6818_ID_69183+ 3261-...  27.7      8.6


>Cv11b_g6418.t1 | NODE_158_length_72943_cov_12.0112_ID_20356+ 45371-45373,45626-45790,46046-46165,46407-46465,46675-46784,47073-47121,47367-47397
Length=537

 Score = 33.9 bits (76),  Expect = 0.067, Method: Compositional matrix adjust.
 Identities = 16/59 (27%), Positives = 29/59 (49%), Gaps = 0/59 (0%)
 Frame = +1

Query  182  RILSIVDYDEDGKLSFSEFSDLMNAFGNVVAANKKEELFKAADLNGDGVVTIDELAALL  240
            +  + D DE GK+SF    +  G  ++  + +E+  AD +GDG V  +E  ++
Sbjct  337  KAFRLFDDDETGKISFKNLKRVAKELGEAISDEELQEMIDEADRDGDGEVDTNEFLRIM  513
```

Figure 3.22: Blast query result of PSD3_ARATH and Cv11b genome: *The verification process of HL sensitive genome mapping with HL-tolerant gene, by using TBLASTN tool, confirmed the absence of PSD3 gene.*

Genes	HL-Tolerant Algal Strains	HL-Sensitive Algal Strains
PSD1	✘	✘
PSD2	✘	✘
PSD3	✓	✘
NDUA1	✓	✘
SUOX	✓	✓
ACD5	✓	✓

Table 3.1: List of the most significant genes and their presence after verification

Conclusively, that NADH dehydrogenase and PSD genes could not be identified in Cv11b in the BLAST searches [Table 3.1].

3.10. Synteny Analysis

To confirm the absence of these genes, the genomic locus where they would be expected based on synteny was inspected. Disruption of synteny could be due to contig rearrangement processes such as:

- Translocation, inversion, contig fusion, and breakage;
- Gene, segment, and contig duplication and loss;
- Polyploidization and return to diploidy;

Synteny analysis assists to identify missing and duplicating genomic regions. It also supports to reveal inversion or deleted segments. The contigs of HL-sensitive strain Cv11b were mapped with the contigs of all HL-tolerant strains Cs8k, Cv264 and CvC1 in search of PSD3 gene. All these depictions are generated by synteny analysis tool SimpleSynteny v1.5 [98]. Where matching genes are linked with same colour lines, while change in gene direction is represented by turning arrow.

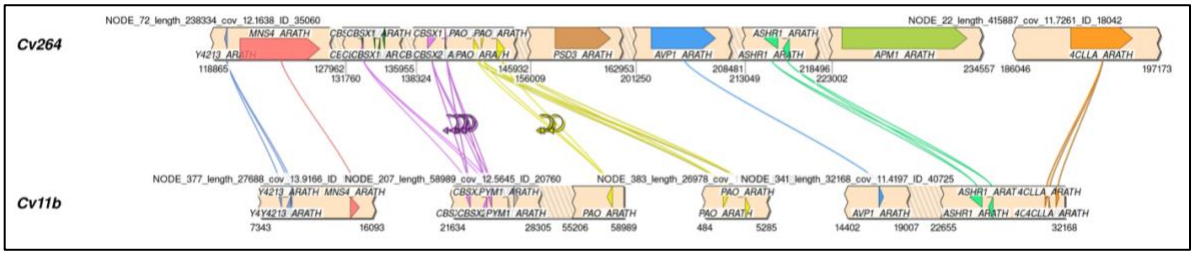


Figure 3.23: Synteny analysis for PSD3 gene: *This contig mapping depiction shows the comparison between two algal strains (i.e., Cv264 and Cv11b) with matching genome contigs. Where PSD3 gene is visible on the Node_72 of Cv264 algal strain while this gene is missing in Cv11b contigs.*

According to this synteny analysis, PSD3 gene was visible on the Node_72 of Cv264 algal strain [Figure 3.23]. While, CvC1 contig showed PSD3 gene location on Node_68 [Figure 3.24]. However, contigs of Cs8k algal strain displayed that PSD3 gene is located on Node_117 [Figure 3.25]. Nonetheless, PSD3 gene could not be mapped on the contigs of Cv11b in all these analytical figures.

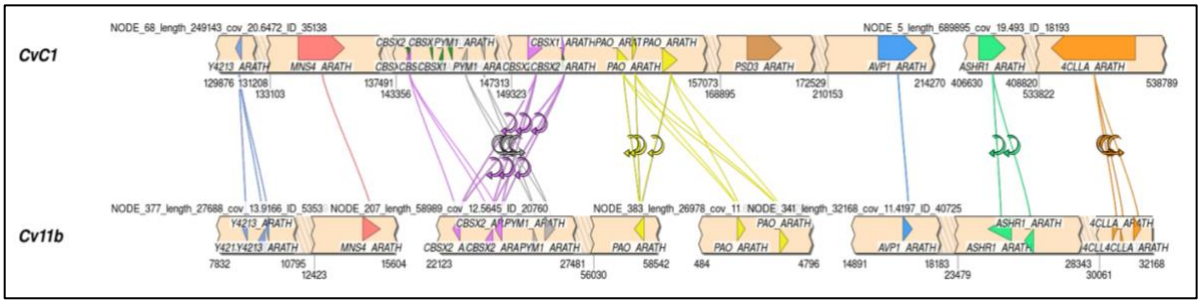


Figure 3.24: Synteny analysis for PSD3 gene: *This contig mapping depiction shows the comparison between two algal strains (i.e., CvC1 and Cv11b) with matching genome contigs. This analysis shows that PSD3 gene location on Node_68 of CvC1 contig. Whereas no mapping of this gene is visible with Cv11b contigs.*

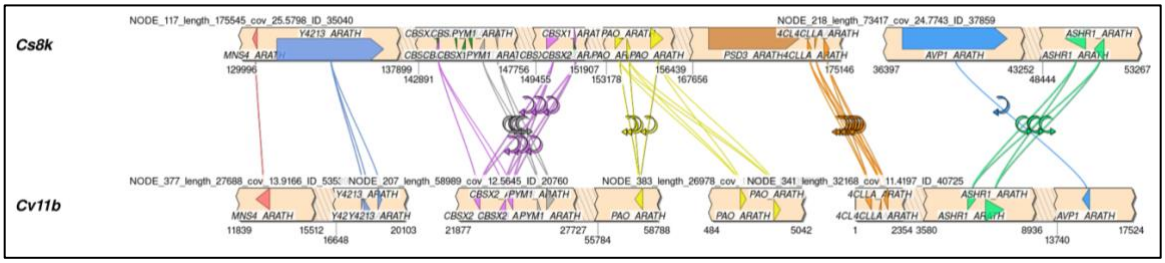


Figure 3.25: Synteny analysis for PSD3 gene: *This contig mapping depiction shows the comparison between two algal strains (i.e., Cs8k and Cv11b) with matching genome contigs. Contigs of Cs8k algal strain displayed that PSD3 gene is located on Node_117. However, no mapping is visible with Cv11b contigs.*

Conversely, contigs of all algal strains Cv264, CvC1, Cs8k and Cv11b were mapped together [Figure 3.26], according to the synteny analysis, PSD3 gene was visible on the contigs of all HL-tolerant algal strains Cv264, CvC1 and Cs8k. While this gene could not be mapped in Cv11b, HL-Sensitive algal genome [Figure 3.26].

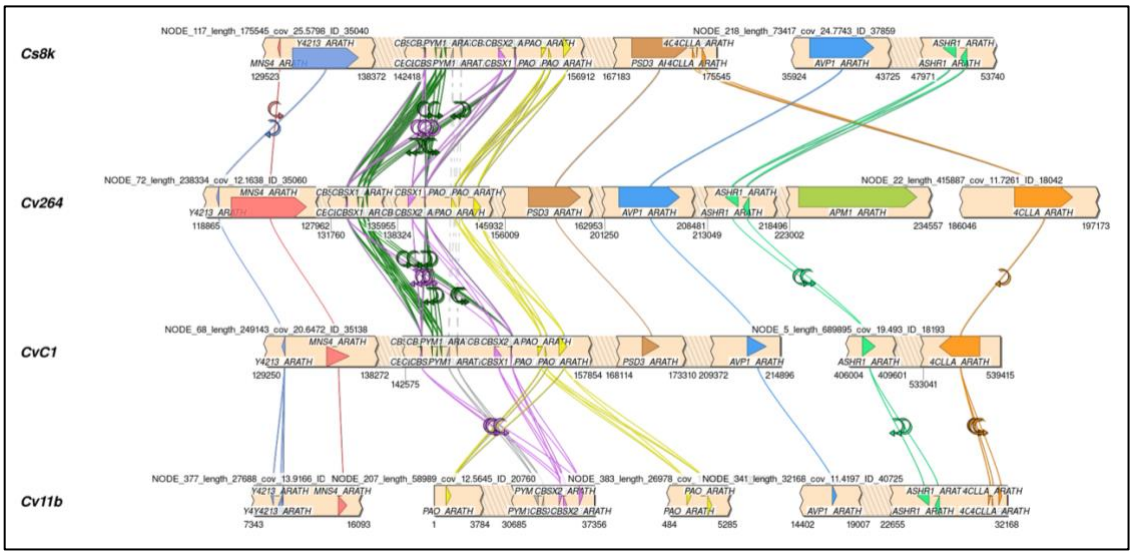


Figure 3.26: Synteny analysis for PSD3 gene: *This contig mapping depiction shows the comparison among four different algal strains (i.e., Cs8k, Cv264, CvC1, and Cv11b) with matching genome contigs. This analysis shows PSD3 gene is visible on the contigs of HL-tolerant Cv264, CvC1 and Cs8k algal strains. While this gene could not be mapped in Cv11b. It also shows the alignment of contiguous genes*

This analysis also revealed that neighbouring genes of PSD3 genomic region are visible on the contig (i.e., Node_72) of Cv264 algal strain. While comparing this contig with CvC1 algal strain showed that adjacent genes are located on a contig (i.e., Node_68) but their direction and positions of some genes are changed.

However, comparing this contig with Cs8k algal strain showed that neighbouring genes are located on two different contigs (i.e., Node_117 and Node_218); moreover, their direction and positions are also changed. While Cv11b contigs showed that contiguous genes are located on four different contigs (i.e., Node_207, Node_341, Node_377 and Node_383); additionally, their direction and positions are also changed. However, PSD3 gene could not mapped in Cv11b, HL-Sensitive algal genome.

3.10.1. Pairwise Synteny Analysis for Selected Nodes of Algal Strains to Explore the PSD3 Gene

This pairwise synteny analysis of Cv264, CvC1, Cs8k and Cv11b was generated according to the sequence order of the genome and then the nodes were reordered and flipped to check and investigate the better-aligned mapping between genome contigs.

3.10.1.1. Cs8k vs Cv264 Contigs

Pairwise synteny analysis between HL-tolerant strains Cs8k and Cv264, conducted without flipping or reordering their contigs, revealed that the genome regions were generally well aligned. However, the orientation of many genes was reversed, although the PSD3 gene aligned well on the corresponding contigs of both HL-tolerant strains.

After reordering the contigs of strain Cv264, the alignment improved further. Matching genes were connected by lines of the same colour, while changes in gene orientation were indicated by curved arrows. Notably, one gene, 4CLLA_ARATH, from strain Cs8k could not be located near its neighbouring genes in strain Cv264 [Figure 3.27].

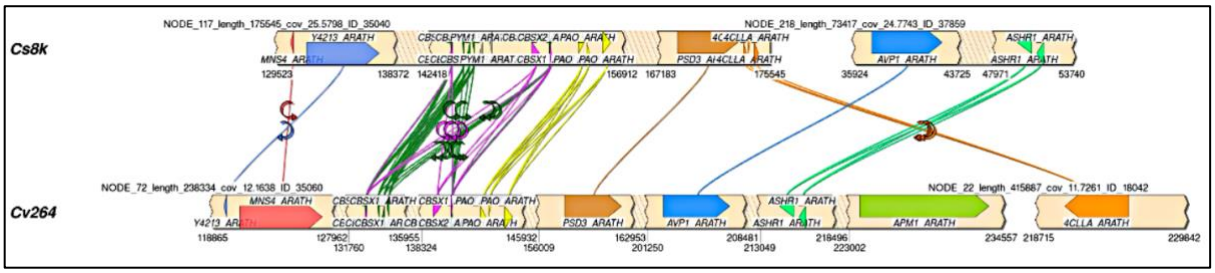


Figure 3.27: Synteny Cs8k vs Cv264 Contigs: *This contig mapping depiction shows the comparison between Cs8k and Cv264, with matching genome contigs, after reordering of Cv264 contigs. PSD3 gene is well aligned and mapped between the contigs of both HL-tolerant strains.*

3.10.1.2. Cs8k vs CvC1 Contigs

Pair-wise synteny analysis between HL-tolerant strains Cs8k and CvC1, without flipping and reordering of their contigs, proposed that contigs were aligned but the direction of multiple genes was reversed, while PSD3 gene was reasonably mapped on both HL-tolerant strains' contigs.

Subsequently, after flipping CvC1 node 68 and reordering of CvC1 contigs, genome region was very well aligned; however, 4CLLA_ARATH and

ASHR1_ARATH genes of Cs8k were located inversely without their vicinity on Cv264 contigs [Figure 3.28].

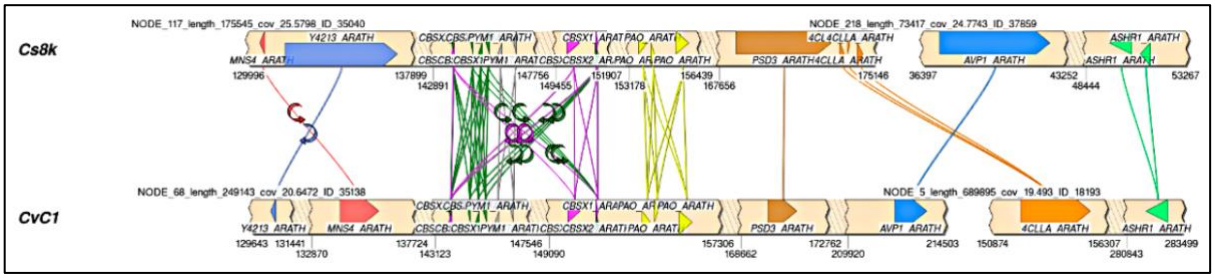


Figure 3.28: Synteny Cs8k vs CvC1 Contigs: *This contig mapping depiction shows the comparison between Cs8k and CvC1, with matching genome contigs, after flipping Node 68 and reordering of CVC1 contigs. Where PSD3 gene was reasonably mapped on both HL-tolerant strains' contigs.*

3.10.1.3. Cs8k vs Cv11b Contigs

Pair-wise synteny analysis between HL-tolerant strain Cs8k and HL-sensitive strain Cv11b, without flipping and reordering of their contigs, anticipated that genes of Cs8k node 117 were dispersed on different contigs of Cv11b and the direction of many genes was reversed; however, it was evident in the synteny representation that PSD3 gene was not mapped between HL-tolerant and HL-sensitive strains. Successively, flipping of Cv11b node 207 and reordering of Cv11b contigs assisted to align these genomic regions in a better way.

Next synteny analysis was conducted by flipping both Cv11b nodes 207 and 377 and reordering of Cv11b contigs; and later on, by flipping three Cv11b

nodes 207, 377 and 341 and reordering of Cv11b contigs [Figure 3.29]. These rearrangements aided to enhance orientation between these genomic regions.

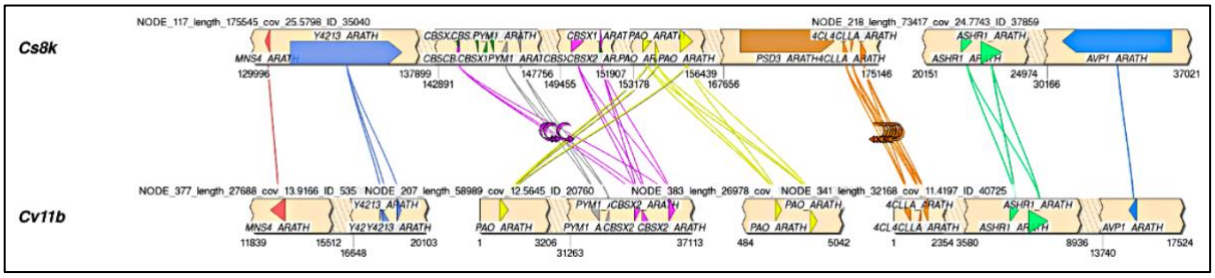


Figure 3.29: Synteny Cs8k vs Cv11b Contigs: *This contig mapping depiction shows the comparison between Cs8k and Cv11b, with matching genome contigs, after flipping Nodes 377 and 341 and reordering of Cv11b contigs. The reorganization aided for better orientation between these genomic regions.*

3.10.1.4. Cv264 vs CvC1 Contigs

A pair-wise synteny analysis between HL-tolerant strains Cv264 and CvC1, without flipping and reordering of their contigs, projected that genes of Cv264 node 72 were very well aligned with CvC1 node 68 except they were aligned in reverse order, while PSD3 gene was reasonably brought into link on both HL-tolerant strains' contigs. Continually, flipping of Cv264 node 72 aided to align these genomic regions in a better way.

However, later synteny analysis was conducted by flipping both CvC1 nodes 5 and 68 and reordering of Cv264 and Cv11b contigs [Figure 3.30]. These

rearrangements revealed a straight mapping orientation of these genomic regions.

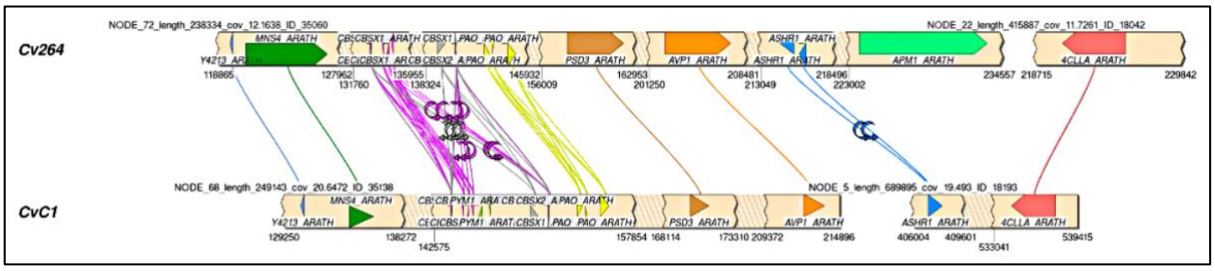


Figure 3.30: Synteny Cv264 vs CvC1 Contigs: *This contig mapping depiction shows the comparison between Cv264 and CvC1, with matching genome contigs, after flipping Nodes 5 and 68 and reordering of Cv264 and CvC1 contigs. The readjustment revealed a straight mapping orientation between the genomic regions of Cv264 and CvC1.*

3.10.1.5. Cv264 vs Cv11b Contigs

Subsequently, pair-wise synteny analysis between HL-tolerant strain Cv264 and HL-sensitive strain Cv11b, without flipping and reordering of their contigs, projected that genes of Cv264 node 72 were dispersed on different contigs of Cv11b and the direction of few genes was reversed, manifestly PSD3 gene's traces were not found on HL-sensitive strain contigs. Continuously, reordering of Cv264 contigs aligned these genomic regions in a better approach.

Afterwards, synteny analysis was managed by aligning Cv11b nodes 341 and 383 and against Cv264 node 72 contigs [Figure 3.31]. These readjustments facilitated to align these genomic regions with better coordination.

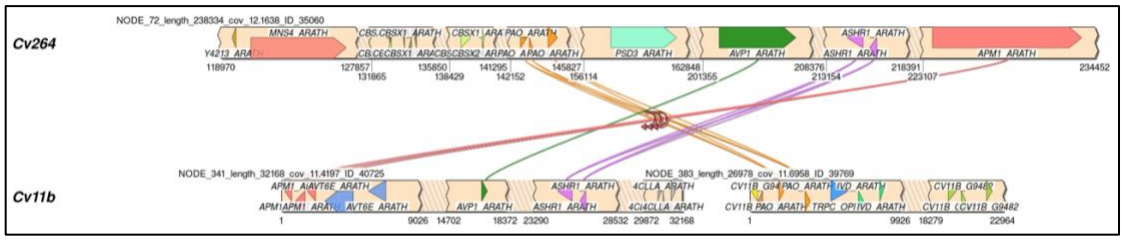


Figure 3.31: Synteny of Cv11b (Nodes 341, and 383) vs Cv264 (Node 72) Contigs: *This contig mapping depiction shows the comparison between Cv264 Node 72 and Cv11b Nodes 341 and 383, with matching genome contigs, after reordering of Cv264 contig. This mapping revealed that contiguous genes of PSD3 gene are dispersed on two different contigs of Cv11b.*

3.10.1.6. <u>CvC1 vs Cv11b Contigs</u>

Consequently, pair-wise synteny analysis between HL-tolerant strain CvC1 and HL-sensitive strain Cv11b, without flipping and reordering of their contigs, presented that genes of CvC1 node 68 were dispersed on multiple contigs of Cv11b and the direction of many genes was reversed, where it is obvious that PSD3 gene was not identified on HL-sensitive strain's contigs. Similarly, flipping of Cv11b nodes 341 and 377 and reordering of Cv11b contigs brought into line these genomic regions with few genes in reverse order.

Then, after flipping Cv11b nodes 383, 207 and 377 and reordering of Cv11b contigs. Successively, Cv11b nodes 341 and 383 were lined up against Cs8k node 68 to explore orientation of these genomic regions entirely [Figure 3.32].

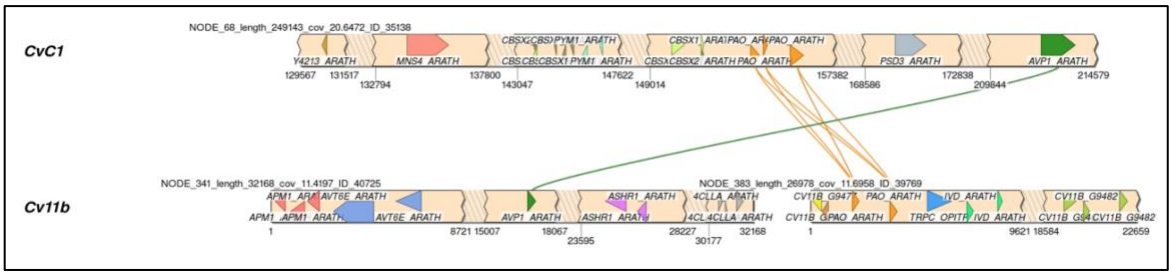


Figure 3.32: Synteny of Cv11b (Nodes 341, and 383) vs CvC1 (Node 68)

Contigs: *This contig mapping depiction shows the comparison between CvC1 Node 68 and Cv11b Nodes 341and 383, with matching genome contigs. This illustration disclosed that neighbouring genes of PSD3 genes are not well aligned between these contigs.*

3.10.2. Mapping of Cv264 with Cv11b Nodes

3.10.2.1. <u>Synteny of Cv11b vs Cv264 Contigs</u>

Neighbouring genomic regions of HL-sensitive strain Cv11b nodes 207, 341, 342, 377, 382 and 383 were further explored by positioning them against Cv264 nodes 22 and 72 contigs [Figure 3.33]. This investigation revealed that Cv11b nodes 342 and 382 were not mapped with genomic region of Cv264. Moreover, no trace of PSD3 gene was discovered on Cv11b contigs.

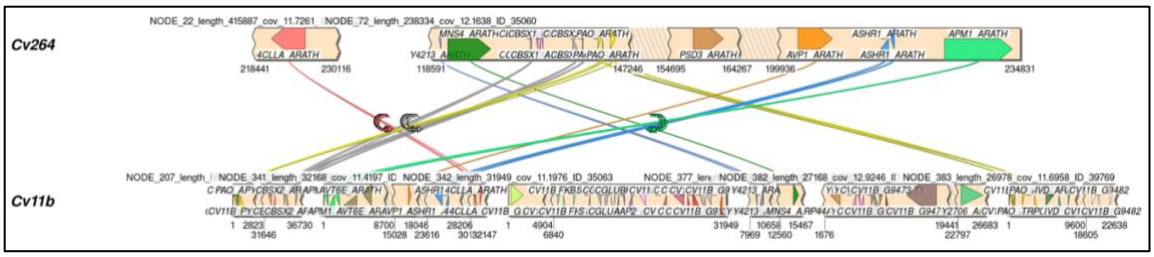


Figure 3.33: Synteny of Cv11b vs Cv264 Contigs: *This contig mapping depiction shows the comparison between Cv264 and Cv11b Nodes 207, 341, 342, 377, 382 and 383. This analysis showed nice alignment of neighbouring genes of PSD3 gene located on Cv264 contig with multiple contigs of Cv11b.*

3.10.2.2. Synteny of Cv264 (Node 72) and Cv11b (Node 207)

Mapping of node 72 of Cv264 and node 207 of Cv11b showed that PSD3 neighbouring genes located on node 72 of Cv264 is mapped in a reverse order with contig node 207 of Cv11b. Moreover, it can be anticipated that the genomic region of PSD3 gene is a part of deleted genomic region. Likewise, a very small genomic region is mapped with HACL_ARATH gene [Figure 3.34].

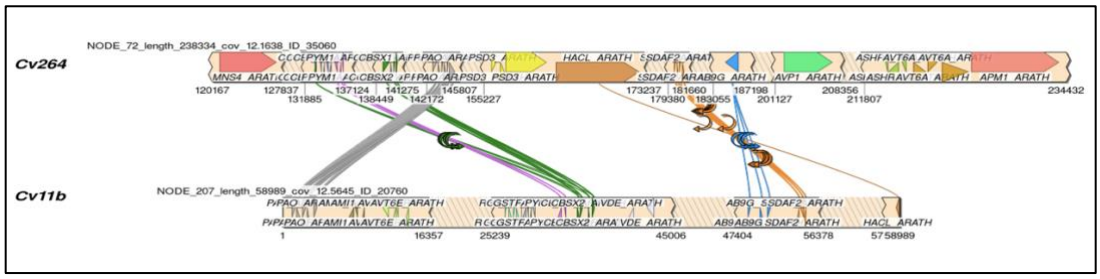
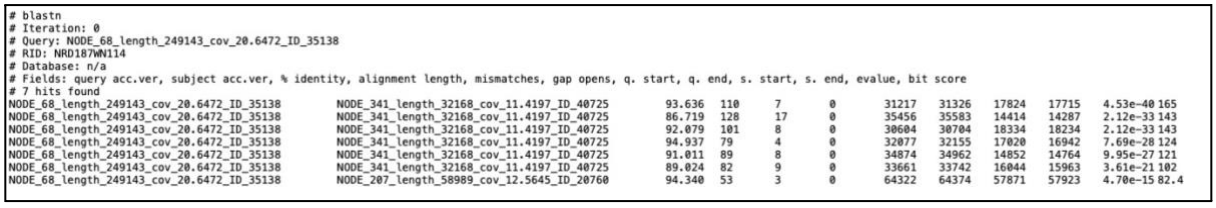


Figure 3.34: Synteny of Node 72 of Cv264 and Node 207 of Cv11b: *This contig mapping depiction shows the comparison between Cv264 Node 72 and Cv11b Node 207. This figure uncovers that PSD3 adjacent genes located on node 72 of Cv264 is mapped in a reverse order with node 207 of Cv11b contig.*

3.11. Re-evaluation of Missing Genomic Region by Blast and Synteny Analyses

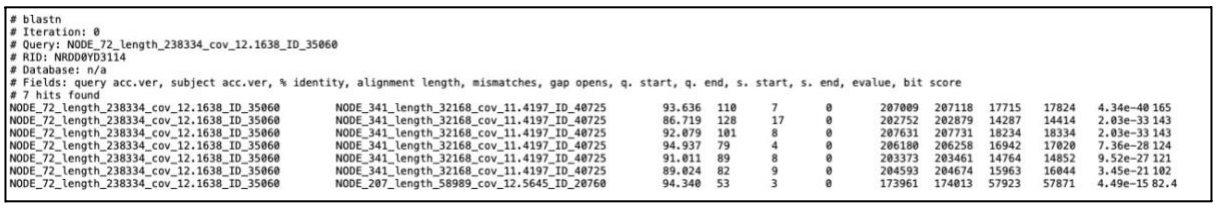
3.11.1. Sequence Alignment by Using TBLASTN and TBLASTX

Sequence alignment between selected nodes of CvC1 strain (i.e., NODE_68) and Cv11b strain (i.e., NODE_207, NODE_383 and NODE_341); with contiguous megablast alignment showed significant results due to high similarity between sequences with very significant e-values and bit scores [Figure 3.35]. Likewise, sequence alignment between nodes of Cv264 strain (i.e., NODE_72) and Cv11b strain (i.e., NODE_207, NODE_383 and NODE_341) by using contiguous megablast also showed high similarity between sequences [Figure 3.36].



```
# blastn
# Iteration: 0
# Query: NODE_68_length_249143_cov_20.6472_ID_35138
# RID: NRD187WN114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 7 hits found
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	93.636	110	7	0	31217	31326	17824	17715	4.53e-40	165
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	86.719	128	17	0	35456	35583	14414	14287	2.12e-33	143
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	92.079	101	8	0	30604	30704	18334	18234	2.12e-33	143
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	94.937	79	4	0	32077	32155	17020	16942	7.69e-28	124
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	91.011	89	8	0	34874	34962	14852	14764	9.95e-27	121
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_341_length_32168_cov_11.4197_ID_40725	89.024	82	9	0	33661	33742	16044	15963	3.61e-21	102
NODE_68_length_249143_cov_20.6472_ID_35138	NODE_207_length_58989_cov_12.5645_ID_20760	94.340	53	3	0	64322	64374	57871	57923	4.70e-15	82.4
```

Figure 3.35: BLASTN Alignment between CvC1 NODE 68 and Cv11b NODES 207, 383, and 341



```
# blastn
# Iteration: 0
# Query: NODE_72_length_238334_cov_12.1638_ID_35060
# RID: NRDD0YD3114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 7 hits found
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	93.636	110	7	0	207009	207118	17715	17824	4.34e-40	165
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	86.719	128	17	0	202752	202879	14287	14414	2.03e-33	143
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	92.079	101	8	0	207631	207731	18234	18334	2.03e-33	143
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	94.937	79	4	0	206180	206258	16942	17020	7.36e-28	124
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	91.011	89	8	0	203373	203461	14764	14852	9.52e-27	121
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_341_length_32168_cov_11.4197_ID_40725	89.024	82	9	0	204593	204674	15963	16044	3.45e-21	102
NODE_72_length_238334_cov_12.1638_ID_35060	NODE_207_length_58989_cov_12.5645_ID_20760	94.340	53	3	0	173961	174013	57923	57871	4.49e-15	82.4
```

Figure 3.36: BLASTN Alignment between Cv264 NODE 72 and Cv11b NODES 207, 383, and 341

While, sequence alignment between nodes of Cs8k strain (i.e., NODE_117) and Cv11b strain (i.e., NODE_207, NODE_383 and NODE_341) by using contiguous megablast indicated that there is no significant similarity found [Figure 3.37]. Then this alignment was done by using discontinuous megablast, which showed moderate result with dissimilarities in alignment. Which revealed that sequence alignment has many dissimilarities between the nodes of Cs8k and Cv11b strains [Figure 3.38].

Figure 3.37: MEGABLAST Alignment between Cs8k NODE 117 and Cv11b NODE 207, 383 and 341 (With contiguous megablast)

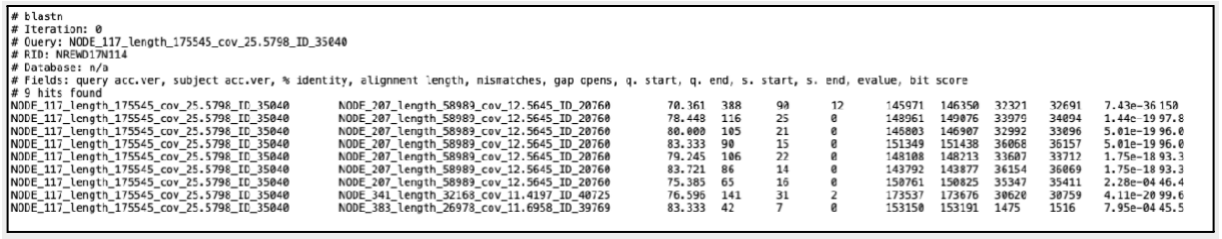


Figure 3.38: BLASTN Alignment between Cs8k NODE 117 and Cv11b NODE 207, 383 and 341 (With discontiguous megablast)

3.11.2. Sequence Alignment by Using TBLASTN and TBLASTX

3.11.2.1. Sequence Alignment with TBLASTN

Sequence alignment between PSD3_ARATH protein sequence and Cs8k NODE_117 by using TBLASTN showed significant e-value 3e-09 [Figure 3.39], while Sequence alignment between PSD3_ARATH protein sequence and Cv264 NODE_72 showed significant e-value 4e-09 [Figure 3.40], and finally sequence alignment between PSD3_ARATH protein sequence and CvC1 NODE_68 by using TBLASTN also showed significant e-value 4e-09 [Figure 3.41]. While bit scores of all these alignments were 53.1.

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cs8kDN
A_NODE_117.fa
           1 sequences; 175,545 total letters



Query= sp|A4GNA8|PSD3_ARATH Phosphatidylserine decarboxylase proenzyme 3
OS=Arabidopsis thaliana OX=3702 PE=1 SV=1

Length=635
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NODE_117_length_175545_cov_25.5798_ID_35040                           53.1    3e-09
```

Figure 3.39: Alignment result of TBLASTN for PSD3_ARATH protein sequence against Cs8k NODE 117

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cv264D
NA_NODE_72.fa
           1 sequences; 238,334 total letters



Query= sp|A4GNA8|PSD3_ARATH Phosphatidylserine decarboxylase proenzyme 3
OS=Arabidopsis thaliana OX=3702 PE=1 SV=1

Length=635
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NODE_72_length_238334_cov_12.1638_ID_35060                            53.1    4e-09
```

Figure 3.40: Alignment result of TBLASTN for PSD3_ARATH protein sequence against Cv264 NODE 72

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/CvC1DN
A_NODE_68.fa
            1 sequences; 249,143 total letters



Query= sp|A4GNA8|PSD3_ARATH Phosphatidylserine decarboxylase proenzyme 3
OS=Arabidopsis thaliana OX=3702 PE=1 SV=1

Length=635
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NODE_68_length_249143_cov_20.6472_ID_35138                            53.1    4e-09
```

Figure 3.41: Alignment result of TBLASTN for PSD3_ARATH protein sequence against CvC1 NODE 68

However, to the contrary sequence alignment between PSD3_ARATH protein sequence and Cv11b NODE_383 by using TBLASTN showed insignificant e-value (i.e., 0.99) and score (bit score = 24.6) [Figure 3.42].

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cv11bD
NA_NODE_207_383_341.fa
           3 sequences; 118,135 total letters



Query= sp|A4GNA8|PSD3_ARATH Phosphatidylserine decarboxylase proenzyme 3
OS=Arabidopsis thaliana OX=3702 PE=1 SV=1

Length=635
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NODE_383_length_26978_cov_11.6958_ID_39769                            24.6    0.99
```

Figure 3.42: Alignment result of TBLASTN for PSD3_ARATH protein sequence against Cv11b NODE 383

3.11.2.2. Sequence Alignment with TBLASTX

Since no similarity match to the PSD3 gene region could be found in Cv11b sequences using TBLASTN, the search was repeated on the protein coding level using TBLASTX. Correspondingly, sequence alignment of CvC1_NODE_68 and Cv264_NODE_72 genome sequences and against Cv11b NODE_207, NODE_383 and NODE_341 by using TBLASTX showed a significant e-value and score (bit score) [Figure 3.43, 3.44], while on the other hand sequence alignment between Cs8k_NODE_117 genome sequence and Cv11b NODE_207, NODE_383 and NODE_341 by using TBLASTX showed relatively moderate e-value and bit score [Figure 3.45].

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cv11bD
NA_NODE_207_383_341.fa
           3 sequences; 118,135 total letters



Query= NODE_68_length_249143_cov_20.6472_ID_35138

Length=249143
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value  N

NODE_341_length_32168_cov_11.4197_ID_40725                            101     0.0    33
NODE_207_length_58989_cov_12.5645_ID_20760                            89.0    1e-73  46
NODE_383_length_26978_cov_11.6958_ID_39769                            62.5    9e-18  8
```

Figure 3.43: Alignment result of TBLASTX CvC1 NODE 68 against Cv11b NODE 207, 383 and 341

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cv11bD
NA_NODE_207_383_341.fa
           3 sequences; 118,135 total letters



Query= NODE_72_length_238334_cov_12.1638_ID_35060

Length=238334
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value  N

NODE_341_length_32168_cov_11.4197_ID_40725                            101     0.0    33
NODE_207_length_58989_cov_12.5645_ID_20760                            89.0    1e-72  46
NODE_383_length_26978_cov_11.6958_ID_39769                            62.5    7e-18  8
```

Figure 3.44: Alignment result of TBLASTX Cv264 NODE 72 against Cv11b NODE 207, 383 and 341

```
Database: /Users/arifsaeed/Documents/ProteinData/New_Analysis_Alignment/Cv11bD
NA_NODE_207_383_341.fa
           3 sequences; 118,135 total letters



Query= NODE_117_length_175545_cov_25.5798_ID_35040

Length=175545
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value  N

NODE_207_length_58989_cov_12.5645_ID_20760                            89.0    6e-76  19
NODE_341_length_32168_cov_11.4197_ID_40725                            101     5e-42  10
NODE_383_length_26978_cov_11.6958_ID_39769                            62.5    8e-16  5
```

Figure 3.45: Alignment result of TBLASTX Cs8k NODE 117 against Cv11b NODE 207, 383 and 341

Consequently, alignment of Cv264 and Cv11b strains showed that PSD3 neighbouring genes were located on sensitive algal strain Cv11b in a reverse order on the contig of node-207, while the PSD3 gene could not be mapped to the Cv11b genome sequence. So, while the genomic neighbourhood of the PSD3 locus can be found in Cv11b and shows synteny, the PSD3 gene itself could not be found and appears to be deleted [Figure 3.46].

3.11.3. Synteny Analysis Diagram of Selected Nodes of All 4 Algal Strains

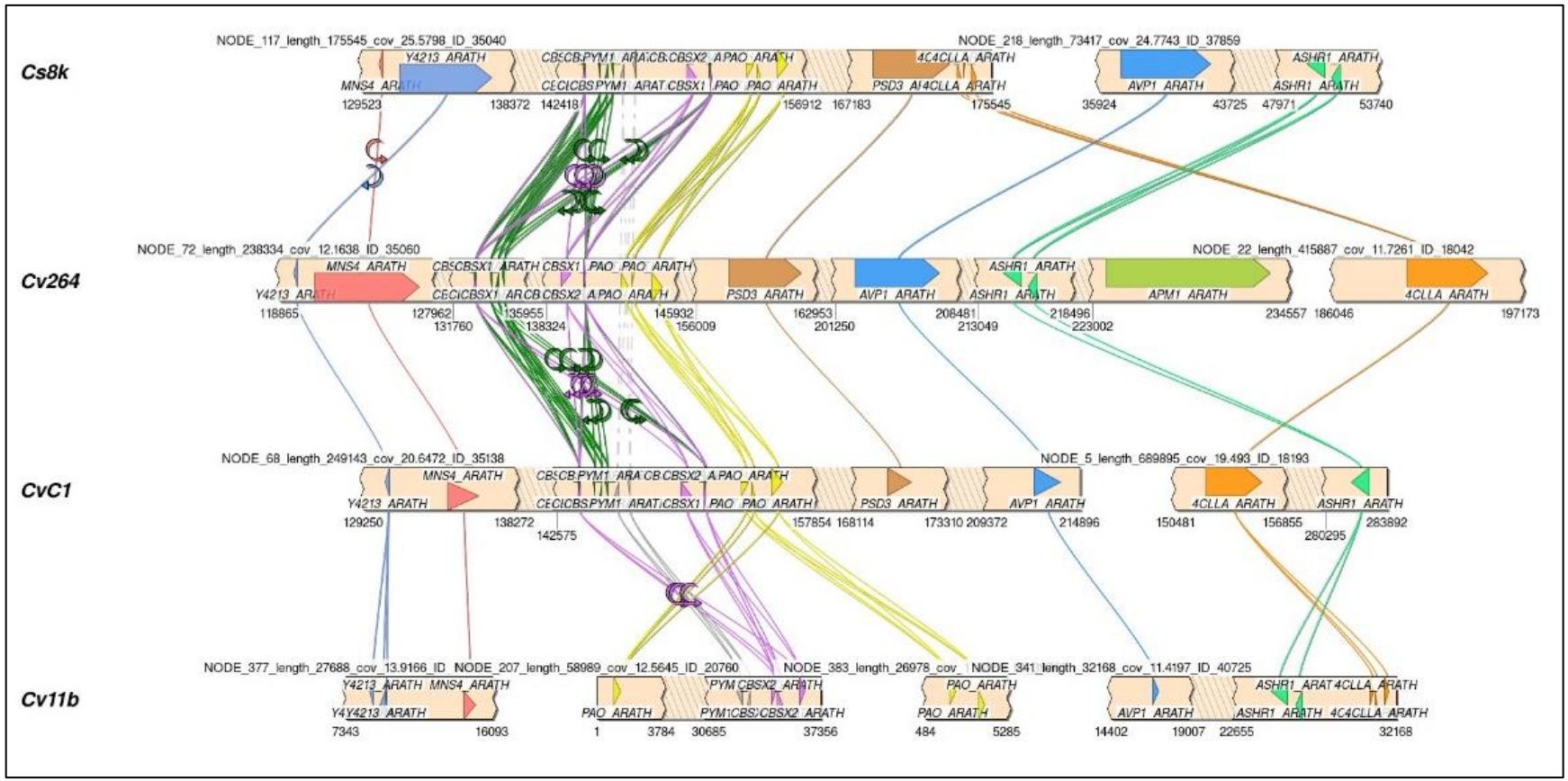


Figure 3.46: Synteny analysis diagram of selected nodes of algal strains after reordering and flipping the nodes to investigate the traces of PSD3 gene: *The Synteny depiction shows the comparison of four different algal strains (i.e., Cv264, CvC1, Cs8k and Cv11b) with multiple contigs*

3.11.4. IGV (Integrative Genomics Viewer) Analysis

The Node117 sequence of Cs8k genome in the region between156912 and 167183 was scrutinised by using IGV (Integrative Genomics Viewer), to discover any sequencing or assembly problems such as NNNN sequence. IGV analysis showed that there is no such evidence found, thus no NNN region identified [Figure 3.47, 3.48].

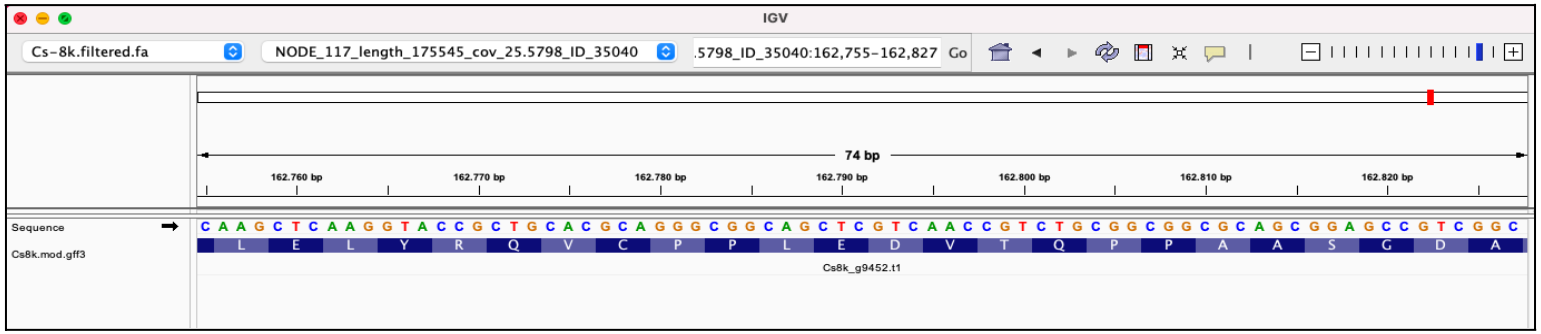


Figure 3.47: Cs8k NODE_117 to check the NNN sequence between 156912 and 167183 base pairs

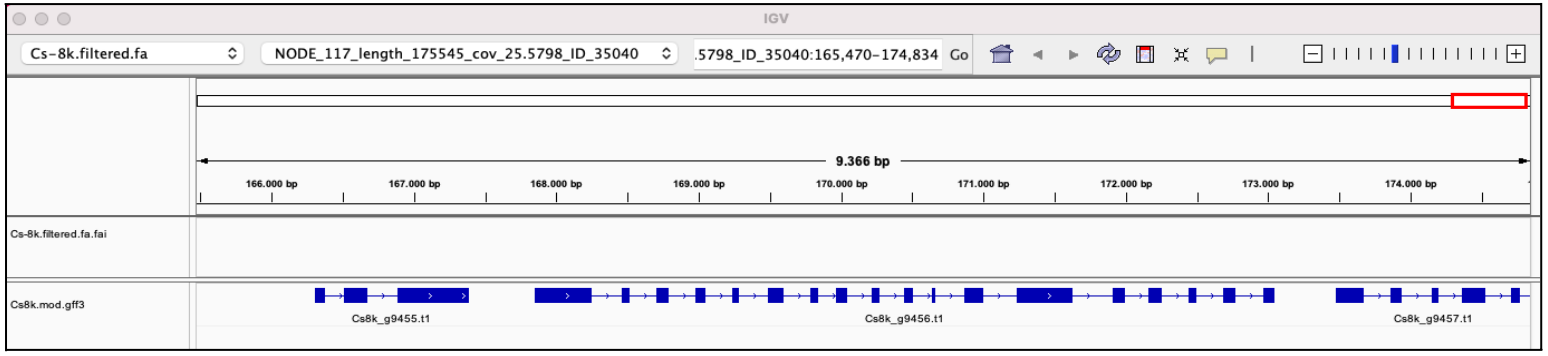


Figure 3.48: Cs8k NODE_117 to check the annotation for PSD3 between 162000 and 173000 base pairs (Cs8k_g9456.t1 → PSD3_ARATH)

3.11.4.1. IGV (Integrative Genomics Viewer) Analysis Between Cv264 and Cv11b

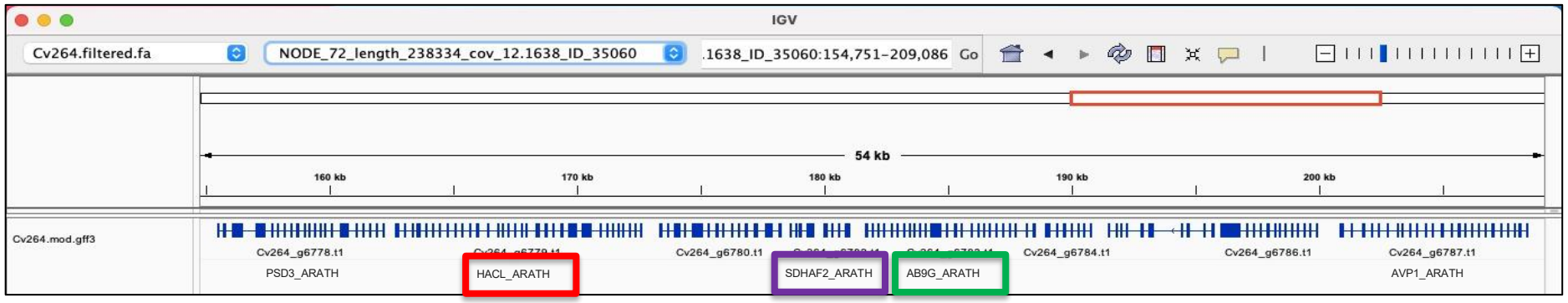


Figure 3.49: IGV-Mapping for Node_72 of Cv264 algal strain

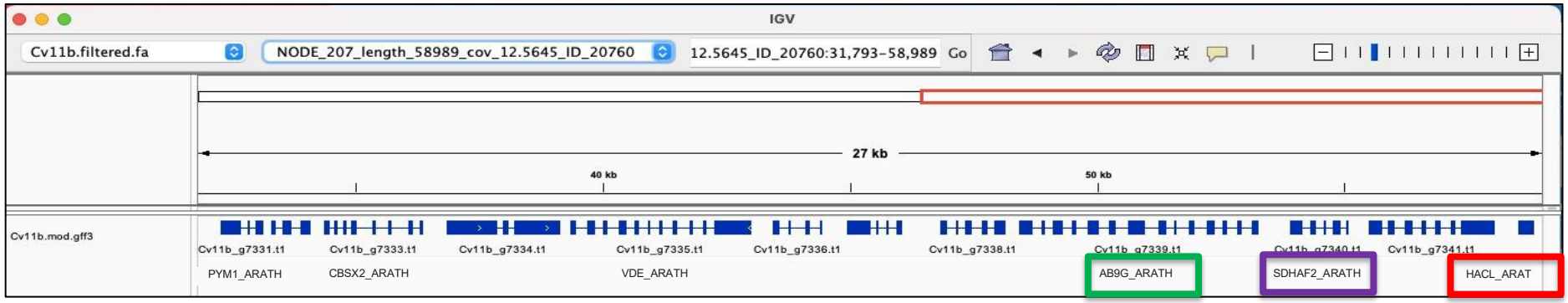


Figure 3.50: IGV-Mapping for Node_207 of Cv11b algal strain

3.11.5. Diagrammatic Representation of IGV Analysis Between Cv264 and Cv11b

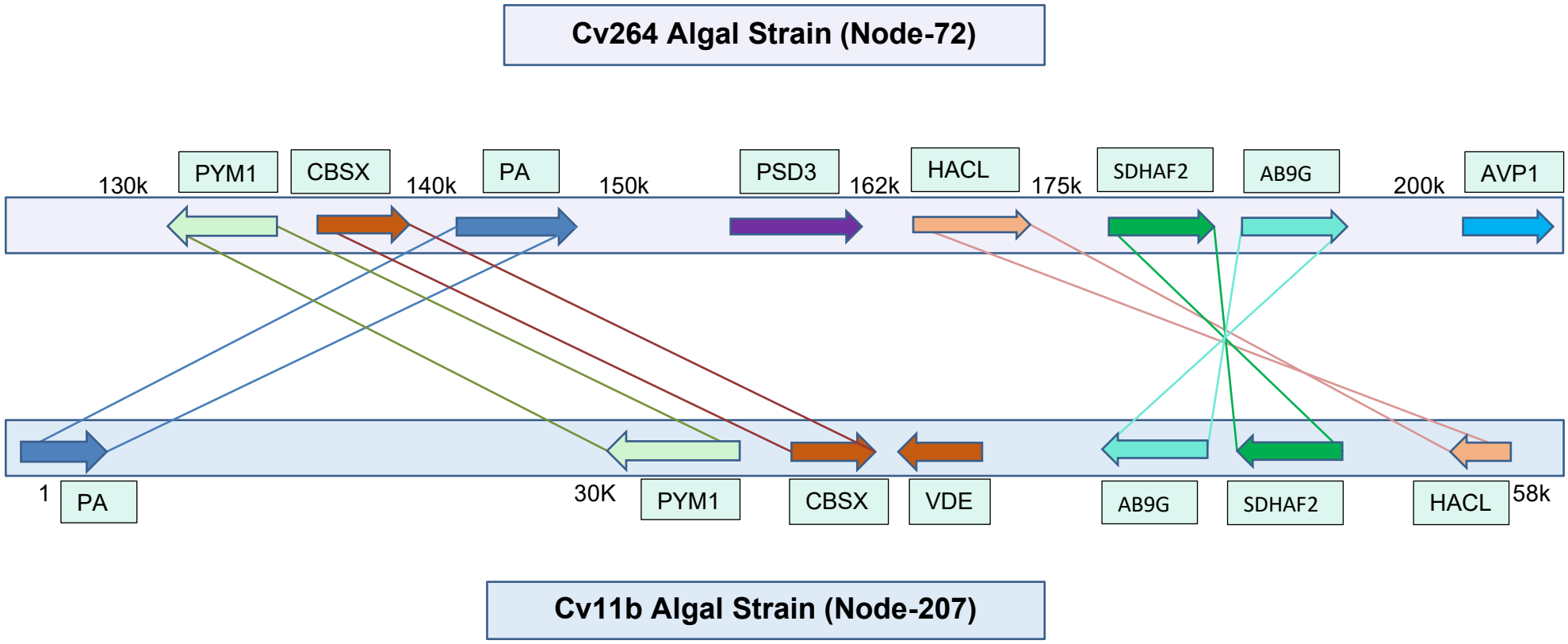


Figure 3.51: Diagrammatic representation of IGV mapping between Nodes of 2 different algal strains Cv264-Node 72 (tolerant) and Cv11b-Node 207 (sensitive)

Diagrammatic representation of the above figure [Figure 3.51], summarizes the findings from synteny, alignment and IGV analyses. Node 207 in Cv11b shows syntenic genes that correspond to the genes flanking PSD3 in Cv264, but no evidence of a PSD3 sequence. Some rearrangements are detected that could be a consequence of the genomic rearrangement that led to deletion of the PSD3 gene. There is no evidence of assembly artefacts, however these cannot be excluded [Figure 3.49, 3.50, and 3.51].

Chapter 4

Conclusion and Discussion

The aim of this research work was to systematically do a comparative genome analysis for four algal strains of *Chlorella* species (including one Highlight sensitive Cv11b and three Highlight tolerant strains Cs8k, Cv264 and CvC1), by applying metabolic network reconstruction methodology; to identify the discrepancies and similarities between tolerant and sensitive algal strains and to figure out the causal genes for these differences. Which could be metabolic processes or relevant candidate genes that allow the tolerant strain to grow in high light stress conditions?

Genome comparison identified that tolerant algal strains have more genes count and gene functions in comparison to the sensitive strain [Figure 3.1]. Moreover, OrthoMCL results revealed that genome of the sensitive strain has much less orthologous groups than the tolerant strains [Figure 3.4].

An evolutionary process, known as regressive evolution, suggests that a species can gradually lose certain genes, typically in response to changes in its environment. Genes that become unnecessary or even disadvantageous may be discarded over time. This idea supports the "less is more" hypothesis, which proposes that gene loss can be beneficial, paving the way for new adaptations and potentially enhancing an organism's overall fitness [109].

Here we can hypothesize that the sensitive strain has been adapted to much milder conditions. In this adaptation process it has lost genes that were necessary to react to specific stresses. After several generations the sensitive strain lost many genes due to the lack of stress or selection pressure. However, additional genes in the pathways are found in all the tolerant strains.

Zhang et al. endorsed in a recent study that a tolerant algal strain has a larger genome and more genes count inferring adaptive alterations under environmental stresses [110]. Thus, this will lead to another hypothesis that likewise a sensitive strain would not only be sensitive to high light stress but it must be more sensitive to other stresses, like salt, heat and cold stress due to loss of gene functions [110].

Sequence similarity searches were carried out with BLASTP, between protein sequences of four algal strains and the database of *Arabidopsis thaliana* as a reference genome. Then KEGG Pathways mapping tool was used to map the KEGG gene identifiers of the algal strains with metabolic pathways; by using reference pathways of *Arabidopsis thaliana* as a template.

There were some enzymes which were missing in the metabolic pathway comparison. Thus, to ensure the completeness of the metabolic network, a more comprehensive sequence similarity analysis was carried out by using TBLASTN, between the algal genome sequences and the protein sequences of *Arabidopsis thaliana*, to compare a protein query sequence against the six-frame translations of nucleotide sequences for finding homologous protein coding regions in

unannotated nucleotide sequences. Subsequently, algal genes were mapped including new found genes with metabolic pathways. Newly extracted genes helped to close the gap between the algal strains and reference genome.

After extracting the similar and diverse gene lists from the algal strain's similarity comparison, we tried to validate the disparity by using Fischer's exact test between shared and strain-specific genes of tolerant and sensitive strains. Fischer's exact test assisted to identify statistically most significant pathways [Figure 3.10].

Next, to ensure the completeness of the metabolic network we mapped lists of shared and strain-specific genes of tolerant and sensitive strains with the KEGG Orthology (KO) IDs to identify the missing gene functions. Results of metabolic network reconstruction had shown significant genetic differences between strains. There were multiple strain-specific genes in algal genomes, for high light tolerant as well as for sensitive algal strains [Figure 3.8]. But after linking these genes to putative functions to identify their roles in metabolism, it was revealed that number of strain-specific functions reduced considerably for high light tolerant as well as for sensitive algal strains [Figure 3.9]. Which established that multiple strain-specific genes associated with the shared functions.

From the top ranked pathway list, four genes were selected which were associated with different gene functions: NDUA1 (NADH dehydrogenase) from Oxidative Phosphorylation pathway [Figure 3.13], SUOX (Sulfite oxidase) from Sulfur metabolism pathway [Figure 3.14], PSD3 (Phosphatidylserine decarboxylase-3)

from Glycerophospholipid metabolism pathway [Figure 3.15] and ACD5 (accelerated cell death 5) from Sphingolipid metabolism pathway [Figure 3.16].

Annotation of these selected genes were again verified after extracting them from tolerant strain, by using TBLASTN against database of sensitive genome. This comparison helped to annotate ACD5 [Figure 3.17] and SUOX genes in sensitive genome [Figure 3.18]. Then genome of 5KB upstream and downstream of PSD3 and NADH dehydrogenase (NDUA1) genes were extracted from tolerant strains by using BEDTOOLS and tried to map it against sensitive algal genome. But, NADH dehydrogenase [Figure 3.19] and PSD3 genes could not be mapped in HL sensitive algal strain [Figure 3.22].

In Energy metabolism, Oxidative Phosphorylation pathway is showing a missing gene NDUA1 in sensitive strain (Cv11b). The NDUA1 gene belongs to the NADH dehydrogenase which plays a significant role for Complex-I functions in the transfer of electrons from NADH to the respiratory chain, in oxidative phosphorylation for the energy metabolism [111]. The immediate electron acceptor for the enzyme is believed to be ubiquinone [101]. However, NADH dehydrogenase (Complex-I) is an accessory subunit of the mitochondrial membrane respiratory chain, that is believed not to be engaged in catalysis. As this gene fulfils a core function for energy metabolism, it appears surprising that the sensitive strain could grow without it. It may be a region of the genome that was not sequenced at sufficient coverage to be included in the assembly.

Finally, we ended up with the selection of PSD3 gene due to its relevancy with lipid production. Altered lipid composition was reported for Cv11b under high light stress [112]. Thus, this gene is very interesting to explain the phenotype.

To verify that PSD3 gene is a deletion in the Cv11b genome, synteny of surrounding genes was analysed in order to identify the precise deletion. Where PSD3 gene was mapped with the contigs of all tolerant algal strains but not with sensitive algal strain, while flanking genes of PSD3 were located on sensitive genome but they were rearranged or inverted [Figure 3.26].

Then, BLASTN [Figure 3.35 – 3.38], TBLASTN [Figure 3.39 – 3.42] and TBLASTX tools [Figure 3.43 – 3.45] were used to identify significant similarity between these selected contigs, and ended up with highly significant similarity. So, the flanking regions where PSD3 would be located in Cv11b, are present, however rearranged and inverted, suggesting that the PSD3 gene was deleted as part of the mutation leading to the rearrangement of this region.

Subsequently, we aimed to exclude an assembly artifact, which might be indicated by gaps (NNNs) in the sequence. Using IGV, no stretches of N were identified in that region of the sensitive genome. So, we come to the conclusion that the PSD3 gene is deleted from the sensitive genome [Figure 3.50].

The PSD3 gene is part of the Phosphatidylserine decarboxylase (PSD) family involved in the lipid metabolism pathway, specifically in glycerophospholipid metabolism. Research shows that PSD3 plays a key role in producing phosphatidylethanolamine mainly in mitochondria, about two-thirds of total PSD

activity. While PSD1 contributes only about one-third of the total PSD activity in leaves of *Arabidopsis thaliana*. However, PSD2 has very low activity [102].

This genomic research work revealed that PSD1 [Figure 3.20] and PSD2 [Figure 3.21] were missing in both tolerant and sensitive algal strains while PSD3 was missing only in the genome of the sensitive algal strain Cv11b.

Consequently, we propose that a rearrangement in the genomic region of the PSD3 gene led to its deletion in Cv11b. PSD3 may have been lost during the repair of the inversion of DNA fragment. However, further laboratory experiments like targeted PCR are necessary for validation, as we cannot exclude e.g. an assembly artifact, even if we did not find hallmarks of an artifact such as assembly gaps or stretches of NNN.

Support for the hypothesis that PSD3 may explain high light tolerance can be found in research work of Widzgowski et al., which showed a difference in phospholipid composition between tolerant and sensitive strains [112]. PSD (Phosphatidyl serine decarboxylase) catalyzes the conversion of phosphatidyl serine to phosphatidyl ethanolamine (PE). We propose that the change in availability of PE due to lack of PSD3 in Cv11b impacts the phospholipid composition under high light conditions, but the mechanism needs to be studied further. Experiments such as knock out studies can help to establish the function of PSD3 under high light conditions and thus demonstrate the *in vivo* relevance of the candidate gene for high light tolerance.

It is important to note that this analysis is limited by the research and publication bias in knowledgebase data. However, the in-silico reconstructed metabolic networks as developed in this work can effectively contribute to a well-informed design of *in vivo* experiments that show the proposed candidate genes are relevant in these environmental conditions.

Summary: Most genes and pathways are shared, but one key gene was identified that is missing and is linked to phospholipid biosynthesis. This is an interesting lead for experiments to establish its role in shaping the phospholipid profile.

References

1. Mostafa, S. S (2012) Microalgal biotechnology: prospects and applications. INTECH Open Access Publisher.

2. Iwamoto, H (2004) Chapter 11. Industrial production of microalgal cell-mass and secondary products–major industrial species: *Chlorella*. Oxford, UK.: Blackwell Publishing Ltd.

3. de Morais, M. G., B. d. S. Vaz, E. G. de Morais & J. A. V. Costa (2015) Biologically active metabolites synthesized by microalgae. BioMed Research International.

4. Skjanes, K., C. Rebours & P. Lindblad (2013) Potential for green microalgae to produce hydrogen, pharmaceuticals and other high value products in a combined process. Critical Reviews in Biotechnology, 33, 172-215.

5. Ip, P.-F. & F. Chen (2005a) Employment of reactive oxygen species to enhance astaxanthin formation in *Chlorella zofingiensis* in heterotrophic culture. Process Biochemistry, 40, 3491-3496.

6. Lemoine, Y. & B. Schoefs (2010) Secondary ketocarotenoid astaxanthin biosynthesis in algae: a multifunctional response to stress. Photosynthesis research, 106, 155-177.

7. Ip, P.-F., K.-H. Wong & F. Chen (2004) Enhanced production of astaxanthin by the green microalga *Chlorella zofingiensis* in mixotrophic culture. Process Biochemistry, 39, 1761-1766.

8. Del Campo, J. A., H. Rodriguez, J. Moreno, M. A. Vargas, J. Rivas & M. G. Guerrero (2004) Accumulation of astaxanthin and lutein in *Chlorella zofingiensis* (Chlorophyta). Applied Microbiology Biotechnology, 64, 848-854.

9. Ip, P.-F. & F. Chen (2005b) Production of astaxanthin by the green microalga *Chlorella zofingiensis* in the dark. Process Biochemistry, 40, 733-738.

10. Cordero, B. F., I. Obraztsova, I. Couso, R. Leon, M. A. Vargas & H. Rodriguez (2011) Enhancement of lutein production in *Chlorella sorokiniana* (Chorophyta) by improvement of culture conditions and random mutagenesis. Marine Drugs, 9, 1607-1624.

11. Fu, X., D. Wang, X. Yin, P. Du & B. Kan (2014) Time course transcriptome changes in Shewanella algae in response to salt stress. PloS one, 9, e96001.

12. Sun, P., Y. Mao, G. Li, M. Cao, F. Kong, L. Wang & G. Bi (2015) Comparative transcriptome profiling of Pyropia yezoensis (Ueda) M.S. Hwang & H.G. Choi in response to temperature stresses. BMC Genomics, 16, 1- 16.

13. Doan, T. T. Y., B. Sivaloganathan & J. P. Obbard (2011) Screening of marine microalgae for biodiesel feedstock. Biomass and Bioenergy, 35, 2534-2544.

14. Perez-Garcia, O., F. M. Escalante, L. E. de-Bashan & Y. Bashan (2011) Heterotrophic cultures of microalgae: metabolism and potential products. Water Research, 45, 11-36.

15. Baba, M. & Y. Shiraiwa (2013) Biosynthesis of lipids and hydrocarbons in algae. In Photosynthesis, ed. Z. Dubinsky, 331-356. Croatia: InTech.

16. Goiris, K., W. Van Colen, I. Wilches, F. León-Tamariz, L. De Cooman & K. Muylaert (2015) Impact of nutrient stress on antioxidant production in three species of microalgae. Algal Research, 7, 51-57.

17. Bartel, D. P. (2004) MicroRNAs-genomics, biogenesis, mechanism, and function. Cell, 116, 281-297.

18. Nichols, N.N., Bothast, R.J. (2008). Production of Ethanol from Grain. In: Vermerris, W. (eds) Genetic Improvement of Bioenergy Crops. Springer, New York, NY, 75–88. https://doi.org/10.1007/978-0-387-70805-8_3

19. Goldemberg, J.; Coelho, S.T.; Guardabassi, P. (2008) The sustainability of ethanol production from sugarcane. Energy Policy, 36 (6), 2086-2097. https://doi.org/10.1016/j.enpol.2008.02.028.

20. Samson, R.A.; Omielan, J.A. (1992) Switchgrass: A Potential Biomass Energy Crop for Ethanol Production. In Proceedings of the Thirteenth North American Prarie Conference, Windsor, ON, Canada, 6–9, 253–258.

21. Schmer, M.R.; Vogel, K.P.; Mitchell, R.B.; Perrin, R.K. (2008) Net energy of cellulosic ethanol from switchgrass. Proc. Natl. Acad. Sci. USA, 105, 464–469.

22. Sheehan, J.; Aden, A.; Paustian, K.; Killian, K.; Brenner, J.; Walsh, M.; Nelson, R. (2003) Energy and environmental aspects of using corn stover for fuel ethanol. J. Ind. Ecol., 7 , 117–146.

23. Ballesteros, I.; Negro, M.J.; Oliva, J.M.; Cabañas, A.; Manzanares, P.; Ballesteros, M. (2006) Ethanol production from steam-explosion pretreated wheat straw. Appl. Biochem. Biotechnol. 129 – 132, 496–508.

24. Kumar, D.; Murthy, G.S. (2010) Pretreatments and enzymatic hydrolysis of grass straws for ethanol production in the Pacific Northwest U.S. Biol. Eng., 3, 97–110.

25. Kumar, D.; Murthy, G.S. (2011) Impact of pretreatment and downstream processing technologies on economics and energy in cellulosic ethanol production. Biotechnol. Biofuels, 4, 27.

26. NREL Web Page. Dynamic Maps, GIS Data, and Analysis Tools—Solar Maps. Available online: http://www.nrel.gov/gis/solar.html (accessed on 9 May 2013).

27. Stockenreiter, M.; Haupt, F.; Graber, A.K.; Seppälä, J.; Spilling, K.; Tamminen, T.; Stibor, H. (2013) Functional group richness: Implications of biodiversity for light use and lipid yield in microalgae.J. Phycol., in press.

28. Sorokin, C.; Krauss, R.W. (1958) The Effects of light intensity on the growth rates of green algae. Plant Physiol., 33, 109–113.

29. Dubinsky, Z.; Matsukawa, R.; Karube, I. (1995) Photobiological aspects of algal mass culture. J. Mar. Biotechnol., 2, 61–65.

30. Fábregas, J.; Maseda, A.; Domínguez, A.; Otero, A. (2004) The cell composition of Nannochloropsis sp. changes under different irradiances in semicontinuous culture. World J. Microbiol. Biotechnol., 20, 31–35.

31. Mock, T.; Kroon, B.M.A. (2002) Photosynthetic energy conversion under extreme conditions—II: The significance of lipids under light limited growth in Antarctic sea ice diatoms. Phytochemistry, 61, 53–60.

32. Gordillo, F.J.L.; Goutx, M.; Figueroa, F.L.; Niell, F.X. (1998) Effects of light intensity, CO2 and nitrogen supply on lipid class composition of Dunaliella viridis. J. Appl. Phycol., 10, 135–144.

33. You, T.; Barnett, S.M. (2004) Effect of light quality on production of extracellular polysaccharides and growth rate of Porphyridium cruentum. Biochem. Eng. J., 19, 251–258.

34. Brody, M.; Vatter, A.E. (1959) Observations on cellular structures of Porphyridium cruentum. J. Biophys. Biochem. Cytol., 5, 289–294.

35. Iqbal, M.; Zafar, S. (1993) Effects of photon flux density, CO2, aeration rate, and inoculum density on growth and extracellular polysaccharide production by Porphyridium cruentum. Folia Microbiol., 38, 509–514.

36. Cuhel, R.L.; Ortner, P.B.; Lean, D.R.S. (1984) Night synthesis of protein by algae. Limnol. Oceanogr., 29, 731–744.

37. Morris, I.; Glover, H.; Yentsch, C. (1974) Products of photosynthesis by marine phytoplankton: The effect of environmental factors on the relative rates of protein synthesis. Mar. Biol., 27 , 1–9.

38. Smith, R.; Cavaletto, J.; Eadie, B.; Gardner, W. (1993) Growth and lipid composition of high Arctic ice algae during the spring bloom at Resolute, Northwest Territories, Canada. Mar. Ecol. Prog. Ser., 97 , 19–29.

39. Cohen, Z. (1999) Porphyridium Cruentum. In Chemicals from Microalgae; Cohen, Z., Ed.; CRC Press: Boca Raton, FL, USA, 1–24.

40. Renaud, S.; Parry, D.; Thinh, L.V.; Kuo, C.; Padovan, A.; Sammy, N. (1991) Effect of light intensity on the proximate biochemical and fatty acid composition of Isochrysis sp. and Nannochloropsis oculata for use in tropical aquaculture. J. Appl. Phycol., 3, 43–53.

41. Orcutt, D.M., Patterson, G.W. (1974) Effect of light intensity upon lipid composition of Nitzschia closterium (Cylindrotheca fusiformis). Lipids, 9, 1000–1003.

42. Sukenik, A.; Carmeli, Y.; Berner, T. (1989) Regulation of fatty acid composition by irradiance level in the eustigmatophyte Nannochloropsis sp. J. Phycol., 25, 686–692.

43. Berner, T.; Dubinsky, Z.; Wyman, K.; Falkowski, P.G. (1989) Photoadaptation and the "package" effect in Dunaliella tertiolecta (chlorophycae). J. Phycol., 25, 70–78.

44. Wu, X.; Merchuk, J.C. (2001) A model integrating fluid dynamics in photosynthesis and photoinhibition processes. Chem. Eng. Sci., 56 , 3527–3538.

45. Long, S.; Humphries, S.; Falkowski, P.G. (1994) Photoinhibition of photosynthesis in nature. Annu. Rev. Plant Biol., 45, 633–662.

46. Pinney, J., Papp, B., Hyland, C., Wambua, L., Westhead, D., and McConkey, G. (2007). Metabolic reconstruction and analysis for parasite genomes. TRENDS in Parasitology, 23(11), 548–554.

47. Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., Caspi, R., (2010). Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in bioinformatics, 11(1), 40–79.

48. Jagadevan, S., Banerjee, A., Banerjee, C. et al. Recent developments in synthetic biology and metabolic engineering in microalgae towards biofuel production. Biotechnol Biofuels11, 185 (2018). https://doi.org/10.1186/s13068-018-1181-1

49. Osterman, A. and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. Current opinion in chemical biology, 7(2), 238–251.

50. Morett, E., Korbel, J., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B., and Bork, P. (2003). Systematic discovery of analogous enzymes in thiamin biosynthesis. Nature biotechnology, 21(7), 790–795.

51. Green, M. and Karp, P. (2004). A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC bioinformatics, 5(1), 76.

52. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., (2000). Gene ontology: tool for the unification of biology. Nature genetics, 25(1), 25.

53. Murzin, A., Brenner, S., Hubbard, T., Chothia, C., (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology, 247(4), 536–540.

54. Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27–30.

55. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. Nucleic acids research, 34(Database issue), D354–7.

56. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. Nucleic acids research, 36(Database issue), D480–4.

57. Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., McNeil, L., Paarmann, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., (2008). The rast server: rapid annotations using subsystems technology. BMC genomics, 9(1), 75.

58. Thiele, I. and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, 5(1), 93–121.

59. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

60. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354-7. doi: 10.1093/nar/gkj102. PMID: 16381885; PMCID: PMC1347464.

61. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008 Jan;36(Database issue):D480-4. doi: 10.1093/nar/gkm882. Epub 2007 Dec 12. PMID: 18077471; PMCID: PMC2238879.

62. Maurice Scheer, Andreas Grote, Antje Chang, Ida Schomburg, Cornelia Munaretto, Michael Rother, Carola Söhngen, Michael Stelzer, Juliane Thiele, Dietmar Schomburg, BRENDA, the enzyme information system in 2011, *Nucleic Acids Research*, Volume 39, Issue suppl_1, 1 January 2011, Pages D670–D676, https://doi.org/10.1093/nar/gkq1089

63. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic

Acids Res. 2002 Jan 1;30(1):268-72. doi: 10.1093/nar/30.1.268. PMID: 11752312; PMCID: PMC99153.

64. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D138-41. doi: 10.1093/nar/gkh121. PMID: 14681378; PMCID: PMC308855.

65. Yeats, T. H., & Rose, J. K. (2008). The biochemistry and biology of extracellular plant lipid-transfer proteins (LTPs). *Protein Science*, *17*(2), 191-198.

66. Karp, P. D., Paley, S., & Romero, P. (2002). The pathway tools software. *Bioinformatics*, *18*(suppl_1), S225-S232.

67. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E Jr, Liolios K, Joukov V, Kaznadzey D, Anderson I, Bhattacharyya A, Burd H, Gardner W, Hanke P, Kapatral V, Mikhailova N, Vasieva O, Osterman A, Vonstein V, Fonstein M, Ivanova N, Kyrpides N. The ERGO genome analysis and discovery system. Nucleic Acids Res. 2003 Jan 1;31(1):164-71. doi: 10.1093/nar/gkg148. PMID: 12519973; PMCID: PMC165577.

68. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008 Feb 8;9:75. doi: 10.1186/1471-2164-9-75. PMID: 18261238; PMCID: PMC2265698.

69. Pinney JW, Shirley MW, McConkey GA, Westhead DR. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Res. 2005 Mar 3;33(4):1399-409. doi: 10.1093/nar/gki285. PMID: 15745999; PMCID: PMC552966.

70. Hyland C, Pinney JW, McConkey GA, Westhead DR. metaSHARK: a WWW platform for interactive exploration of metabolic networks. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W725-8. doi: 10.1093/nar/gkl196. PMID: 16845107; PMCID: PMC1538829.

71. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003 Nov 15;31(22):6633-9. doi: 10.1093/nar/gkg847. PMID: 14602924; PMCID: PMC275543.

72. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res. 1999 Jun 1;27(11):2369-76. doi: 10.1093/nar/27.11.2369. PMID: 10325427; PMCID: PMC148804.

73. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

74. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40. doi: 10.1093/bioinformatics/btu031. Epub 2014 Jan 21. PMID: 24451626; PMCID: PMC3998142.

75. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014 Jan;42(Database issue):D222-30. doi: 10.1093/nar/gkt1223. Epub 2013 Nov 27. PMID: 24288371; PMCID: PMC3965110.

76. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol. 2009;563:123-40. doi: 10.1007/978-1-60761-175-2_7. PMID: 19597783; PMCID: PMC6608593.

77. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Romá-Mateo C, Theodosiou A, Mitchell AL. The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database (Oxford). 2012 Apr 15;2012:bas019. doi: 10.1093/database/bas019. PMID: 22508994; PMCID: PMC3326521.

78. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 2000 Jan 1;28(1):231-4. doi: 10.1093/nar/28.1.231. PMID: 10592234; PMCID: PMC102444.

79. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 2009 Jan;37(Database issue):D380-6. doi: 10.1093/nar/gkn762. Epub 2008 Nov 26. PMID: 19036790; PMCID: PMC2686452.

80. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003 Jan 1;31(1):371-3. doi: 10.1093/nar/gkg128. PMID: 12520025; PMCID: PMC165575.

81. Francke C, Siezen RJ, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol. 2005 Nov;13(11):550-8. doi: 10.1016/j.tim.2005.09.001. Epub 2005 Sep 19. PMID: 16169729.

82. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9. doi: 10.1038/75556. PMID: 10802651; PMCID: PMC3037419.

83. GO: http://geneontology.org

84. interpro2go: http://current.geneontology.org/ontology/external2go/interpro2go

85. Griesemer M, Kimbrel JA, Zhou CE, Navid A, D'haeseleer P. Combining multiple functional annotation tools increases coverage of metabolic annotation. BMC Genomics. 2018 Dec 19;19(1):948. doi: 10.1186/s12864-018-5221-9.

86. ec2go: http://current.geneontology.org/ontology/external2go/ec2go

87. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003 Sep;13(9):2178-89. doi: 10.1101/gr.1224503. PMID: 12952885; PMCID: PMC403725.

88. OrthoMCL DB, Ortholog Groups of Protein Sequences: https://orthomcl.org/orthomcl/app

89. VennDiagram: Generate High-Resolution Venn and Euler Plots. https://cran.r-project.org/web/packages/VennDiagram/index.html

90. BLAST Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/ Blast.cgi

91. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

92. KEGG Mapper – Convert ID: https://www.genome.jp/kegg/tool/conv_id

93. Kanehisa, M., Sato, Y., and Kawashima, M.; KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 31, 47-53 (2022)

94. KEGG Mapper A suite of KEGG mapping tools: https://www.genome.jp/kegg/kegg3a.html

95. Giorgi, F.M.; Ceraolo, C.; Mercatelli, D. The R Language: An Engine for Bioinformatics and Data Science. *Life***2022**, *12*, 648. https://doi.org/10.3390/life12050648

96. tblastn: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

97. blastx: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

98. Veltri D, Wight MM, Crouch JA. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. Nucleic Acids Res. 2016 Jul 8;44(W1):W41-5. doi: 10.1093/nar/gkw330. Epub 2016 May 3. PMID: 27141960; PMCID: PMC4987899.

99. James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).

100. Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

101. The respiratory complex I of bacteria, archaea and eukarya and its module common with membrane-bound multisubunit hydrogenases. Friedrich T, Scheide D. FEBS Lett. 479, 1-5, (2000).

102. Nerlich A, von Orlow M, Rontein D, Hanson AD, Dörmann P. Deficiency in phosphatidylserine decarboxylase activity in the psd1 psd2 psd3 triple mutant of *Arabidopsis* affects phosphatidylethanolamine accumulation in mitochondria. Plant Physiol. 2007 Jun;144(2):904-14. doi: 10.1104/pp.107.095414. Epub 2007 Apr 20.

103. Moroney, J.V.; Ynalvez, R.A. Algal Photosynthesis. *Encycl. Life Sci.* **2018**.

104. Sreenikethanam A, Raj S, J RB, Gugulothu P, Bajhaiya AK. Genetic Engineering of Microalgae for Secondary Metabolite Production: Recent Developments, Challenges, and Future Prospects. Front Bioeng Biotechnol. 2022 Mar 23;10:836056. doi: 10.3389/fbioe.2022.836056. PMID: 35402414; PMCID: PMC8984019.

105. FARM-ENERGY. Algae for Biofuel Production. APRIL 3, 2019 https://farm-energy.extension.org/algae-for-biofuel-production/

106. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W182-5. doi: 10.1093/nar/gkm321. Epub 2007 May 25.

107. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol. 2008 Oct;26(10):1155-60. doi: 10.1038/nbt1492.

108. Qi Q, Li J, Cheng J. Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods. BMC Proc. 2014 Oct 13;8(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer):S5. doi: 10.1186/1753-6561-8-S6-S5.

109. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016 Jul;17(7):379-91. doi: 10.1038/nrg.2016.39. Epub 2016 Apr 18. PMID: 27087500.

110. Zhang Z, Qu C, Zhang K, He Y, Zhao X, Yang L, Zheng Z, Ma X, Wang X, Wang W, Wang K, Li D, Zhang L, Zhang X, Su D, Chang X, Zhou M, Gao D, Jiang W, Leliaert F, Bhattacharya D, De Clerck O, Zhong B, Miao J. Adaptation to Extreme Antarctic Environments Revealed by the Genome of a Sea Ice Green Alga. Curr Biol. 2020 Sep 7;30(17):3330-3341.e7. doi: 10.1016/j.cub.2020.06.029. Epub 2020 Jul 2.

111. Maldonado M, Padavannil A, Zhou L, Guo F, Letts JA. Atomic structure of a mitochondrial complex I intermediate from vascular plants. Elife. 2020 Aug 25;9:e56664. doi: 10.7554/eLife.56664.

112. Widzgowski J, Vogel A, Altrogge L, Pfaff J, Schoof H, Usadel B, Nedbal L, Schurr U, Pfaff C. High light induces species specific changes in the membrane lipid composition of Chlorella. Biochem J. 2020 Jul 17;477(13):2543-2559. doi: 10.1042/BCJ20200160. PMID: 32556082.