Integration of transcriptomics data in a genome-scale metabolic model of *Nannochloropsis gaditana* to study nitrogen and phosphorus deprivation

José Diogo Moura[1] and Emanuel Cunha[1]

[1]Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal
pg45965@alunos.uminho.pt

Abstract. *Nannochloropsis gaditana* is a microalgae known to accumulate high amounts of lipids, although its autotrophic growth is not compatible with the requirements for industrial exploitation. Consequently, it is of interest to research and develop heterotrophic growth alternatives. Genome-scale metabolic models are useful systems biology tools as they accelerate growth media optimization and strain design. More accurate information about metabolic states can be obtained through the integration of transcriptomics data into these models. In this project, transcriptomics data was integrated into a generic model of *N. gaditana* to assess metabolic alterations under nitrogen and phosphorus deprivation. This algae seem to use the polyketide synthase system to produce polyunsaturated fatty acids when nitrogen is absent. Moreover, phosphorus deprivation induced the mitochondrial energy production, reducing the activity of photosynthetic pathways.

Keywords: Genome-scale metabolic model · Systems biology · Data integration · Transcriptomics · Microalgae · *Nannochloropsis gaditana*

1 Context and Motivation

Nannochloropsis gaditana is recognized as a species with biotechnological interest mainly due to its ability to accumulate large amounts of lipids, in particular, polyunsaturated fatty acids (PUFAs) [18,24,60,1,41,56]. However, the growth of this microalgae is rather slow and often incompatible with the requirements associated with industrial utilization. Thus, it is of interest to research approaches that provide significant improvements regarding the biomass yield, while maintaining or increasing the lipid production. The traditional approaches to overcome this problem are based on changing the availability of certain nutrients and varying the light conditions, introducing environmental stress that favours the accumulation of lipids. However, as many factors can influence its growth, the experimental screening is complex, expensive, and time-consuming. Furthermore, this microalgae reveals promising features that may allow strain optimization, as it presents a well-established genetic toolbox [31].

The development of sequencing methods and biological databases allowed a huge availability of genome sequences that can be very useful to research less-studied organisms at the systems level [19]. Hence, it is possible to accelerate procedures like strain optimization, reducing the cost of experimental procedures. The reconstruction of genome-scale metabolic models (GEMs) is one of the most useful approaches to apply metabolic engineering tasks, as well to optimize the growth media of microorganisms. These models can be used in conjunction with omics data to improve the model's predictions in different environmental and genetic conditions [37].

2 State of Art

2.1 *Nannochloropsis gaditana*

Nannochloropsis is a genus of microalgae that belongs to the *Stramenopiles* group, and is divided into six species: *Nannochloropsis gaditana*, *N. australis*, *N. granulata*, *N. maritima*, *N. oculata*, and *N. oceanica*. These species can be found in both fresh and marine water with spherical and non-motile cells, presenting an average diameter of 2 to 3 µm.

One of these species with greatest industrial potential is *N. gaditana*, due to its ability to accumulate high amounts of lipids. Usually, its lipid content is around 37% to 60% of its biomass dry weight, mostly in the form of triacylglycerol [48] containing polyunsaturated fatty acids [18,24,60,52,80]. PUFAs are one of the most important constituents in the structure of the cell membranes, presenting well-known health-promoting effects, like playing a vital role in the treatment of non-alcoholic fatty liver, autoimmune reactions and various chronic diseases [39] and also to prevent cardiovascular diseases [64]. Among PUFAs, ω-3 and ω-6 fatty acids are known to be essential fatty acids, as the human body cannot produce them and must be obtained through diets [33].

Currently, there are four genome sequences of *N. gaditana* strains available at the NCBI Assembly database (accessions ASM24072v1, we3730_nuc, NagaB31_1.0, and ASM161421v1), and diverse reports of gene editing, quantification and analysis [31,29,5]. The potential industrial applications of *N. gaditana* have been extended to the production of biofuels, cosmetics, and food [27].

2.2 Genome-scale metabolic models

Metabolic models are important tools to understand the functioning of complex biological systems. Such models are partial or full mathematical representations of the metabolism of an organism [14,66], describing a whole set of stoichiometry-based metabolic reactions of a target organism based on information retrieved from the genome sequence and experimentally data [26]. These models take into account hundreds or thousands of genes, reactions and metabolites, providing information regarding the physiological and metabolic properties of organisms under different environmental and genetic conditions [19].

The reconstruction of a high-quality GEM is a complex and time-consuming process that may take months to years to finish, depending on the complexity of the target organism, the desired model curation level, and the tools used [67]. The reconstruction process is well described in literature [19,67], and is usually divided into four main stages: genome annotation, assembly of a draft metabolic network, conversion into a stoichiometric model and validation of the model.

At the end of this process, different methodologies and algorithms can be applied to the validated model, depending on the purpose of the work. Flux Balance Analysis (FBA) [70] is one of the most used methods to analyse biochemical networks. This method calculates the flux of reactions in a metabolic model through an optimization problem, where a given objective function (usually biomass production) is maximized. However, this method offers a solution that may not be unique. To avoid the degenerated solutions provided by FBA, it is possible to use parsimonious FBA (pFBA) [42] - this one aims at minimizing the total flux of the metabolic network after maximizing the main objective, thus offering a unique optimal solution. Another useful method to evaluate the behavior of a metabolic model is the Flux Variability Analysis (FVA) [49], which manages to determine a range of fluxes for each reaction in the model.

The first GEMs were applied to design metabolic engineering approaches to enhance the production of certain products of interest, such as lycopene and sesquiterpene [7,10]. Nevertheless, metabolic models have also been used for other purposes, including drug targeting, prediction of enzyme functions, modeling interactions between different cells/organisms, metabolic pathway analysis, and media optimization [26,23,36,43,45].

2.3 Processing of transcriptomics

RNA sequencing data is often used for differential gene expression analysis [65]. However, transcriptomics data has also been used for metabolic modeling purposes enhancing the accuracy and predictive capabilities of metabolic models. The computational processing of the raw data obtained from high-throughput platforms is usually divided into four phases: alignment and/or assembly of sequencing reads in a transcriptome; quantification of reads that overlap the transcripts; filtration and normalization between samples; and statistical modeling of significant changes in the expression levels of individual genes and/or transcripts between sample groups [65]. These steps require using diverse bioinformatics tools as represented in Table 1.

Phase 1 - Alignment and assembly of sequencing reads The raw data retrieved from high throughput sequencing pipelines contain base-called sequencing reads, typically in FASTQ format [16]. The quality of the data must be evaluated using tools like FastQC [76] to ensure the reliability of the experimental procedure. Then, the reads are mapped to a known transcriptome or annotated genome, which can be performed using alignment tools (Table 1 - Phase 1). These tools already perform splicing, allowing gaps in the readings

when compared to the reference genome (which contains introns and exons). In case there is no genome annotation available, *de novo* transcription assembly can be performed using tools like StringTie [55] and SOAPdenovo-Trans [77].

Phase 2 — Quantification of transcript abundance The expression quantification can be performed at two different levels: gene-level, which only accounts for genes regardless of alternative splicing events, and isoform-level quantification, which allows measuring different isoforms of the same gene [78]. For quantification at the gene level, featureCounts and Enhanced Read Analysis of Gene Expression (ERANGE) are state-of-the-art tools [65]. Isoform quantification methods can be based on alignments of reads using the transcriptome (e.g., RSEM), the whole genome sequence (e.g., StringTie), or any of them (e.g., Sailfish), as a reference.

Table 1. State-of-the-art tools used for processing of transcriptomics data, and respective plataform availability.

Tools for processing of transcriptomics data				
Phase	Workflow	Tool Name	Availability	Reference
Phase 1	Quality check	FASTQC	CLI[a], Docker, Galaxy[b], GUI[c]	[76]
	Alignment	STAR	CLI, Docker, Galaxy	[20]
		HISAT	CLI, Docker, Galaxy	[34]
		TOPHAT	CLI, Docker, Galaxy	[35]
		Bowtie 2	CLI, Docker, Galaxy	[38]
Phase 2	Gene level quantification	ALEXA-seq	CLI, R	[25]
		NEUMA	CLI	[40]
		featureCounts	CLI, Docker, Galaxy, R	[46]
		ERANGE	CLI, Python	[50]
	Isoform level quantification	RSEM	CLI, Docker	[44]
		Sailfish	CLI, Docker, Galaxy	[53]
		StringTie	CLI, Docker, Galaxy	[55]
Phase 3	Filtering and normalization	edgeR	Galaxy, R	[58]
		DESeq2	CLI, Docker, Galaxy, R	[47]
Phase 4	Differential gene expression	edgeR	Galaxy, R	[58]
		DESeq2	CLI, Docker, Galaxy, R	[47]
	Differental isoform expression	Ballgown	CLI, Docker, R	[8]
		CuffDiff	CLI, Docker, Galaxy	[68]
		MMSEQ	CLI, Docker	[69]

[a]Command Line Interface; [b]Galaxy portal at `usegalaxy.org`; [c]Graphical User Interface.

Phase 3 — Filtering and normalization Quantified genes or transcripts need to be filtered and normalized. Different parameters can be accounted for, including reading depth, expression patterns, and biases [65,57,73]. The selection

of the normalization method can have a big impact on the biological interpretation of transcriptomics data [61]. Most computational normalization methodologies are supported by two important assumptions: expression levels of most genes remain the same across replicated groups [57], and different sample groups do not exhibit a significant difference in overall mRNA levels. However, these assumptions may not always be true, thus the method must be selected by accounting for the experiment's specificity [65]. The most used normalization methods are the M-value trimmed mean method (TMM) [59] (which is incorporated in edgeR) and Median Ratio Normalization, available in DESeq2 [47].

Phase 4 - Differential expression analysis Once the expression matrix is filtered and normalized, the data can be used for different purposes [65], including differential expression analysis. Although this step is not required for the integration of transcriptomics data in GEMs, it can provide additional information regarding the expression of non-metabolic genes.

2.4 Integration of transcriptomics

In the last decade, there has been an effort to integrate omics data into GEMs, as this information allows obtaining models with better performances [37]. The methods designed for this purpose differ essentially in the reaction's constraining approach, and requirement for defining an objective function (Table S1). The integration can be faced on a "switch" approach, where a reaction is considered active or not based on the expression associated with certain gene [71]. Another methodology is the "valve" approach, which determines flux constraints of reactions using quantitative gene expression data [71].

Regarding the need to define objective functions, the algorithms are divided into three groups: (1) Gene Inactivation Moderated by Metabolism and Expression (GIMME) [11], which allows the creation of models under specific conditions, taking into account an objective function, transcriptomics, and metabolomics data; (2) Integrative Metabolic Analysis Tool (iMAT) [81], which maximizes the agreement between the flux of reactions and gene expression data; (3) Model-Building Algorithm (MBA) [30], that builds two sets of reactions (core and non-core reactions), specific for a given tissue, and iteratively removes the non-core reactions [32,75].

For studies relying on organisms with a fast-growing profile, where the assumption of biomass flux maximization has successfully predicted metabolic behavior, methods belonging to the GIMME group, like E-Flux [17] should be used. However, if the main study focus relies on a broad range of systems (for example, microorganisms with variable biomass composition or cells of a multicellular organism), where the objective functions could be universally applied to different conditions, iMAT and MBA-like algorithms are more adequate [37].

The integration of transcriptomics into metabolic models is simplified by using tools that include several of these algorithms, like *troppo* [22], COBRA Toolbox v.3.0 [28], and *MEWpy* [54].

3 Problem analysis and Objectives

N. gaditana grows very poorly in heterotrophy, despite producing large amounts of lipids. To use this species industrially it is required improving its biomass production yields while optimizing the production of PUFAS (especially ω-3).

The processing and integration of RNAseq data into GEMs implies using a high number of tools, which can be a complex and time-consuming process. Designing and implementing an appropriate workflow will assist and accelerate this procedure.

The main goal of this project is to process and analyze transcriptomics data of *N. gaditana*, obtained under nitrogen and phosphorus deprivation conditions, as well as integrate the data into a GEM of the same organism.

4 Methodology

Several bioinformatics tools and methods will be used to achieve the designed goals. A pipeline will be implemented in Python to manage the whole process, as described below.

Workflow Implementation Methods to process and integrate transcriptomics data were implemented in Python. Since several tools useful for transcriptomics processing are integrated on Galaxy [2], this portal was used remotely via BioBlend API. Methods to upload/download data, run alignments and gene expression quantification, and generate reports were be implemented.

Pre-processing of transcriptomics data Transcriptomics data for *N. gaditana* obtained from two independent studies [5,29] were used (EBI-ENA accessions PRJNA360152 and PRJNA589063). These studies were carried out under nitrogen, and nitrogen and phosphorus deprivation, respectively.

The quality of raw sequence data was assessed with FastQC. The paired-end read sequences were aligned against the genome of *N. gaditana* CCMP1894 (GenBank Assembly GCA_002838785.1) using RNA STAR. An *in house* genome structural annotation was used as gene model template for the alignment. Default parameters were used except for the length of the SA pre-indexing string 'genomeSAindexNbases', which was changed to 11, according with the RNA STAR user guide recommendation. FeatureCounts was used to quantify the gene expression in *CPM*, which was then converted into *TPM*, using the bioinfokit python package [12].

Integration of transcriptomics data In this stage, the gene expression dataset was integrated in a *in house* developed GEM for *N. gaditana*. The integration was performed with MEWpy using the GIMME and E-Flux algorithms, resulting in a set of context-specific GEMs. Both algorithms provide a simulation solution obtained with pFBA using the CPLEX solver.

Condition-specific GEM analysis Finally, the condition-specific GEMs were used to perform *in silico* simulations - pFBA and FVA - to evaluate potential changes in the metabolism of the microalgae according with the different nitrogen and phosphorus availability. Standard environmental conditions were used: the uptake of photons (restricted to 200 $mmol \cdot g_{DW} \cdot h^{-1}$), CO_2, O_2, H_2O, H_2SO_4, Mg^{2+}, NH_4, and orthophosphate (unrestricted conditions). The pFBA and FVA simulations were performed using the biomass production as objective function.

5 Results and Discussion

5.1 Workflow Implementation

The workflow described above was implemented in python. A class was created to implement methods to run different tools across the Galaxy platform through the BioBlend API, namely FASTQC, RNAStar, and featureCounts. The calculation of the TPM was imlpemented using the package bioinfokit [13].

A class *Integration* was implemented to allow an easy integration of transcriptomics data, using the *TPM* counts obtained previously. *COBRApy* [21] and *MEWpy* [54] were used to handle the model and run the GIMME and E-flux algorithms.

The code implemented in this work, as well as the results obtained are available in github: *https://github.com/beach-fossils/bioinformatics-project/* .

5.2 Pre-processing of transcriptomics data

FASTQC - FASTQC offers an *html* file including several information regarding the characterization of the quality of the raw sequences. To follow the next analysis, it is recommended the accompaniment of plot visualization for each module, and this is possible here (report 1) and here (report 2). It offers, at first, a table revealing some basic statistical information about the sequences. Both files reveal identical characteristics, according to the FASTQC reports. The raw data of `toy_SRR5152513_1_fastq` and `toy_SRR5152513_2_fastq` have a total of 15000 sequences each, in which 0 are flagged as poor quality. The length of sequences varies between 26-76 bp and the %GC is 55. The report offers other modules, among these, the focus will rely if a flag or an alert was raised. The per base sequence content module was the only module where a flag was raised. It usually creates a fail for RNA-Seq data, as the first bases (10-12) in the library preparation result from random hexamer priming. Although this module often fails in RNA-Seq libraries, redoing the process using a different library preparation may solve the problem. Regarding the remaining modules, both files reveal no additional problem, passing with success in any of the other modules. It is possible to infer that both files are of good quality and ready to be used at next steps.

RNA STAR The read sequences were aligned against the genome of *N. gaditana* CCMP1894 (GenBank Assembly GCA 002838785.1), using a gene model template developed *in house*. The overall statistics of the alignment can be seen at Table 2. Full details of the alignments are available here. As noticeable the rate of mismatch, deletion, insertion per base is very low (all lower than 1%), indicating an alignment with a high read coverage. The number of *reads unmapped: too short* is higher at the conditions related to the study PRJNA589063 (EBI-ENA accessions) [29].

Table 2. Main results obtained with RNA STAR for alignment against *N. Gaditana* genome

Results obtained with RNA STAR					
Condition	wt[a]	nd[b]	c5a[c]	n5a[d]	p5a[e]
Number of input reads	10238117	8294347	17695722	18765398	17934875
Average input read length	149	192	300	300	300
Uniquely mapped reads (%)	95.08	93.35	75.42	72.97	79.05
Number of splices Annotated	1628911	2179997	3827735	3819448	4762281
Mismatch rate per base	0.27%	0.10%	0.83%	0.82%	0.82%
Deletion rate per base	0.00%	0.00%	0.05%	0.06%	0.03%
Insertion rate per base	0.00%	0.00%	0.02%	0.02%	0.02%
% of reads mapped to multiple loci	3.10%	3.17%	5.26%	2.80%	6.27%
% of reads mapped to too many loci	0.21%	0.21%	0.10%	0.10%	0.15%
Number of reads unmapped: too many mismatches	0	0	0	0	0
% of reads unmapped: too short	1.51%	3.14%	18.89%	23.82%	14.17%
% of reads unmapped: other	0.10%	0.13%	0.34%	0.31%	0.37%
Number of chimeric reads	0	0	0	0	0

[a]Wild type [5]; [b]N deprived [5]; [c]Control [29]; [d]N deprived [29]; [e]N and P deprived [29]

Gene counts The gene counts in TPM were determined for both studies under analysis using featureCounts and the bioinfokit package. The output can be found here. The overall statistical information regarding the gene expression data for all conditions under study are demonstrated at Table 3.

Table 3. Results and statistical characterization of *tpm* expression data for different conditions

Statistical results obtained with Gene counts					
Condition	wt[a]	nd[b]	c5a[c]	n5a[d]	p5a[e]
Number of genes	11261	11261	11261	11261	11261
Mean	88.8	88.8	88.8	88.8	88.8
Standard deviation	426.1	402.0	638.2	349.3	640.5
Minimum value	0	0	0	0	0
First quartile (25%)	4.9	4.7	4.8	7.8	5.3
Median (50%)	25.8	27.2	16.7	28.8	16.8
Third quartile (75%)	69.0	71.0	50.7	74.4	50.7
Maximum value	14076.9	12720.7	52815.8	19040.1	52459.0

[a]Wild type [5]; [b]N deprived [5]; [c]Control [29]; [d]N deprived [29]; [e]N and P deprived [29]

The TPM mean has always the same value since it is equal to one million divided by the number of transcripts annotated [79]. As expected, the minimum value of TPM is zero for all conditions, while the maximum value is in the same order of magnitude (10^4). With respect to the other parameters, the values for each condition are very similar between them.

5.3 Integration of transcriptomics data

The transcriptomics data from both studies were integrated into an *in house* developed model of *N. gaditana*, using GIMME and E-Flux. The results of both methods for each condition can be found here. The results obtained with GIMME did not allowed observing any differences between the control and nitrogen/phosphorus deprivation conditions. This might occur because the reactions turned off by GIMME have no impact on the optimizations in both conditions. A more restrictive approach with this algorithm could allow observing differences. On the other hand, E-Flux provided different results for the different conditions considered in both studies, allowing obtaining different condition-specific GEMs.

5.4 Condition-specific GEM analysis

FVA and pFBA simulations were performed for each condition-specific GEM. The results of the pFBA were analyzed and compared to identify reactions whose flux increased (overexpressed) or decreased (underexpression) by the nitrogen/phosphorus depletion. For the first study [5], 195 reactions were identified as overexpressed and 213 reactions as underexpressed.

Table 4. Number of reactions per metabolic pathway for each condition against *wild type/control* after pFBA simulations

Number of reactions related to specific pathway			
	wt[a] vs nd[b]	c5a[c] vs n5a[d]	c5a[c] vs p5a[e]
Overexpressed reactions			
Transporters pathway	36	22	19
Biosynthesis of amino acids	25	9	11
Fatty acid metabolism	20	2	4
Fatty acid biosynthesis	20	2	4
Biosynthesis of unsaturated fatty acids	19	0	0
PKS system	18	0	0
Glycolysis / Gluconeogenesis	15	6	10
Citrate cycle (TCA cycle)	12	3	2
Pentose phosphate pathway	7	2	8
Pyruvate metabolism	13	5	10
Underexpressed reactions			
Transporters pathway	27	27	50
Biosynthesis of amino acids	16	12	17
Fatty acid metabolism	49	2	4
Fatty acid biosynthesis	38	2	4
Biosynthesis of unsaturated fatty acids	10	0	0
PKS system	0	0	0
Glycolysis / Gluconeogenesis	16	4	9
Citrate cycle (TCA cycle)	5	11	15
Pentose phosphate pathway	12	1	3
Pyruvate metabolism	10	7	0

[a]Wild type [5]; [b]N deprived [5]; [c]Control [29]; [d]N deprived [29]; [e]N and P deprived [29]

The number of reactions over and underexpressed by pathway in the different conditions is summarized in Table 4. The reactions identified as overexpressed are often related to transporters pathway (36), biosynthesis of amino acids (25), fatty acid metabolism (20), fatty acid biosynthesis (20) and biosynthesis of unsaturated fatty acids (19). Regarding the underexpressed reactions, 59 of them were related to transporters pathway, 49 with the fatty acid metabolism, 38 with fatty acid biosynthesis and 16 with glycolysis/gluconeogenesis.

The pathways and reactions associated with the fatty acid and lipid metabolism were deeply analyzed. Almost all of the underexpressed reactions connected to fatty acid biosynthesis are located in the chloroplast, and a few in the endoplasmic reticulum. The *wild type* algae is probably demonstrating typical fatty acid production at the chloroplast (from *de novo* until C18 fatty acids), which then are forwarded to the endoplasmatic reticulum to be insaturated and elongated [74]. This pathway is present in the wild-type, but not in the nitrogen deprivation conditions Here, the reactions belonging to this pathway were underexpressed, while reactions associated with the polyketide synhase system (PKS) were iden-

tified as overexpressed. This enzyme is able to produce fatty acids *de novo* in the cytoplasm, as well as elongating them and introducing double bounds [74]. Thus, nitrogen absence might induce a shift in the production of fatty acids in *N. gaditana*. This observation was already reported for *N. oceanica* [74]. However, this behavior was not observed in the second study. Nitrogen deprivation did not increased the activity of polyketide synthase system, as confirmed by the authors of the study [29].

The accuracy of *in silico* simulations applied to the condition-specific GEMs could be improved by defining condition specific biomass compositions. Such approach would be relevant for studying lipid metabolism since *N. gaditana* is known to produce less polar lipids and more triacylglycerols in nutrient starvation conditions. Thus, a biomass composition representing such change would allow a more rigorous analysis.

Both reaction expressions demonstrate high amount of transporters pathway, which is well described as lipid accumulation is accompanied by the regulation of inorganic phosphate transport across membranes [6]. Transport reactions were explored together with the data from the two studies and a few reactions were found to be overexpressed in all conditions with the following ids: `TR2000009_PLAS__mitmem`, `T_Acetate__chlomem`, `T_CoA_AMP__mitmem` and `TR2000131_PLAS__mitmem`. Notice these reactions refer to transport related to the mitochondria, having different metabolites associated with, namely: Orthophosphate, Acetate, CoA and ATP. A similar pattern was found for the reactions being underexpressed: `T_MALICITtm__mitmem`, `TR2000105_PLAS__mitmem`, `T_OAAICITtm__mitmem`, `TR2300023_PLAS__mitmem`, `TR2900000_PLAS__mitmem` and `TZ2900004_PLAS__chlomem`. Notice, again, these reactions refer to transportation related to mitochondria, which have different metabolites associated with, such as: Malate, Isocitrate, Oxaloacetate, L-Aspartate, L-Glutamate.

Regarding the transcriptomics data related to the second study [29], phosphorus deprivation lead to an increase in the activity of glycolysis, phosphorylative oxidation and pentose phosphate pathway. Surprisingly, several reactions associated with the citrate cycle were identified as underexpressed in phosphorus deprivation. After analyzing the reaction's fluxes, a change in the carbon route in mitochondria was observed. Instead of following the usual citrate cycle, the isocitrate was converted directly into succinate and glyoxylate. The glyoxylate was recycled while the succinate was used to feed the overexpressed oxidative phosphorylation (Fig.S.1). Moreover, the reactions of the photosynthetic pathway were also underexpressed in phosphorus deprivation conditions, as reported by the authors of the study [29].

6 Conclusion

The data obtained through RNA-Seq has to be processed and for that it is necessary to use a series of tools, which Galaxy [2] can provide. The combination of transcriptomics with metabolic models, is extremely powerful to draw various conclusions about a certain topic in study. For the project's case it shown to

be useful to explore reaction expression for different conditions holistically with constraints.

N. Gaditana revealed to have several reactions overexpressed related to the polyketide synhase system (PKS) under nitrogen deprivation, which may be a good pointer for unsaturated lipid production at industrial scale. Regarding the availability of phosphorus for the algae, it does not seem to have an impact on its performance producing lipids, instead it seems to affect its energy metabolism.

In perspective for possible future studies, it could be advantageous to select data reflecting other kinds of conditions, such as: salinity, light regimes and light intensity, because these greatly make the metabolism of algae to adjust. Having these data, the conclusions can be even more interesting and more accurate.

References

1. Adamczyk, M., Lasek, J., Skawińska, A.: Co2 biofixation and growth kinetics of chlorella vulgaris and nannochloropsis gaditana. Applied biochemistry and biotechnology **179**(7), 1248–1261 (2016)
2. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research **46**(W1), W537–W544 (2018)
3. Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., Nielsen, J.: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. PLoS computational biology **8**(5), e1002518 (2012)
4. Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., Nielsen, J.: Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology **10**(3), 721 (2014)
5. Ajjawi, I., Verruto, J., Aqui, M., Soriaga, L.B., Coppersmith, J., Kwok, K., Peach, L., Orchard, E., Kalb, R., Xu, W., et al.: Lipid production in nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. Nature biotechnology **35**(7), 647–652 (2017)
6. Alboresi, A., Perin, G., Vitulo, N., Diretto, G., Block, M., Jouhet, J., Meneghesso, A., Valle, G., Giuliano, G., Maréchal, E., et al.: Light remodels lipid biosynthesis in nannochloropsis gaditana by modulating carbon partitioning between organelles. Plant Physiology **171**(4), 2468–2482 (2016)
7. Alper, H., Jin, Y.S., Moxley, J., Stephanopoulos, G.: Identifying gene targets for the metabolic engineering of lycopene biosynthesis in escherichia coli. Metabolic engineering **7**(3), 155–164 (2005)
8. Anders, S., Pyl, P.T., Huber, W.: Htseq—a python framework to work with high-throughput sequencing data. bioinformatics **31**(2), 166–169 (2015)
9. Angione, C., Lió, P.: Predictive analytics of environmental adaptability in multi-omic network models. Scientific reports **5**(1), 1–21 (2015)
10. Asadollahi, M.A., Maury, J., Patil, K.R., Schalk, M., Clark, A., Nielsen, J.: Enhancing sesquiterpene production in saccharomyces cerevisiae through in silico driven metabolic engineering. Metabolic engineering **11**(6), 328–334 (2009)
11. Becker, S.A., Palsson, B.O.: Context-specific metabolic networks are consistent with experiments. PLoS computational biology **4**(5), e1000082 (2008)

12. Bedre, R.: bioinfokit 0.6
13. Bedre, R.: Project links
14. Benner, P., Findeisen, R., Flockerzi, D., Reichl, U., Sundmacher, K., Benner, P.: Large-scale networks in engineering and life sciences. Springer (2014)
15. Chandrasekaran, S., Price, N.D.: Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. Proceedings of the National Academy of Sciences **107**(41), 17845–17850 (2010)
16. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M.: The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. Nucleic acids research **38**(6), 1767–1771 (2010)
17. Colijn, C., Brandes, A., Zucker, J., Lun, D.S., Weiner, B., Farhat, M.R., Cheng, T.Y., Moody, D.B., Murray, M., Galagan, J.E.: Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. PLoS computational biology **5**(8), e1000489 (2009)
18. Converti, A., Casazza, A.A., Ortiz, E.Y., Perego, P., Del Borghi, M.: Effect of temperature and nitrogen concentration on the growth and lipid content of nannochloropsis oculata and chlorella vulgaris for biodiesel production. Chemical Engineering and Processing: Process Intensification **48**(6), 1146–1151 (2009)
19. Dias, O., Rocha, I.: Systems biology in fungi. Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi pp. 69–92 (2015)
20. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
21. Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R.: Cobrapy: constraints-based reconstruction and analysis for python. BMC systems biology **7**(1), 1–6 (2013)
22. Ferreira, J., Vieira, V., Gomes, J., Correia, S., Rocha, M.: Troppo-a python framework for the reconstruction of context-specific metabolic models. In: International Conference on Practical Applications of Computational Biology & Bioinformatics. pp. 146–153. Springer (2019)
23. Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., Shlomi, T.: Predicting selective drug targets in cancer through metabolic networks. Molecular systems biology **7**(1), 501 (2011)
24. Gouveia, L., Oliveira, A.C.: Microalgae as a raw material for biofuels production. Journal of industrial microbiology and biotechnology **36**(2), 269–274 (2009)
25. Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D., Corbett, R., Tang, M.J., Hou, Y.C., Pugh, T.J., et al.: Alternative expression analysis by rna sequencing. Nature methods **7**(10), 843–847 (2010)
26. Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., Lee, S.Y.: Current status and applications of genome-scale metabolic models. Genome biology **20**(1), 1–18 (2019)
27. Halim, R., Danquah, M.K., Webley, P.A.: Extraction of oil from microalgae for biodiesel production: A review. Biotechnology Advances **30**, 709–732 (5 2012). https://doi.org/10.1016/j.biotechadv.2012.01.001, `https://linkinghub.elsevier.com/retrieve/pii/S0734975012000031`
28. Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al.: Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. Nature protocols **14**(3), 639–702 (2019)

29. Hulatt, C.J., Smolina, I., Dowle, A., Kopp, M., Vasanth, G.K., Hoarau, G.G., Wijffels, R.H., Kiron, V.: Proteomic and transcriptomic patterns during lipid remodeling in nannochloropsis gaditana. International journal of molecular sciences **21**(18), 6946 (2020)
30. Imam, S., Schäuble, S., Valenzuela, J., López García de Lomana, A., Carter, W., Price, N.D., Baliga, N.S.: A refined genome-scale reconstruction of chlamydomonas metabolism provides a platform for systems-level analyses. The Plant Journal **84**(6), 1239–1256 (2015)
31. Janssen, J.H., Spoelder, J., Koehorst, J.J., Schaap, P.J., Wijffels, R.H., Barbosa, M.J.: Time-dependent transcriptome profile of genes involved in triacylglycerol (tag) and polyunsaturated fatty acid synthesis in nannochloropsis gaditana during nitrogen starvation. Journal of Applied Phycology **32**, 1153–1164 (4 2020). https://doi.org/10.1007/S10811-019-02021-2/FIGURES/7, `https://link.springer.com/article/10.1007/s10811-019-02021-2`
32. Jerby, L., Shlomi, T., Ruppin, E.: Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Molecular systems biology **6**(1), 401 (2010)
33. Kapoor, B., Kapoor, D., Gautam, S., Singh, R., Bhardwaj, S.: Dietary polyunsaturated fatty acids (pufas): Uses and potential health benefits. Current Nutrition Reports **10**(3), 232–242 (2021)
34. Kim, D., Langmead, B., Salzberg, S.L.: Hisat: a fast spliced aligner with low memory requirements. Nature methods **12**(4), 357–360 (2015)
35. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L.: Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology **14**(4), 1–13 (2013)
36. Kim, H.U., Kim, S.Y., Jeong, H., Kim, T.Y., Kim, J.J., Choy, H.E., Yi, K.Y., Rhee, J.H., Lee, S.Y.: Integrative genome-scale metabolic analysis of vibrio vulnificus for drug targeting and discovery. Molecular systems biology **7**(1), 460 (2011)
37. Kim, M.K., Lun, D.S.: Methods for integration of transcriptomic data in genome-scale metabolic models. Computational and structural biotechnology journal **11**(18), 59–65 (2014)
38. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357–359 (2012)
39. Lee, J.M., Lee, H., Kang, S., Park, W.J.: Fatty acid desaturases, polyunsaturated fatty acid regulation, and biotechnological advances. Nutrients **8**(1), 23 (2016)
40. Lee, S., Seo, C.H., Lim, B., Yang, J.O., Oh, J., Kim, M., Lee, S., Lee, B., Kang, C., Lee, S.: Accurate quantification of transcriptome from rna-seq data by effective length normalization. Nucleic acids research **39**(2), e9–e9 (2011)
41. Letsiou, S., Kalliampakou, K., Gardikis, K., Mantecon, L., Infante, C., Chatzikonstantinou, M., Labrou, N.E., Flemetakis, E.: Skin protective effects of nannochloropsis gaditana extract on h2o2-stressed human dermal fibroblasts. Frontiers in Marine Science **4**, 221 (2017)
42. Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., Weitz, K.K., Eils, R., König, R., Smith, R.D., Palsson, B.Ø.: Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology **6**(1), 390 (jul 2010). https://doi.org/10.1038/msb.2010.47, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925526/http://www.ncbi.nlm.nih.gov/pubmed/20664636`

43. Lewis, N.E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M.P., Cheng, J.K., Patel, N., Yee, A., Lewis, R.A., Eils, R., et al.: Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nature biotechnology **28**(12), 1279–1285 (2010)
44. Li, B., Dewey, C.N.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics **12**(1), 1–16 (2011)
45. Li, L., Zhou, X., Ching, W.K., Wang, P.: Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in nci-60 cell lines. BMC bioinformatics **11**(1), 1–16 (2010)
46. Liao, Y., Smyth, G.K., Shi, W.: featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**(7), 923–930 (2014)
47. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**(12), 1–21 (2014)
48. Ma, X.N., Chen, T.P., Yang, B., Liu, J., Chen, F.: Lipid production from nannochloropsis. Marine drugs **14**(4), 61 (2016)
49. Mahadevan, R., Schilling, C.H.: The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic engineering **5**(4), 264–276 (2003)
50. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by rna-seq. Nature methods **5**(7), 621–628 (2008)
51. Navid, A., Almaas, E.: Genome-level transcription data of yersinia pestis analyzed with a new metabolic constraint-based approach. BMC systems biology **6**(1), 1–18 (2012)
52. Pal, D., Khozin-Goldberg, I., Cohen, Z., Boussiba, S.: The effect of light, salinity, and nitrogen availability on lipid production by nannochloropsis sp. Applied microbiology and biotechnology **90**(4), 1429–1441 (2011)
53. Patro, R., Mount, S.M., Kingsford, C.: Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. Nature biotechnology **32**(5), 462–464 (2014)
54. Pereira, V., Cruz, F., Rocha, M.: Mewpy: a computational strain optimization workbench in python. Bioinformatics **37**(16), 2494–2496 (2021)
55. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L.: Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nature biotechnology **33**(3), 290–295 (2015)
56. Radakovits, R., Jinkerson, R.E., Darzins, A., Posewitz, M.C.: Genetic engineering of algae for enhanced biofuel production. Eukaryotic cell **9**(4), 486–501 (2010)
57. Risso, D., Ngai, J., Speed, T.P., Dudoit, S.: Normalization of rna-seq data using factor analysis of control genes or samples. Nature biotechnology **32**(9), 896–902 (2014)
58. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
59. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of rna-seq data. Genome biology **11**(3), 1–9 (2010)
60. Rodolfi, L., Chini Zittelli, G., Bassi, N., Padovani, G., Biondi, N., Bonini, G., Tredici, M.R.: Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnology and bioengineering **102**(1), 100–112 (2009)

61. Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P.T., Au, K.F., Bani Asadi, N., Gerstein, M.B., Wong, W.H., Snyder, M.P., et al.: Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum rna-seq analysis. Nature communications **8**(1), 1–15 (2017)
62. Schmidt, B.J., Ebrahim, A., Metz, T.O., Adkins, J.N., Palsson, B.Ø., Hyduke, D.R.: Gim3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics **29**(22), 2900–2908 (2013)
63. Schultz, A., Qutub, A.A.: Reconstruction of tissue-specific metabolic networks using corda. PLoS computational biology **12**(3), e1004808 (2016)
64. Serini, S., Fasano, E., Piccioni, E., Cittadini, A.R., Calviello, G.: Dietary n-3 polyunsaturated fatty acids and the paradox of their health benefits and potential harmful effects. Chemical research in toxicology **24**(12), 2093–2105 (2011)
65. Stark, R., Grzelak, M., Hadfield, J.: Rna sequencing: the teenage years. Nature Reviews Genetics **20**(11), 631–656 (2019)
66. Terzer, M., Maynard, N.D., Covert, M.W., Stelling, J.: Genome-scale metabolic networks. Wiley Interdisciplinary Reviews: Systems Biology and Medicine **1**(3), 285–297 (2009)
67. Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols **5**(1), 93–121 (2010)
68. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols **7**(3), 562–578 (2012)
69. Turro, E., Su, S.Y., Gonçalves, Â., Coin, L.J., Richardson, S., Lewin, A.: Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. Genome biology **12**(2), 1–15 (2011)
70. Varma, A., Palsson, B.O., Varma, A., Palsson, B.O.: Metabolic capabilities of escherichia coli. II. Optimal growth patterns. Journal of Theoretical Biology **165**(4), 503–522 (dec 1993). https://doi.org/10.1006/jtbi.1993.1203, `https://ui.adsabs.harvard.edu/abs/1993JThBi.165..503V/abstract`
71. Vivek-Ananth, R., Samal, A.: Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. Biosystems **147**, 1–10 (2016)
72. Vlassis, N., Pacheco, M.P., Sauter, T.: Fast reconstruction of compact context-specific metabolic network models. PLoS computational biology **10**(1), e1003424 (2014)
73. Wagner, G.P., Kin, K., Lynch, V.J.: Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. Theory in biosciences **131**(4), 281–285 (2012)
74. Wang, Q., Feng, Y., Lu, Y., Xin, Y., Shen, C., Wei, L., Liu, Y., Lv, N., Du, X., Zhu, W., et al.: Manipulating fatty-acid profile at unit chain-length resolution in the model industrial oleaginous microalgae nannochloropsis. Metabolic Engineering **66**, 157–166 (2021)
75. Wang, Y., Eddy, J.A., Price, N.D.: Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. BMC systems biology **6**(1), 1–16 (2012)
76. Wingett, S.W., Andrews, S.: Fastq screen: A tool for multi-genome mapping and quality control. F1000Research **7** (2018)
77. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al.: Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. Bioinformatics **30**(12), 1660–1666 (2014)

78. Yang, I.S., Kim, S.: Analysis of whole transcriptome sequencing data: workflow and software. Genomics & informatics **13**(4), 119 (2015)

79. Zhao, S., Ye, Z., Stanton, R.: Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. Rna **26**(8), 903–909 (2020)

80. Zou, N., Zhang, C., Cohen, Z., Richmond, A.: Production of cell mass and eicosapentaenoic acid (epa) in ultrahigh cell density cultures of nannochloropsis sp.(eustigmatophyceae). European Journal of Phycology **35**(2), 127–133 (2000)

81. Zur, H., Ruppin, E., Shlomi, T.: imat: an integrative metabolic analysis tool. Bioinformatics **26**(24), 3140–3142 (2010)

Supplementary Material

Appendix 1

Table S1. State-of-the-art algorithms for integration of transcriptomics data into GEMs, and the respective requirements for multiple transcriptomic dataset, definition of a gene expression threshold, definition of a required metabolic function, and ability to perform flux prediction tasks.

Methods for integrating transcriptomics data into GEMs						
Group	Method	Multiple transcriptomic dataset	Gene expression threshold	RMF[a]	Flux prediction	Reference
GIMME	GIMME	No	Yes	Yes	No	[11]
	GIM3E	No	Yes	Yes	No	[62]
	E-flux	No	No	Yes	Yes	[17]
	METRADE	No	No	Yes	Yes	[9]
	GX-FBA	No	Yes	Yes	Yes	[51]
	PROM	Yes	Yes	Yes	No	[15]
iMAT	iMAT	No	No	No	Yes	[81]
	iNIT	No	No	No	Yes	[3]
	tINIT	No	No	No	Yes	[4]
MBA	MBA	No	No	No	No	[32]
	mCADRE	No	No	No	No	[75]
	FastCORE	No	No	No	No	[72]
	CORDA	No	No	No	Yes	[63]

[a]RMF: Required Metabolic Function.

Appendix 2

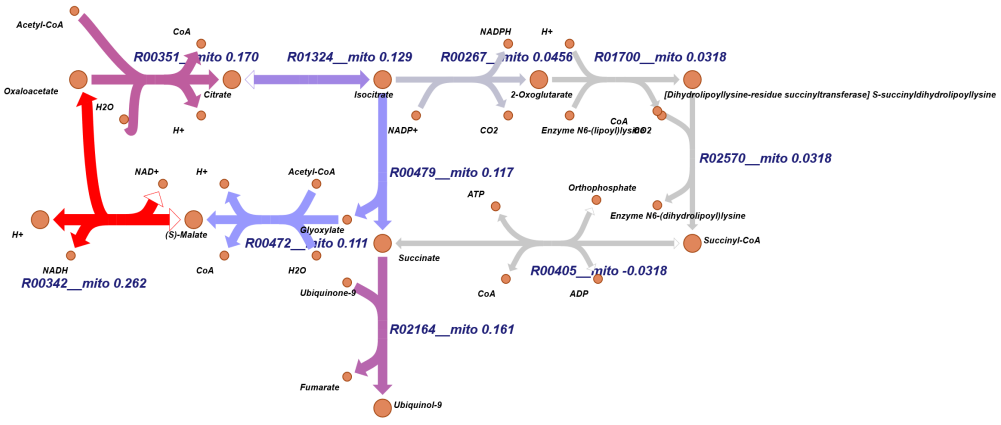

Fig. S.1. Mitochondrial carbon metabolism for energy production in nitrogen deprivation conditions.