UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

ÉTUDE DES LACUNES DE L'EXPRESSION ÉPISOMALE DANS LA DIATOMÉE *PHAEODACTYLUM TRICORNUTUM* POUR LA PRODUCTION DE PROTÉINES HÉTÉROLOGUES

FILLING GAPS OF EPISOMAL EXPRESSION IN THE DIATOM PHAEODACTYLUM TRICORNUTUM FOR THE PRODUCTION OF HETEROLOGOUS PROTEINS

THÈSE PRÉSENTÉE

COMME EXIGENCE PARTIELLE DU

DOCTORAT EN BIOLOGIE CELLULAIRE ET MOLECULAIRE

PAR

ARACELY MARIBEL DIAZ GARZA

SEPTEMBRE 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

DOCTORAT EN BIOLOGIE CELLULAIRE ET MOLÉCULAIRE

Direction de recherche :

Isabel Desgagné-Penix — Directrice de recherche

Brian Ingalls — Codirecteur de recherche

Jury d'évaluation de la thèse :

Isabel Desgagné-Penix — Directrice de recherche

Brian Ingalls — Codirecteur de recherche

Hugo Germain — Président de jury

Jörg Behnke — Évaluateur externe

Bogumil Karas — Évaluateur externe

Thèse soutenue le 6 décembre 2024

A mis papás, la mejor herencia es mi educación,
gracias.

AKNOWLEDGEMENTS

The work presented in this thesis would not have been possible without the support of numerous people. I would like to start thanking my supervisor Prof. Isabel Desgagné-Penix, my co-supervisor Prof. Brian Ingalls, and Dr. Natacha Mérindol for helping me give direction to this thesis project and providing their expertise to guide me. To Melodie B. Plourde, Andrew Diamond, Karen Gonçalves, and Elisa Fantino for their patience, for helping me improve my lab techniques (wet and dry lab), and transferring their knowledge. To the friends and colleagues that became like my family in Canada current and former members of the lab of Prof. Isabel Desgagné-Penix and Prof. Hugo Germain. It is indispensable to have support from the people in your surroundings, and the completion of this Ph.D. could not have been possible without them.

Me gustaría agradecer a mis papás, mi hermano, Lety, mis tíos y primos que siguen pensando que la gente inteligente hace un doctorado y están orgullosos de mí, en mi opinión se necesita resiliencia y perseverancia más que inteligencia la mayoría de las veces. Sin mis padres no hubiera podido tener la oportunidad de venir a Canadá a realizar mis estudios de grado, gracias por permitirme lograr mis metas y apoyarme. Quiero agradecer especialmente a mi mamá, hay muchas personas que me han dicho que no puedo hacer algo, pero tú nunca has sido una de ellas.

To my life partner Benoit Delcroix who has listened to my molecular biology “challenges” without having much idea of what I was talking about when I needed to give sense to my results, and to all of the

people that listened to my new ideas. To all my friends from the “latinas” group, I consider all of you as my closest friends. Gracias por darme la apertura de mostrar mi personalidad completa con ustedes. Karen Gonçalves quien fue mi mentora en la estancia de investigación clave para formar parte de este laboratorio y quien me convenció a regresar a hacer mis estudios de grado dándome argumentos y su amistad. *Obrigada* por todo el apoyo profesional y personal a lo largo de los años. A Ingrid con quien puedo tener conversaciones “telepáticas”, me enseña a ser una mejor persona y científica. Valeria por ayudarme a encontrar maneras en las cuales pensar en cosas fuera del doctorado. Eli con quien compartí viajes y experiencias que parecen sacadas de un sketch. Gracias a Ilse a quien también conocí en mi estancia, por las carreras en la nieve, en bici, en patines y en los días soleados.

To Nicolas Sene who started the Ph.D. at the same time as me and became one of my best friends to the point of inventing another language (Spanishos). To Snehi Gazal who has been the best partner of the office with whom I can have two minutes discussions and continue working. To Théo that understands the struggles of dry lab and cares a lot for his friends. To Fatima Awwad and Sameera Liyanage who have the most amazing ideas, great sense of humor, and a very caring heart. We had fruitful and not so fruitful discussions of science and life.

RÉSUMÉ

La synthèse hétérologue de protéines et de produits naturels offre des solutions aux défis liés à la croissance rapide de la population. Le développement de systèmes d'expression utilisant des hôtes hétérologues nécessite une caractérisation préalable des gènes et des enzymes, mais aussi de l'hôte. Bien que les principaux organismes hôtes comme *E. coli* et la levure soient bien connus, leur utilisation nécessite dans certains cas un génie génétique important. Les hôtes hétérologues émergents tels que les microalgues ont attiré l'attention pour leur utilisation dans la synthèse de produits naturels en raison de leur taux de croissance relativement rapide et des conditions peu coûteuses de leur croissance. En particulier, la diatomée *Phaeodactylum tricornutum* a été utilisée pour la production de protéines complexes et de produits naturels tels que les monoterpénoïdes, les caroténoïdes et les cannabinoïdes. Un aspect qui la différencie des autres microalgues est sa capacité à utiliser des vecteurs d'expression épisomale. Malgré ses succès, ce système d'expression présente encore quelques limites.

L'objectif de ce travail était d'utiliser le système d'expression épisomale de *P. tricornutum* pour caractériser la localisation subcellulaire de la protéine végétale vanilline synthase qui n'a pas d'activité enzymatique dans les micro-organismes. Nous avons utilisé des signaux caractérisés pour diriger la protéine vers quatre compartiments cellulaires différents de la diatomée. Bien que les lignées transgéniques présentent une réplication stable du plasmide sans mutations, et une localisation confirmée dans les compartiments ciblés, elles présentent des phénotypes hétérologues individuels dans une seule lignée cellulaire et une faible production de la protéine recombinante malgré l'enrichissement des cellules productrices de protéines.

Ces résultats nous ont amenés à étudier plus en profondeur le système d'expression épisomale en utilisant une lignée cellulaire exprimant l'eGFP qui présentait des sous-populations stables. Nous avons étudié la variation du nombre de copies du plasmide, les mutations et, à l'aide d'analyses transcriptomiques, les gènes candidats co-exprimés avec l'eGFP dans des cultures enrichies par tri cellulaire, afin de caractériser les différences entre trois sous-populations présentant des intensités de fluorescence distinctes. Nos résultats révèlent que le faible nombre de copies et la délétion de la cassette d'expression dans une partie de la population ont contribué à la faible expression du transgène. Toutefois, les mécanismes à l'origine de cette aberration de copies du plasmide ne sont pas

encore clairs. Nous avons également identifié une mutation dans la sous-population non fluorescente, qui n'était pas suffisamment abondante pour être détectée dans la culture non triée. Notre étude ne limite pas le phénotype non fluorescent à des mutations dans la cassette d'expression, puisque le séquençage des plasmides provenant de la population majoritairement non fluorescente des lignées exprimant le *Vp*VAN était dépourvu de mutations dans les séquences génétiques des cassettes. Nous pouvons donc conclure qu'il existe différents mécanismes régissant la réduction de l'expression du transgène et que la mutation des cassettes d'expression peut être l'un d'entre eux. D'autres études sont nécessaires pour décrypter les mécanismes qui modifient le nombre de copies de plasmides et établir des seuils pour la production stable de protéines hétérologues chez *P. tricornutum*.

Mots clés : diatomée, vanilline synthase, cytométrie en flux, compartimentation des organelles, sous-populations hétérologues.

ABSTRACT

Heterologous synthesis of proteins and natural products offers solutions to challenges related to the rapid population growth. The development of expression systems using heterologous host requires prior characterization of genes and enzymes, as well as of the host. While organisms like *E. coli* and yeast are well understood, production using these main hosts requires in some cases extensive genetic engineering. Emerging heterologous hosts like microalgae have attracted attention for their use for synthesis of natural products due to their relative fast growth rates and the cheap requirements for their culture. In particular, the diatom *Phaeodactylum tricornutum* has been used to produce complex proteins, and natural products such as monoterpenoids, carotenoids, and cannabinoids. An aspect that differentiates it from other microalgae is the ability to support episomal expression vectors. Despite its success stories the episomal expression system still presents some limitations.

The objective of this work was to use *P. tricornutum* episomal expression system to characterize the subcellular localization of the plant protein vanillin synthase which lacks enzymatic activity in microorganisms. We used characterized signals to direct the protein to four different cell compartments in the diatom. Even though transgenic lines presented stable plasmid replication without mutations, and confirmed localization in the targeted compartments, they had heterologous phenotypes of individuals in a single cell line and low production of the recombinant protein despite enrichment of protein producing cells.

These results led us to study deeper the episomal expression system using a cell line expressing *eGFP* which was exhibiting stable subpopulations. We studied plasmid copy number variation, mutations, and, using transcriptomic analyses, candidate genes co-expressed with *eGFP* in enriched cultures by cell sorting, to characterize the differences between three subpopulations with distinct fluorescence intensities. Our findings revealed that lower copy numbers and the deletion of the expression cassette in part of the population contributed to low transgene expression. However, the mechanisms causing this plasmid copy aberration are still unclear. The deletion identified in the non-fluorescent subpopulation, was not abundant enough to be detected in the non-sorted culture. Our study does not restrict the non-fluorescent phenotype to mutations in the expression cassette, since sequencing of plasmids coming from mostly non-fluorescent population of *Vp*VAN expressing lines were absent of mutations in gene sequences in the cassettes. Therefore,

we could conclude that there are different mechanisms governing the reduction of transgene expression and that the mutation of the expression cassettes can be one of them. Further investigations are necessary to decipher mechanisms to alter plasmid copy number and establish thresholds for stable production of heterologous proteins in *P. tricornutum*.

Key words: diatom, vanillin synthase, flow cytometry, organelle compartmentalization, heterologous subpopulations.

TABLE OF CONTENT

AKNOWLEDGEMENTS i
RÉSUMÉ iii
ABSTRACT v
TABLE OF CONTENT i
LIST OF TABLES 1
LIST OF FIGURES 2
LIST OF ACRONYMNS AND ABBREVIATIONS 5
CHAPTER I: INTRODUCTION 9
1.1. Uses of heterologous production 9
1.1.1. Heterologous synthesis of natural products, the example of vanillin 11
1.1.2. Development of heterologous expression systems 13
1.1.3. Impact of heterologous production on the host organism 17
1.2. Main host organisms for heterologous production 18
1.2.1. E. coli and other pioneer bacterial hosts. 19
1.2.2. S. cerevisiae and non-conventional yeast 23
1.3. Emerging heterologous host for characterization of NPs biosynthetic pathways. 26
1.3.1. Filamentous fungi 26
1.3.2. Cyanobacteria and eukaryotic microalgae 28
1.4. *Phaeodactylum tricornutum* episomal expression system 35
1.5. *P. tricornutum* as potential host organism for production of phenolic compounds 38
1.6 Research objectives 40
CHAPTER II: What is true for plants may not be true for *Phaeodactylum tricornutum*: The case of *Vanilla planifolia* vanillin synthase targeted to four subcellular compartments of the diatom. 42
Résumé 43
Abstract 45
Introduction 46
Results 49
Neither vanillin nor ferulic acid affect P. tricornutum growth at pH 8. 49
Selection and characterization of P. tricornutum exconjugants. 50

Successful localization of the recombinant proteins in four different cell compartments.....52
Enrichment in fluorescent cells, protein purification, and plastid isolation failed as strategies to detect heterologous protein by western blot in positive clones.....56
Discussion.....58
Conclusions.....64
Materials and Methods.....65
Microbial Strains and Growth Conditions.....65
Cytotoxicity assays of P. tricornutum with ferulic acid and vanillin.....65
Plasmid construction and transformation of P. tricornutum by bacterial conjugation from E. coli cells.....66
Screening positive exconjugants by fluorescence microscopy.68
Episome rescue, whole plasmid sequencing and in silico analysis.....68
Assessing subcellular localization by confocal microscopy. 69
Fluorescence-activated cell sorting (FACS).....69
Protein extraction and purification. 70
Plastid isolation.....71
Western blot. 71
Data Availability Statement.....72
Acknowledgments.....72
CHAPTER III: No two clones are alike: Characterization of heterologous subpopulations in a transgenic cell line of the model diatom *Phaeodactylum tricornutum.*.....74
Authors contributions.....74
Résumé.....75
Abstract.....77
Background.....78
Material and methods.....80
Microbial Strains and Growth Conditions.....80
Transformation of P. tricornutum by bacterial conjugation from E. coli cells.....80
Flow cytometry and fluorescence-activated cell sorting (FACS) 82
Episome DNA isolation and sequencing.....83
DNA extraction and copy number quantification.....84

RNA extraction and sequencing 84
Bioinformatic analyses 85
Results 87
Subpopulations of a heterogenic episomal expression cell line were efficiently enriched using Fluorescence Activated Cell Sorting (FACS). 87
The stability of the GFP^{-} subpopulation could be due to differences in sequence. 89
Plasmid copy number differs between GFP^{++} and GFP^{-}. 90
eGFP is selectively expressed at higher level in GFP positive subpopulations. 90
Differences at transcriptomic level highlight enriched protein families. 91
Coordinated expression patterns reveal candidate genes to increase recombinant protein expression in P. tricornutum. 92
Discussion 94
Availability of data and materials 102
Funding 102
Acknowledgements 103
Authors' information 103
Tables 104
Figures 110
CHAPTER IV: DISCUSSION AND CONCLUSIONS 117
4.1 Perspectives 120
4.2 Conclusion 123
REFERENCES 126
ANNEX A: Supplementary Information from Chapter II 157
ANNEX B: Supporting Information from Chapter III 167
Supplementary Figures 167
Supplementary Tables 175
ANNEX C: Clustering algorithms for flow cytometry data 177
ANNEX D: Collaborations and contributions 179

LIST OF TABLES

CHAPTER II:

Table 2. 1. Fluorescent colonies per construction. 52

CHAPTER III:

Table 3. 1. Plasmid copy number (PCN) of subpopulations is significantly different. 104

Table 3. 2. Doubling time of enriched cultures is only significantly higher in GFP^{++}..... 105

Table 3. 3. Complete list of genes following similar expression pattern with *eGFP* clustered in module 4.1..... 106

ANNEX A:

Table S2. 1. Primer sequences used for Gibson assembly. 157

Table S2. 2. Sequences used in this study..... 160

ANNEX B:

Table S3. 1. Primer sequences used for Sanger sequencing and qPCR to measure copy number. 175

Table S3. 2. Highly interconnected genes from module 73 of co-expression analysis. 176

LIST OF FIGURES

CHAPTER I:

Figure 1. 1 Overview of key factors governing gene expression. (cyanobacteria) 11

Figure 1. 2. Important elements of expression plasmid in bacteria.14

Figure 1. 3. Schematic of metabolic burden. 16

Figure 1. 4. Design of a fusion expression vector to aid in solubilizing the expressed recombinant protein. 21

Figure 1. 5. A review of engineering strategies for improved protein production by *S. cerevisiae* 24

Figure 1. 6. Heterologous synthesis of natural products in microalgae. 29

Figure 1. 7. Micrographs of *P. tricornutum*'s morphotypes. 33

Figure 1. 8. Conjugative transfer of plasmids from *E. coli* to *P. tricornutum*. 36

Figure 1. 9. Simplified biosynthetic pathway of vanillin and its potential intra-cellular localization. 39

CHAPTER II

Figure 2. 1. Assessing cytotoxicity of ferulic acid and vanillin to wildtype *P. tricornutum* cells. 51

Figure 2. 2. Stable episome propagation without mutations in the expression cassettes with localization signals for three different organelles and the cytosol. 53

Figure 2. 3. The fusion protein VpVAN:Xa:eGFP localizes in the targeted cell compartment of *P. tricornutum*. 56

Figure 2. 4. Enrichment for fluorescent cells by FACS fails to yield a high percentage of fluorescent cells. 58

Figure 2. 5. Enrichment of protein fraction and purification failed to increase fusion protein to detectable levels. 59

CHAPTER III:

Fig 3. 1. Subpopulations are efficiently enriched through cell sorting.. 110

Fig 3. 2. Enriched GFP$^-$ subpopulation exhibits episome rearrangement.. 112

Fig 3. 3. *eGFP,* but not *sh ble,* is expressed at higher level in GFP positive subpopulations. 113

Fig 3. 4. Most of the protein families enriched in the down-regulated genes are shared between the three subpopulations. 114

Fig 3. 5. Nested co-expression analysis highlights specific expression patterns in the subpopulations.. 115

Fig 3. 6. Speculative mechanisms of *P. tricornutum* cells in a clonally propagated culture creating differences in genetic content and gene expression..... 116

CHAPTER IV:

Figure 4. 1. Summarized findings of heterologous gene expression in cell lines originated from exconjugants of *P. tricornutum*. 125

ANNEX A:

Figure S2. 1. Proteins transferred to the blotting membrane..... 165

Figure S2. 2. Constructions of controls for subcellular localization.166

ANNEX B:

Fig. S3. 1. Initial analysis of GFP cell lines (pDMi8) do not present three stable subpopulations. 167

Fig. S3. 2. Fluorescence intensity of eGFP and mCherry is highly correlated.. 168

Fig. S3. 3. *Sh ble* primer pair validation for copy number quantification. 169

Fig. S3. 4. Subpopulations are detected in a clonally propagated *P. tricornutum* cell culture. 170

Fig. S3. 5. Enrichment of subpopulations through fluorescence activated cell sorting. 171

Fig. S3. 6. PCR amplification reveals two different versions of expression cassettes are present in the GFP- subpopulation.175

Fig. S3. 7. Deletion of 707 bp in GFP$^-$ sorted clones. 173

Fig. S3. 8. Commonly deregulated genes between subpopulations.174

Fig. S3. 9. A single GFP cell line presents three subpopulations. 178

ANNEX C:

Figure S4. 1. A single GFP cell line presents three subpopulations178

LIST OF ACRONYMNS AND ABBREVIATIONS

ΔMFI^{GFP}	Delta mean fluorescence intensity subtracting the background fluorescence from empty vector strain
40SRPS8	40S ribosomal protein S8
4-COMT	caffeic acid *O*-methyltransferase
AAR	acyl-ACP carrier protein reductase
ARS	*autonomously replicating sequence*
ATPC	plastid ATP synthase gamma chain
BGCs	biosynthetic gene clusters
C3'H	coumaric acid-4-hydroxylase
C4H	cinnamic acid-4-hydroxylase
CAH	*CEN6-ARSH4-HIS3*
CEN	*centromeric sequence*
CENH3	centromeric histone 3 variant
CO_2	carbon dioxide
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CS	plastid (chloroplast) stroma
CT	cytosolic tail and transmembrane domain
Ctl	control
CUB	codon usage bias
Cy	cytosol
DBS	double-strand breaks

Dub-seq	dual-barcoded shotgun expression library sequencing
Ech	enoyl-CoA hydratase/aldolase
EE	Extrachromosomal expression
eGFP	enhanced green fluorescent protein
ER	endoplasmic reticulum
EV	empty vector
FACS	fluorescent activated cell sorting
Fcs	trans-feruloyl-CoA synthase
Flp	flipase
FSC-A	forward scatter area
gDNA	genomic DNA
GFP^-	cells inside the gate defined using the negative control with no fluorescence
GFP^+	cells outside of the gate defined using the negative control with higher values of GFP intensity
GMM	gaussian mixture model
GO	gene ontology
Go	Golgi apparatus
GOI	gene of interest
GRAS	Generally Recognized as Safe
HASP1	highly abundant secreted protein 1
HIS3	imidazoleglycerol-phosphate dehydratase
IPTG	β-D-1-thiogalactopyranoside

JUMP	Joint Universal Modular Plasmids
LB	Luria broth
MCS	multiple cloning site
mRNA	messenger RNA
NPs	Natural products
OriT	origin of transfer
PAL	phenylalanine ammonia lyase
PCN	plasmid copy number
Pe	peroxisome
PEG	polyethylene glycol
PI	propidium iodide
PTS1	peroxisome targeting signal 1
RBS	ribosome binding site
rDNA	recombinant DNA
RNAi	RNA interference
RNA-seq	RNA-sequencing
rRNA	ribosomal RNA
RB-TnSeq	TnSeq with random DNA barcoding
RuBisCo	ribulose bisphosphate carboxylase oxigenase
sh ble	resistant marker to zeocin
SSC-A	side scatter area
TALE	transcription activator-like effector
TMD	transmembrane domain
TnSeq	Transposon mutagenesis followed by sequencing

TPM	transcripts per million
tRNAs	transfer RNAs
U.S. FDA	United States Food and Drug Administration
UTRs	untranslated regions
*Vp*VAN	*Vanilla planifolia* vanillin synthase
Xa	factor Xa cleavable peptide
XylT	β1,2-xylosyltransferase

CHAPTER I: INTRODUCTION

The world's population is expected to increase in the next 50 to 60 years and peak at 10.3 billion people in the mid-2080s from the current 8.2 billion (United Nations 2024). The rising global population generates concerns about food security, sustainability, sufficient drugs and medicines, species conservation, and so on, motivating developers to look for solutions (Henchion et al. 2017; Zhang et al. 2017). While chemical manufacturing was efficient in meeting the growing demand during the Industrial Revolution, their traditional model to reduce cost by increasing the production scale limits their adaptability to change in order to minimize environmental impact, creates an access barrier for developing economies and companies, and limits the potential for rapid innovative solution to tackle challenges (Clomburg et al. 2017). Alternatively, using living organisms as bio-factories could solve these issues.

1.1. Uses of heterologous production

Although the use of living organisms in the production of foods is centuries old (bread, wine, cheese), producing proteins and metabolites of high commercial interest in living organisms that naturally are not able to produce the molecule of interest (heterologous) only started several decades ago (Watts et al. 2021). Recombinant DNA (rDNA) technologies made it possible in the early 1970s (Mertz and Davis 1972). rDNA technologies involve assembling pieces of DNA from different sources containing a gene of interest (GOI) in a vector for its transfer to the host organism (Gill et al. 2023). Since the United States Food and Drug Administration (U.S. FDA) approval of recombinant insulin produced in *Escherichia coli* in 1982, more

than 300 therapeutical recombinant proteins are available in the market and around half of the industrial enzymes are produced in heterologous hosts (Brasil et al. 2017; Kastberg et al. 2022). Some other first successes for heterologous production of therapeutic molecules are the recombinant human growth hormone (somatropin), recombinant human coagulation factor IX, and penicillin (Bedford et al. 1995; Ashikaga et al. 2004; Lyseng-Williamson 2017). The first industrial enzyme approved by the U.S. FDA produced in a heterologous expression system was chymosin, used for milk curdling (Harris et al. 1982). Since then, monoclonal antibodies, complex natural products, and vaccines produced in heterologous expression systems have been commercialized (Rockman et al. 2020; Tiedge 2022; Strazza et al. 2024). For instance, the Oxford-AstraZeneca COVID-19 vaccine was developed using synthetic biology techniques to introduce the full-length SARS-CoV-2 spike glycoprotein gene to a replication-defective chimpanzee adenovirus vector (Ramasamy et al. 2020).

Synthetic biology applies engineering approaches to genetic components (parts) to design novel biological networks and was built using a small set of laboratory organisms, mainly *E. coli* and *Saccharomyces cerevisiae* (Adams 2016). Molecular biologists were familiar with these organisms since they were adapted to laboratory conditions, have fast growth rates, and accumulate a high amount of proteins, leading to a deep understanding of their biology and physiology. Thus, they became the legacy of chassis in synthetic biology (Adams 2016). Moreover, synthetic biology has allowed scientists to perturbate natural systems in a controlled manner to identify and validate mechanistic models and build genetic circuits using characterized components in new configurations, leading to significant

breakthroughs in biology (Khakhar and Voytas 2021).

The production of heterologous proteins has been simplified into two major stages: introducing foreign DNA into a host cell and expressing foreign DNA for protein synthesis (summarized in Figure 1. 1), which is affected by the chosen host (Rai and Padh 2001). Thus, the choice of host organisms must be based on the target product. The metabolic chassis of the host is a factor, especially when the target of the heterologous production is a metabolite. The host organism's physiological nature, such as its tolerance to heat and high concentration of product, abundance of key precursors, and expression conditions of heterologous pathway enzymes, must be considered for practical applications (Calero and Nikel 2019; Choi et al. 2019).

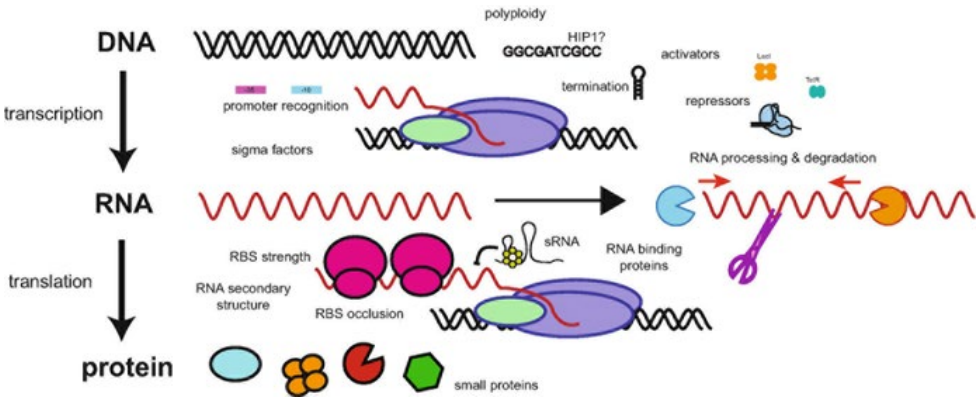


Figure 1. 1. Overview of key factors governing gene expression. Source Gordon and Pfleger, 2018

1.1.1. Heterologous synthesis of natural products, the example of vanillin

Natural products (NPs) with therapeutic, agro-industrial, antimicrobial, or nutritional properties are metabolites naturally synthesized by plants and microorganisms. NPs' complex chemical structures often make them difficult to produce through chemical synthesis (Park et al. 2020). Extraction

from whole plants is limited by seasonal constraints, the content of the desired product *in planta*, growth conditions, and rate (Wawrosch and Zotchev 2021).

In particular, the phenylpropanoid vanillin (4-hydroxy-3-methoxybenzaldehyde) is used in food and beverages as a flavor enhancer and, thanks to its antimicrobial and antioxidant properties, as a preservative (Kaur and Chakraborty 2013). Additionally, it is used in the pharmaceutical industry as raw material in the production of dopamine, L-3,4-dihydroxyphenylalanine, L-α-methyl-3,4-dihydroxyphenylalanine, papaverine, and an antifoaming agent (Kaur and Chakraborty 2013). Natural vanillin is mainly found in the beans of the orchid *Vanilla planifolia*, in lower quantities in V. tahitienis as well as V. pompona (Kundu 2017), and in trace amounts in other plants, including food crops (Sinha et al. 2008). This phenolic aldehyde extraction requires a curing process that lasts 6 months (Kundu 2017). The curing process uses green vanilla beans from plants 6-8 months old that are produced by hand pollination (Kundu 2017) and yields one kg of natural vanillin per 5,000 kg of vanilla orchid (Gallage et al. 2014). This process covers only 0.25% of the annual global market of vanillin (Brochado et al. 2010). The rest of the market demand is met by producing synthetic vanillin from lignin (Fache et al. 2016) and fossil fuels by chemical synthesis, sold for 0.3% of the price of natural vanillin (Ni et al. 2015). Alternatively, heterologous production of vanillin using microorganisms can serve as a solution.

In the case of microorganism NPs, the growth rate and seasonal constraints are not an issue, but not all microbes are culturable in laboratory conditions. Microorganisms generally present elements involved in NPs production in

biosynthetic gene clusters (BGCs) (Zhang et al. 2019). BGCs can be transferred to heterologous hosts for their characterization without the need to develop new genetic tools for each pathway (Zhang et al. 2019).

The discovery of the enzymes involved in biosynthetic pathways of the desired plant and microorganisms NPs is the basis for heterologous production. Currently, many NP biosynthetic pathways still need to be fully elucidated (Zhu et al. 2021). Developing *in vivo* systems to functional characterize novel enzymes bypasses problems associated with protein purification, substrate availability, and *in vitro* analysis of recombinant proteins (Facchini et al. 2012). Upon optimization, heterologous production of natural products accumulation can exceed the content from whole plants (Wawrosch and Zotchev 2021). For instance, developing a yeast cell factory to produce high titers of vanillin was accomplished by doing 24 genetic modifications (Mo and Yuan 2024).

1.1.2. Development of heterologous expression systems

Developing systems for synthetic biology in natural products requires optimized hosts to produce specialized metabolites, libraries of synthetic DNA parts (promoter, ribosome binding sites, terminators, selection markers), and assembly methods (Facchini et al. 2012; Aubry et al. 2019). Before introducing the foreign DNA into the host organisms, it is assembled (cloned) in a vector, which can be replicative when it stays as independent chromosomes or integrative in the case where it inserts into the host's genome (Pouresmaeil and Azizi-Dargahlou 2023). Elements of the vector will influence the synthesis and accumulation of the recombinant enzyme.

Replicative vectors such as expression plasmids and artificial

chromosomes typically contain an origin of replication, a promoter, affinity tags, coding sequences, multiple cloning site or MCS (in traditional cloning), a terminator sequence, and a selectable marker gene (Georgiou and Segatori 2005) (Figure 1. 2). A factor in selecting a vector for protein expression is the copy number of the plasmid controlled by the origin of replication. Promoters and terminators are regulatory sequences involved in the transcription. Promoters control the affinity and conditions to which the RNA polymerase binds to DNA and can be regulated using inhibitors or activators in inducible systems. Terminator sequences will promote the detachment of the RNA polymerase to terminate the transcription; they have regulatory roles in mRNA stability, silencing, and tuning other transcription functions (Kocaoglan et al. 2024). In integrative vectors, the GOI is typically flanked by homology regions to the host genome that will be used to transfer the genes by recombination (Taton et al. 2014).

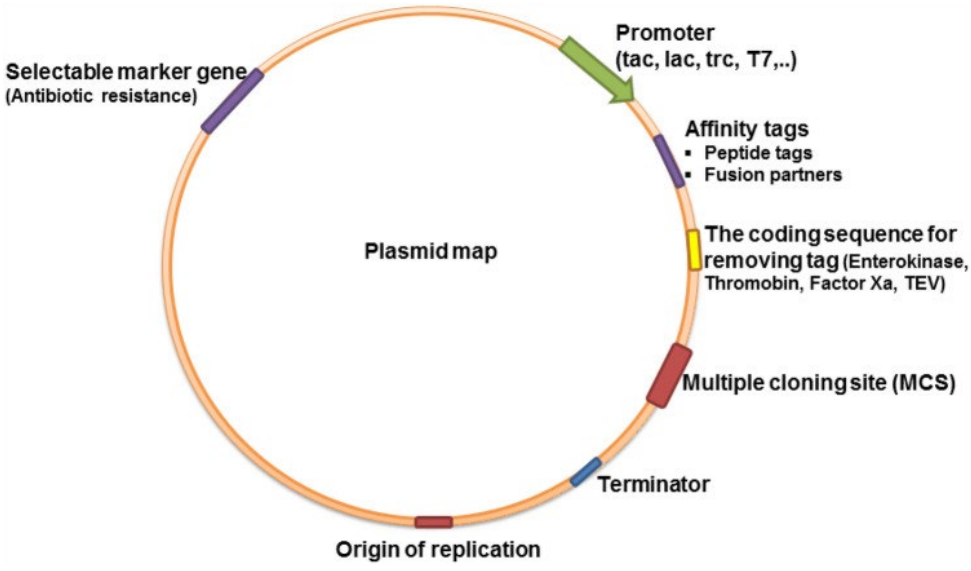


Figure 1. 2. Important elements of expression plasmid in bacteria. The distance between different elements is not to scale. Source: Pouresmaeil and Azizi-Dargahlou, 2023.

Traditional approaches for molecular cloning involve using type II endonucleases (restriction enzymes) to insert the GOI in the MCS by incorporating restriction sites flanking the GOI (Hillson 2011). Inserting genes through this approach is less favorable when we must insert multiple genes into the same vector (Hillson 2011). Standardized assembly methods to overcome this difficulty include BioBrick, SLIC, Gibson, CPEC, and Golden Gate (Li and Elledge 2007; Engler et al. 2008; Shetty et al. 2008; Gibson et al. 2009; Quan and Tian 2009). They are easier to use in parallel and automate than traditional cloning. Efforts have also been made to facilitate the assembly of multiple elements in a vector, creating standards for their architecture that make cloning easier. An example is Joint Universal Modular Plasmids (JUMP), a vector standard to improve the flexibility of current modular cloning systems, which is compatible with BioBricks and PhytoBricks with backbones based on the Standard European Vector architecture (Valenzuela-Ortega and French 2021).

The regulatory elements that work in the main expression systems *of E. coli* and *S. cerevisiae* cannot be shared with other species in most cases since they tend to be strain-specific. Even when they work in a broad host range, their efficiency will likely vary between strains (Taton et al. 2014). This means that only a few elements are available for heterologous production in other hosts. The shortage of well characterized elements results in repeating sequences when multiple expression cassettes are necessary. However, this can reduce plasmid stability and increase the frequency of homology dependent gene silencing (Kocaoglan et al. 2024). Therefore, the availability of characterized elements is a crucial factor in choosing the expression host for heterologous production.

Other factors to consider for the choice of the host include the ones that would affect the expression of foreign DNA, such as the size of the protein, codon usage bias (CUB), protein or metabolite toxicity to the cell, mRNA secondary structure, the protein structure such as the presence of transmembrane domains and post-translational modifications (Kaur et al. 2018). Some of these factors contribute to the metabolic burden that represents producing the heterologous product to the host. An explanation of the term metabolic burden is presented in Figure 1. 3.

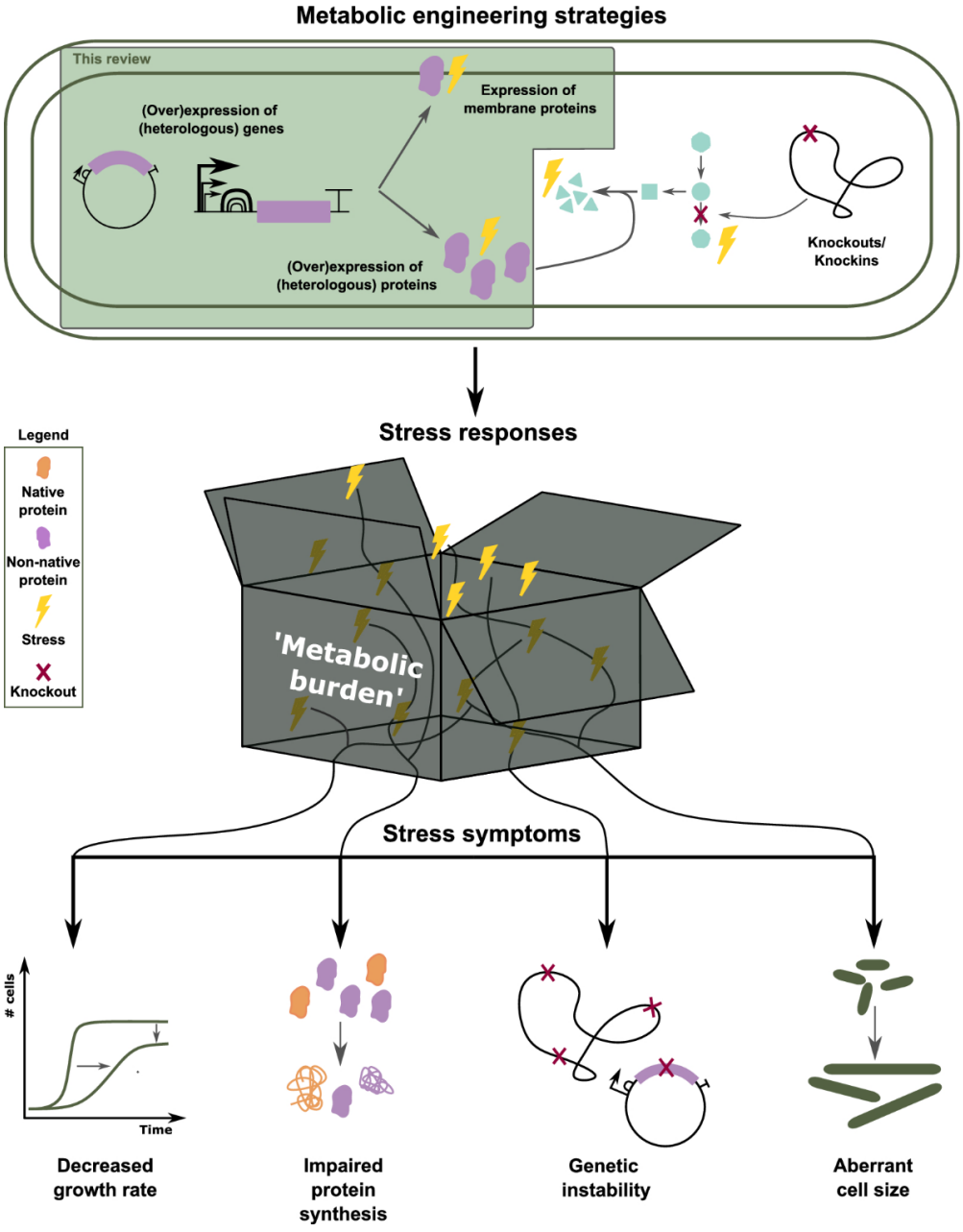


Figure 1. 3. Schematic of metabolic burden. Different metabolic engineering strategies often render stressed cells (top). This translates into different stress symptoms observed in the cells (bottom). However, the nature of the stress responses, their interconnections, and their translation into stress symptoms remains a black box (middle). This box is often summarized as “metabolic burden” without specifying the associated responses. Source: Snoeck *et al.*, 2024.

1.1.3. Impact of heterologous production on the host organism

Unlike cell toxicity that needs to be addressed for certain proteins and metabolites, metabolic burden represents a universal challenge in the synthesis of heterologous proteins since it sequesters resources of the host cells and can be presented as a slower growth rate, impaired protein synthesis, genetic instability or changes in cell size (Kastberg et al. 2022). In traditional approaches, recombinant protein levels are kept as high as possible, with the aim of maximizing the yield (Snoeck et al. 2024). These high expression levels can activate multiple stress responses. Understanding the consequences of heterologous protein synthesis is needed to alleviate or prevent stress in engineered cells (Snoeck et al. 2024). Regulation of plasmid copy number and promoter strength bypasses the problem of the burden of producing the recombinant proteins to the host metabolism (Pouresmaeil and Azizi-Dargahlou 2023).

CUB refers to the preference in the microorganism for a specific pattern of codons for an amino acid that associates with the tRNA pool (Pouresmaeil and Azizi-Dargahlou 2023). For example, *E. coli* does not synthesize all tRNAs found in plants. Therefore, researchers change the codons in the

GOI for the preferred codons in the host organism. Alternatively, strains can be optimized to translate all codons, as previously done for *E. coli* (Pouresmaeil and Azizi-Dargahlou 2023). CUB will also contribute to the translation speed and increase the metabolic burden because of incorrect protein folding due to fast translation, increasing the error rate and, in turn, triggering the cell stress response (Drummond and Wilke 2008; Yang et al. 2014). To avoid this, less optimal codons, for which tRNAs are still present in the cell but in lower abundance, can help correct protein folding (Mellitzer et al. 2012; Yang et al. 2014). Changing codons will also impact the mRNA secondary structure. Similar to CUB, the mRNA secondary structure influences the speed of translation and the half-life of the mRNA. The 5' end of the mRNA strongly influences translation initiation because it contains the ribosome binding site (RBS) (Snoeck et al. 2024).

Another way to decrease the metabolic burden is by uncoupling the synthetic system from the host by designing circuits that operate independently, orthologous to the host metabolism (Adams 2016). An example is co-expressing the T7 bacteriophage RNA polymerase and the GOI under its promoter in *E. coli* strains, dedicating an RNA polymerase exclusively to the GOI expression (Durbin 1999).

1.2. Main host organisms for heterologous production

In addition to all the factors mentioned above, in choosing the host for heterologous production, aspects such as ease of cultivation and genetic manipulation, resistance to stress, and substrate utilization need to be considered (Porro et al. 2011). The advantages of *E. coli* and *S. cerevisiae* include the availability of large libraries of genetic parts to engineer cells

and a deep understanding of cell biology. They are, therefore, considered the main chassis for heterologous production (Adams 2016). Chassis selection for heterologous production can be genetically optimized to achieve functional expression of enzymes, supply enough precursors and cofactors, enhance product transport, etc. Synthetic biology and emerging technologies such as next-generation sequencing, functional genomics, genome editing, and gene circuits provide new ways to engineer chassis to produce NPs (Liu et al. 2020).

1.2.1. E. coli and other pioneer bacterial hosts.

E. coli is the most extensively studied Gram-negative bacteria and used expression system (Pouresmaeil and Azizi-Dargahlou 2023). Different strains of *E. coli* have been genetically engineered to acquire features such as higher mRNA stability, ability to form disulfide bonds in the cytoplasm, absence of proteases, production of non-canonical tRNAs, harboring stable plasmids, and stringent control of heterologous expression (Terpe 2006; Baeshen et al. 2015; Hayat et al. 2018). Routine protein expression is most frequently done in *E. coli* BL21 and K12 and their derivative strains (Terpe 2006).

Many molecular tools are available for high level expression of heterologous proteins in *E. coli*. It has been used to produce low and high molecular weight proteins with desired yields (Gupta and Shukla 2016). Some examples include simultaneous manipulation of genes across different chromosomal loci and the establishment of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) Prime Editing toolkit for E. coli (Tong et al. 2021). In addition, multienzyme clustering to enrich local

enzyme concentrations using transcription activator-like effector (TALE) fused enzymes accelerated the heterologous metabolic system to produce indole-3-acetic-acid in this bacterial system (Zhu et al. 2016). Improved cloning pipelines using modular assembly are available for pathway reconstruction. ePathBrick vectors compatible with BioBrick standards allow fine-tunning gene expression in *E. coli* by integrating multiple transcriptional activation or repression signals into the operator region. They support modular assembly of parts, and a seven genes pathway has been successfully assembled in a single vector (Xu et al. 2012).

However, the production of recombinant proteins in *E. coli* still has some challenges and limitations. Producing therapeutic proteins requires extra steps of purification because it accumulates the lipopolysaccharide endotoxin, which triggers immune response in humans and other mammals (Petsch and Anspach 2000). High levels of recombinant protein synthesis often result in accumulating insoluble aggregates of misfolded proteins in the cytoplasm as inclusion bodies (Kaur et al. 2018). Proteins in inclusion bodies do not have biological activity and thus account for 15-25% less yield of bioactive protein (Kaur et al. 2018). The formation of inclusion bodies is related to strongly induced promoters; therefore, inducer concentration, such as Isopropyl β-D-1-thiogalactopyranoside (IPTG), can be optimized to prevent it (Baneyx and Mujacic 2004). Optimization of culture conditions has proven effective in preventing the formation of inclusion bodies (Sørensen and Mortensen 2005). Correct folding can be promoted by fusing the target protein with an endogenous protein that can be removed after purification to affinity tags or by plasmid display (Figure 1. 4) (Tolia and Joshua-Tor 2006).

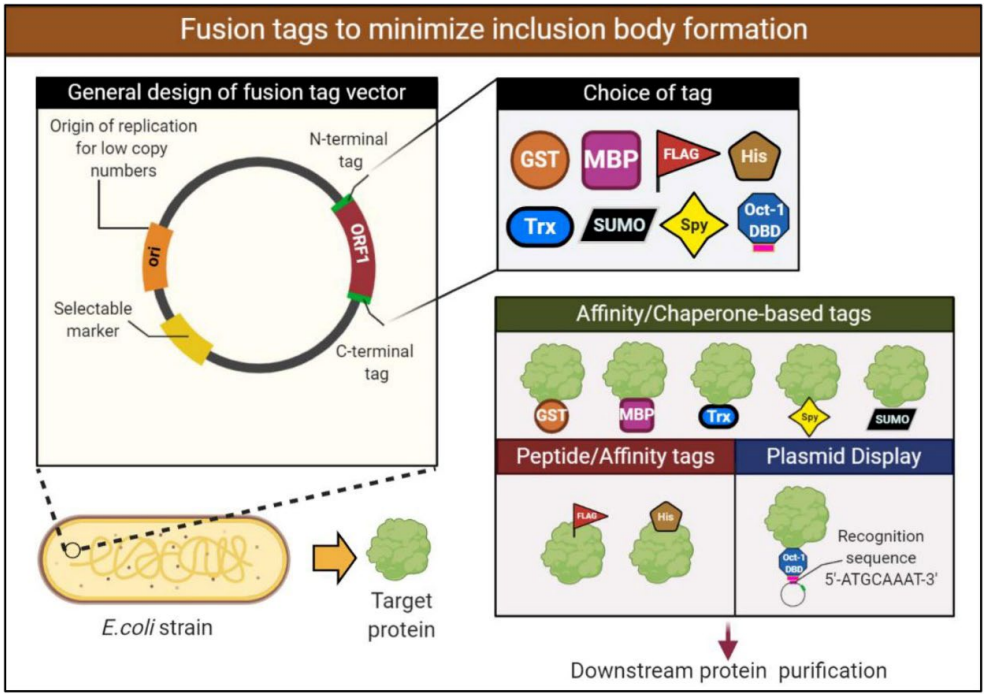


Figure 1. 4. Design of a fusion expression vector to solubilize the expressed recombinant protein. Examples of protein fusion tags include glutathione-S-transferase (GST) and mannose binding protein (MBP), while peptide tags include commonly utilized histidine (His). Plasmid display system technology uses fusion partners to attach the fusion and target protein to the plasmid to improve the stability of the expressed protein. For example, the transcription factor Oct-1 is utilized in the following figure, which possesses a DNA binding domain that recognizes and attaches to the recognition sequence. Source: Bhatwa *et al.*, 2021

However, complex natural products are barely produced in *E. coli* without in-depth strain engineering (Liu et al. 2022). Limitations in *E. coli* production have encouraged the investigation of other bacterial expression systems. *Bacillus subtilis* natural competence, easy DNA integration into chromosomes, natural two components' systems, protein secretion system, and quorum-sensing systems that can be used as biosensors are attractive

features to use this Gram-positive bacteria as an expression system (Adams 2016). In addition, it has the “Generally Recognized as Safe” (GRAS) certification from the FDA. *B. subtilis* can use a wide range of carbon sources and has been used to produce complex proteins and natural products (Gu et al. 2018; Kaspar et al. 2019). Stable transformation of *B. subtilis* has been demonstrated (Zhang and Zhang 2011; Vojcic et al. 2012; Wozniak and Simmons 2022)**.** These characteristics converted it into another pioneer bacterial expression system.

Streptomyces spp. is another Gram-positive bacterium that can produce bioactive NPs. It is excellent at providing precursor pools and has the potential to characterize metabolic pathways (Alam et al. 2022). Refactoring BGCs in monocistronic units resulted in similar levels of heterologous protein production compared to the natural host, and the system was validated using the BGCs for landomycin and mithramycin aromatic polyketides in the host strain of *Streptomyces coelicolor* M1146 (Javorova et al. 2024).

Unfortunately, *Streptomyces* spp. requires longer periods to generate genetically engineered strains than *E. coli* or *S. cerevisiae*. Also, *Streptomyces* has the limitation of exhibiting weak DNA homologous recombination (Zhao et al. 2020). Therefore, several genome editing technologies have been developed, especially CRISPR/Cas-based tools (Lee et al. 2024). The CRISPR/Cas system reduced the time and effort needed for screening compared to conventional double-crossover events (Lee et al. 2024). This system has expanded and can now be used in more than six *Streptomyces* species, including *S. lividans, S. viridochromogenes*, *S. albus*, *S. rimosus, S. ambofaciens*, and *S.*

roseosporus (Cobb et al. 2015; Jia et al. 2017; Li et al. 2018; Najah et al. 2019; Yeo et al. 2019; Jiang et al. 2021).

1.2.2. S. cerevisiae and non-conventional yeast

Bacteria lack of some post-translational modifications and intracellular membranes led to the development of eukaryotic expression systems for heterologous production. *As a heterologous production platform, S. cerevisiae (budding yeast) offers several advantages*. It has bacteria's advantages of high growth rates and easy cultivation but with the capacity to produce complex proteins since they can have typical eukaryotic protein processing (folding, assembly, post-translational modification) (Branduardi *et al.*, 2004). It has the GRAS certification from the U.S. FDA, a simple genome structure, and the ability to grow as a haploid organism (Schindler 2020). Standardized transformation methods in *S. cerevisiae* that can be automated make it an appropriate host for heterologous production (Nora et al. 2019). In addition, secretion in yeast allows disulfide bond formation by passing by the endoplasmic reticulum (ER)-Golgi secretory pathway, proteolytic maturation, *N*- and *O*- linked glycosylation and other post-translational modifications (Tang et al. 2016).

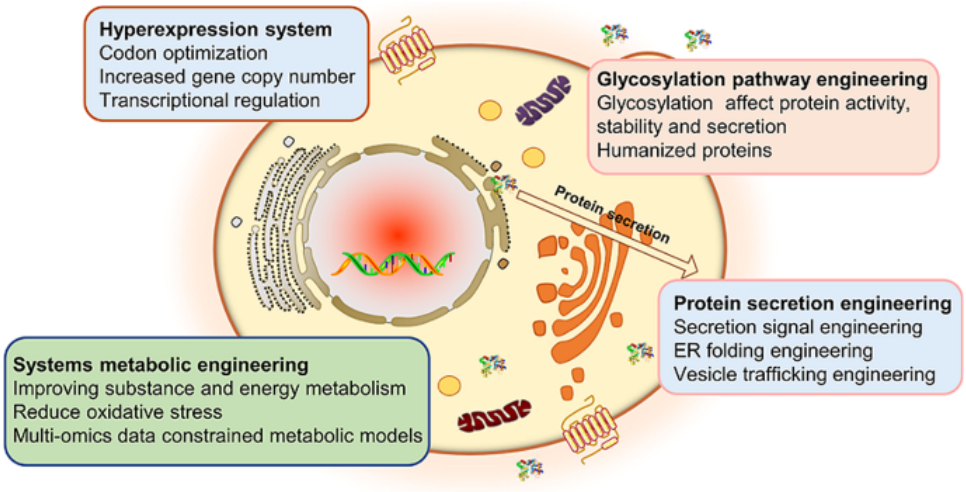


Figure 1. 5. A review of engineering strategies for improved protein production by *S. cerevisiae*, including constructing a hyperexpression system, secretion engineering, glycosylation pathway engineering, and systems metabolic engineering. Source: Zhao *et al.*, 2024

However, heterologous production in *S. cerevisiae* has limited product yield and plasmid stability, difficulties scaling up, hyper-glycosylation, and low secretion capabilities (Kastberg et al. 2022). *N*-glycan hyper-mannosylation renders some glycoproteins nonbioactive (Jung and Kim 2018). Engineering strategies to overcome limitations of heterologous protein production in yeast are summarized in Figure 1. 5. Yeast cell factories can be used to produce some glycosylated proteins when adding downstream processing steps or by generating strains with more humanized glycosylation patterns (Liu et al. 2018). Other strategies to overcome the limitations of the *S. cerevisiae* system are hyperexpression systems to increase protein production, metabolic engineering, and improvement of protein secretion (Kim et al. 2013; Zhao et al. 2024). The use of yeast secretion signal sequences, gene integration in its genome, and mutation in specific host cell genes proved to increase the secretion of calf prochymosin

as a fully activable zymogen (Smith et al. 1985). Display-enhancing genes from *S. cerevisiae* increased the secretion of active single-chain T-cell receptors and single-chain antibodies (Wentz and Shusta 2007).

Work has been designated to characterize alternative yeast systems named “non-conventional” yeast for heterologous production. Examples of this group of yeast are *Schizosaccharomyces pombe*, methylotrophs like *Pichia pastoris*, and alkane-utilizer *Yarrowia lipolytica* (Madzak et al. 2004). Translational studies from *S. cerevisiae* beneficiated the development of non-conventional yeast systems. Besides having advantages in yield, they have reduced hyper-glycosylation and secretion efficiency, even for high molecular weight proteins (Madzak et al. 2004). *P. pastoris* is a highly successful expression system for producing a variety of heterologous proteins using methanol as a carbon source (Karbalaei et al. 2020). However, overexpression of heterologous proteins can lead to the accumulation of unfolded proteins (Ingram et al. 2021). Another yeast under investigation due to its exceptional resistance to environmental stresses is *Zygosaccharomyces bailii*, which is evolutionary close to *S. cerevisiae* (Branduardi *et al.*, 2004). It can tolerate high sugar concentrations, acidic environments, high temperatures, and survive the presence of high concentrations of chemical preservatives. Simplifying the fermentation process by surviving restrictive conditions and preventing bacterial contaminations (Branduardi et al. 2004b). Despite the success of heterologous production in yeast, challenges remain in this system. For instance, yeast cells easily adapt to changing environmental conditions and thus may down-regulate protein production to match intracellular demands (Gasch 2003).

1.3. Emerging heterologous host for characterization of NPs biosynthetic pathways.

Conventional microbial expression systems (i.e., E. coli, S. cerevisiae, and P. pastoris) exhibit significant drawbacks mainly related to protein processing, solubility, and production costs (Brasil et al. 2017). Identifying challenges in the model organisms has promoted the search for novel heterologous hosts to exploit the diversity of organisms and develop new expression systems (Branduardi et al. 2004). Extensive manipulation is needed for *E. coli* and *S. cerevisiae* to produce certain heterologous products, and even though we know more about them, the potential of more suitable naturally resistant hosts would require fewer bioengineering efforts.

1.3.1. Filamentous fungi

Filamentous fungi, mainly Trichoderma and Aspergillus spp., have been investigated for heterologous production of industrial enzymes and pharmaceuticals due to their high levels of protein synthesis. *T. reesei* has been studied for its capability to produce biomass degrading enzymes already commercialized and applied in biofuel plants (Bischof et al. 2016). Filamentous fungi can be grown in large-scale fermentation in small clumps of swollen filaments; this culture morphology was shown to produce the highest yield of citric acid (Demain 2006). Modifying culture conditions (low pH, high aeration, use of additives, pulse feeding the carbon source) or genetic manipulation of strains can result in this desirable morphology (Meyer et al. 2021). While genome sequence is available for several species of *Aspergillus* and of *Trichoderma*, advances in bioengineering filamentous fungi have been done mainly in *Aspergillus* species (Baker 2006; Soanes

et al. 2008; Nora et al. 2019). *A. nidulans* is a model species of filamentous fungi and has been used for the study of gene clusters of other species (Alberti et al. 2017). The development of a smart system for selectable marker recycling in this host has been done to allow the stepwise transfer of 13 genes of *A. terreus* to *A. nidulans* for the heterologous biosynthesis of the NP geodin (Nielsen et al. 2013). Fisch *et al*. reprogrammed polyketide synthases by doing rational domain swaps between polyketide synthases that synthesize closely related compounds and expressed them in *A. oryzae* (Fisch et al. 2011).

Genetic manipulation methods to transform filamentous fungi use polyethylene glycol (PEG)-mediated protocols but mainly rely on *Agrobacterium tumefaciens*-mediated transformation since it presents higher transformation frequencies and allows for multiple or single-copy gene integrations (Storms et al. 2005; Steiger et al. 2011; Jørgensen et al. 2014; Arentshorst et al. 2015; Hooykaas et al. 2018). Constitutive promoters have been characterized for the expression of heterologous proteins in *Aspergillus*, as well as an inducible system using pH changes in *A. niger* (Hamer and Timberlake 1987; Gressler et al. 2015; Yin et al. 2017).

Improvements in heterologous protein production should focus on characterizing strong promoters, translation efficiency, glycosylation, and protein folding (Wakai et al. 2017). The latter can be improved by the overexpression of chaperones and the addition of fused signals (Wakai et al. 2017). The fusion of highly expressed endogenous proteins using five different linker peptides were tested in *A. niger* to develop strategies to improve the efficiency of heterologous expression in this system (Wu et al. 2023).

1.3.2. Cyanobacteria and eukaryotic microalgae

Microalgae are another emerging platform for genetic engineering and potential solutions for environmentally sustainable manufacturing (Butler et al. 2020). They consist of a broad group of photosynthetic microorganisms, including cyanobacteria (prokaryotic), green algae, and diatoms, and are found in different aquatic ecosystems of sea and freshwater (Dehghani et al. 2020). They produce NPs such as carotenoids (pigments), vitamins, and polyunsaturated fatty acids (Kaye et al. 2015). Besides presenting advantages common to the main expression systems, i.e., higher growth rates than multicellular systems, they can use atmospheric CO2 as a carbon source, therefore having low and cheap requirements to grow. Eukaryotic microalgae species are attractive for therapeutic protein engineering since they can glycosylate proteins with patterns closer to humans. Thus, glycoengineering may require less effort compared to yeast hosts (Banerjee and Ward 2022). Sequencing new algal genomes and developing genetic tools opened the path for engineering them as platforms for heterologous production (Brasil et al. 2017). Some species require relatively short times to be bioengineered for heterologous production and have been granted the “GRAS” certification by the U.S. FDA (Su et al. 2023). Even though microalgae-based recombinant proteins are not yet available in the market, they have been envisioned as promising production platforms of biotechnological drugs in the near future (Dehghani et al. 2020).

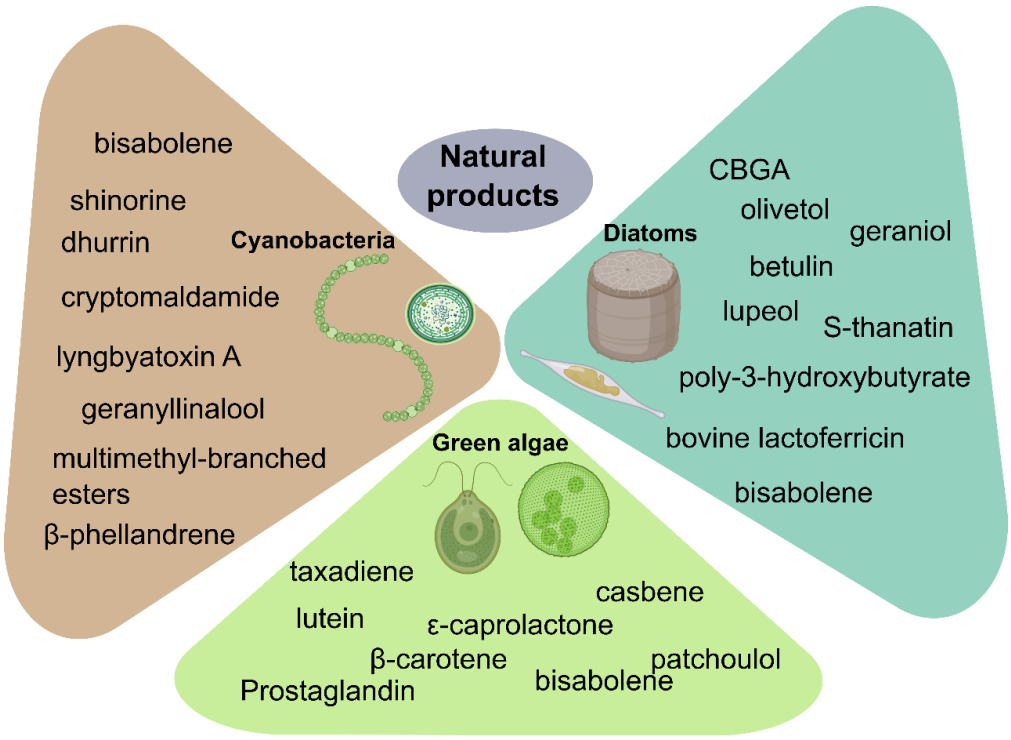


Figure 1. 6. Heterologous synthesis of natural products in microalgae. Metabolites produced in cyanobacteria (Chaves et al. 2016; Formighieri and Melis 2016, 2017; Vavitsas et al. 2017; Betterle and Melis 2018; Roulet et al. 2018; Yang et al. 2018a; Sebesta and Peebles 2020; Taton et al. 2020), green algae (Lauersen et al. 2016, 2018; Wichmann et al. 2018; Rathod et al. 2020; Maeda et al. 2021; Perozeni et al. 2023), and diatoms (D'Adamo et al. 2019; Fabris et al. 2020; Tanaka et al. 2021; Awwad et al. 2023; Baiden et al. 2023; Fantino et al. 2024; Windhagauer et al. 2024). Created in Biorender.com.

Cyanobacteria metabolic versatility to produce different metabolites, their fast growth rates, salt tolerance, ability to use waste streams as nutrients, together with the fact that they can be genetically engineered faster than plant expression systems, make them interesting autotroph hosts for heterologous expression (Gordon and Pfleger 2018; Dhakal et al. 2021). *Synechocystis* sp. strain PCC 6803 use as model organism to study

photosynthesis started decades ago (Ikeuchi and Tabata 2001). Likewise, *Synechococcus* sp. strain PCC 7942 has been extensively studied for circadian clock model, and strain *Synechococcus* sp. PCC 7002 short doubling times and halotolerance have attracted attention for its genetic manipulation (Bernstein et al. 2016). Heterologous genes have been inserted in what are known as neutral sites in the chromosomes of several strains of cyanobacteria, including *Synechococcus* spp., without affecting their phenotype (Taton et al. 2014). Additionally, expression vectors for integration of genes in *S. elongatus* genome have been developed SyneBrick, with inducible expression systems including IPTG and TetR-regulated gene expression (Kim et al. 2017).

Production of NPs has been accomplished in cyanobacteria (Figure 1. 6). Heterologous production of monoterpene hydrocarbons in *Synechocystis* was accomplished by co-expression of β-phellandrene synthase fused to the endogenous protein cyanobacteria phycocyanin β-subunit with enzymes from the mevalonic acid pathway and geranyl-diphosphate synthase, increasing the carbon flux towards the terpenoid pathway (Formighieri and Melis 2016). Additionally, *S. elongatus* was engineered for heterologous production of multimethyl-branched esters by expression of the the mycobacterial PKS-based mycocerosic biosynthetic pathway (Roulet et al. 2018).

Cyanobacteria bioengineering major drawbacks are the oligoploidy or polyploidy, presence of restriction/modification systems, and their circadian rhythms (Liu et al. 1995; Elhai et al. 1997; Griese et al. 2011). They comprise a group highly diverse microorganisms in genome content, size, and CUB, affecting their ability to express the inserted GOI (Taton et al.

2014). Adaptation of protocols for genetic engineering are necessary since they also differ in their natural ability to undergo transformation.

Chlamydomonas reinhardtii is a model green alga capable of sexual reproduction, easily transformed nuclear, chloroplast, and mitochondrial genomes (Perozeni et al. 2023). It has recently received the “GRAS” certification by the U.S. FDA for food applications (Perozeni et al. 2023). *C. reinhardtii* has been extensively researched for heterologous production (Figure 1. 6). Some examples include the expression of zeolin targeting and accumulating in *C. reinhardtii* ER, systematic investigation of bottlenecks in astaxanthin accumulation, expression of the phytoene-β-carotene synthase from *Xanthophyllomyces dendrorhous* increasing the content of β-carotene and lutein in the green alga, and production of high levels of functional human growth hormone (Wannathong et al. 2016; Rathod et al. 2020; Amendola et al. 2023; Perozeni et al. 2023).

High GC content (65%) in *C. reinhardtii* genome and high CUB require most heterologous genes to be codon optimized for its expression in the algae (Tran and Kaldenhoff 2020). Random integration in its nuclear genome generates DNA deletions in the inserted region and causes low heterologous protein production (Harris 2001). Strategies to overcome gene silencing consist in the use of endogenous ribulose bisphosphate carboxylase oxygenase (RuBisCo) introns (Tran and Kaldenhoff 2020). In addition, transformation of the chloroplast genome represents a better alternative, it is possible to do homologous recombination, which simplifies genetic modification strategies and can accumulate more proteins (Engel et al. 2015; Arias et al. 2023). Several studies have identified the use of untranslated regions (UTRs) for heterologous gene expression in the

chloroplast of *C. reinhardtii* as crucial (Sun et al. 2003; Yang et al. 2006; He et al. 2007). Also, modifying the medium by adding diselenide selenocystamine increased the accumulation of proteins with disulfide bonds (Ferreira-Camargo et al. 2015).

Major bottlenecks for microalga production are cost and limitations of scale production, since high growth rate and productivity at lab scale is difficult to replicate at larger scale (Su et al. 2023). Variability of outdoor environments in terms of temperature and light intensities decreases algae productivity (Su et al. 2023).

Diatoms, a diverse group of stramenopiles, can have a fast growth rate and short generation times, and they are able to adapt quickly to fluctuating environmental conditions such as low iron availability (Leterme 2015). Therefore, diatoms have attracted interest for their use as green biofactories. They have adapted to many environments colonizing seas, freshwaters, and terrestrial habitats (Kuwata and Jewson 2015). Diatoms are important in ocean ecology since they are responsible for over 20% of global production of oxygen (Kuczyńska et al. 2024). They are distantly related to green algae and land plants and evolved by secondary endosymbiosis of a red alga and a heterotrophic protist (de Grahl and Reumann 2021). Model organisms from this group include the centric diatom *Thalassiosira pseudonana* and the pennate diatom *Phaeodactylum tricornutum*, whose genomes were the first sequenced from this group (Armbrust et al. 2004; Bowler et al. 2008). The latest has a genome of 27.4 Mb, which relatively smaller than *T. pseudonana* genome (32.4 Mb), however more than half of the genes seem to be not present in other eukaryots (Armbrust et al. 2004; Bowler et al. 2008). In addition telomere to

telomere assembly revealed that *P. tricornutum* has 25 chromosomes similar to the model centric diatom with 24 (Armbrust et al. 2004; Giguere et al. 2022).

1.3.2.1. Phaeodactylum tricornutum host for heterologous production.

P. tricornutum is not a common pennate diatom since it does not require silica to grow with a frustule slightly or not silicified. This feature gives plasticity to the cell wall, escaping from successive cell size reduction triggering sexual reproduction in most diatoms (Bartual et al. 2011). Its marked pleomorphism is one of the most distinctive characteristics of the diatom, and it may confer them adaptation to various environments (Ovide et al. 2018). *P. tricornutum* occurs naturally in more than one morphotype: oval, fusiform, and triradiate (Figure 1. 7), and to a lower extend in cruciform (Galas et al. 2021). Fusiform (Figure 1. 7a) is the more frequent morphotype with poorly silicified cell wall making it easier for transformation. Oval is the only morphotype that possess a full silicified frustule (Figure 1. 7c), it has been regarded as a response to environmental stressors, and it has been particularly interesting for the secretion of proteins based on transcriptomics and proteomic studies (Chuberre et al. 2022).

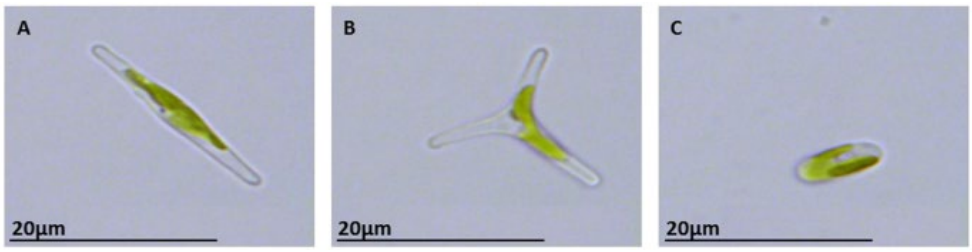


Figure 1. 7. Micrographs of *P. tricornutum*'s morphotypes. (**a–c**) Micrographs obtained by light microscopy of *P. tricornutum* cells alive. (**a**)

fusiform morphotype; (**b**) triradiate morphotype and (**c**) Oval morphotype. Source: Ovide *et al.*, 2018

P. tricornutum potential for industrial applications has been evaluated (Sharma et al. 2021). Biomass can be valorized to enhance the economic viability of biorefineries using *P. tricornutum*, with this purpose Branco-Vieira and colleagues biochemically characterized the biomass and scaled up the production. Biomass percentage of proteins, lipids, fucoxanthin, and biosilica could be valorized to promote circular economy (Branco-Vieira et al. 2020).

In addition, to have economically competitive heterologous production in diatoms, high expression rates are required (de Grahl and Reumann 2021). Protocols for efficient transformation have been developed for diatoms (Doron et al. 2016). Transformation methods in *P. tricornutum* include particle bombardment, electroporation, and more recently, bacterial conjugation using *E. coli* (Seo et al. 2015a; Windhagauer et al. 2021).

The development of transformation methods enabled the characterization of constitutive and inducible endogenous promoters in diatoms. For example, Erdene-Ochir *et al.* (2019) identified the most abundant secreted proteins and characterized the promoter of a highly abundant secreted protein 1 (HASP1) and its secretion signal. Reference genes for quantifying gene expression have been identified, and a series of diatom expression vectors based on Gateway cloning for high-throughput protein tagging and overexpression studies are available for *P. tricornutum* (Siaut et al. 2007).

Gene editing tools have been used in the diatom, including TALE nucleases mediated gene knockouts and CRISPR/Cas9, as well as gene silencing

using RNA interference (RNAi) (Bielinski et al. 2017; Serif et al. 2017a). Gene knock-down of *ligIV* enabled specific integration of DNA at desired loci providing a new way to specifically alter *P. tricornutum* genome (Angstenberger et al. 2019). Also, a library of genetic elements for standardized modular cloning (MoClo) for combination of multiple transcriptional units in a single vector, is available for *P. tricornutum* (Patron et al. 2015). In addition, a PCR-based and a precloned assembly strategy for cloning the *P. tricornutum* chloroplast genome was recently developed (Walker et al. 2024).

The characterization of endogenous promoters and molecular tools allowed for its use as a cell factory for the production of eicosapentaenoic acid, fucoxanthin, neutral lipids, and crysolaminarin, endogenously produced by the diatom (Butler et al. 2020). Heterologous production of recombinant proteins and natural products has been done in *P. tricornutum* (Figure 1. 6). A fully-assembled functional antibody against the Hepatitis B Virus surface protein, and the antigen have been produced in this diatom (Hempel et al. 2011).

1.4. *Phaeodactylum tricornutum* episomal expression system

One of the aspects that differentiates *P. tricornutum* from other algal heterologous platforms is the development of extrachromosomal (episomal) replication of expression vectors (Windhagauer et al. 2021). In 2015, Karas and colleagues developed a conjugation based transformation method for *P. tricornutum*, which was more efficient that PEG, electroporation, biolistic (routine transformation methods for the diatom) (Karas et al. 2015). Previously, trans-kingdom conjugation had only been done in yeast and

mammalian cells using *E. coli* host harboring pTA-MOB plasmid without origin of transfer (OriT). Karas *et al.* first tested several regions from chromosome 25 from *P. tricornutum* in a plasmid containing the *CEN6-ARSH4-HIS3* (*CAH*) sequence from *S. cerevisiae* to see if they could maintain episomes in the diatom, proven by consecutive transformations (Figure 1. 8). Surprisingly while the endogenous sequences from chromosome 25 increased the number of exconjugants in 2.5-fold they were not responsible for episome maintenance. This led to the investigation of non-diatom sequences in this plasmid and found that the *CAH* sequence from *S. cerevisiae* was able to keep the stable circular episome at ploidy equivalent to native chromosomes. This episome can be as big as 50 kb, therefore, it has potential to harbor full heterologous pathways (Karas et al. 2015).

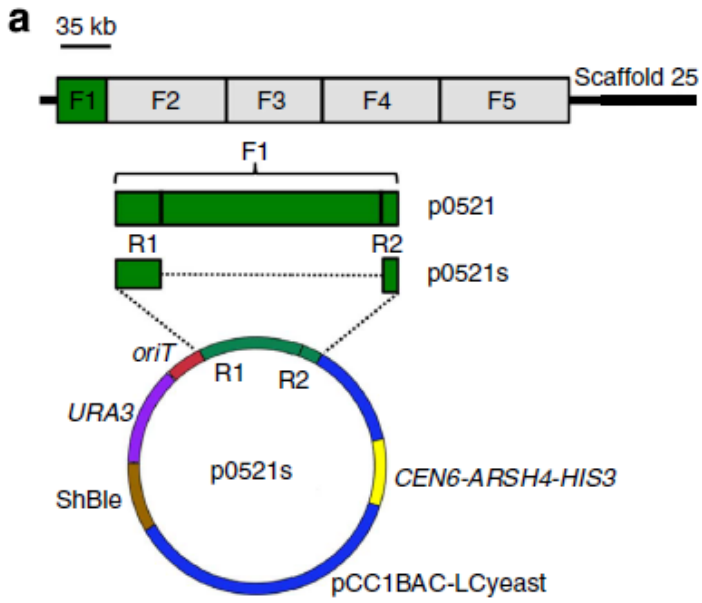


Figure 1. 8. Conjugative transfer of plasmids from *E. coli* to *P. tricornutum*. (a) Map of the plasmids p0521 and p0521s and their derivation from *P. tricornutum* scaffold 25. *OriT*, origin of transfer; *URA3*, gene encoding orotidine 50-phosphate decarboxylase from *S. cerevisiae*,

and *ShBle*, phleomycin resistance cassette with *P. tricornutum* FcpF promoter and FcpA terminator. Source: Karas *et al.*, 2015

Further investigation of the *CAH* sequence by Diner and colleagues led to the identification of the *CEN6* and *ARSH4* as essential for plasmid replication and maintenance, and HIS3 low GC content in the 3' contributed to generating high number of exconjugant colonies (Diner et al. 2016). They also minimized the length of the vector and improved the conjugation protocol with high throughput options that enhanced exconjugant yields, reducing time and resources needed for the conjugation. They identified that low GC content 500 bp sequences are potential indicators of a functional maintenance sequence (Diner et al. 2016). Later on, it was confirmed that the CENH3 *P. tricornutum* centromeric histone protein is recruited by the *CAH* sequence from yeast (Diner et al. 2017). Suggesting that foreign DNA sequences can use native diatom machinery for DNA replication and segregation functioning as a centromere (Diner et al. 2017). This means that *P. tricornutum* is permissive for nuclear gene acquisition.

The establishment of the episomal expression system, with the advantage of reducing possible gene disruptions by random integration, generated a way for determining reproducible measurements of promoter and terminator strengths since the episomes copy numbers seem to be stable at chromosome level, as well as for testing inducible promoters with greater consistency in their control (Diner et al. 2016). On this note, promoter elements with tunable, reversible dose and time-dependent transcriptional levels have been characterized with the aid of this system (Windhagauer et al. 2021; Kassaw et al. 2022; Garza et al. 2023). In addition, delivery of

Cas9 by conjugation with *P. tricornutum* and extrachromosomal expression, minimizes unwanted changes in the genome (Russo et al. 2018; Slattery et al. 2018; Moosburner et al. 2020; Gao et al. 2022; Taparia et al. 2022).

Heterologous production in *P. tricornutum* episomal expression system include the receptor-binding domain of the SARS-CoV-2 spike protein (Slattery et al. 2022) and plant NPs of pharmaceutical interest, i.e., monoterpenoids (Fabris et al. 2020), cannabinoids (Fantino et al. 2024), and cannabinoid precursors (Awwad et al. 2023).

1.5. *P. tricornutum* as potential host organism for production of phenolic compounds

Phenolic compounds are a diverse group of natural products, including phenylpropanoids, and thanks to their bioactivity they are widely used in the food, pharmaceutical, and cosmetic industries (Albuquerque et al. 2021). They are mainly produced by either extraction from the natural source or chemical synthesis, each of them presenting limitations (Sun et al. 2021). Particularly, vanillin has high market demand since it is the major organoleptic compound of the vanilla extract and is of great importance for several industries (Yan et al. 2016). The biosynthesis of vanillin starts from L-phenylalanine in *V. planifolia* involving the enzymes phenylalanine ammonia lyase (PAL), cinnamic acid-4-hydroxylase (C4H), coumaric acid-4-hydroxylase (C3H), caffeic acid *O*-methyltransferase (4-COMT), and vanillin synthase (*Vp*VAN) (Figure 1. 9) (Gallage et al. 2014).

Recently enzymes involved in the biosynthesis of phenyl compounds have been investigated by *in silico* analysis of 47 algae taxa with potential for bioengineering including *P. tricornutum*. Out of the five enzymes that have

been proposed to integrate the pathways to synthetize vanillin, *P. tricornutum* possess homologous genes for *C4H*, *4CL*, and *C3'H*, lacking the first gene (*PAL*) like the majority of the algae, *4-COMT* used to convert to ferulic acid, and *VpVAN* (Del Mondo et al. 2022). The latter has high homology in amino acid sequence to cysteine proteases, and its activity has not been reproduced when expressed in microorganisms (Yang et al. 2017). It has been stated that lack of a good experimental model system limits the characterization of the last step of the vanillin pathway, since the re-differentiated plastid changing environment where vanillin is synthesized may be difficult to recreate *in vivo* (Havkin-Frenkel and Belanger 2018).

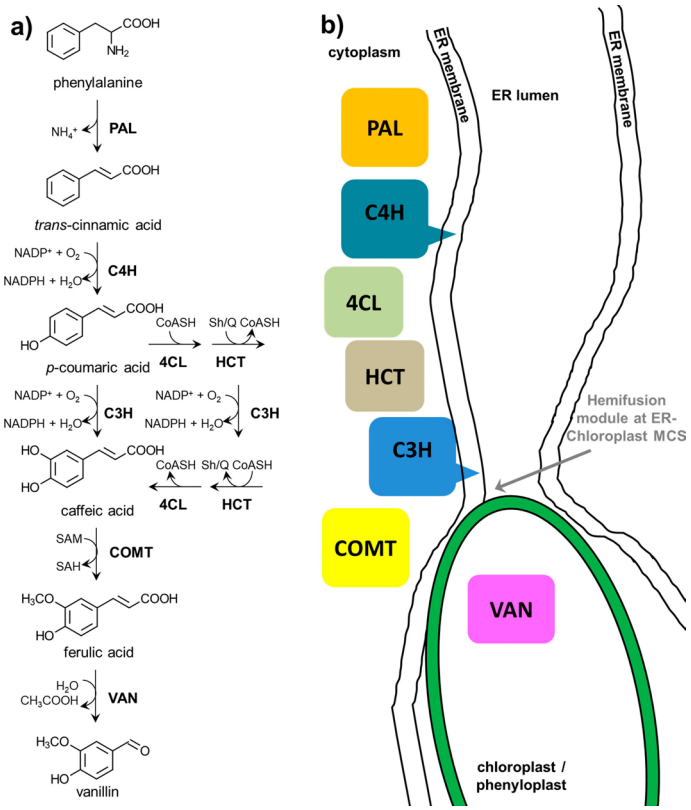


Figure 1. 9. Simplified biosynthetic pathway of vanillin and its potential intra-cellular localization. (**a**) Enzymes proposed by Gallage *et al.* (Gallage et al. 2014, 2018) to convert vanillin from L-phenylalanine. (**b**) Schematic representation of proposed intra-cellular

localization of enzymes involved in vanillin biosynthesis. The basis of the diagram showing the continuity of chloroplast and ER membranes used freeze-fracturing techniques, complementation, and regulation of enzyme activities. Overlapping black and green lines depict the two outer leaflets merging on each side. Source: Diamond *et al.*, 2023.

1.6 Research objectives

Although *P. tricornutum* episomal expression system has had many success stories, few reports underline its limitations. It was highlighted by George *et al.* (2020) that dynamics of episome segregation in a population are not known for *P. tricornutum.* It was hypothesized that it could cause variations in episomal expression cell lines. In a previous report we demonstrated that successful transformation with unaltered expression cassette could lead to *P. tricornutum* exconjugants without producing the recombinant protein (Diamond *et al.*, 2023). *P. tricornutum* episomal expression system was further used to characterize *Vp*VAN subcellular localization providing the enzyme with a specific environment for its activity. This study is described in chapter II. Heterogeneity of phenotype of individuals in a single cell line and low production of recombinant protein led us to study deeper the episomal expression system using a cell line expressing *eGFP* which was exhibiting stable subpopulations is presented in chapter III. We studied plasmid copy number variation, mutations, and, using transcriptomic analyses, candidate genes co-expressed with *eGFP*, to characterize the differences between three subpopulations with distinct fluorescence intensities. Discussion, general conclusion, and perspectives of *P. tricornutum* episomal expression system characterization are

presented in chapter IV. Supplementary information of chapter II and III are presented as annexes A and B, respectively. Further explanation about the clustering methodology that supported the gating strategy for chapter III is in annex C.

In addition, an article which I was co-first author, “Instability of extrachromosomal DNA transformed into the diatom *Phaeodactylum tricornutum*", is presented in annex D. Also in this annex, other articles to which I contributed are included: a review entitled “Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow”, and two research articles.

CHAPTER II: What is true for plants may not be true for *Phaeodactylum tricornutum*: The case of *Vanilla planifolia* vanillin synthase targeted to four subcellular compartments of the diatom.

Authors

Aracely Maribel Diaz-Garza[1], Félix Lavoie-Marchand[1], Natacha Merindol[1], Andrew Diamond[1], and Isabel Desgagné-Penix[1,2,†]

[1] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

[2] Plant Biology Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

[†] Corresponding author Isabel.Desgagne-Penix@uqtr.ca

Author Contributions:

ID-P, AD, and AMD-G conceived the idea of the project; FL-M, NM, and AMD-G. designed methodology and performed experiments; AMD-G wrote the first draft and created the figures; All authors have read and corrected the manuscript.

Résumé

La vanilline est le principal composé organoleptique de l'arôme de vanille. Comme le processus d'extraction à partir de sa source naturelle (*Vanilla planifolia*) est limité par un faible rendement, la demande du marché en vanilline est satisfaite en la produisant par synthèse chimique ou biosynthèse en utilisant des micro-organismes pour convertir différents substrats en vanilline. Notre objectif était d'utiliser la diatomée *Phaeodactylum tricornutum*, un organisme modèle photosynthétique déjà utilisé pour produire des protéines eucaryotes et des composés de grande valeur, pour caractériser la vanilline synthase de *V. planifolia* (*Vp*VAN). Cette enzyme fait l'objet de résultats contradictoires et a été rapportée comme étant localisée dans le plaste des cellules de *V. planifolia*, ce qui pourrait jouer un rôle dans son activité. Dans ce but, nous avons ciblé *Vp*VAN liée par une séquence clivable *in vitro* (Xa) à la protéine fluorescente verte améliorée (eGFP en anglais) dans le cytoplasme, les peroxysomes, le plaste et l'appareil de Golgi. Nous avons utilisé un système d'expression épisomale transformant *P. tricornutum* avec *VpVAN:Xa:eGFP* par conjugaison bactérienne. Les lignées cellulaires positives ont été sélectionnées par fluorescence et caractérisées par séquençage des épisomes, confirmant une réplication stable dans la diatomée, ainsi qu'une localisation subcellulaire dans les compartiments cellulaires ciblés. Des stratégies telles que l'enrichissement des cellules positives à l'eGFP par tri cellulaire, la purification des protéines et l'isolement des organites n'ont pas permis d'obtenir des niveaux détectables de la protéine de fusion. Une meilleure compréhension du système d'expression des diatomées est nécessaire pour compléter la caractérisation de cette enzyme et révéler si

la localisation subcellulaire pourrait être un aspect limitant pour son expression dans les micro-organismes.

Mots-clés : vanilline ; diatomée ; compartimentation des organelles ; ingénierie métabolique; système d'expression épisomal.

Abstract

Vanillin is the major organoleptic compound of the vanilla flavor. Since the extraction process from its natural source (*Vanilla planifolia*) is limited by a low yield, the market demand for vanillin is met by producing it through chemical synthesis or biotechnologically using microorganisms to convert ferulic acid, lignin, and guaiacol into vanillin. Our aim was to use the diatom *Phaeodactylum tricornutum*, a photosynthetic model organism already used to produce eukaryotic proteins and high-value compounds, to characterize *V. planifolia* vanillin synthase (*Vp*VAN). This enzyme is the subject of conflicting results and was reported to be localized in the plastid of *V. planifolia* cells, which may play a role in its activity. For this purpose, we targeted *Vp*VAN linked by an *in vitro* cleavable sequence (Xa) to the enhanced green fluorescent protein (eGFP) to the cytoplasm, peroxisomes, plastid, and Golgi apparatus. We used an episomal expression system transforming *P. tricornutum* with *VpVAN:Xa:eGFP* by bacterial conjugation. Positive cell lines were selected by fluorescence and characterized by sequencing the episomes, confirming stable replication in the diatom, as well as assessing the subcellular localization in the targeted cell compartments. Strategies such as enriching GFP-positive cells by cell sorting, protein purification, and organelle isolation did not yield detectable levels of the fusion protein. A better understanding of the diatom expression system is needed to complete the characterization of this enzyme and reveal whether the subcellular localization could be a missing aspect of *Vp*VAN expression in microorganisms.

Keywords: vanillin; diatom; organelle compartmentalization; metabolic engineering; episomal expression system.

Introduction

Natural vanilla flavor source is the orchid *Vanilla planifolia* native to Mexico and some countries from Central America (Martău et al. 2021). The main constituent of the 200 compounds that make up vanilla flavor is vanillin (4-hydroxy-3-methoxy benzaldehyde) (Anand et al. 2019). The low yield of extraction of vanillin from its natural source, *e. g.* one kg of vanillin requires approximately 500 kg of vanilla pods, has led to the development of alternative production using biotechnology to meet the market demand (Gallage and Møller 2015).

Efforts to elucidate the complete biosynthetic pathway from *V. planifolia* have led to the controversial discovery of vanillin synthase (*Vp*VAN) an enzyme with hydratase/lyase activity, which was described to catalyze the conversion of ferulic acid and its glucoside into vanillin (Gallage et al. 2014). Gallage *et al.* found *Vp*VAN to localize in the inner part of the vanilla pod with high transcripts levels in the inner epidermis (Gallage et al. 2018). Enzymatic activity was tested producing the enzyme in a coupled transcription/translation assay. Transient expression in tobacco and stable expression in barley resulted in the production of vanillyl alcohol glucoside. *Vp*VAN controversy arose from the high sequence similarity to cysteine proteases, and its identical amino acid sequence to a protein previously discovered to participate in the conversion of coumaric acid to 4-hydroxylbenzaldehyde (Yang et al. 2017). In addition, Yang *et al.* demonstrated that the recombinant *Vp*VAN produced *in vitro*, in *Escherichia coli*, yeast or plant systems was unable to synthesize vanillin from ferulic acid. Added to the lack of tissue-specific gene expression, Yang and colleagues concluded that this enzyme cannot by itself directly convert

ferulic acid to vanillin and that the pathway of vanillin biosynthesis in *V. planifolia* was yet to be determined (Yang et al. 2017).

However, heterologous production of *Vp*VAN in other plant systems, which already produce high amounts of ferulic acid and trace amounts of vanillin, such as *Capsicum frutescens* (Chee et al. 2017) and rice (*Oryza sativa* L.) calli (Arya et al. 2022), succeeded to increase the production of vanillin. This reinforced that *Vp*VAN plays a role in the production of vanillin and that there may be missing key elements in expression systems which do not produce it endogenously. Along this line, Gallage *et al.* characterized the subcellular localization of *Vp*VAN demonstrating that it localized in the chloroplast and the re-differentiated chloroplasts named phenyloplast. In addition, they isolated chloroplast and proved their biosynthetic capacity to produce vanillin from phenylalanine (Gallage et al. 2014).

Even though plant systems were successful in expressing *Vp*VAN, expression of this enzyme in microorganisms has not been well established, leading to the use of alternative bacterial pathways. Ni *et al.* generated *E. coli* strains able to *de novo* synthesize vanillin from inexpensive substrates (xylose and glycerol) by introducing five genes encoding for the enzymes tyrosine ammonia-lyase, 4-coumarate 3-hydroxylase, caffeate *O*-methyltransferase, trans-feruloyl-CoA synthase (Fcs), and enoyl-CoA hydratase/aldolase (Ech) (Ni et al. 2015). The high reactivity of the aldehyde functional group of vanillin that causes damage to the cell membrane of conventional host for metabolic engineering (i. e. *Saccharomyces cerevisiae*, and *E. coli*) limits production in microorganism cell factories (Nguyen et al. 2014; Pattrick et al. 2019). Since accumulation of vanillin is toxic, *V. planifolia* cell strategies such as the co-expression of a UDP-

glycosyltransferase has been implemented for its accumulation in yeast (Hansen et al. 2009). Also, in *S. cerevisiae* after strain optimization using a combination of engineering strategies including the blocking of the metabolic pathways responsible for the oxidation and reduction of vanillin, while the strains transformed with the *Fcs* and *Ech* from *Streptomyces*, produced up to 10.4 mg of vanillin per g of carbon source, vanillin was not detected in transgenic strains expressing *Vp*VAN (Xin et al. 2024). A possible explanation is that compartmentalization of its production may be required to synthetize vanillin recreating the environment of phenyloplasts. Therefore, our hypothesis was that *Vp*VAN enzymatic activity depended on its subcellular localization. Thus, localizing the enzyme in the correct cellular compartment would have an impact in the production of vanillin.

Phaeodactylum tricornutum, a diatom which possesses a plastid, the organelle where vanillin is stored and produced in the vanilla pod (Gallage et al. 2018), is already successfully used for recombinant protein expression attributed to its high growth rates and high efficiency expressing complex eukaryotic proteins (Hempel and Maier 2012; Taparia et al. 2019). Metabolically engineered *P. tricornutum* successfully accumulated docosahexaenoic acid (DHA) (Hamilton et al. 2014), produced geraniol (George et al. 2020b), and showed an increase in biomass, lipid content, and growth (Seo et al. 2020). In addition, the sequencing and annotation of the diatom's genome in 2008 (Yang et al. 2018b), and the development of a variety of genetic tools, have enable its use in biotechnology (Daboussi et al. 2014; Slattery et al. 2018). Several genetic tools have been developed, including episomal expression systems (Kassaw et al. 2022), TALEN technology (Serif et al. 2017b), and CRISPR/Cas9 for gene knock out

(Stukenberg et al. 2018). Promoters (Seo et al. 2015b; Chu et al. 2016; Adler-Agnon et al. 2018; Watanabe et al. 2018; Erdene-Ochir et al. 2019b) and characterized subcellular localization signals were successfully identified to direct proteins to different compartments of the cell such as the plastid (Kilian and Kroth 2005), central vacuole (Schreiber et al. 2017), peroxisome (Gonzalez et al. 2011), and Golgi apparatus (Liu et al. 2016). Therefore, the aim of this study was to identify if the subcellular localization of *Vp*VAN could influence its enzymatic activity in a microorganism cell. Our objectives were to test three subcellular compartments of the diatom cell and characterize the protein production by increasing the percentage of positive cells. We used *P. tricornutum* to express *VpVAN* linked to the enhanced green fluorescence protein (*eGFP*) with three different subcellular localization signals including bipartite signal of the chloroplast ATPase, the conserved peroxisome targeting signal 1 (PTS1), and the sequence encoding the first 23 amino acid of the xylose transferase localized in the lumen of Golgi. Subcellular localization in the targeted organelles was confirmed by confocal microscopy. Even though fluorescence was detected, *Vp*VAN recombinant protein was not. Organelle isolation for plastid constructions was used as strategy to enrich in protein fraction with *Vp*VAN as well as purification using GFP-trap without success to detect protein. This indicates that the fusion protein was produced at lower amount than the detection levels and characterization of the diatom is needed to ensure recombinant protein production.

Results

Neither vanillin nor ferulic acid affect P. tricornutum growth at pH 8.

A limitation of phenyl aldehydes heterologous production is their antimicrobial property. Therefore, prior to choosing *P. tricornutum* as a candidate organism to test *Vp*VAN's activity, its growth was assessed under the presence of ferulic acid and vanillin. The phenolic compounds tested delayed the growth of *P. tricornutum* (Figure 2. 1a-b). Cell morphology was assessed by light microscopy at day five (120h), revealing no observable differences between cells grown in L1 media supplemented with 1 mM of vanillin and the negative control with L1 media alone (Figure S2. 1).

After adjusting pH of the media to 8, neither vanillin nor ferulic acid had a negative effect in *P. tricornutum* growth (Figure 2. 1c-d). In fact, growth was slightly higher at 144h of cells cultured with 2.5 mM of vanillin (Figure 2. 1d). Cell viability was also assessed in the presence of 2.5 mM of vanillin at pH 8 using propidium iodide (PI), showing no cytotoxic effect at 24, 48, or 72h (Figure 2. 1e).

Selection and characterization of P. tricornutum exconjugants.

To target and assess *Vp*VAN localization in specific compartments, several constructions were assembled, *VpVAN* was linked to *eGFP* through the cleavable peptide by factor Xa. To target the chloroplast, *VpVAN* was preceded by a signal specific to the plastid stroma in pCS*Vp*VAN:eGFP (Figure 2. 2). To avoid hampering the normal function of *P. tricornutum* plastid, we assembled alternative constructions p*GoVpVAN:eGFP* and p*PeVpVAN:eGFP* to direct the enzyme to Golgi and peroxisome, respectively, as well as one construction without any signal to target the cytosol (p*CyVpVAN:eGFP*). The signal peptide to target the Golgi was

placed before *VpVAN*, while it was placed after e*GFP* for localization in peroxisomes.

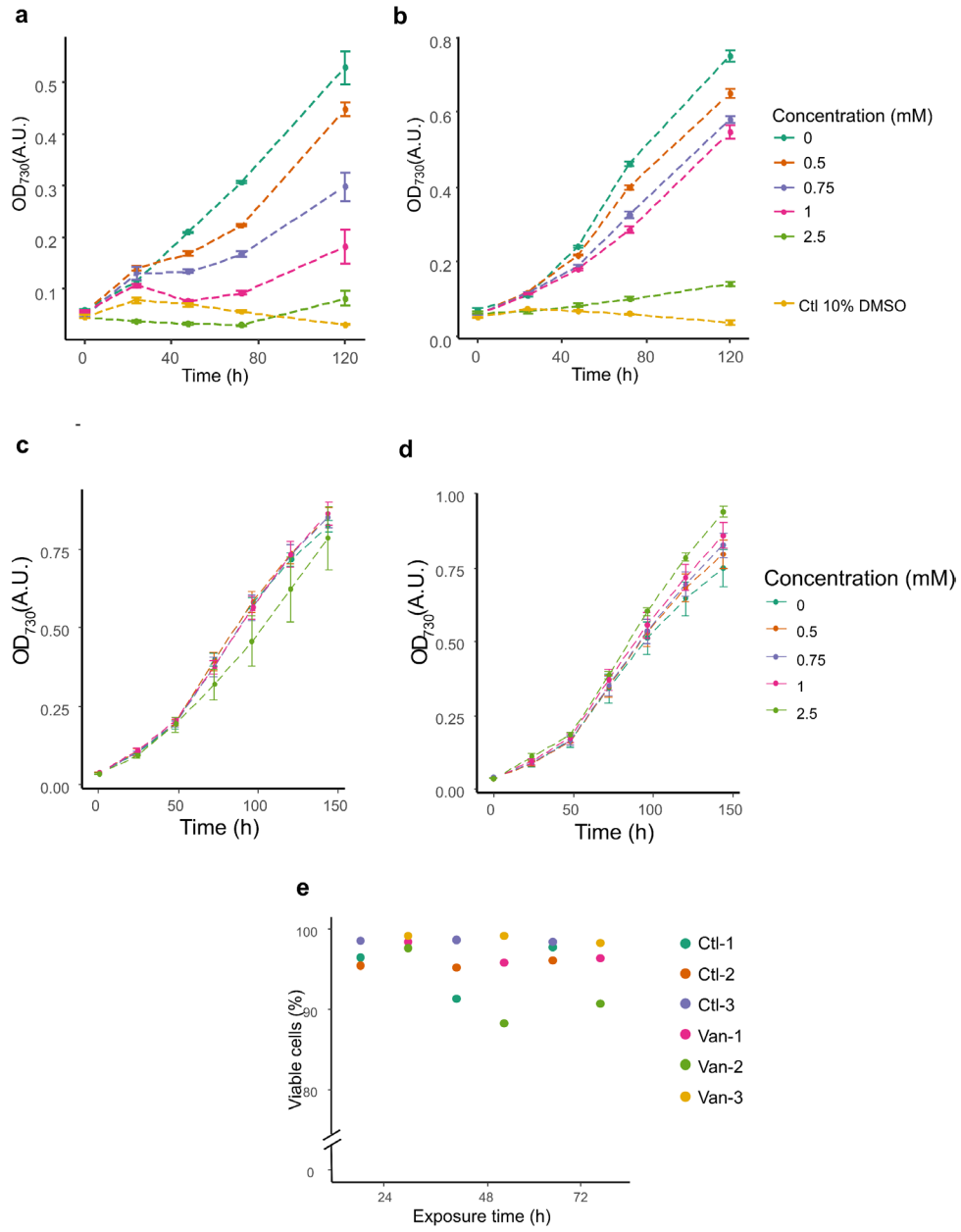


Figure 2. 1. Assessing cytotoxicity of ferulic acid and vanillin to wildtype *P. tricornutum* cells. Growth of *P. tricornutum* with different concentrations of **a**) ferulic acid and **b**) vanillin, for 5 days (120h), without pH adjustment. Growth of *P. tricornutum* at pH 8 with different concentrations of **c**) ferulic acid and **d**) vanillin, for 6 days (144h). **e**. Percentage of viable cells of *P. tricornutum* incubated with 2.5 mM of vanillin

for 24, 48 and 72h. n= 3; Ctl: L1 media alone; Van: 2.5 mM of vanillin; numbers indicate biological replicates.

Following transformation and a two-week selection in *P. tricornutum in* ½L1 agar plates with zeocin*,* at least 36 colonies of each construction were screened for fluorescence using an epifluorescence microscope (Table 2. 1). The percentage of fluorescent colonies for each construction is presented in Table 2. 1. p*CSVpVAN:eGFP* transformants exhibited the lowest percentage of positive colonies with 53.7%. The peroxisome construction had the highest percentage with 80.5% positive colonies. Three different positive colonies per construction were picked for further analyses. To confirm the integrity of the gene of interest, whole plasmid sequencing was performed, revealing no mutations in the expression cassettes in any of the clones (Figure 2. 2).

Table 2. 1. Fluorescent colonies per construction.

Construction	Percentage of fluorescent colonies
p*CyVpVAN:eGFP*	24/36 (66.7%)
p*PeVpVAN:eGFP*	29/36 (80.5%)
p*CSVpVAN:eGFP*	29/54 (53.7%)
p*GoVpVAN:eGFP*	25/36 (69.4%)

Successful localization of the recombinant proteins in four different cell compartments.

The assessment of the subcellular localization was done for three different transformants for each construction by confocal microscopy. Fluorescent

signal from the clones carrying *VpVAN:Xa:eGFP* was compared to the empty vector (EV) strain as negative control (Figure 2. 3a). Clones were only considered positive for protein production if the fluorescent pattern was different from the EV strain, including higher signal intensity. Localization in the cytosol, visible as a uniform signal throughout the cell, was confirmed in Cy*Vp*VAN:eGFP clones (Figure 2. 3b). In the case of Pe*Vp*VAN:eGFP and Pe:eGFP clones, a punctuated pattern of three to five dots surrounding the plastid was observed (Figure 2. 3c). The signal for these constructs was also detected in the cytosol.

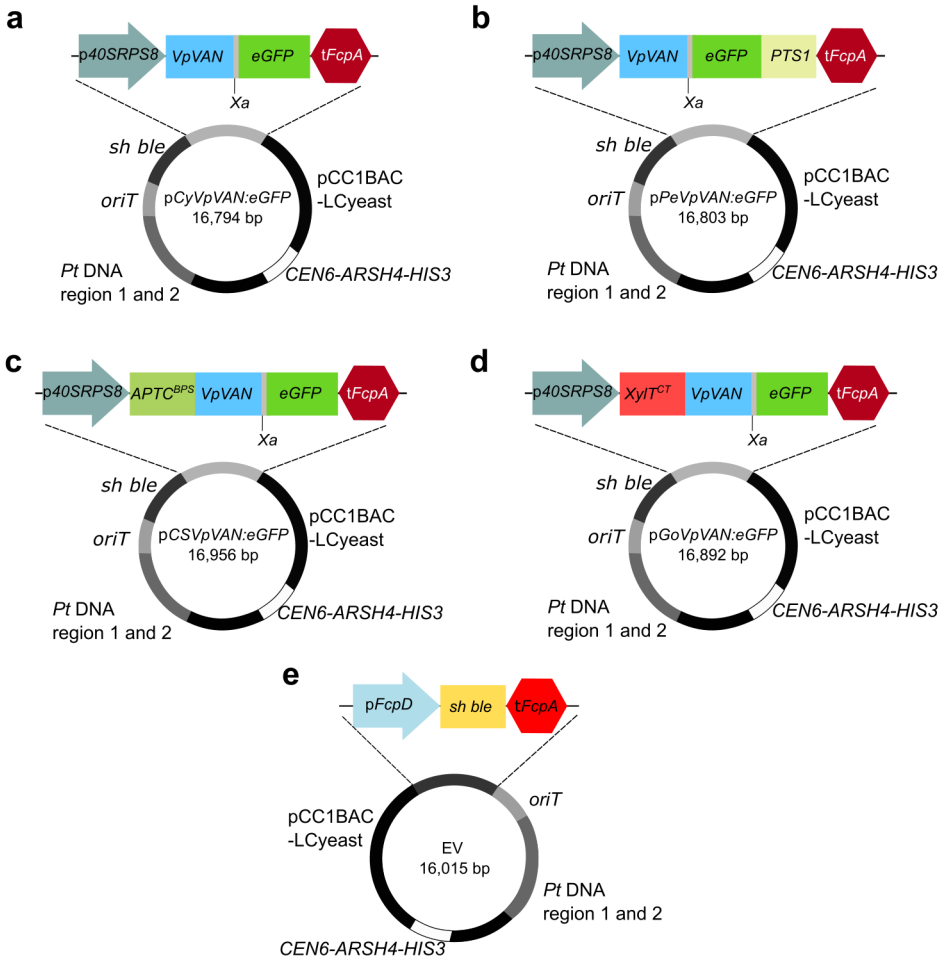


Figure 2. 2. Stable episome propagation without mutations in the expression cassettes with localization signals for three different organelles and the cytosol. a. Cytosolic construction of *VpVAN:Xa:eGFP*

without any localization signal; **b**) peroxisome construction harboring the peroxisomal targeting signal 1 (*PST1*); **c**) plastid construction with the bipartite signal of the chloroplast ATPase gamma subunit (*ATPC*^{BPS}); **d**) β1,2-xylosyltransferase cytosolic tail and transmembrane domain (CT) for medial Golgi (*XylT*^{CT}). *Vp*VAN expression cassette is antisense from *sh ble*. **e.** Empty vector strain harboring the pPtGE30 plasmid. Distances are not at scale.

Interestingly, differences between controls and clones with *Vp*VAN constructions were observed for plastid and Golgi subcellular localization. Co-localization of the recombinant protein and plastid autofluorescence was observed in CS*Vp*VAN clones (Figure 2. 3d), as well as aggregation of proteins seen as high intensity dots together with cytosolic fluorescence. Clones harboring the Golgi retention signal, with and without *Vp*VAN, presented a characteristic “C”-shape in the middle of the plastid (Figure 2. 3e). In addition, Go:eGFP clones also presented fluorescence in the endoplasmic reticulum (ER). This pattern was not observed in Go*Vp*VAN:eGFP clones, which instead had overlapping fluorescence with the plastid. The discrepancy between control and the localization of *Vp*VAN clones could be attributed to the amino acids located after the predicted transmembrane domain (TMD) from XylT. *In silico* analysis predicted the three first amino acids of eGFP as part of the TMD of Go:eGFP, not present in either the full length XylT or Go*Vp*VAN:eGFP.

Since most of the localization signals are in the N-terminal of the fusion protein and sequencing results confirmed intact cassettes, we could deduce that the fusion protein is present in the positive clones. However, the main

cause of *Vp*VAN controversy is the inability to reproduce vanillin synthesis by itself. Therefore, it is vital that we confirm the production of the full fusion protein in our desired cellular compartments.

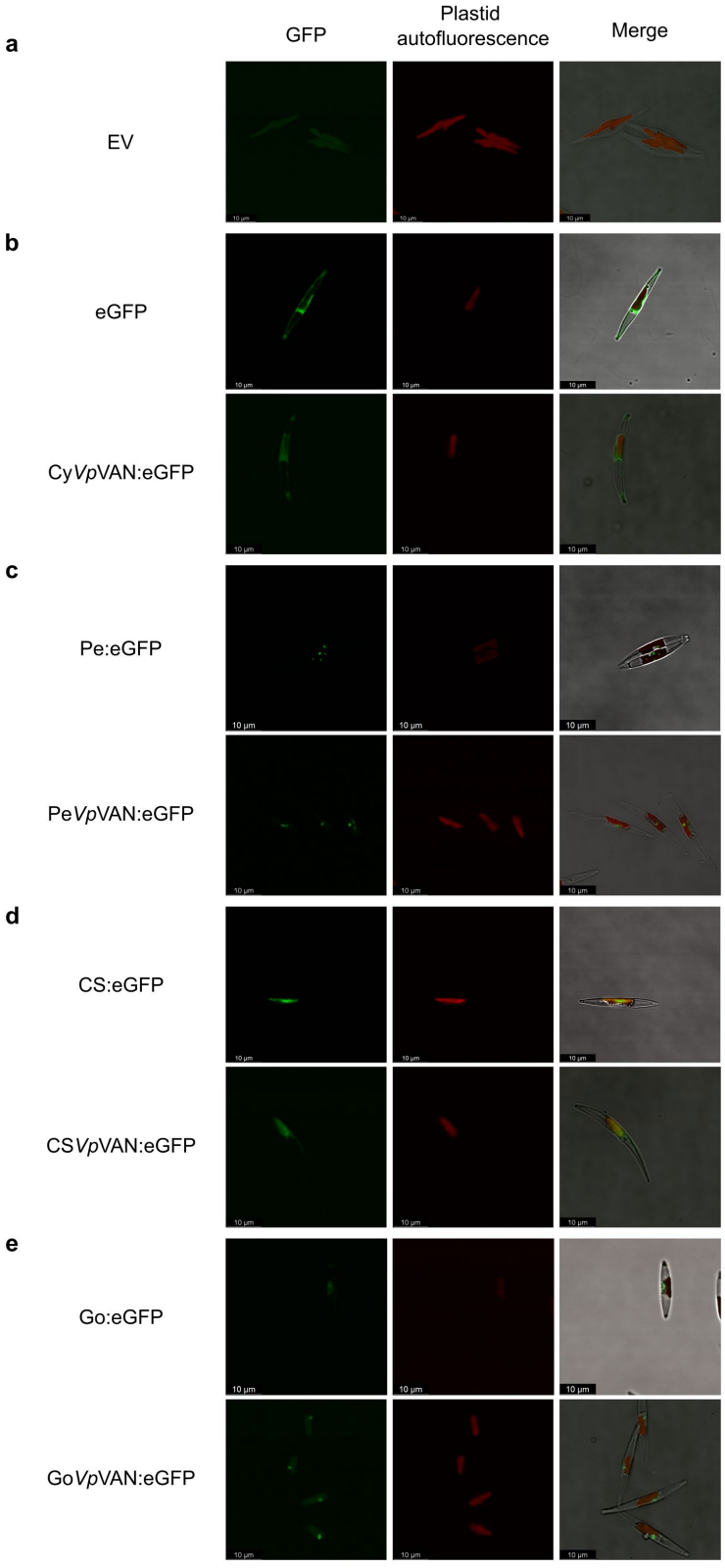


Figure 2. 3. The fusion protein VpVAN:Xa:eGFP localizes in the targeted cell compartment of *P. tricornutum*. a. EV: empty vector strain. **b.** cytosolic constructions: eGFP and Cy*Vp*VAN:eGFP. **c.** Pe:eGFP and Pe:*Vp*VAN:eGFP clones. **d.** plastid stroma clones with the bipartite signal of ATPC linked to eGFP alone (CS:eGFP) and the fusion protein (CS*Vp*VAN:eGFP). **e.** Clones with Golgi retention signal: Go:eGFP and Go*Vp*VAN:eGFP. Merge images represent the combination with plastid autofluorescence, GFP channel and bright field. Scale bars 10 µm.

Enrichment in fluorescent cells, protein purification, and plastid isolation failed as strategies to detect heterologous protein by western blot in positive clones.

We aimed to enrich in fluorescent cells by FACS with the purpose of increasing the percentage of cells in one clone that produce the fusion protein. We used an empty vector strain to gate the GFP+ cells (Figure 2. 4a). However, after sorting GFP+ cells and growing them for four weeks we were not able to obtain high percentage of fluorescent cells (Figure 2. 4b-e).

*Vp*VAN production was not detected in any of the transformed cell lines by western blot, mainly attributed to the low percentage of fluorescent cells. Since we were not able to increase the number of fluorescent cells in the population by cell sorting, we decided to enrich the protein fraction to detect the protein by western blot. We used two different strategies: plastid isolation for the CS*Vp*VAN clones and protein purification by GFP-trap for Cy*Vp*VAN clones. Unlike for peroxisomes and Golgi, plastid isolation protocol for *P. tricornutum* is available in the literature. Besides, the

CS*Vp*VAN:eGFP had the highest percentage of fluorescent cells. Purification using GFP-trap was done only for the cytosolic clones since it is the only one where the protein is not compartmentalized in another membranous structure. However, western blots failed to reveal bands specific to *Vp*VAN, thus protein production of the fusion protein was not detected in any of the tested clones (Figure 2. 5). The controls of protein transfer to the membrane are present in Figure S2. 2. We succeeded to purify eGFP:T2A and eGFP:T2A:mCherry from a pDMi8 cell line as a positive control (Figure 2. 5a), the detected bands at 36kDa and 58 kDa correspond to the cleaved (eGFP:T2A) and uncleaved (eGFP:T2A:mCherry). The percentage of GFP+ cells for the control was 35.2 %.

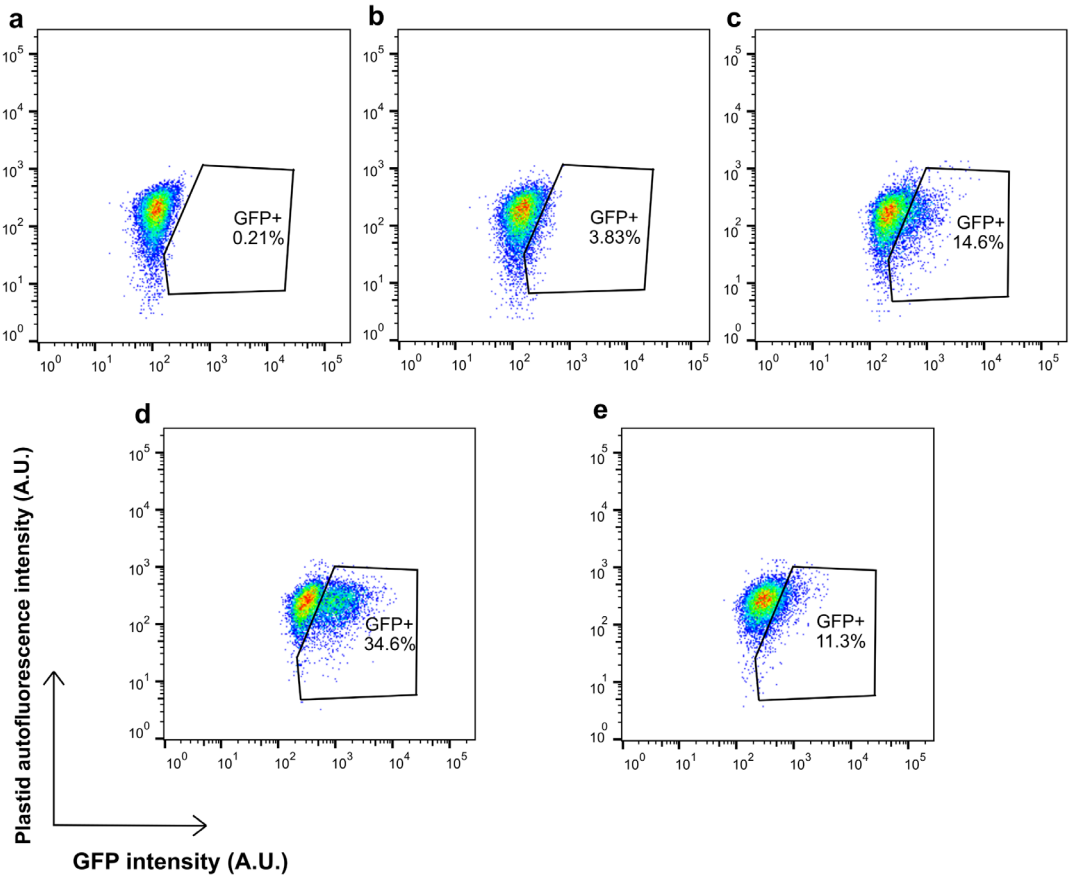


Figure 2. 4. Enrichment for fluorescent cells by FACS fails to yield a high percentage of fluorescent cells. A single cell line is presented per targeted cellular compartment, four weeks after sorting. **a**. empty vector (EV); **b**. Cy*Vp*VAN:eGFP; **c**. Pe*Vp*VAN:eGFP; **d**. CS*Vp*VAN:eGFP; **e**. Go*Vp*VAN:eGFP.

Discussion

To first assess if *P. tricornutum* was a suitable platform to produce vanillin we observed the diatom growth in 2.5 mM of vanillin. Vanillin has been reported to be toxic for *E. coli*, *Lactobacillus plantarum* and *Listeria innocua*, since it is a membrane-active compound able to dissipate ion gradients and inhibits respiration (Fitzgerald et al. 2004). *P. tricornutum* growth was delayed when cultured with ferulic acid and vanillin without adjusting the pH of the media. However, after we adjusted the pH to 8, we did not observe any delay in growth, we even observed a higher growth when culturing *P. tricornutum* at 2.5 mM of vanillin. Studies show that vanillin is a transient intermediate of ferulic acid degradation and can be used as carbon source by 42 strains of Actinobacteria (Brink et al. 2019). It has been proven that by increasing the pH of the media, ferulic acid occurs in its anionic form and can be transported better into the cell with less membrane damage increasing both, ferulic acid consumption rate and vanillin yield in an engineered *E. coli* strain harboring *Ech* and *Fcs* (Luziatelli et al. 2019). We validated the fitness of the cells at 2.5mM of vanillin pH 8 by assessing membrane integrity. We did not observe significant cytotoxic effect after 24, 48, or 72h adding vanillin to the media. While our results support the use of *P. tricornutum* as a platform to produce phenylpropanoids, they also

indicate that metabolic optimization is needed since phenyl compounds could be metabolized by the cell.

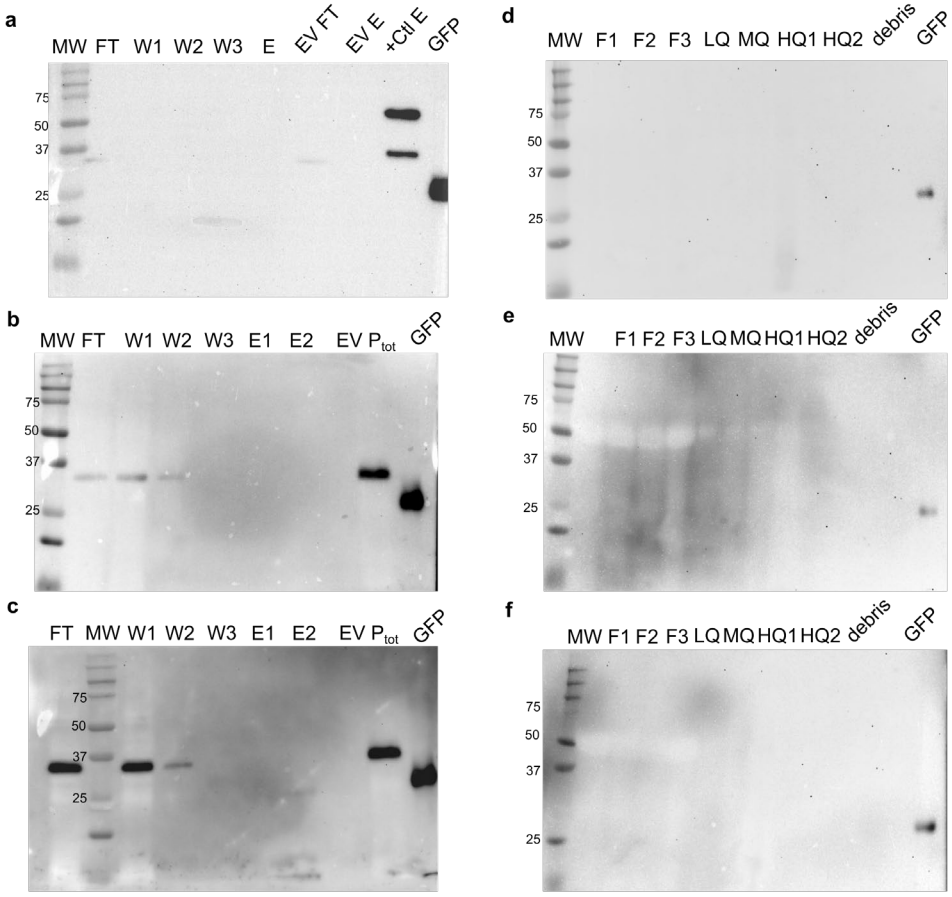


Figure 2. 5. Enrichment of protein fraction and purification failed to increase fusion protein to detectable levels. **a**-**c**. Western blot of protein extracts from three different clones with cytosol constructions after purification using GFP-trap. **d**-**f**. Western blot using protein extract of plastid purification of three different clones harboring the plastid construction. MW: molecular weight marker; FT: flow through; W1-3: washes; E1-2: elution fraction; P_{tot}: total protein in the soluble fraction; +Ctl: eGFP:T2A:mCherry clone; F1: total protein extract; F2: thylakoids fragments and mitochondria; F3: plastids; LQ: low quality plastids; MQ: medium quality plastids.

We constructed four different expression cassettes to target *Vp*VAN to cytosol, peroxisomes, plastid stroma, and Golgi apparatus. While in *V. planifolia* this enzyme is localized in the plastid, we decided to target the production of vanillin to other cellular compartments in *P. tricornutum* since previous studies have shown that the accumulation of phenylpropanoids in a plastid can cause a re-differentiation, stopping the production of chlorophyll and thus the photosynthesis (Brillouet et al. 2014a). The Golgi apparatus is an organelle of interest since glycosyltransferases (Baïet et al. 2011) reside in this compartment. Therefore, targeting the production of vanillin to Golgi could yield production of vanillin glycoside, which can be accumulated in the cell without presenting toxicity. Peroxisomes have a single-layer membrane where low molecular weight compounds can cross passively or by protein channels (Dusséaux et al. 2020). These organelles are in charge of detoxification which means they can sequester and handle molecules that are toxic for the rest of the cell, such as vanillin (Dusséaux et al. 2020). In addition, a re-evaluation of vanillin biosynthetic machinery has been proposed since Gallage *et al.* only confirmed the subcellular localization of *Vp*VAN (2018). It could be possible to have cytosolic and ER-anchored enzymes that were also isolated with the chloroplast tested as co-isolation of the fragments of other organelles with chloroplasts is common to happen (Diamond et al. 2023a).

After selecting three positive colonies of each construction, we verified the integrity of the episomes by sequencing. While plasmid re-arrangements have been described to happen in *P. tricornutum* (Diamond et al. 2023b), our results are consistent with previous reports showing stable plasmid propagation in the diatom (Slattery et al. 2018; Awwad et al. 2023).

The subcellular localization patterns matched what has been previously described for the localization signals used in this study (Kilian and Kroth 2005; Gonzalez et al. 2011; Liu et al. 2016). However, we observed differences in eGFP signal comparing clones carrying *VpVAN* and the control clones with eGFP linked to the localization signals. The CS*Vp*VAN:eGFP clones presented besides the chloroplast localized fluorescent signal, aggregation of proteins observed as high intensity dots and cytosolic fluorescence not seen in the control without *Vp*VAN (p*CSeGFP*). This distinct fluorescent pattern could be caused by a cleavage of the Xa peptide by native *P. tricornutum* peptidases, as it has been hypothesized by Messaabi and colleagues producing free eGFP, localizing in the cytosol (Messaabi et al. 2024).

Optimization of protein compartmentalization may be necessary for the peroxisome and Golgi localizations. Clones harboring PTS1 presented cytosolic fluorescence most likely due to the presence of the recombinant protein prior to its import. An enhanced PTS1 sequence in *S. cerevisiae* has been reported to increase the rate of cytosolic protein import by the addition of the linker LGRGRR, which has a higher affinity to the import receptor (DeLoache et al. 2016). A similar strategy could be implemented for the diatom to effectively compartmentalize the enzymatic activity into the peroxisome. There is evidence in other organisms that a single TMD, arginine-based motifs residues used as sorting sequences, and protein conformation are responsible in the localization of glycosyltransferases within Golgi (Banfield 2011). Moreover, it has been shown that Golgi-resident enzymes possess multiple mechanisms to effectively be retained in this compartment, which can be complementary or additive (Welch and

Munro 2019). Therefore, a characterization is needed to identify the elements in XylT sequence responsible of targeting this enzyme to Golgi and obtain reproducible results using different proteins. CTS regions of glycosyltransferases from other organisms contain a luminal stem region separating the catalytic domain from the TMD that plays a role in sub-localization and oligomerization of the protein, in addition to the cytosolic tail and a single short TMD (Grabenhorst and Conradt 1999; De Graffenried and Bertozzi 2004). Identification of the minimal stem region necessary to retain enzymes in Golgi could replace the use of the full amino acid sequence to target proteins to this compartment. Alternatively, the fragment tested in this study could provide a shared localization in Golgi and ER, which may be beneficial for accumulation of vanillin glycoside since both compartments harbor glycosyltransferases (Schmitz et al. 2008; Tannous et al. 2015).

Detection of the *Vp*VAN:Xa:eGFP protein by western blot was not possible in this study. A plausible cause of the lack of signal could be very low amounts of produced protein that despite purification was below the detection limits of our assay. This led us to use different strategies to increase the protein fraction of our protein of interest including, purification, organelle enrichment and increasing the fraction of GFP+ cells by FACS. However, these three strategies failed to yield detectable levels of our protein of interest. In addition, we obtained low percentages of GFP+ cells after growing the sorted cells from our positive clones for four weeks. These results indicate that we selected more than one phenotype when sorting the cells and/or *P. tricornutum* regulated the production of the fusion protein in days post-sorting. Our results are consistent with George and colleagues

who were not able to enrich in phenotypic populations by cell sorting of *P. tricornutum* cells episomally expressing *mVenus* (George et al. 2020b). Studies done in heterologous subpopulations in cell lines expressing *mVenus* integrated into the genome of *P. tricornutum* suggest that differential DNA methylation patterns and histone modification (H3K4me2) are not responsible for transgene silencing (Faessler 2024). In addition, in another study, we reported differences in plasmid copy number, a large deletion, and co-expression of a putative zeocin resistant gene with the transgene, as factors contributing to the formation of subpopulations in a single clonally propagated episomal cells line of *P. tricornutum* (Diaz-Garza et al. 2024).

The decrease in heterologous protein production may be specific to cell lines harboring *Vp*VAN, as a response for the high production of an enzyme with high amino acid sequence homology to a cysteine protease. Identification of protein degradation mechanisms in *P. tricornutum*, by using protease or proteasome inhibitors, could provide insight into the accumulation of heterologous proteins, although this could affect *Vp*VAN cysteine protease function. Additionally, this information could lead to the engineering of strains with lower protein degradation mechanisms. Similar work has been done in yeast strains, generating mutant variants with disrupted protease genes, efficiently increasing heterologous protein accumulation (Cho et al. 2010; Tomimoto et al. 2013).

An alternative hypothesis regarding the low accumulation of *Vp*VAN is that the trace amount of ferulate in *P. tricornutum* (Rico et al. 2013) could be used by the enzyme to produce vanillin, and the cell would in turn downregulate its production. The diatom cells could be adapting to reduce

the production of a toxic enzyme or product, as it has been suggested in our previous study, where re-arrangements occur in high frequency in an episome harboring *Ech* and *Fcs* genes, able to convert ferulic acid into vanillin (Diamond et al. 2023b). Diatoms have evolved to rapidly adjust to a changing environment; therefore, mechanisms lighten the metabolic burden of producing a toxic enzyme may be in place to decrease the synthesis of *Vp*VAN. Better understanding of dynamics of *P. tricornutum* expression system is needed to successfully characterize possibly toxic proteins such as *Vp*VAN.

Conclusions

Synthesis of vanillin from ferulic acid in one-step using *Vp*VAN remains a challenge, since many characteristics of this enzyme are still unknown. However, by establishing requirements for the expression of *Vp*VAN in microorganisms, such as targeting a specific compartment of the cell, we could contribute to answer questions about its activity. In this study we were able to introduce *VpVAN*:*Xa:eGFP* in an episome by bacterial conjugation into *P. tricornutum*. Plasmid rescue and sequencing showed that episomes remained extrachromosomal and were stably propagated. Besides, we confirmed that previously characterized localization signals of the diatom effectively directed our fusion protein into the plastid, peroxisome, and Golgi. However, we also observed that optimization of this sequences to increase the rate of transport of the desired protein may be needed to avoid pre-import enzymatic activity. Finally, we were not able to detect either truncated or full protein by western blot. Enrichment strategies used by previous studies failed to yield detectable levels of the fusion protein, which

questions if this effect is due to the alternative enzymatic activity proposed of *Vp*VAN as cysteine protease exhibiting toxicity for the diatom cell. These results strongly suggest that *P. tricornutum* expression system needs to be studied in more details to be able to characterize “toxic” enzymes.

Materials and Methods

Microbial Strains and Growth Conditions.

Escherichia coli (Epi300, Epicenter) was grown in Luria Broth (LB) supplemented with appropriate antibiotics (gentamicin (20 mg L^{-1}) or chloramphenicol (15 mg L^{-1}) and gentamicin (20 mg L^{-1})). *Phaeodactylum tricornutum* (CCAP 1055/1, Culture Collection of Algae and Protozoa) was grown in modified L1 medium without silica at 18 °C under cool white fluorescent lights (75 µE m^{-2} s^{-1}) and a photoperiod of 16 h light:8 h dark with an agitation of 130 rpm for liquid cultures (Diamond et al. 2023b).

Cytotoxicity assays of P. tricornutum with ferulic acid and vanillin.

To observe the effect of phenylpropanoids in the diatom’s growth, 2.0 x 10^6 *P. tricornutum* cells were cultured in 200 µL in a 96-well plate with L1 media for 5 days with 0, 0.5, 0.75, 1, 2 mM of either ferulic acid or vanillin, and its growth was assessed by measuring each 24h the optical density at 730 nm. The experiment was repeated adjusting the pH of the media to 8 and assessing growth for 6 days. Cells grown in 1mM of vanillin were assessed for changes in their morphology at day five by light microscopy (OMAX, USA) using the 40X objective. Finally, the percentage of viable cells of cultures with 2.5 mM of vanillin after 24, 48, and 72 hours of treatment was measured by incubating for 10 min with propidium iodide (PI) final

concentration of 3 µg/mL. PI emission was measured by BD FACS Melody (BD Biosciences, La Jolla, CA, USA) in the ECD channel (610/20 nm). The percentages of live cells were calculated relative to the untreated control wells.

Plasmid construction and transformation of P. tricornutum by bacterial conjugation from E. coli cells.

Plasmids were constructed by Gibson Assembly using the NEBuilder® HiFi DNA Assembly Bundle for Large Fragments (New England Biolabs, Canada). All fragments used in the assemblies were amplified by PCR using PrimeSTAR GXL DNA Polymerase (Takara Bio, Japan) following the manufacturer's protocol. Episomes were completed by replacing the coding sequence in the expression cassette of pDMi7 (Diamond et al. 2023b) containing the *40SRPS8* (40S ribosomal protein S8) promoter and the *FcpA* terminator, with the backbone of pPtGE30 harboring the *CAH* region (*CEN6-ARSH4-HIS3*) and the *sh ble* gene that confers resistance to zeocin. All expression cassettes consisted in *VpVAN* without the predicted signal peptide linked to the enhanced green fluorescent protein gene (*eGFP*) by the factor Xa cleavable peptide (*Xa*). Depending on the targeted cellular compartment *P. tricornutum* localization signals were added to each construction: cytosol without any signal, peroxisome construction harboring the peroxisomal targeting signal 1 (PST1) tripeptide SKL, plastid construction with the bipartite signal of the plastid ATP synthase gamma chain (Phatr3_J20657) to direct the recombinant protein to the stroma, and Golgi apparatus with the N-terminal (31 amino acids) of the β1,2-xylosyltransferase (Phatr3_J45496). The latest contains a predicted

transmembrane domain as well as and a triple arginine motif (RRR) in the cytosol tail. In addition, four constructs harboring the localization signals and *eGFP* were assembled as controls for subcellular localization (Figure S2. 3). The empty vector used in this study was the plasmid pPtGE30 (Slattery et al. 2018), harboring no expression cassette besides *sh ble*.

Sequences were synthesized and optimized for *P. tricornutum* codon usage by Bio Basic (Markham, Ontario, Canada). All primers used in the assembly are listed in Table S2. 1. DNA sequences from this study are available in Table S2. 2.

Successful cloning and transformation of all four constructions was confirmed by screening 10 chloramphenicol resistant *E. coli* NEB 10-beta colonies by colony PCR. A single colony for each construction was chosen for plasmid extraction and the integrity of the episomes was verified by Sanger sequencing (Génome Québec). Plasmids with intact expression cassettes were transformed into *E. coli* Epi300 pTA-MOB, since this strain harbors the plasmid that allows to do conjugation with *P. tricornutum*.

Episomes were transformed into wild type *P. tricornutum* by bacterial conjugation as previously described (Diamond et al. 2023b). Briefly, 250 µL of *P. tricornutum* culture, concentration of 1.0×10^8 cells mL^{-1}, were plated in ½L1 agar plates and grown under the conditions mentioned above for 4 days. Prior to transformation, cells were harvested by scraping the plates and adding 1 mL of L1 media, cell concentration was then adjusted to 5.0×10^8 cells mL^{-1}. A culture of 25 mL of *E. coli* EPI300 pTA-MOB containing either the empty vector or pDMi8 plasmids was grown at 37°C and 220 rpm until reaching an OD_{600} of 0.7, then centrifuged at 3000 g for 10 min and resuspended in 250 µL of SOC media. To initiate the conjugation, 200 µL

of *P. tricornutum* cells were mixed with 200 µL *E. coli* cells, plated in ½L1 5% LB agar plates, and incubated at 30°C for two hours in the dark. After conjugation, plates were kept at standard growth conditions for *P. tricornutum* for the recovery period of 2 days. Cells were collected by scraping with 1 mL of L1 media and plated in selective ½L1 agar plates with zeocin 50 µg/mL. Colonies appeared after 10 to 14 days of growth at 18°C and photoperiod of 16 h light:8 h dark.

Screening positive exconjugants by fluorescence microscopy.

Two weeks following the conjugation colonies were screened by eGFP fluorescence using a Fluorescent Stereo Microscope Leica M165 FC with GFP filter for eGFP fluorescence detection. Colonies were observed with a magnification of 80 to 120x. Three positive colonies for each construction were selected to start 25 mL cultures.

Episome rescue, whole plasmid sequencing and in silico analysis.

Episome isolation from P. tricornutum was conducted as described in Diamond et al. (Diamond et al. 2023b), using Large Plasmid Mini Kit (Geneaid Biotech Ltd., Taiwan), with approximately 1 x 10^8 cells of *P. tricornutum*. Plasmid extracts were used to transformed NEB 10-beta chemically competent E. coli cells following the manufacturer's protocol. After overnight incubation of the plates a single colony was used to inoculate 10 mL LB cultures to proceed for miniprep using Large Plasmid Mini Kit (Geneaid Biotech Ltd., Taiwan). Plasmids were sequenced by CCIB DNA Core (Massachusetts General Hospital, United States of America) through their Next-Generation sequencing Illumina MiSeq platform.

Amino acid sequences obtained from in silico translation of the sequenced episomes targeting Golgi were analyzed using TMHMM 2.0 as well as the full XylT sequence to predict the transmembrane domains (Sonnhammer et al. 1998).

Assessing subcellular localization by confocal microscopy.

Live cell images of the five-day-old culture were captured using a Leica SP8 confocal laser microscope (Leica, Wetzlar, Germany) with an HCX PL APO 60×/1.25–0.75 Oil CS objective. The excitation of eGFP and chlorophyll fluorescence occurred at 488 nm with a 65 mW argon laser. eGFP fluorescence emission was detected between 499nm and 531nm, whereas chlorophyll fluorescence was detected at a bandwidth of 640nm - 710nm. Bright-field light microscopy images were also taken. The images were analyzed using LAS X software (version) (Leica).

Fluorescence-activated cell sorting (FACS).

The BD FACS Melody (BD Biosciences, La Jolla, CA, USA) equipped with blue (488 nm), red (640 nm) and violet (405 nm) lasers was used to analyze and sort *P. tricornutum* transconjugants according to eGFP production as described by Diamond *et al.*(Diamond et al. 2023b). Five-day-old *P. tricornutum* cultures were filtered with Falcon™ Cell Strainers (Fisher Scientific, USA) prior to sorting. Events were acquired at a fixed flow rate and at least 10,000 events were analyzed. Cells were gated according to FSC-A (forward scatter area) and SSC-A (side scatter area) parameters and doublets were excluded according to further gating on homogeneous FSC-H (height) *vs.* FSC-W (width) and SSC-H *vs.* SSC-W populations. Chloroplast autofluorescence was gated in the PerCP channel (700/54 nm,

665 LP). Cells with non-specific autofluorescence detected in the PE channel (582/15 filter, 582 long pass filter mirror – LP) were excluded from sorting. eGFP was further analyzed on the 527/32 nm band-pass filter channel. Sorting was set on purity parameter. Sorted cells were collected in 1.5 mL microcentrifuge tubes containing 250 µL of L1 media with ampicillin (100 µg/mL) and zeocin (50 µg/mL). Cells were centrifuged for 10 min at 3,500 g, supernatant was removed to avoid toxicity from the FACS sheath fluid, and fresh L1 media with antibiotics was added. Cultures were grown for two weeks and used as inoculum for 50 mL cultures. Data analysis was conducted with BD FlowJo version 10 software (BD Biosciences, La Jolla, CA, USA, 2020).

Protein extraction and purification.

Protein was extracted using 20 mL seven-days-old cultures as described by Fantino *et al.* (2024). Cells were harvest by centrifugation at 3,500 x*g* for 10 min at 4°C. Pellets were weighted and 500µL of lysis buffer (51.4 mM Tris pH 8, 0.75 mM SDS, 10 % Glycerol, 0.02 mM EDTA, 10 mM PMFS and 2 µL of protease inhibitor cocktail) per 200 mg of wet biomass was added. Cells were lysed by sonication with 35% of amplitude and 6 pulses of 30 s with a rest period of 30 s between pulses using Fisherbrand Model 505 Sonic Dismembrator (Thermo Fisher Scientific). The soluble fraction was separated by 30 min of centrifugation at 20,000 x*g*. The supernatant of clones harboring the cytosol construction was used for purification with GFP-trap beads following the manufacturer protocol, keeping aliquots of each fraction for posterior analyses. Protein quantification was performed using the RC DC Protein Assay Kit I (Bio-Rad).

Plastid isolation.

Plastid isolation from clones carrying the construction targeting this organelle was performed according to the protocol of Schober *et al.* (Schober et al. 2018) with the modification of using sonication as mentioned above to lyse the cells. After cell disruption, intact cells and debris were removed by centrifugation at 500 x*g*. The supernatant containing mainly plastids and mitochondria was centrifuged at 2,000 *xg* to concentrate the plastids. The supernatant was discarded, and the pellet was resuspended in isolation buffer (0.5 M Sorbitol, 50 mM HEPES-KOH pH 8, 6 mM EDTA, 5 mM $MgCl_2$, 10 mM KCl, 1mM $MnCl_2$, 1% polyvinylpyrrolidone, 0.5% BSA, and 0.1% L-cysteine). High quality plastids were isolated by applying the plastid-containing solution into a layered Percoll density gradient and centrifuged at 10,000 *x g*. Plastid were washed by resuspending them in washing buffer (0.5 M Sorbitol, 30 mM HEPES-KOH pH 8, 6 mM EDTA, 5 mM $MgCl_2$, 10 mM KCl, 1mM $MnCl_2$, and 1% polyvinylpyrrolidone) and centrifugated at 4,000 x*g*. Proteins were extracted from the final plastid fraction by freeze and thaw (using liquid nitrogen), and resuspending on protein extraction buffer containing 10% glycerol, 5mM EDTA, 10 mM DTT, and 100 mM HEPES (pH 7.2).

Western blot.

Proteins were detected by western blot, loading 50 µg for soluble fraction. SDS-PAGE of 12% polyacrylamide. Regarding the washing and elution fractions, 25 µL of sample were loaded since the amount protein was either too low or protein quantification was not possible due to interaction with the reagents of the elution buffer. Also, a 3 µL of positive control for western

blot (purified GFP) were loaded in the gel. Protein transfer to a 0.2 µm PVDF membrane was done by semi-wet transfer at constant amperage of 1 A for 30 min and maximum voltage of 25 V. The membrane was blocked using 3% BSA in Tris-buffered saline, 0.1 % Tween 20 (TBST) solution for one hour at room temperature. Afterwards it was incubated overnight with anti-GFP antibody (CLH106AP) purchased from Cedarlane (Ontario, Canada). Then, the membrane was washed using TBST and incubated for 1 h in a 1:20,000 dilution in 5 % milk of Immun-Star Goat Anti-Mouse (GAM)-HRP conjugate from Bio-Rad (Ontario Canada). Three washes were done, and protein detection was performed using Clarity Max Western ECL Substrate-Luminol solution from Bio-Rad. Chemiluminescence detection and Ponceau S stained (Glacial Acetic Acid 5 % *v*/v, Ponceau Red dye 0.1 % m/v) of the blots were visualized using ChemiDoc Imaging System with Image Lab™ Software (Bio-Rad). The molecular weights of the proteins corresponding to the detected bands were confirmed with protein markers (Precision Plus Protein Dual Color Standards).

Data Availability Statement

The original contributions presented in the study are included in the article and supplementary information, further inquiries can be directed to the corresponding author.

Acknowledgments

Warm thanks to Elisa Fantino for the lab training in molecular biology techniques. This research was funded by Canada Research Chair on plant specialized metabolism Award No CRC-2018-00137 to I.D-P. Thanks are extended to the Canadian taxpayers and to the Canadian government for

supporting the Canada Research Chairs Program. Additional support in the form of scholarship to A.M.D-G. was provided by Mitacs-Acceleration program grant #IT12310.

CHAPTER III: No two clones are alike: Characterization of heterologous subpopulations in a transgenic cell line of the model diatom *Phaeodactylum tricornutum.*

Aracely Maribel Diaz-Garza[1], Natacha Merindol[1], Karen Cristine Gonçalves dos Santos[1], Félix Lavoie-Marchand[1], Brian Ingalls[3] and Isabel Desgagné-Penix[1,2,†]

[1] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

[2] Plant Biology Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

[3] Department of applied mathematics, University of Waterloo, Waterloo, ON, Canada.

[†] Corresponding author Isabel.Desgagne-Penix@uqtr.ca

Authors contributions

A.M.D.G., F.L.M., and N.M. performed experiments. A.M.D.G. analyzed the data with help of K.C.G.d.S and N.M., A.M.D.G., N.M., B.I., and I.D.P. conceived the research project. All authors wrote and revised the manuscript.

Résumé

La délivrance d'épisomes par conjugaison est une méthode très efficace utilisée pour transférer l'ADN dans la diatomée *Phaeodactylum tricornutum*, facilitant ainsi la production de protéines recombinantes et de métabolites de grande valeur. Cependant, des rapports précédents ont indiqué une hétérogénéité phénotypique entre les cellules individuelles des lignées cellulaires exconjugantes propagées par clonage, ce qui pourrait affecter la stabilité de la production de protéines recombinantes dans la diatomée. Ici, nous avons caractérisé les différences entre les sous-populations présentant des phénotypes d'intensité de fluorescence distincts dérivés d'une seule colonie exconjugante de *P. tricornutum* exprimant la protéine fluorescente verte améliorée (eGFP en anglais). Nous avons analysé l'intégrité de la séquence de la cassette d'expression, le nombre de copies du plasmide et l'expression globale du gène. Nos résultats révèlent que le faible nombre de copies et la suppression de la cassette d'expression dans une partie de la population ont contribué à la faible expression du transgène. L'analyse de la co-expression des gènes a permis d'identifier un ensemble de gènes ayant un profil d'expression similaire à celui de l'eGFP, y compris un gène codant pour une recombinase Flp putative, ce qui peut être lié aux variations de l'intensité de la fluorescence. Ces gènes se présentent donc comme des candidats potentiels pour augmenter la production de protéines recombinantes dans le système d'expression épisomal de *P. tricornutum*. Globalement, notre étude élucide les différences génétiques et transcriptomiques entre des sous-populations distinctes dans une culture à propagation clonale, et contribue à une

meilleure compréhension de l'hétérogénéité dans les systèmes d'expression des diatomées pour des applications de biologie synthétique.

Mots-clés

Diatomée, sous-populations hétérologues, cytométrie en flux, nombre de copies de plasmides, transcriptomique.

Abstract

Background

Conjugation-based episome delivery is a highly efficient method used to transfer DNA into the diatom *Phaeodactylum tricornutum*, facilitating the production of recombinant proteins and high-value metabolites. However, previous reports have indicated phenotypic heterogeneity among individual cells from clonally propagated exconjugant cell lines, potentially affecting the stability of recombinant protein production in the diatom.

Results

Here, we characterized the differences between subpopulations with distinct fluorescence intensity phenotypes derived from a single exconjugant colony of *P. tricornutum* expressing the enhanced green fluorescent protein (*eGFP*). We analyzed the expression cassette sequence integrity, plasmid copy number, and global gene expression. Our findings reveal that lower copy numbers and the deletion of the expression cassette in part of the population contributed to low transgene expression. Gene co-expression analysis identified a set of genes with similar expression pattern to *eGFP* including a gene encoding a putative Flp recombinase, which may be related to variations in fluorescence intensity. These genes thus present themselves as potential candidates for increasing recombinant proteins production in *P. tricornutum* episomal expression system.

Conclusions

Overall, our study elucidates genetic and transcriptomic differences between distinct subpopulations in a clonally propagated culture, contributes to a better understanding of heterogeneity in diatom expression systems for synthetic biology applications.

Keywords

Diatom, heterologous subpopulations, flow cytometry, plasmid copy number, transcriptomics.

Background

Diatoms' robustness to grow at industrial-scale along with their capacity to produce specialized metabolites have attracted scientific interest for their use in biotechnology (Huang and Daboussi 2017; Sharma et al. 2021). *Phaeodactylum tricornutum* has been established as model organism because of its ease of cultivation and cryopreservation, as well as its ability to be reproducibly genetically transformed, characteristics that made it possible to unravel molecular mechanisms in diatoms (Russo et al. 2023). In addition, the availability of the complete and annotated genome assembly from telomere to telomere (Rastogi et al. 2018; Giguere et al. 2022) has positioned *P. tricornutum* as one of the leading photosynthetic eukaryotic chassis for metabolic engineering (Pampuch et al. 2022).

Conjugation-based episome delivery is the most efficient method to transform diatoms (Karas et al. 2015). It has facilitated the characterization of new promoter elements (Windhagauer et al. 2021; Garza et al. 2023), the development of tunable and reversible dose- and time-dependent transcriptional systems (Kassaw et al. 2022), and the delivery of gene editing elements (Slattery et al. 2018; Gao et al. 2022; Taparia et al. 2022). For instance, *P. tricornutum* episomal expression system has been used to produce the receptor-binding domain of the SARS-CoV-2 spike protein (Slattery et al. 2022) and specialized metabolites of pharmaceutical interest, i.e., monoterpenoids (Fabris et al. 2020), cannabinoids (Fantino et al. 2024),

and cannabinoid precursors (Awwad et al. 2023). While episomal expression provides advantages over genomic integration, such as avoiding the disruption of essential genes by random integration of DNA and insertion position-specific expression effects on the transgenes (Diner et al. 2016), this system is still not fully characterized. Even though it is rarely characterized, the heterogeneity among the clonal bacterial population after DNA transfer is known to be an issue, raising doubts about this concept in other unicellular genetically engineered organisms (Tomoiaga et al. 2022). There have been few reports of *P. tricornutum* cells originated from a single exconjugant colony showing different phenotypes, generating inconsistencies when producing recombinant proteins. George *et al.* (George et al. 2020a) reported cell lines of episomal expression (EE) of mVenus with high mean fluorescent signal that seemed to be constituted of cells greatly different from each other, since the distribution profiles of fluorescence per cell within individual cell lines was diverse. Moreover, in a previous study we investigated episomal rearrangements in *P. tricornutum* exconjugants showing that individuals from a single colony are not identical in genetic material, but may harbor different versions of episomes (Diamond et al. 2023b). This raises several questions regarding the dynamics of episomal copy number and segregation across cells within a population.

Therefore, the aim of our study was to characterize the population dynamics of individuals coming from a single exconjugant colony of *P. tricornutum* exhibiting different phenotypes grouped in subpopulations. For this purpose, we enriched by fluorescent activated cell sorting (FACS) three subpopulations of an EE cell line harboring the episome pDMi8 (Diamond et al. 2023b) with the enhanced green fluorescent protein (GFP) and

mCherry, with a broad fluorescent profile. These subpopulations differed in GFP and mCherry intensities and were characterized by sequencing of the transgene cassettes, RNA-sequencing, and plasmid copy number quantification.

Material and methods

Microbial Strains and Growth Conditions.

Escherichia coli (Epi300, Epicenter) was grown in Luria Broth (LB) supplemented with appropriate antibiotics (gentamicin (20 mg L^{-1}) alone or chloramphenicol (15 mg L^{-1}) and gentamicin (20 mg L^{-1})). *Phaeodactylum tricornutum* (CCAP 1055/1, Culture Collection of Algae and Protozoa) was grown in modified L1 medium without silica (Diamond et al. 2023b) at 18 °C under cool white fluorescent lights (75 µE m^{-2} s^{-1}) and a photoperiod of 16 h light:8 h dark with an agitation of 130 rpm for liquid cultures.

Transformation of P. tricornutum by bacterial conjugation from E. coli cells.

GFP cell line was generated by introducing in *P. tricornutum* the plasmid pDMi8 which was previously constructed by Diamond *et al*. (2023). This plasmid contains the *sh ble* gene that confers resistance to zeocin, with the backbone of pPtGE30 harboring the CAH region (*CEN6-ARSH4-HIS3*) and an expression cassette containing the *40SRPS8* (40S ribosomal protein S8) promoter, a coding sequence composed of *eGFP* linked to *mCherry* by the T2A self-cleaving peptide, and the *FcpA* terminator. The empty vector cell line was generated as described by Fantino *et al.* with a plasmid harboring no expression cassette besides *sh ble* [16]. Episomes were transformed into wild type *P. tricornutum* by bacterial conjugation as previously

described (Diamond et al. 2023b). Briefly, 250 µL of *P. tricornutum* culture, concentration of 1.0 × 10^8 cells mL^{-1}, were plated in ½L1 agar plates and grown under the conditions mentioned above for 4 days. Prior to transformation, cells were harvested by scraping the plates and adding 1 mL of L1 media, cell concentration was then adjusted to 5.0 × 10^8 cells mL^{-1}. A culture of 25 mL of *E. coli* EPI300 pTA-MOB containing either the empty vector or pDMi8 plasmids was grown at 37°C and 220 rpm until reaching an OD_{600} of 0.7, then centrifuged at 3000 g for 10 min and resuspended in 250 µL of SOC media. To initiate the conjugation, 200 µL of *P. tricornutum* cells were mixed with 200 µL *E. coli* cells, plated in ½L1 5% LB agar plates, and incubated at 30°C for two hours in the dark. After conjugation, plates were kept at standard growth conditions for *P. tricornutum* for the recovery period of 2 days. Cells were collected by scraping with 1 mL of L1 media and plated in selective ½L1 agar plates with zeocin 50 µg mL^{-1}. Colonies appeared after 10 to 14 days of growth at 18°C and photoperiod of 16 h light:8 h dark. Four colonies were randomly picked for starting 200 µL liquid cultures in 96-well microplate and were analyzed by flow cytometry after six weeks of subculturing (Fig. S3.1). Clone 2 was chosen for 25 mL culture and was kept for a year with frequent subculturing (once every 14 days) before sorting.

Doubling times were calculated in the enriched cultures using the following equation:

$t_d = \frac{Ln\,2}{Ln(N_1/N_0)}$ where N_1 is the number of cells at t1 and N_0 is the number of cells at t_0.

Flow cytometry and fluorescence-activated cell sorting (FACS)

The BD FACSMelody (BD Biosciences, La Jolla, CA, USA) equipped with blue (488 nm), red (640 nm) and violet (405 nm) lasers was used to sort *P. tricornutum* transformed transconjugants according to eGFP production as described by Diamond *et al.* (Diamond et al. 2023b). GFP and empty vector cell lines were grown in semicontinuous batch culture for two weeks. *P. tricornutum* cultures were filtered with Falcon™ Cell Strainers (Fisher Scientific, USA) prior to sorting. Events were acquired at a fixed flow rate and at least 10,000 events were analyzed. Cells were gated according to FSC-A (forward scatter area) and SSC-A (side scatter area) parameters and doublets were excluded according to further gating on homogeneous FSC-H (height) *vs.* FSC-W (width) and SSC-H *vs.* SSC-W populations. Chloroplast autofluorescence was gated in the PerCP channel (700/54 nm, 665 LP). Cells with non-specific autofluorescence detected in the PB450 channel (448/45 nm filter mirror) were excluded from sorting by gating on $PB450^-$ events. eGFP was further analyzed on the 527/32 nm band-pass filter channel. Sorting was set on purity parameter. Sorted cells were collected in 5 mL round bottom tubes containing 250 µL of L1 media with ampicillin (100 µg mL^{-1}) and zeocin (50 µg mL^{-1}). Cells were centrifuged for 10 min at 1,500 g, supernatant was removed to avoid toxicity from the FACS sheath fluid, and fresh L1 media with antibiotics was added. Cultures were grown for two weeks and used as inoculum for 50 mL semicontinuous batch cultures kept in early exponential phase (OD_{730} ~ 0.6) for 7 days by subculturing.

Fluorescence intensity per cell was quantified before and after sorting using a CytoFLEX S flow cytometer (Beckman Coulter Life Sciences) equipped

with violet (405 nm), blue (488 nm), yellow-green (561 nm) and red (638 nm) lasers. Chlorophyll autofluorescence was detected in the PerCP channel (690/50 nm), while GFP fluorescence was detected in the FITC channel (525/40 nm). mCherry fluorescence was detected in the ECD channel (610/20 nm). Since we observed a Pearson's correlation coefficient of ~0.92 between GFP and mCherry mean fluorescence intensities (MFI) of the non-sorted cultures (p=0.0002; Fig. S3.2), all analyses are only presented based on GFP intensity.

Data analysis was conducted with BD FlowJo version 10 software (BD Biosciences, La Jolla, CA, USA, 2020) and python 3.11.2 using FlowCytometryTools package (v. 0.5.1) (Yurtsev and Friedman 2015). GMM-based clustering was done by scikit learn package (v. 1.5.1) setting the number of components to three and covariance type as "full" (Pedregosa et al. 2011). Violin and scatter plots were designed with the package matplotlib (v. 3.8.0) (Hunter 2007).

Episome DNA isolation and sequencing.

Episome isolation from *P. tricornutum* was conducted as described in Diamond *et al.* (Diamond et al. 2023b), using Large Plasmid Mini Kit (Geneaid Biotech Ltd., Taiwan), with approximately 9 x 10^7 cells of *P. tricornutum*. Expression cassettes were amplified by PRIMESTAR GXL (Takara Bio) using the primer pair pPtGE30 Bb F and pPtGE30 Bb R (Table S3.1). The integrity of the episomes was verified by Sanger sequencing (Génome Québec).

DNA extraction and copy number quantification.

To quantify the plasmid copy number (PCN) in each subpopulations, DNA was extracted according to Filloramo *et al.* (2021). Briefly, cell pellets were resuspended in SDS lysis buffer (200 Mm Tris-HCl Ph 8, 250 Mm NaCl, 25 Mm EGTA, 0.5% w/v SDS) and lysed through 10 cycles of freeze and thaw. Proteinase K was added to the lysed cells and samples were incubated at 50 °C for 60 min. RNA contamination was removed by incubating the samples at 37°C for 30 min with RNase I, then DNA was extracted by adding phenol:chloroform:isoamyl alcohol (25:24:1) and recovering the aqueous phase by centrifugation. Subsequently, chloroform:isoamyl alcohol (24:1) was added and the aqueous phase was recovered again by centrifugation. DNA was precipitated with room temperature 100% isopropanol and washed using 70% ethanol. Pellets were dried in a SpeedVac Concentrator SPD1010 (Thermo Scientific, USA) and resuspended in 50µl of nuclease free water. DNA was quantified using the Nanophotometer (IMPLEN).

PCN was quantified by qPCR using the Luna Universal Master Mix (New England Biolabs) according to the manufacturer's protocol. PCN relative quantification was performed according to Lee *at al*. (Lee et al. 2006) using as reference gene Ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit *N*-methyltransferase I (Phatr3_J46871). Primers pairs binding in the *sh ble* cassette were validated using a standard curve with pDMi8 DNA (Fig. S3.3). All primers used in the qPCR are listed in supplementary table 1.

RNA extraction and sequencing

For RNA extraction, three replicates of each subpopulation were used, including a negative control, containing only the antibiotic resistance gene (*sh ble*), which went through the same process of sorting and culture. RNA was extracted using NucleoSpin RNA Plus XS (TAKARA) with minor modifications from the manufacturer's protocol as follows. Briefly, the volume of lysis buffer 1 and 2 was tripled and then split into three columns to avoid clumping the columns. After removing genomic DNA (gDNA), the technical replicates from the same biological replicates were pooled together and divided into two RNA isolation columns. RNA was eluted using 30 µL of nuclease free water and the remaining DNA was removed by using Turbo DNase (Invitrogen) with the manufacturer's protocol. The RNA integrity and the absence of gDNA was verified by migrating 200 ng of RNA per sample in an agarose gel. Samples were sequenced at the Centre d'expertise et de services Génome Québec using Illumina NovaSeq PE 100bp 25M reads with a poly-A enriched library.

Bioinformatic analyses

To analyze the transcriptome sequencing results, first adapters were trimmed and the reads were filtered according to quality, amount of uncalled (N) bases, and length using fastp (v0.23.4) with the default parameters (Chen et al. 2018). Good quality reads were mapped using HISAT2 (v 2.2.1) (Zhang et al. 2021) to the latest version of the genome of *Phaeodactylum tricornutum* (Phatr3) (Rastogi et al. 2018) downloaded from Ensembl Protists (Kinsella et al. 2011) and modified adding the transgenes *FcpC::shble::fcpC* and *40SRP8::eGFP:T2A:mCherry::fcpA*. The alignments

files were sorted and indexed using samtools (v 1.17) (Li et al. 2009). Posterior analyses were carried in R (v 4.3.1) (R Core Team 2023).

Read counts were used to calculate the transcript expression levels in transcripts per million (TPM) and genes with expression lower than the minimum expression level, calculated with the function DAFS from CustomSelection R package (v 1.1) (Dos Santos et al. 2021), were excluded from further analysis. Raw read counts of retained genes were then used for differential expression analysis using DESeq2 (v 1.40.2), using the default parameters, comparing each subpopulation against the *sh ble* sample. Genes with $|\log_2$ fold change$| > 2$ and adjusted p-value < 0.05 were considered as deregulated.

InterPro protein families annotation was extracted from Ensembl Protists (Kinsella et al. 2011) and used to identified enriched protein families in the deregulated genes using clusterProfiler (v. 4.8.3) (Wu et al. 2021). Results of different subpopulations were compared visually with Venn diagrams using ggvenn (v. 0.1.10) (Linlin 2023).

Gene co-expression analysis, without bait genes, was done to highlight specific expression patterns following the workflow from Li *et al.* (Li et al. 2023) powered by tidyverse (v2.0.0) (Wickham et al. 2019) and igraph (v2.0.2) (Csárdi et al. 2024) packages. The three biological replicates TPMs were averaged for each subpopulation, and the expression pattern was standardized by z-score. The genes with the highest variance were selected to perform gene-wise correlation, and only the statistically significant correlations with $r > 0.7$ (edges) were selected. Groups of highly correlated genes were clustered using the Leiden algorithm with a resolution of 2.5 to

detect the modules. Nested co-expression analysis was done by reducing the universe of genes to the ones previously grouped in the selected modules and repeating the analysis pipeline.

Results

Subpopulations of a heterogenic episomal expression cell line were efficiently enriched using Fluorescence Activated Cell Sorting (FACS).

We transformed wild type *P. tricornutum* by bacterial conjugation, with a construction composed of the 40SRPS8 promoter, the *eGFP* linked by the sequence encoding for the self-cleavable peptide *T2A* to *mCherry* with the *FcpA* terminator. The empty vector strain (EV), which only had the *sh ble* cassette, served as a negative control for gating to exclude cells that were not fluorescent (GFP^-) having a background of GFP positive cells (cells outside of the gate GFP negative with higher values of GFP intensity) of less than 1%. A cell line episomally expressing eGFP isolated from a single colony was phenotypically characterized by flow cytometry (Fig. 3.3.1a) showing a broad distribution of fluorescence per cell, with GFP intensity grouping in three subpopulations identified as different components by Gaussian mixture model (GMM) based clustering (Fig. 3.3.1b).

To gate the positive subpopulations, named GFP^+ and GFP^{++}, we used the naturally clustered distribution of the cells according to GFP intensity levels, background: GFP^-, medium: GFP^+, and high: GFP^{++}. These three subpopulations differed in mean fluorescence intensity (subtracting the background fluorescence from EV, ΔMFI^{GFP}) with a p-value < 0.001 (Fig. 3.3.1c). Clones contained a mixed population of 11.8-14.0% GFP^-, 39.9-42.4% of GFP^+, and 44.6-48.6% of GFP^{++} cells (Fig. S3.4).

We proceeded to enrich the subpopulations in individual cultures by FACS, as shown in Fig. 3.3.1d and e. Three weeks following sorting, GFP^{+} sorted cultures were composed of 75-88.2% of cells gated in this subpopulation with a ΔMFI^{GFP} ranging between 1659-2045 (Fig. S3.5). In the case of GFP^{++} sorted cells, the corresponding subpopulation represented 89-92% of the total population with an ΔMFI^{GFP} of 4603-4972, while 99% of GFP^{-} cultures remained GFP^{-} with ΔMFI^{GFP} of 4.3-28. The percentage of GFP^{++} subpopulations fluctuated between 5.35 to 18.2% in the GFP^{+} enriched cultures, while the ΔMFI^{GFP} did not vary with the same magnitude. This variation in percentages of GFP^{+} cells can be attributed to the use of fixed gates to analyze flow cytometry data. After sorting, the distribution of the enriched cultures was more homogeneous (Fig. 3.3.1e) compared to the non-sorted culture. GFP^{-} sorted cells were the purest subpopulation, while the GFP^{+} and GFP^{++} still contained around 5% of cells gated as GFP^{-}. Sorting accuracy could have influenced the distribution of the phenotype in the enriched cultures. Alternatively, sorted cells may revert to the non-sorted culture phenotype after several generations, or there may be dynamic fluctuations in GFP levels in the subpopulations. Our results denote that populations with distinct GFP levels can be significantly enriched using this method. The loss of GFP fluorescence was observed to occur more frequently than shifts between variation between the active GFP expressing states (GFP^{+} and GFP^{++}) (Fig. 3.3.1e). This is intriguing considering the selective pressure for the persistence of the episome, containing both *eGFP* and *sh ble,* imposed by antibiotic addition in the culture media. Thus, we successfully isolated three populations with distinct levels of GFP fluorescence with stable heterogenous phenotypes from a single colony.

The stability of the GFP$^-$ subpopulation could be due to differences in sequence.

To assess whether the differences in fluorescence were due to changes in gene sequence, we extracted the episomes from sorted cultures and amplified the expression cassettes. Amplification by PCR yielded a single band for GFP$^+$ enriched cultures and non-sorted GFP cell line, while for GFP$^-$ there were one faint band at the expected size of around 4kb and a lower band of 3.3 kb (Fig. S3.6).

Sequencing of the episomal expression cassettes showed that both GFP$^+$ and GFP^{++} enriched cultures had identical sequences to the non-sorted culture (Fig. 3.3.2), while GFP$^-$ 3.3kb sequence was the product of a deletion of 707 bp starting in *eGFP* which caused a frame shift of the *T2A* and *mCherry* sequences (Fig. S3.7). This is consistent with the GFP$^-$ phenotype and its persistence following culture after sorting. Interestingly the GFP$^-$ enriched culture also contained the 4kb band corresponding to the full construct. However, we were not able to characterize this sequence due to insufficient amount of amplified fragment and the small difference in molecular weight causing co-purification of the shorter sequence presenting the deletion. The presence of a sequence with the expected length despite having no more GFP positive cells than the EV strain (<1%), suggested that some episomes with the intact cassette did not produce the protein.

Taken together these results indicate that the GFP$^-$ phenotype was partially caused by a deletion and frame shift in the expression cassette, while differences in fluorescence intensity in GFP$^+$ were not due to modifications in the sequence. This led us to investigate other factors that could be responsible for the changes in phenotype.

Plasmid copy number differs between GFP^{++} and GFP^{-}.

To test whether the differences in GFP fluorescence intensity were due to differences in the number of copies of the episome in each subpopulation, we quantified the plasmid copy number (PCN) of the enriched cultures by qPCR. Relative PCN was significantly different between GFP^{++} compared GFP^{+} and GFP^{-} enriched cultures (Table 3. 1). GFP^{++} had an average of 1.8 copies of plasmid, which was twice the amount found in GFP^{-} (0.83) and GFP^{+} (0.89). The average PCN in the GFP^{+} subpopulation was not statistically different from GFP^{-}. Thus, asymmetrical segregation of episomes during mitosis is an additional mechanism that could explain the presence of cells with low fluorescence (GFP^{-}) or medium fluorescence (GFP^{+}) and cells that are highly fluorescent (GFP^{++}) in the same sample.

In addition, we assessed if the double copy number and the production of the expression of eGFP imposed a burden by reducing the growth rate by calculating the doubling times of each enriched culture and comparing it to the EV strain (Table 3. 2). GFP++ enriched cultures had a significantly higher doubling time (p value < 0.05) than the EV strain. However, GFP- and GFP+ enriched cultures doubling times were not significantly different.

eGFP is selectively expressed at higher level in GFP positive subpopulations.

To further characterize the enriched cultures, we analyzed differences in gene expression between the transcriptomes of the subpopulation-enriched cultures compared to the EV strain. In total we identified 388 differentially expressed genes (Fig. S3.8), among them only a few showed an increase in expression in specific subpopulations, namely 20, 12 and 20 genes were

up-regulated in GFP^-, GFP^+, and GFP^{++}, respectively. We detected 246, 241 and 226 genes that were down-regulated in GFP^-, GFP^+ and GFP^{++} subpopulations, respectively, which represented almost 2% of the expressed genes in *P. tricornutum* in our study (11, 734 expressed genes in total).

As expected, *eGFP* was up-regulated in all the subpopulations and the log_2 fold-change varied between the three subpopulation-enriched cultures. The *eGFP* expression, measured in transcript per million (TPM), correlated (ρ = 0.85 and *p*-value < 0.001) with the fluorescence intensity observed in flow cytometry (Fig. 3.3). GFP^- enriched cultures presented a lower abundance of *eGFP* transcripts compared to GFP^+ populations, but a higher level compared to EV (*eGFP* absent), possibly coming from the intact expression cassette. GFP^+ and GFP^{++} subpopulations showed higher expression of *eGFP* than GFP^- and EV. However, the difference between GFP^+ and GFP^{++} was not statistically significant.

Differences at transcriptomic level highlight enriched protein families.

Out of the 388 deregulated genes, 176 were annotated with gene ontology (GO) terms, 226 with InterPro protein families, and 382 with functional domains. Because so few GO annotations were possible, we focused on the InterPro and functional domain annotation rather than on GO terms for posterior analyses. Deregulated genes with their respective annotations are presented in Data1. We identified over-represented protein families among the down- and up-regulated genes compared to the EV strain in the enriched cultures as follows. Six protein families were over-represented across the down-regulated genes of the three enriched cultures, including

integrase, histone H2A/H2B/H3, chromo-like domain, the ribonuclease H superfamily, SWIM type zinc fingers, and reverse transcriptase (Fig. 3.4). Histone fold and HSP20-like chaperons were over-represented in the downregulated genes exclusively in the GFP^{++} subpopulation. In the case of the GFP^{+} subpopulation, the enoylreductase domain of polyketide synthase was the only over-represented protein family. The genes belonging to this family were annotated as alcohol dehydrogenases. Finally, DNA/RNA polymerase superfamily were mostly downregulated in the GFP^{-} culture, the genes Phatr3_J50124 and Phatr3_J52678 are associated to DNA repair processes and annotated with an exonuclease DNA polymerase family B domain, RNase H superfamily, and C-4 type zinc fingers.

Among the up-regulated genes, there were between 10 and 20 genes annotated with InterPro protein families per subpopulation. Therefore, protein families with a single gene annotated in a gene set were identified as enriched (Data S2). These protein families were mostly enriched in specific subpopulations. A single family, ribonuclease H superfamily, enriched for the presence of a single gene (Phatr3_EG0075), was identified in the three subcultures. The only protein family shared between GFP^{+} and GFP^{++} subpopulations was the alpha/beta hydrolase fold superfamily. However, a single gene per subpopulation was identified: Phatr3_EG00439 in GFP^{++} and Phatr3_J45633 in GFP^{+}. While Phatr3_EG00439 functional domain annotation also describes it as an acetyl xylan esterase, Phatr3_J45633 annotation is limited to the protein family level.

Coordinated expression patterns reveal candidate genes to increase recombinant protein expression in P. tricornutum.

Gene co-expression analysis was used to identify patterns of expression that could reveal factors involved in the mechanism behind distinct fluorescent phenotypes. Out of 11,734 expressed genes that we identified in our analysis, 3,908 were identified as highly interconnected and grouped in modules based on their expression patterns in each subpopulation (Fig. 3.5a). We detected 699 genes co-expressed with eGFP (module 4) which may be related to the difference in phenotype, as well as 882 genes highly expressed in GFP^{++} (module 1), 10 genes with a peak of expression in GFP^{+} (module 73), and 327 in GFP^{-} subpopulation (module 5). To break down expression patterns increasing the resolution between genes grouped in the same modules, we performed nested co-expression analysis using modules 1, 4, and 5 (Fig. 3.5b-d). We selected these modules since module 4 showed differential co-expression between GFP positive and GFP negative cells, and *eGFP* was identified among these genes, and modules 1 and 5 because they included genes with higher expression in GFP^{++} and GFP^{-}, respectively. Module 73 was not used in for this analysis because it was composed of only 10 genes (Table S3.2).

Nested analyses allow us to identify candidate genes correlated with the difference in phenotype of the subpopulations. In module 1, we found protein families related to gene expression, including winged helix DNA-binding domain superfamily, heat-shock transcription factor, and peptidase S8 like proteases (Data S3). By nesting the analysis from module 4 (Data S4), we found genes co-expressed with *eGFP* in module 4.1 (Table 3. 3), including Phatr3_J4423 annotated as histone deacetylase domain protein family. Additionally, we discovered that genes involved in fitness, ribosome biogenesis, plasmid mobilization and protein folding shared the same

expression pattern. Among them, there were two genes upregulated in GFP^{++} enriched culture, encoding for maintenance of telomere capping protein 1 (Phatr3_J45944) and DnaJ domain containing protein (Phatr3_J15138). Other genes that were in the same cluster, but not significantly up-regulated in the GFP^{++} culture, were the MbeD/MobD like protein (Phatr3_J48608), and the SAM-dependent RNA methyltransferase (Phatr3_J34157). Also, in module 4.1, we found a predicted recombinase Flp protein (Phatr3_J36840) showing increased expression in GFP^{+} enriched cultures (log_2 fold change > 1). In addition, Phatr3_J7801, a gene that encodes for a predicted nuclear transcription factor Y, gamma and belongs to the protein families Histone H2A/H2B/H3 and Histone fold, was found among the genes highly expressed in both GFP^{+} and GFP^{++} (module 4.3). Finally, in module 5 (Data S5), consisting of genes highly expressed only in the the GFP^{-} subpopulation, we found a gene annotated as glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase.

Discussion

Recently, the heterogeneity of clonally propagated cultures after transformation has been addressed in bacteria (Tomoiaga et al. 2022). Here, we showed that DNA and RNA content differs between *P. tricornutum* cells in a clonally propagated culture, giving rise to phenotypically distinct subpopulations. We enriched cultures in individual subpopulations with different GFP intensities to investigate the mechanisms of difference among cells from the same EE cell line. Our enriched GFP positive subpopulation cultures exhibited a percentage of cells outside of their gate, which can be

due to sorted cells reverting to their original phenotype after several generations. Suggesting that growing the sorted cultures for a longer time lapse could potentially yield similar phenotype to the unsorted culture. In this regard, George *et al.* compared the distribution of a *P. tricornutum* EE line expressing mVenus, before and six months after sorting and did not observe an enrichment of phenotypic populations using cell sorting (George et al. 2020a). The percentage of non-fluorescent cells (GFP^-) in both GFP^+ and GFP^{++} enriched cultures was around 5%, even though GFP^{++} carries double amount of the episome. This percentage of GFP^- cells may represent cells that downregulated the protein production, which have been described in previous studies (Diamond et al. 2023b; Faessler 2024; Fantino et al. 2024). The findings suggest that factors other than copy number, such as transcriptional regulation, episome instability, or post-transcriptional silencing, could be responsible for this shutdown.

Recombination events in *P. tricornutum* are known to accumulate over time to increase variability in diatom clonal population during mitosis (Bulankova et al. 2021). An increase in the number of haplotypes was detected by Bulankova *et al*. (2021) over a period of six months from cultures coming from a single founder cell of the diatom. Although the recombination was observed between homologous chromosomes, double-strand breaks (DSBs) occurring before the S phase are believed to be the cause of mitotic recombination in other unicellular organisms (LaFave and Sekelsky 2009). Since DSBs could also occur in episomal DNA, mitotic recombination could cause episome rearrangements explaining the deletion harbored by part of the GFP^- subpopulation, which was not detected in the unsorted culture. Mechanism of DNA repair by microhomology end joining (MHEJ) requires

the presence of small (2-70 bp) identical sequences at the junction of the deletion (Ottaviani et al. 2014). Sequence analysis revealed three identical nucleotides (GGA) at each extremity of the deletion, however only the two guanines were conserved after the repair (Fig. S3.7). MHEJ has not been characterized in the diatom, however, Matsui et al. used this mechanism in *P. tricornutum* to repair DSB induced by CRISPR-Cas9 (Matsui et al. 2024). Further characterization of MHEJ mechanism in the diatom is needed. The truncated plasmid may have been present before sorting at a very low abundance, as most cells were GFP positive, with the deletion occurring during the diatoms cell division. Interestingly, part of GFP^- subpopulation may harbor an intact expression cassette, suggesting that episomes with the intact cassette did not produce the protein, as we previously reported [21]. Alternatively, there could be mutations causing a frame shift or truncated versions of GFP which would not emit fluorescence that we were unable to identify due to the low abundance of the 4kb sequence after amplification.

While based on the literature, big deletions are not common in episomes, rearrangements have been reported by Diamond *et al.* (Diamond et al. 2023b). High strength of the promoter and plasmid copy number have been associated with the activation of multiple stress responses in the host cell, decreasing the growth rate, impairing protein production, and promoting genetic instability (Snoeck et al. 2024). Therefore, we assessed doubling time of the enriched cultures, and detected that GFP^{++} enriched cultures had significantly higher doubling times compared to the EV strain (Table 3. 2). This suggests a metabolic burden caused by producing a higher amount of eGFP. Since the GFP^+ cultures had similar doubling time to the EV strain

we cannot hypothesize that the strength of the 40SRPS8 promoter forced mutations to occur. Nevertheless, Diamond and colleagues showed instability in episomes in *P. tricornutum* where the expression of the gene of interest was driven by the same promoter sequence (Diamond et al. 2023b).

It has been reported that the T2A self-cleavable peptide efficiency is not 100% in *P. tricornutum,* which could lead to subpopulations with varying proportions of free eGFP (Slattery et al. 2018; Baiden et al. 2023; Diamond et al. 2023b). The self-cleavage mechanism is known to involve ribosome skipping, with the possibility of ribosome fall-off and discontinued translation after both successful and unsuccessful skipping events (Liu et al. 2017). In this study, we did not analyze the efficiency of T2A cleavage as potential factor contributing to the clustering of three distinct subpopulations. A difference in T2A cleavage efficiency would be expected to affect mCherry fluorescence differently than eGFP, but this was not observed. In addition, the GFP cell lines were maintained in suspension culture for a year, during which mutation events likely gave rise to phenotypic subpopulations varying in PCN and sequence.

In addition, GFP^- and GFP^+ subpopulations contained two-fold less plasmid copy number compared to GFP^{++} subpopulation, indicating that alterations in episome replication or asymmetric segregation of the episomes during mitosis could contribute to phenotypic differences. Our results of average PCN are relative to a single copy gene from *P. tricornutum*, which is a diploid organism, thus having a PCN equal to one represents two copies of plasmid per cell. A PCN of less than one in GFP^- and GFP^+ subpopulations may be due to asymmetric segregation causing part of the subpopulation to have

only one copy per cell. Segregation of episomes has not been yet characterized in *P. tricornutum*. However, plasmid maintenance and distribution mechanisms have been described for other model organisms used to produce recombinant proteins and metabolites, such as *E. coli* and *Saccharomyces cerevisiae*, the latest being more likely to mirror the diatom's mechanisms because of their eukaryotic nature. In bacteria plasmid segregation may occur by different mechanisms, low copy number plasmids transfer to daughter cells by partitioning systems, high copy number plasmids by random segregation, and post-segregational killing eliminates plasmid-free progeny (Million-Weaver and Camps 2014). Yeast episomal plasmids, such as the 2-micron, segregate by chromosome "hitchhiking": localizing at the telomeres of sister chromatids during mitosis to utilize spindle forces and localize to opposite cell poles (Sau et al. 2019). In *S. cerevisiae*, centromeric plasmids containing centromeric (CEN) sequence and autonomously replicating sequence (ARS) elements segregate using chromosome machinery (Gnügge and Rudolf 2017). Centromeric plasmid copy number can vary in a population, averaging one copy per haploid genome (Tschumper and Carbon 1983). The variation primarily results from asymmetric plasmid segregation and, less frequently, plasmid replication failure. Although the pDMi8 plasmid used in this study harbors centromeric and autonomous replicating sequences (*CEN6-ARSH8*) from yeast, little is known about how these sequences function in *P. tricornutum*. It has been reported that the *CEN6-ARSH4-HIS3* region of the episome associates to the centromeric histone 3 variant in *P. tricornutum* (Diner et al. 2017). This suggests that foreign DNA sequences can recruit native diatom machinery for DNA replication and episome

maintenance. If centromeric plasmid pDMi8 copy number dynamics are similar to yeast (Tschumper and Carbon 1983; Karas et al. 2015), we could hypothesize that GFP^{+} was the original population (PCN ~ 1) after conjugation, and that mutation events caused aberrations in copy number gave rise to the GFP^{-} and GFP^{++} subpopulations. Studies are needed to test this hypothesis using different cell lines to characterize the frequency of this event. Additionally, identifying the mechanisms of plasmid segregation in *P. tricornutum* could expand the synthetic biology toolkit to control gene expression by creating tunable plasmid copy number systems, as developed for *E. coli* (Rouches et al. 2022).

Transcriptomic analysis revealed that only a small proportion (3%) of the total number of genes were deregulated across the three subpopulations. While the expression levels of *sh ble* did not vary between samples, *eGFP* expression levels were higher in GFP^{+} and GFP^{++} subpopulations compared to GFP^{-} and EV (Fig. 3.3.2). This discrepancy is likely due to the sample point being selected based on GFP phenotype during the early exponential phase. Since the expression of *sh ble* gene is driven by a different promoter than *eGFP*, its expression may differ at other stages of the culture. Additionally, *FcpC* promoter (*sh ble*) is known to have a lower relative expression compared to *40SRPS8* (*eGFP*) [11].

Among the over-represented protein families (Fig. 3.4) shared between the down-regulated genes of the three enriched cultures, those associated with integrase, chromo-like domain, ribonuclease H superfamily, SWIM type zinc fingers and reverse transcriptase are linked to transposon elements (Novikova 2009; Liu and Wessler 2017). Moreover, HSP20-like chaperones were enriched in GFP^{++} down-regulated genes. These chaperones

accumulate in *P. tricornutum* under stressful culture conditions, such as nitrogen depletion or activation of stress response pathways mediated by the signaling nucleotides guanosine penta- and tetraphosphate ((p)ppGpp) (Longworth et al. 2016; Avilan et al. 2021). However, down-regulation of these chaperones has not been linked to a specific condition in *P. tricornutum*. The DNA/RNA polymerase superfamily, associated to DNA repair processes, was over-represented in downregulated genes in the GFP^- enriched culture. Thus, down-regulation of these genes could be related to the mutated expression cassette in GFP^- enriched cultures.

Co-expression analysis identified genes with expression patterns similar to that of *eGFP*, potentially related to the high intensity fluorescent phenotype. Two genes were upregulated in GFP^{++} enriched culture: one annotated as ‘’maintenance of telomere capping protein 1’’ and another as ‘’DnaJ domain containing protein’’. The former has been shown to increase fitness in telomere capping (*Cdc13*) protein yeast mutants (Addinall et al. 2008), while the latter is known to play a role in bacteria, yeast and mammals in protein translation, folding, translocation, and degradation by stimulating the ATPase activity of chaperone proteins (Qiu et al. 2006). A MbeD/MobD-like protein encoding gene clustered with *eGFP* but was not significantly up-regulated in GFP^{++} subpopulations. This protein is involved in entry exclusion mechanism of the ColE1 plasmid family during plasmid transfer through conjugation in *Citrobacter*, decreasing plasmid transmissibility (Zharikova et al. 2018). Also in this cluster was the SAM-dependent RNA methyltransferase, involved in rRNA methylation during ribosome biogenesis in humans (Shen et al. 2020). A predicted recombinase Flp protein, responsible for plasmid amplification and

preventing copy number drop of the 2-micron plasmid in yeast (Sau et al. 2015), was co-expressed with *eGFP*. Plasmid amplification in *S. cerevisiae* is triggered by a Flp-mediated recombination event during bi-directional replication, reconfiguring the replication mode into double uni-directional forks that produce plasmid copies in tandem (Ma et al. 2019). To restore the replication fork movement and complete amplification, a second recombination event is needed, with the plasmid separating into monomers by either Flp-mediated or homologous recombination (Ma et al. 2019). It remains unclear if *Flp* overexpression could be related to the higher copy number found in GFP^{++}, since pDMi8 does not contain the recombination elements of the yeast 2-micron plasmid. Further investigation is needed to determine whether *Flp* overexpression in *P. tricornutum* could increase episome copy number as well as to characterize the recombinase recognition sites and its transcriptional regulators. Additionally, episomal constructs and CRISPR-Cas technologies could be used to overexpress or knock-out the gene annotated as *Flp recombinase* in *P. tricornutum* cell lines to analyze its impact on plasmid copy number.

Finally, for genes highly expressed only in the GFP^{-} subpopulation, we identified a gene annotated as “glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase”. This gene could be related to resistance to zeocin, since it is a member of the pleomycin/bleomycin antibiotic family (Buck et al. 2018), potentially compensating for lower copy number of *sh ble* in GFP^{-} compared to GFP^{++} subpopulations.

Taken together, these results highlight candidate genes for enhancing recombinant proteins production in *P. tricornutum* episomal expression system. The speculative mechanisms based on these observations are

shown in Fig. 3.6. This study contributes to diatom synthetic biology by elucidating genetic differences between cells in clonally propagated cultures and linking to phenotype. A more in-depth characterization of plasmid dynamics in *P. tricornutum* is needed to manipulate and expand the genetic toolbox of this model organism.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. Raw reads from RNA-seq are available at NCBI Sequence Read Archive BioProject ID PRJNA1108718.

Competing interests

The authors declares that they have no competing interests.

Funding

We acknowledge that financial support for this study was funded by the Canada Research Chair on plant specialized metabolism Award No CRC-2018-00137 to I.D-P. Thanks are extended to the Canadian taxpayers and to the Canadian government for supporting the Canada Research Chairs Program. Additional support in the form of scholarship to A.M.D-G. was

provided by Mitacs-Acceleration program grant #IT12310 and to K.C.G.d.S by Mitacs-Elevate fellowship.

Acknowledgements

Warm thanks to Andrew Diamond for lab training and helpful discussions. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and the Digital Research Alliance of Canada (www.alliancecan.ca/en).

Authors' information

Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

Aracely Maribel Diaz-Garza, Natacha Merindol, Karen Cristine Gonçalves dos Santos, Félix Lavoie-Marchand, Isabel Desgagné-Penix

Department of applied mathematics, University of Waterloo, Waterloo, ON, Canada.

Brian Ingalls

Tables

Table 3. 1. Plasmid copy number (PCN) of subpopulations is significantly different. For normalization we used the EV strain's single copy gene of the ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit *N*-methyltransferase I and the *sh ble* gene. Statistical comparisons were made using one-way ANOVA and Tukey post-hoc test. Letters denote statistically significant differences between samples with p value < 0.001.

Sample	**PCN*** **$2.04^{-\Delta\Delta Ct}$**
GFP $^{-}$	0.83 ± 0.038^{a}
GFP $^{+}$	0.89 ± 0.066^{a}
GFP $^{++}$	1.80 ± 0.147^{b}

* Average ± S.D. (n = 3)

Table 3. 2. Doubling time of enriched cultures is only significantly higher in GFP^{++}. * denote statistically significant differences between samples with p value < 0.05 by Student's t-test. Doubling times are presented in as average ± S.D. (n = 3)

Sample	**Doubling time (h)**
EV	27.74 ± 4.97
GFP^{-}	33.71 ± 2.65
GFP^{+}	27.67 ± 1.07
GFP^{++}	39.58 ± 4.93*

Table 3. 3. Complete list of genes following similar expression pattern with *eGFP* clustered in module 4.1.

Gene ID	**Functional domain**
EPrPhatr3G00000013123	NA
Phatr3_EG00208	Exostosin family;
Phatr3_EG00439	Acetyl xylan esterase (AXE1);
Phatr3_EG00504	Fibrinogen binding protein;
Phatr3_EG01949	Putative glutamine amidotransferase ;
Phatr3_EG01955	Carboxyl transferase domain; Biotin-requiring enzyme; Carbamoyl-phosphate synthase L chain, ATP bi nding domain;Biotin carboxylase, N-terminal domain; Acetyl-CoA carboxylase, central region;
Phatr3_EG01993	Endomembrane protein 70;
Phatr3_EG02214	P-loop containing dynein motor region D3; Radical SAM superfamily; AAA domain (dynein-related subfamily); AAA domain (dynein-related subfamily);

Phatr3_EG02258	GET complex subunit GET2; Steroid receptor RNA activator (SRA 1);
Phatr3_EG02486	Protein of unknown function (DUF1295);
Phatr3_J14176	PQQ enzyme repeat;WD domain, G-beta repeat;PQQ-like domain;
Phatr3_J14327	NLI interacting factor-like phosphatase;
Phatr3_J15138	X-domain of DnaJ-containing; DnaJ domain;
Phatr3_J19329	Carboxyl transferase domain;
Phatr3_J228	haloacid dehalogenase-like hydrolase; Cation transporting ATPase, C-terminus; Cation transporter/ATPase, N-terminus;
Phatr3_J23658	Flavodoxin;
Phatr3_J29658	Oxidoreductase FAD-binding domain; Oxidoreductase NAD-binding domain ;
Phatr3_J29660	Oxidoreductase NAD-binding domain;

	Oxidoreductase FAD-binding domain;
Phatr3_J34157	SAM-dependent RNA methyltransferase;
Phatr3_J36840	Recombinase Flp protein;
Phatr3_J37425	ZIP Zinc transporter;
Phatr3_J39006	Haloacid dehalogenase-like hydrolase;
Phatr3_J39019	tRNA (Guanine-1)-methyltransferase;
Phatr3_J43348	GyrI-like small molecule binding domain;
Phatr3_J4423	Ankyrin repeat; Ankyrin repeats (many copies);Histone deacetylase domain;
Phatr3_J44262	Protein of unknown function (DUF3619);
Phatr3_J44680	THRAP3/BCLAF1 family;
Phatr3_J45031	P-loop ATPase protein family;
Phatr3_J45324	Transport protein Avl9;
Phatr3_J45341	ST7 protein;
Phatr3_J45944	Maintenance of telomere capping protein 1;
Phatr3_J46275	Fatty acid desaturase;
Phatr3_J46345	Fibronectin type I domain;

Phatr3_J47842	Glycosyl transferase 1 domain A;
Phatr3_J48565	Protein of unknown function (DUF4551);
Phatr3_J48608	MbeD/MobD like;
Phatr3_J49986	Low iron-inducible periplasmic protein;
Phatr3_J49991	Sulfotransferase domain;
Phatr3_J50093	Domain of unknown function (DUF3402);
Phatr3_J50187	STAS domain;
gfp_t2a_mcherry	eGFP:T2A:mCherry

Figures

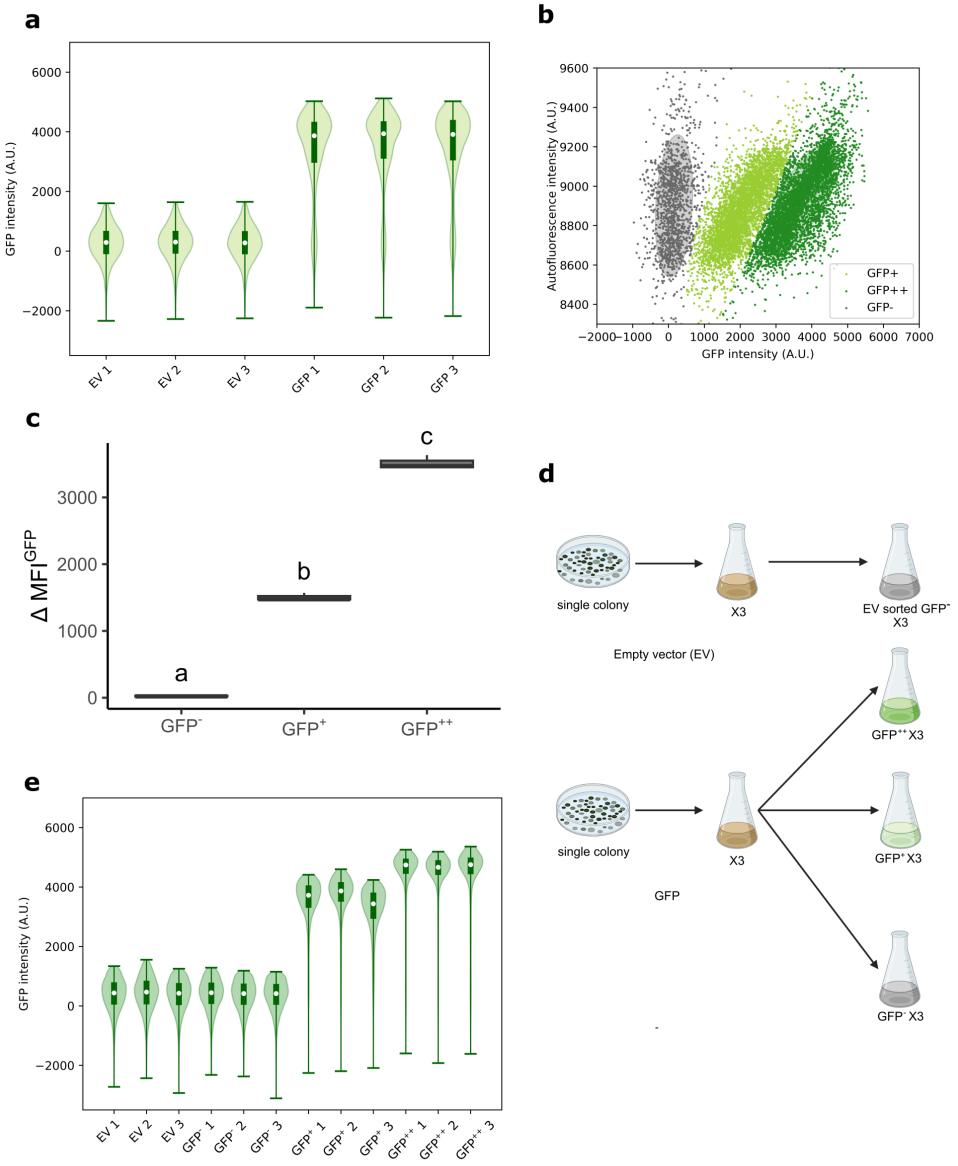


Fig 3. 1. Subpopulations are efficiently enriched through cell sorting. **(a)** Violin plots of per cell GFP fluorescence intensity of exconjugants harboring *eGFP* (GFP 1-3) and empty vector (EV 1-3) analyzed by flow cytometry. **(b)** Scatter plot of GFP cell line (GFP 1) before sorting with three populations grouping separately in GMM-based clustering (GFP^-, GFP^+, and GFP^{++}); ellipsoids indicate the confidence region for each cluster. **(c)** Box plot of the mean fluorescence intensity of eGFP subtracting the background fluorescence from the EV strain (ΔMFI^{GFP}) for each

subpopulation, letters denote distinct significance with a p-value < 0.001.**(d)** Scheme showing the sorting strategy (created with BioRender.com). **(e)** Violin plots after sorting in GFP^- and GFP^+ and GFP^{++}. The median (n= 10,000 cells) is indicated by a white dot. Fluorescence intensity is presented as a corrected measurement in arbitrary units (A.U.).

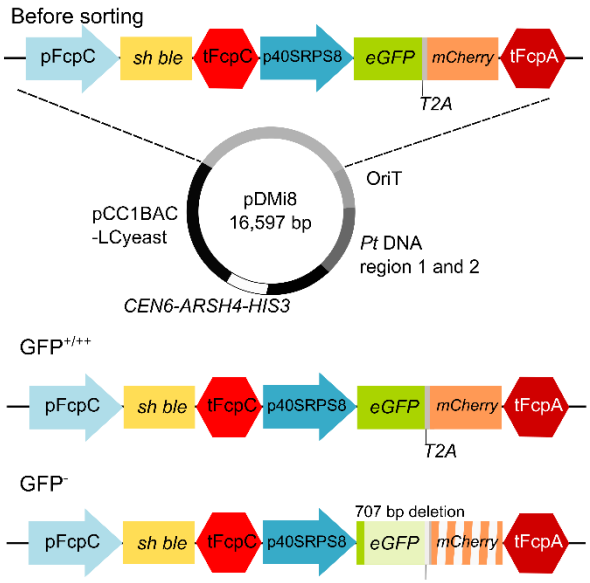


Fig 3. 2. Enriched GFP$^-$ subpopulation exhibits episome rearrangement. Schematic representation of expression cassettes sequenced from the pDMi8 before and after sorting. The 707 bp deletion causes a frame shift of the expression cassette.

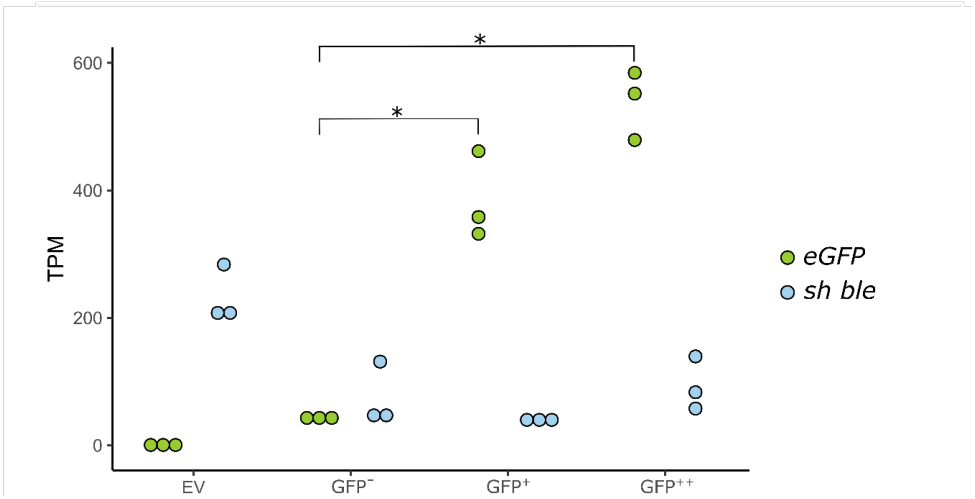


Fig 3. 3. *eGFP,* but not *sh ble,* is expressed at higher level in GFP positive subpopulations. Dot plot showing differences in *eGFP* expression between sorted subpopulations measured in transcript per million (TPM). Statistically significant differences were calculated using Kruskal-Wallis and Dunn's *post-hoc* test; * denotes *p*-value < 0.05.

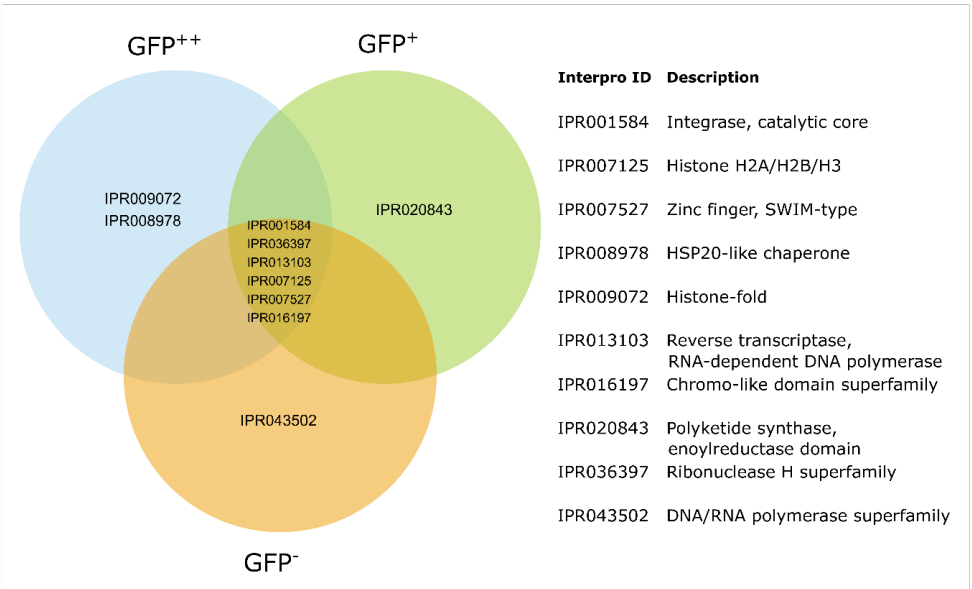


Fig 3. 4. Most of the protein families enriched in the down-regulated genes are shared between the three subpopulations. Venn diagram representing the overlap between subpopulations of InterPro protein families enriched in down-regulated genes compared to empty vector strain (EV). Genes down-regulated correspond to log fold-change < -2 and an adjusted *p*-value < 0.05.

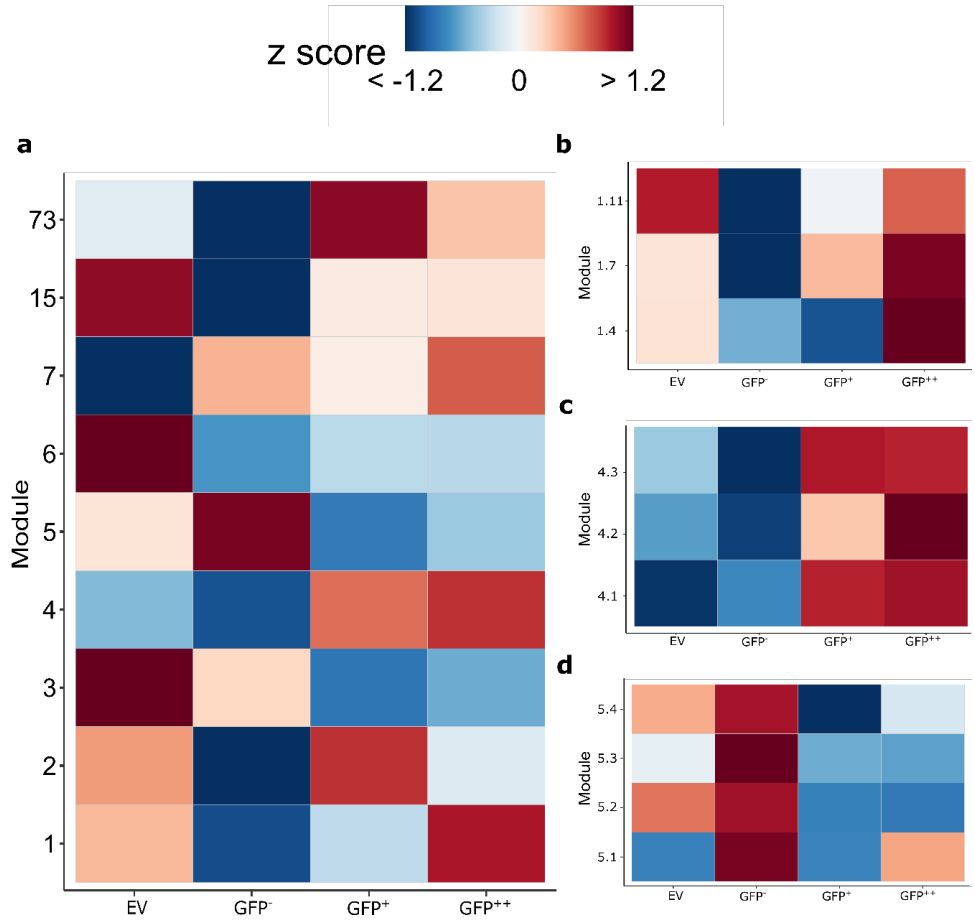


Fig 3. 5. Nested co-expression analysis highlights specific expression patterns in the subpopulations. Co-expression analysis of (**a**) all the highly correlated genes expressed in *P. tricornutum* enriched cultures, (**b**) genes from module 1, (**c**) genes from module 4, and (**d**) genes from module 5.

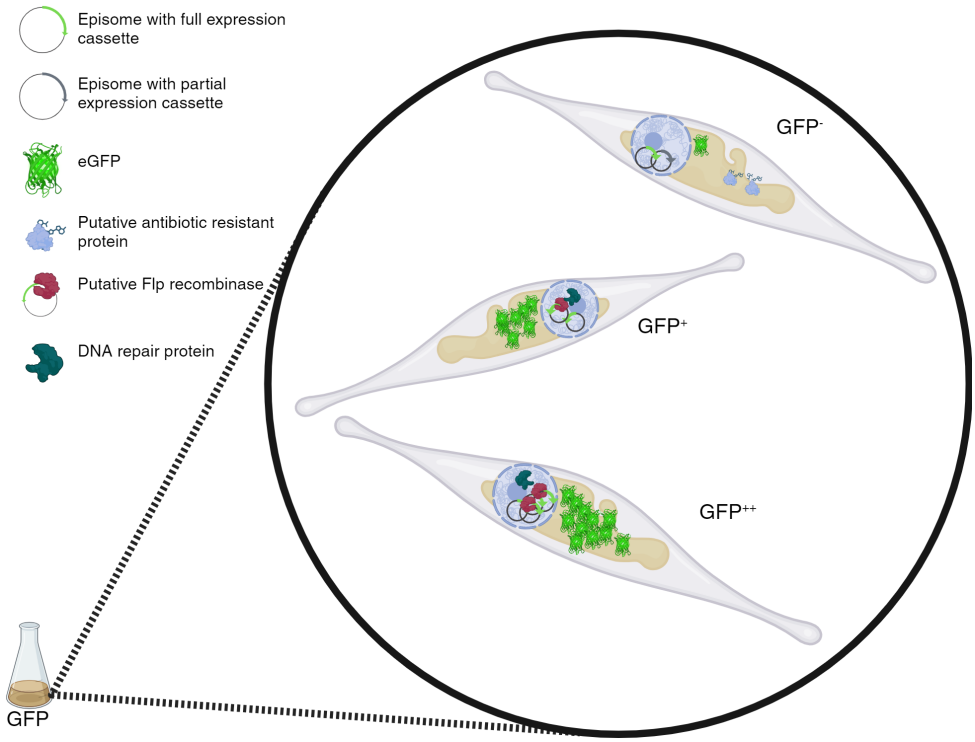


Fig 3. 6. Speculative mechanisms of *P. tricornutum* cells in a clonally propagated culture creating differences in genetic content and gene expression. Schematic representation of the proposed mechanisms causing phenotypic subpopulation in a single culture: GFP^- down-regulation of DNA repair proteins may lead to a partial loss due to mutation, differences in PCN could be caused by prevention of copy number drop with higher expression of *Flp recombinase*, GFP^- may compensate lower copy number of *sh ble* by overexpressing a native protein conferring antibiotic resistance. Created with Biorender.com.

CHAPTER IV: DISCUSSION AND CONCLUSIONS

Investigation of the long-term stability of transgene expression in microalgae is still needed since robustness, reproducibility, and reliability are required for large-scale production. Here, we maintained stable replication of plasmids and confirmed subcellular localization in the targeted compartment of *P. tricornutum* cells with low accumulation of *Vp*VAN, which was under the detection limits of western blot. Flow cytometry analysis showed that even after enrichment by cell sorting, cells downregulated the production of recombinant protein. A possible explanation for low levels of recombinant *Vp*VAN is that its production may be a metabolic burden for the cell, and it reduces its expression to avoid hampering growth. The impact on the host metabolism when producing a heterologous enzyme has been assessed in *E. coli*. Rajacharya and colleagues expressed the acyl-ACP carrier protein reductase (AAR) of cyanobacteria in *E. coli* and used it as a reference recombinant protein (Rajacharya et al. 2024). They used a proteomic and metabolomic approach to gather insights into the dynamic response of the metabolism to produce recombinant AAR. They identified transcription as the primary factor contributing to metabolic burden with an effect of growth retardation (Rajacharya et al. 2024).

The permissibility of *P. tricornutum* to acquire genes in the nucleus may require the diatom to downregulate their transcription or translation when they represent a burden. In our other study, we mentioned that the transcript levels between two distinct fluorescent populations (GFP^{++} and GFP^{+}) with different plasmid copies were not statistically significant. These results support the hypothesis of regulation of transcript in *P. tricornutum*. A study in yeast revealed that upon the induction of expression of a

phosphoribulokinase (PRKp), the plasmid copy number of CEN-ARS-constructs stayed identical. However, the enzyme activity was lower at higher inducer concentrations (Rouzeau et al. 2018). Rouzeau and colleagues did not measure the amount of recombinant protein. Still, since their enzymatic assays were done with total protein extract, PRKp may have been produced less even after induction.

Our approach to compartmentalize the *Vp*VAN reaction to identify the effect in the activity and increase vanillin accumulation was made to take advantage of organelle barriers. Organelles in eukaryotes provide a manner to sequester specialized processes incompatible with the rest of intracellular metabolism (Tan et al. 2021). Compartmentalization increases the efficiency of metabolic pathways by providing close proximity to the functional components of the pathway and separating them from competing processes (Zecchin et al. 2015; Bar-Peled and Kory 2022). Successful compartmentalization of the amino acid biosynthesis in the peroxisome *Schizosaccharomyces japonicus* facilitated the adaptation of its metabolism to nutrient-poor conditions (Gu et al. 2023). In addition, glycosylation allows the accumulation of phenolic compounds to maintain phenol homeostasis in the cell (Brillouet et al. 2014b).

Regarding the controversy of *Vp*VAN, we must acknowledge that the only evidence of activity provided by authors other than Gallage and colleagues is in plant cells that already produce the compound. Therefore, the cysteine protease-like protein may assist the chain-shortening reaction as part of a complex (Havkin-Frenkel and Belanger 2018). Another possible role for *Vp*VAN proposed by Havkin-Frenkel and Belanger is that it could trigger proteolysis activating enzymes for vanillin synthesis.

Using co-expression analyses, our study of different phenotypic subpopulations in a single culture aimed to identify genes involved in the mechanism originating the subpopulations. However, we encountered difficulties due to a lack of protein characterization and genes missing annotations. Therefore, our analyses were limited to the available information. We were interested in finding insight into how plasmid copy number doubled in a population to take advantage of native mechanisms of *P. tricornutum* to tune plasmid copy number and heterologous expression. It was previously reported that episomes are kept at the native chromosomal number in P. tricornutum (Karas et al. 2015). This supports our hypothesis that the GFP++ subpopulation originated from GFP+ by an aberration in copy number maintenance mechanisms. In our analysis, we identified a co-expressed gene annotated as “Flp recombinase,” which prevents copy number drop in yeast. Yeast diploid cells adapt to burden by plasmid loss, while haploid strains maintain a constant plasmid copy number in a minimal medium (Kastberg et al. 2022).

The acquisition of zeocin resistance in *P. tricornutum* GFP^- subpopulation and posterior silencing of the plasmid would require the levels of transgene expression of the heterologous sh ble protein to be significantly lower in the GFP^- subpopulation. In the enriched cultures, we identified a gene annotated as “glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase” that could be related to resistance to zeocin, highly expressed only in the GFP^- subpopulation. This result should be taken cautiously since this predicted protein lacks characterization. While antibiotic resistance occurs as a global response of a population, antibiotic persistence in bacteria is characterized by the survival of a subpopulation to high

concentrations of antibiotics (Huemer et al. 2020). Stochastic expression of genes that could confer resistance to a subpopulation has been hypothesized to play a role in persistent and non-persistent states (Avery 2006). However, since this subpopulation was not susceptible to zeocin once it was sorted, we ruled out a state of persistence involved in the heterologous phenotypes, leaving unclear the involvement of the endogenous candidate gene.

Diatoms epigenetic processes have been proposed to mediate responses in highly dynamic environments significantly (Tirichine et al. 2017). While in our study, we did not explore epigenetic marks differentiating the subpopulations, DNA methylation and H3K4me2 marks in transgene insertion sites have been studied as causes for the apparition of subpopulations in genome-integrated transgenic *P. tricornutum* cell lines. *P. tricornutum* exhibits low methylation levels in transgene insertion sites, with no difference between fluorescent and non-fluorescent subpopulations (Faessler 2024). Likewise, H3K4me2 mark levels were not found to be different between fluorescent and non-fluorescent subpopulations (Faessler 2024). However, it does not rule out the possibility of involvement of other histone marks.

4.1 Perspectives

P. tricornutum has the potential to be a good platform to produce phenolic aldehydes since neither vanillin nor ferulic acid slowed its growth. However, efforts should be directed to understand the metabolism of these compounds in the diatom. Ferulic acid, vanillic acid, caffeic acid, and protocatechuic acid have been identified at trace levels in the diatom

(Haoujar et al. 2019; Rico et al. 2013). Protocatechuate is likely to come from the degradation of shikimate, but the formation of ferulate is unclear. Investigating the vanillin biosynthetic pathway will provide a clearer picture of vanillin production. It will improve its commercial production to meet the global market demand of the pharmaceutical and food industry (Kundu 2017).

To understand how heterologous protein production is governed in *P. tricornutu*m, a study of the metabolic burden imposed by the transgene expression should be conducted. Using media with lower contents of nitrogen and phosphate could allow the identification of factors that contribute to the burden. In addition, a study following the growth of the transgenic strains after multiple subcultures could establish a threshold where downregulation of recombinant protein could occur. This would be based on the hypothesis that the diatom is able to downregulate the production of recombinant protein to accelerate its growth in presence of antibiotic. The pleomorphism of *P. tricornutum* cell is thought to be the reason they can easily adapt to changing environments. Through-out the conjugation, the cell goes from oval to fusiform. This change of shape could change the production of the recombinant genes. Oval morphotype has been investigated for the secretion of recombinant proteins by transcriptomic and proteomic analysis, even though it is a morphotype believed to help the cell cope with stresses and it could be possible that the protein production of non-indispensable proteins is downregulated. Further studies could be done to maintain the diatom in oval morphotype by media limitation and observe if it impacts the production of the recombinant protein.

As stated before, we encountered limitations in our analysis caused by the insufficient number of genes characterized from the genome *P. tricornutum* genome. Efforts should be directed to the characterization of genes and proteins of the model diatom. Large-scale tools have been developed to bridge the gap between sequencing and functional characterization. Loss-of-function is a common approach to assess the phenotype of an unknown protein. High throughput methods of this approach include the use of Transposon mutagenesis followed by sequencing (TnSeq), measuring mutant phenotypes genome-wide in one experiment (van Opijnen et al. 2009). The coupling of TnSeq with random DNA barcoding (RB-TnSeq) has been used to identify the mutant phenotypes of 11,779 protein-coding genes of unknown function (Price et al. 2018). However, since *P. tricornutum* is a diploid organism, after performing transposon mutagenesis the mutant strains will likely recover the wild-type phenotype unless homologous end joining based DNA repair mechanism is impaired.

In addition, the generation of gain-of-function overexpression libraries is a complementary approach to analyze the phenotype of increasing the gene copy number. A high-throughput technology was developed to avoid the necessity of organism-specific genetic tools, called dual-barcoded shotgun expression library sequencing (Dub-seq) (Mutalik et al. 2019). This technique was validated by overexpressing nearly all genes on *E. coli* fitness under 52 experimental conditions, showing consistent results with known biological functions and provided novel insights. This high-throughput approach could be implemented to investigate the function of *P. tricornutum* poorly annotated genes, because it only requires DNA and a suitable host organism for assessing the fitness. Due to the complex

evolutionary history of the diatoms, a combination of experiments in multiple host organisms could provide validation to gene functions, as well as implementation of loss-of-function and overexpression in *P. tricornutum* using a fewer number of genes.

Another aspect that requires further studies is the aberration in plasmid copy number. We could study their occurrence comparing with different cell lines and identify mutations in the host genome of *P. tricornutum* subpopulations to pinpoint genes involved in this. CRISPR/Cas could be used to knock out the gene annotated as “glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase” for its characterization. Alternatively, we could use auxotrophic strains and markers instead of zeocin, since uracil and histidine auxotroph strains are available for *P. tricornutum* (Slattery et al. 2020). In order for Flp to be responsible of increasing plasmid copy number in *P. tricornutum*, specific recognition sites would need to be located in the plasmid. Validation of this hypothesis would require overexpression and/or knock-out of Flp, and *in vitro* characterization of the recombinase. Identification of the mechanism causing copy number increase in subpopulations of *P. tricornutum* could expand the genetic toolbox of the model diatom, allowing to tune plasmid copy number.

4.2 Conclusion

While *Vp*VAN was stably replicated in *P. tricornutum* episome and localized in targeted compartments, low protein production levels did not allow detection by western blot. Flow cytometry analysis revealed low percentage of fluorescent cells even after enrichment. This led to the study of differences between stable subpopulations using a cell line expressing GFP

episomally. Our study of phenotypic subpopulations showed that plasmid copy numbers can vary between them. However, the mechanisms causing this plasmid copy aberration are still unclear. We also identified a specific mutation in the episome of the non-fluorescent subpopulation. Nonetheless, sequencing of plasmids coming from a majoritarian non-fluorescent population of *Vp*VAN expressing lines was absent of mutations in gene sequences in the cassettes. We could conclude that there are different mechanisms governing the reduction of transgene expression and that the mutation of the expression cassettes can be one of them. However, our results also indicate that post-transcriptional regulation is likely to be involved in the apparition of the non-fluorescent subpopulations. Findings and hypotheses of causes of heterogeneity in *P. tricornutum* EE cell lines generated by conjugation are summarized in Figure 4.1. This research improves our understanding of synthetic biology of diatoms and highlights how crucial it is to characterize the cell response after introduction of foreign DNA in *P. tricornutum*.

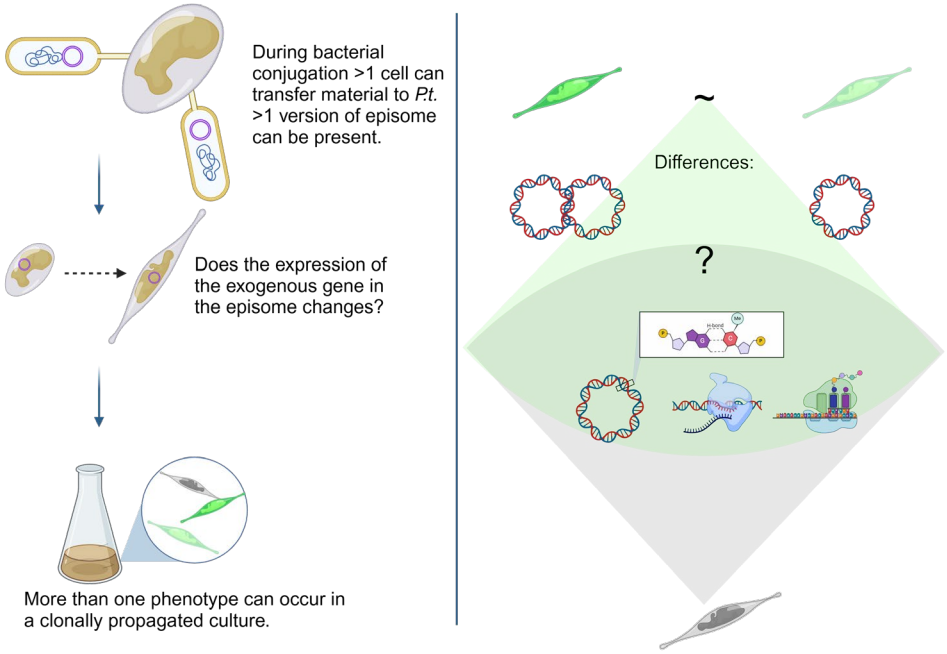


Figure 4. 1. Summarized findings of heterologous gene expression in cell lines originated from exconjugants of *P. tricornutum*. Different intensities of green are used to represent levels of eGFP accumulation. Created in Biorender.com.

REFERENCES

Adams BL. The Next Generation of Synthetic Biology Chassis: Moving Synthetic Biology from the Laboratory to the Field. ACS Synth Biol. 2016:**5**(12):1328–1330. https://doi.org/10.1021/acssynbio.6b00256

Addinall SG, Downey M, Yu M, Zubko MK, Dewar J, Leake A, Hallinan J, Shaw O, James K, Wilkinson DJ, et al. A Genomewide Suppressor and Enhancer Analysis of cdc13-1 Reveals Varied Cellular Processes Influencing Telomere Capping in Saccharomyces cerevisiae. Genetics. 2008:**180**(4):2251–2266. https://doi.org/10.1534/genetics.108.092577

Adler-Agnon Z, Leu S, Zarka A, Boussiba S, and Khozin-Goldberg I. Novel promoters for constitutive and inducible expression of transgenes in the diatom Phaeodactylum tricornutum under varied nitrate availability. J Appl Phycol. 2018:**30**(5):2763–2772. https://doi.org/10.1007/s10811-017-1335-8

Ahmed M, Seraj R, and Islam SMS. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics. 2020:**9**(8):1295. https://doi.org/10.3390/electronics9081295

Alam K, Mazumder A, Sikdar S, Zhao Y-M, Hao J, Song C, Wang Y, Sarkar R, Islam S, Zhang Y, et al. Streptomyces: The biofactory of secondary metabolites. Front Microbiol. 2022:**13**. https://doi.org/10.3389/fmicb.2022.968053

Alberti F, Foster GD, and Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Appl Microbiol Biotechnol. 2017:**101**(2):493–500. https://doi.org/10.1007/s00253-016-8034-2

Albuquerque BR, Heleno SA, Oliveira MBP, Barros L, and Ferreira ICF. Phenolic compounds: current industrial applications, limitations and future challenges. Food Funct. 2021:**12**(1):14–29. https://doi.org/10.1039/D0FO02324H

Amendola S, Kneip JS, Meyer F, Perozeni F, Cazzaniga S, Lauersen KJ, Ballottari M, and Baier T. Metabolic Engineering for Efficient Ketocarotenoid Accumulation in the Green Microalga Chlamydomonas reinhardtii. ACS Synth Biol. 2023:**12**(3):820–831. https://doi.org/10.1021/acssynbio.2c00616

Anand A, Khurana R, Wahal N, Mahajan S, Mehta M, Satija S, Sharma N, Vyas M, and Khurana N. VANILLIN: A COMPREHENSIVE REVIEW OF PHARMACOLOGICAL ACTIVITIES. Plant Arch. 2019:**19 (2)**:1000–1004.

Angstenberger M, Krischer J, Aktaş O, and Büchel C. Knock-Down of a ligIV Homologue Enables DNA Integration via Homologous Recombination in the Marine Diatom Phaeodactylum tricornutum.

ACS Synth Biol. 2019:**8**(1):57–69. https://doi.org/10.1021/acssynbio.8b00234

Arentshorst M, Lagendijk EL, and Ram AF. A new vector for efficient gene targeting to the pyrG locus in Aspergillus niger. Fungal Biol Biotechnol. 2015:**2**(1):2. https://doi.org/10.1186/s40694-015-0012-4

Arias CAD, Oliveira CFM de, Molino JVD, Ferreira-Camargo LS, Matsudo MC, and Carvalho JCM de. Production of Recombinant Biopharmaceuticals in Chlamydomonas reinhardtii. Int J Plant Biol. 2023:**14**(1):39–52. https://doi.org/10.3390/ijpb14010004

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science. 2004:**306**(5693):79–86. https://doi.org/10.1126/science.1101156

Arya SS, Mahto BK, Sengar MS, Rookes JE, Cahill DM, and Lenka SK. Metabolic Engineering of Rice Cells with Vanillin Synthase Gene (VpVAN) to Produce Vanillin. Mol Biotechnol. 2022:**64**(8):861–872. https://doi.org/10.1007/s12033-022-00470-8

Ashikaga H, Ben-Yehuda O, and Chien KR. chapter 1 - Biotechnology and Cardiovascular Medicine: Recombinant Protein Therapy. . In. Molecular Basis of Cardiovascular Disease (Second Edition), KR Chien, ed. (W.B. Saunders: Philadelphia), pp. 1–15. https://doi.org/10.1016/B978-0-7216-9428-3.50006-8

Aubry C, Pernodet J-L, and Lautru S. Modular and Integrative Vectors for Synthetic Biology Applications in Streptomyces spp. Appl Environ Microbiol. 2019:**85**(16):e00485-19. https://doi.org/10.1128/AEM.00485-19

Avery SV. Microbial cell individuality and the underlying sources of heterogeneity. Nat Rev Microbiol. 2006:**4**(8):577–587. https://doi.org/10.1038/nrmicro1460

Avilan L, Lebrun R, Puppo C, Citerne S, Cuiné S, Li-Beisson Y, Menand B, Field B, and Gontero B. ppGpp influences protein protection, growth and photosynthesis in Phaeodactylum tricornutum. New Phytol. 2021:**230**(4):1517–1532. https://doi.org/10.1111/nph.17286

Awwad F, Fantino EI, Héneault M, Diaz-Garza AM, Merindol N, Custeau A, Gélinas S-E, Meddeb-Mouelhi F, Li J, Lemay J-F, et al. Bioengineering of the Marine Diatom Phaeodactylum tricornutum with Cannabis Genes Enables the Production of the Cannabinoid Precursor, Olivetolic Acid. Int J Mol Sci. 2023:**24**(23):16624. https://doi.org/10.3390/ijms242316624

Baeshen MN, Al-Hejin AM, Bora RS, Ahmed MMM, Ramadan HAI, Saini KS, Baeshen NA, and Redwan EM. Production of

Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives. J Microbiol Biotechnol. 2015:**25**(7):953–962. https://doi.org/10.4014/jmb.1412.12079

Baiden N, Gandini C, Goddard P, and Sayanova O. Heterologous expression of antimicrobial peptides S-thanatin and bovine lactoferricin in the marine diatom *Phaeodactylum tricornutum* enhances native antimicrobial activity against Gram-negative bacteria. Algal Res. 2023:**69**:102927. https://doi.org/10.1016/j.algal.2022.102927

Baïet B, Burel C, Saint-Jean B, Louvet R, Menu-Bouaouiche L, Kiefer-Meyer M-C, Mathieu-Rivet E, Lefebvre T, Castel H, Carlier A, et al. N-Glycans of Phaeodactylum tricornutum Diatom and Functional Characterization of Its N-Acetylglucosaminyltransferase I Enzyme. J Biol Chem. 2011:**286**(8):6152–6164. https://doi.org/10.1074/jbc.M110.175711

Baker SE. Aspergillus niger genomics: past, present and into the future. Med Mycol. 2006:**44 Suppl 1**:S17-21. https://doi.org/10.1080/13693780600921037

Banerjee A and Ward V. Production of recombinant and therapeutic proteins in microalgae. Curr Opin Biotechnol. 2022:**78**:102784. https://doi.org/10.1016/j.copbio.2022.102784

Baneyx F and Mujacic M. Recombinant protein folding and misfolding in Escherichia coli. Nat Biotechnol. 2004:**22**(11):1399–1408. https://doi.org/10.1038/nbt1029

Banfield DK. Mechanisms of Protein Retention in the Golgi. Cold Spring Harb Perspect Biol. 2011:**3**(8):a005264. https://doi.org/10.1101/cshperspect.a005264

Bar-Peled L and Kory N. Principles and functions of metabolic compartmentalization. Nat Metab. 2022:**4**(10):1232–1244. https://doi.org/10.1038/s42255-022-00645-2

Bartual A, Villazán B, and Brun FG. Monitoring the long-term stability of pelagic morphotypes in the model diatom *Phaeodactylum tricornutum*. Diatom Res. 2011:**26**(2):243–253. https://doi.org/10.1080/0269249X.2011.619365

Bedford DJ, Schweizer E, Hopwood DA, and Khosla C. Expression of a functional fungal polyketide synthase in the bacterium Streptomyces coelicolor A3(2). J Bacteriol. 1995:**177**(15):4544–4548.

Bernstein HC, McClure RS, Hill EA, Markillie LM, Chrisler WB, Romine MF, McDermott JE, Posewitz MC, Bryant DA, Konopka AE, et al. Unlocking the Constraints of Cyanobacterial Productivity: Acclimations Enabling Ultrafast Growth. mBio. 2016:**7**(4):e00949-16. https://doi.org/10.1128/mBio.00949-16

Betterle N and Melis A. Heterologous Leader Sequences in Fusion Constructs Enhance Expression of Geranyl Diphosphate Synthase and Yield of β-Phellandrene Production in Cyanobacteria (Synechocystis). ACS Synth Biol. 2018:**7**(3):912–921. https://doi.org/10.1021/acssynbio.7b00431

Bhatwa A, Wang W, Hassan YI, Abraham N, Li X-Z, and Zhou T. Challenges Associated With the Formation of Recombinant Protein Inclusion Bodies in Escherichia coli and Strategies to Address Them for Industrial Applications. Front Bioeng Biotechnol. 2021:**9**. https://doi.org/10.3389/fbioe.2021.630551

Bielinski VA, Bolt TM, Dupont CL, and Weyman PD. Episomal tools for RNAi in the diatom Phaeodactylum tricornutum (PeerJ Inc.). https://doi.org/10.7287/peerj.preprints.2907v1

Bischof RH, Ramoni J, and Seiboth B. Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei. Microb Cell Factories. 2016:**15**(1):106. https://doi.org/10.1186/s12934-016-0507-6

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008:**456**(7219):239–244. https://doi.org/10.1038/nature07410

Branco-Vieira M, San Martin S, Agurto C, Freitas MAV, Martins AA, Mata TM, and Caetano NS. Biotechnological potential of *Phaeodactylum tricornutum* for biorefinery processes. Fuel. 2020:**268**:117357. https://doi.org/10.1016/j.fuel.2020.117357

Branduardi P, Valli M, Brambilla L, Sauer M, Alberghina L, and Branduardi P, Valli M, Brambilla L, Sauer M, Alberghina L, and Porro D. The yeast Zygosaccharomyces bailii: a new host for heterologous protein production, secretion and for metabolic engineering applications. FEMS Yeast Res. 2004:**4**(4–5):493–504. https://doi.org/10.1016/S1567-1356(03)00200-9

Brasil B dos SAF, de Siqueira FG, Salum TFC, Zanette CM, and Spier MR. Microalgae and cyanobacteria as enzyme biofactories. Algal Res. 2017:**25**:76–89. https://doi.org/10.1016/j.algal.2017.04.035

Brillouet J-M, Verdeil J-L, Odoux E, Lartaud M, Grisoni M, and Conéjéro G. Phenol homeostasis is ensured in vanilla fruit by storage under solid form in a new chloroplast-derived organelle, the phenyloplast. J Exp Bot. 2014:**65**. https://doi.org/10.1093/jxb/eru126

Brink DP, Ravi K, Lidén G, and Gorwa-Grauslund MF. Mapping the diversity of microbial lignin catabolism: experiences from the eLignin database. Appl Microbiol Biotechnol. 2019:**103**(10):3979–4002. https://doi.org/10.1007/s00253-019-09692-4

Brochado AR, Matos C, Møller BL, Hansen J, Mortensen UH, and Patil KR. Improved vanillin production in baker's yeast through in silico design. Microb Cell Factories. 2010:**9**(1):84. https://doi.org/10.1186/1475-2859-9-84

Buck JM, Río Bártulos C, Gruber A, and Kroth PG. Blasticidin-S deaminase, a new selection marker for genetic transformation of the diatom Phaeodactylum tricornutum. PeerJ. 2018:**6**:e5884. https://doi.org/10.7717/peerj.5884

Bulankova P, Sekulić M, Jallet D, Nef C, Van Oosterhout C, Delmont TO, Vercauteren I, Osuna-Cruz CM, Vancaester E, Mock T, et al. Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. Curr Biol. 2021:**31**(15):3221-3232.e9. https://doi.org/10.1016/j.cub.2021.05.013

Butler T, Kapoore RV, and Vaidyanathan S. Phaeodactylum tricornutum: A Diatom Cell Factory. Trends Biotechnol. 2020:**38**(6):606–622. https://doi.org/10.1016/j.tibtech.2019.12.023

Calero P and Nikel PI. Chasing bacterial chassis for metabolic engineering: a perspective review from classical to non-traditional microorganisms. Microb Biotechnol. 2019:**12**(1):98–124. https://doi.org/10.1111/1751-7915.13292

Chaves JE, Romero PR, Kirst H, and Melis A. Role of isopentenyl-diphosphate isomerase in heterologous cyanobacterial (Synechocystis) isoprene production. Photosynth Res. 2016:**130**(1):517–527. https://doi.org/10.1007/s11120-016-0293-3

Chee MJY, Lycett GW, Khoo T-J, and Chin CF. Bioengineering of the Plant Culture of Capsicum frutescens with Vanillin Synthase Gene for the Production of Vanillin. Mol Biotechnol. 2017:**59**(1):1–8. https://doi.org/10.1007/s12033-016-9986-2

Chen S, Zhou Y, Chen Y, and Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:**34**(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cho EY, Cheon SA, Kim H, Choo J, Lee D-J, Ryu HM, Rhee SK, Chung B-H, Kim J-Y, and Kang HA. Multiple-yapsin-deficient mutant strains for high-level production of intact recombinant proteins in *Saccharomyces cerevisiae*. J Biotechnol. 2010:**149**(1):1–7. https://doi.org/10.1016/j.jbiotec.2010.06.014

Choi KR, Jang WD, Yang D, Cho JS, Park D, and Lee SY. Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. Trends Biotechnol. 2019:**37**(8):817–837. https://doi.org/10.1016/j.tibtech.2019.01.003

Chu L, Ewe D, Río Bártulos C, Kroth PG, and Gruber A. Rapid induction of GFP expression by the nitrate reductase promoter in the diatom

Phaeodactylum tricornutum. PeerJ. 2016:**4**:e2344. https://doi.org/10.7717/peerj.2344

Chuberre C, Chan P, Walet-Balieu M-L, Thiébert F, Burel C, Hardouin J, Gügi B, and Bardor M. Comparative Proteomic Analysis of the Diatom Phaeodactylum tricornutum Reveals New Insights Into Intra- and Extra-Cellular Protein Contents of Its Oval, Fusiform, and Triradiate Morphotypes. Front Plant Sci. 2022:**13**. https://doi.org/10.3389/fpls.2022.673113

Clomburg JM, Crumbley AM, and Gonzalez R. Industrial biomanufacturing: The future of chemical production. Science. 2017:**355**(6320):aag0804. https://doi.org/10.1126/science.aag0804

Cobb RE, Wang Y, and Zhao H. High-efficiency multiplex genome editing of Streptomyces species using an engineered CRISPR/Cas system. ACS Synth Biol. 2015:**4**(6):723–728. https://doi.org/10.1021/sb500351f

Csárdi G, Nepusz T, Müller K, Horvát S, Traag V, Zanini F, and Noom D. igraph for R: R interface of the igraph library for graph theory and network analysis. 2024. https://doi.org/10.5281/ZENODO.7682609

Daboussi F, Leduc S, Maréchal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M, et al. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. Nat Commun. 2014:**5**(1):1–7. https://doi.org/10.1038/ncomms4831

D'Adamo S, Schiano di Visconte G, Lowe G, Szaub-Newton J, Beacham T, Landels A, Allen MJ, Spicer A, and Matthijs M. Engineering the unicellular alga Phaeodactylum tricornutum for high-value plant triterpenoid production. Plant Biotechnol J. 2019:**17**(1):75–87. https://doi.org/10.1111/pbi.12948

De Graffenried CL and Bertozzi CR. The roles of enzyme localisation and complex formation in glycan assembly within the Golgi apparatus. Curr Opin Cell Biol. 2004:**16**(4):356–363. https://doi.org/10.1016/j.ceb.2004.06.007

Dehghani J, Adibkia K, Movafeghi A, Maleki-Kakelar H, Saeedi N, and Omidi Y. Towards a new avenue for producing therapeutic proteins: Microalgae as a tempting green biofactory. Biotechnol Adv. 2020:**40**:107499. https://doi.org/10.1016/j.biotechadv.2019.107499

Del Mondo A, Sansone C, and Brunet C. Insights into the biosynthesis pathway of phenolic compounds in microalgae. Comput Struct Biotechnol J. 2022:**20**:1901–1913. https://doi.org/10.1016/j.csbj.2022.04.019

DeLoache WC, Russ ZN, and Dueber JE. Towards repurposing the yeast peroxisome for compartmentalizing heterologous metabolic

pathways. Nat Commun. 2016:**7**(1):11152. https://doi.org/10.1038/ncomms11152

Demain AL. From natural products discovery to commercialization: a success story. J Ind Microbiol Biotechnol. 2006:**33**(7):486–495. https://doi.org/10.1007/s10295-005-0076-x

Dhakal D, Chen M, Luesch H, and Ding Y. Heterologous production of cyanobacterial compounds. J Ind Microbiol Biotechnol. 2021:**48**(3–4):kuab003. https://doi.org/10.1093/jimb/kuab003

Diamond A, Barnabé S, and Desgagné-Penix I. Is a spice missing from the recipe? The intra-cellular localization of vanillin biosynthesis needs further investigations. Plant Biol. 2023a:**25**(1):3–7. https://doi.org/10.1111/plb.13465

Diamond A, Diaz-Garza AM, Li J, Slattery SS, Merindol N, Fantino E, Meddeb-Mouelhi F, Karas BJ, Barnabé S, and Desgagné-Penix I. Instability of extrachromosomal DNA transformed into the diatom Phaeodactylum tricornutum. Algal Res. 2023b:**70**:102998. https://doi.org/10.1016/j.algal.2023.102998

Diaz-Garza AM, Merindol N, dos Santos KCG, Lavoie-Marchand F, Ingalls B, and Desgagné-Penix I. No two clones are alike: characterization of heterologous subpopulations in a transgenic cell line of the model diatom Phaeodactylum tricornutum. Microb Cell Factories. 2024:**23**(1):286. https://doi.org/10.1186/s12934-024-02559-y

Diner RE, Bielinski VA, Dupont CL, Allen AE, and Weyman PD. Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods. Front Bioeng Biotechnol. 2016:**4**:65. https://doi.org/10.3389/fbioe.2016.00065

Diner RE, Noddings CM, Lian NC, Kang AK, McQuaid JB, Jablanovic J, Espinoza JL, Nguyen NA, Anzelmatti MA, Jansson J, et al. Diatom centromeres suggest a mechanism for nuclear DNA acquisition. Proc Natl Acad Sci U S A. 2017:**114**(29):E6015–E6024. https://doi.org/10.1073/pnas.1700764114

Dorfman DM, LaPlante CD, and Li B. FLOCK cluster analysis of plasma cell flow cytometry data predicts bone marrow involvement by plasma cell neoplasia. Leuk Res. 2016:**48**:40–45. https://doi.org/10.1016/j.leukres.2016.07.003

Doron L, Segal N, and Shapira M. Transgene Expression in Microalgae—From Tools to Applications. Front Plant Sci. 2016:**7**. https://doi.org/10.3389/fpls.2016.00505

Dos Santos KCG, Desgagné-Penix I, and Germain H. Correction to: Custom selected reference genes outperform pre-defined reference

genes in transcriptomic analysis. BMC Genomics. 2021:**22**(1):607. https://doi.org/10.1186/s12864-021-07743-7

Drummond DA and Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008:**134**(2):341–352. https://doi.org/10.1016/j.cell.2008.05.042

Durbin R. GENE EXPRESSION SYSTEMS BASED ON BACTERIOPHAGE T7 RNA POLYMERASE. . In. Gene Expression Systems. (Elsevier), pp. 9–44. https://doi.org/10.1016/B978-012253840-7/50002-X

Dusséaux S, Wajn WT, Liu Y, Ignea C, and Kampranis SC. Transforming yeast peroxisomes into microfactories for the efficient production of high-value isoprenoids. Proc Natl Acad Sci. 2020:**117**(50):31789–31799. https://doi.org/10.1073/pnas.2013968117

Elhai J, Vepritskiy A, Muro-Pastor AM, Flores E, and Wolk CP. Reduction of conjugal transfer efficiency by three restriction activities of Anabaena sp. strain PCC 7120. J Bacteriol. 1997:**179**(6):1998–2005. https://doi.org/10.1128/jb.179.6.1998-2005.1997

Engel BD, Schaffer M, Kuhn Cuellar L, Villa E, Plitzko JM, and Baumeister W. Native architecture of the Chlamydomonas chloroplast revealed by in situ cryo-electron tomography. eLife. 2015:**4**:e04889. https://doi.org/10.7554/eLife.04889

Engler C, Kandzia R, and Marillonnet S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. PLOS ONE. 2008:**3**(11):e3647. https://doi.org/10.1371/journal.pone.0003647

Erdene-Ochir E, Shin B-K, Kwon B, Jung C, and Pan C-H. Identification and characterisation of the novel endogenous promoter HASP1 and its signal peptide from Phaeodactylum tricornutum. Sci Rep. 2019a:**9**(1):9941. https://doi.org/10.1038/s41598-019-45786-9

Erdene-Ochir E, Shin B-K, Kwon B, Jung C, and Pan C-H. Identification and characterisation of the novel endogenous promoter HASP1 and its signal peptide from Phaeodactylum tricornutum. Sci Rep. 2019b:**9**(1):9941. https://doi.org/10.1038/s41598-019-45786-9

Fabris M, George J, Kuzhiumparambil U, Lawson CA, Jaramillo-Madrid AC, Abbriano RM, Vickers CE, and Ralph P. Extrachromosomal Genetic Engineering of the Marine Diatom Phaeodactylum tricornutum Enables the Heterologous Production of Monoterpenoids. ACS Synth Biol. 2020:**9**(3):598–612. https://doi.org/10.1021/acssynbio.9b00455

Facchini PJ, Bohlmann J, Covello PS, De Luca V, Mahadevan R, Page JE, Ro D-K, Sensen CW, Storms R, and Martin VJJ. Synthetic biosystems for the production of high-value plant metabolites. Trends

Biotechnol. 2012:**30**(3):127–131. https://doi.org/10.1016/j.tibtech.2011.10.001

Fache M, Boutevin B, and Caillol S. Vanillin Production from Lignin and Its Use as a Renewable Chemical. ACS Sustain Chem Eng. 2016:**4**(1):35–46. https://doi.org/10.1021/acssuschemeng.5b01344

Faessler AC. Optimising tools for metabolic engineering in the marine diatom Phaeodactylum tricornutum. 2024.

Fantino E, Awwad F, Merindol N, Diaz Garza AM, Gélinas S-E, Gajón Robles GC, Custeau A, Meddeb-Mouelhi F, and Desgagné-Penix I. Bioengineering *Phaeodactylum tricornutum*, a marine diatom, for cannabinoid biosynthesis. Algal Res. 2024:**77**:103379. https://doi.org/10.1016/j.algal.2023.103379

Ferreira-Camargo LS, Tran M, Beld J, Burkart MD, and Mayfield SP. Selenocystamine improves protein accumulation in chloroplasts of eukaryotic green algae. AMB Express. 2015:**5**(1):39. https://doi.org/10.1186/s13568-015-0126-3

Filloramo GV, Curtis BA, Blanche E, and Archibald JM. Re-examination of two diatom reference genomes using long-read sequencing. BMC Genomics. 2021:**22**(1):379. https://doi.org/10.1186/s12864-021-07666-3

Fisch KM, Bakeer W, Yakasai AA, Song Z, Pedrick J, Wasil Z, Bailey AM, Lazarus CM, Simpson TJ, and Cox RJ. Rational Domain Swaps Decipher Programming in Fungal Highly Reducing Polyketide Synthases and Resurrect an Extinct Metabolite. J Am Chem Soc. 2011:**133**(41):16635–16641. https://doi.org/10.1021/ja206914q

Fitzgerald DJ, Stratford M, Gasson MJ, Ueckert J, Bos A, and Narbad A. Mode of antimicrobial action of vanillin against Escherichia coli, Lactobacillus plantarum and Listeria innocua. J Appl Microbiol. 2004:**97**(1):104–113. https://doi.org/10.1111/j.1365-2672.2004.02275.x

Formighieri C and Melis A. Sustainable heterologous production of terpene hydrocarbons in cyanobacteria. Photosynth Res. 2016:**130**(1):123–135. https://doi.org/10.1007/s11120-016-0233-2

Formighieri C and Melis A. Heterologous synthesis of geranyllinalool, a diterpenol plant product, in the cyanobacterium Synechocystis. Appl Microbiol Biotechnol. 2017:**101**(7):2791–2800. https://doi.org/10.1007/s00253-016-8081-8

Galas L, Burel C, Schapman D, Ropitaux M, Bernard S, Bénard M, and Bardor M. Comparative Structural and Functional Analyses of the Fusiform, Oval, and Triradiate Morphotypes of Phaeodactylum tricornutum Pt3 Strain. Front Plant Sci. 2021:**12**:638181. https://doi.org/10.3389/fpls.2021.638181

Gallage NJ, Hansen EH, Kannangara R, Olsen CE, Motawia MS, Jørgensen K, Holme I, Hebelstrup K, Grisoni M, and Møller BL. Vanillin formation from ferulic acid in Vanilla planifolia is catalysed by a single enzyme. Nat Commun. 2014:**5**. https://doi.org/10.1038/ncomms5037

Gallage NJ, JØrgensen K, Janfelt C, Nielsen AJZ, Naake T, Duński E, Dalsten L, Grisoni M, and MØller BL. The Intracellular Localization of the Vanillin Biosynthetic Machinery in Pods of Vanilla planifolia. Plant Cell Physiol. 2018:**59**(2):304–318. https://doi.org/10.1093/pcp/pcx185

Gallage NJ and Møller BL. Vanillin–Bioconversion and Bioengineering of the Most Popular Plant Flavor and Its *De Novo* Biosynthesis in the Vanilla Orchid. Mol Plant. 2015:**8**(1):40–57. https://doi.org/10.1016/j.molp.2014.11.008

Gao S, Zhou L, Yang W, Wang L, Liu X, Gong Y, Hu Q, and Wang G. Overexpression of a novel gene (Pt2015) endows the commercial diatom Phaeodactylum tricornutum high lipid content and grazing resistance. Biotechnol Biofuels Bioprod. 2022:**15**(1):131. https://doi.org/10.1186/s13068-022-02221-y

Garza EA, Bielinski VA, Espinoza JL, Orlandi K, Alfaro JR, Bolt TM, Beeri K, Weyman PD, and Dupont CL. Validating a Promoter Library for Application in Plasmid-Based Diatom Genetic Engineering. ACS Synth Biol. 2023:**12**(11):3215–3228. https://doi.org/10.1021/acssynbio.3c00163

Gasch AP. The environmental stress response: a common yeast response to diverse environmental stresses. . In. Yeast Stress Responses, S Hohmann and WH Mager, eds. (Springer: Berlin, Heidelberg), pp. 11–70. https://doi.org/10.1007/3-540-45611-2_2

Ge Y and Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics. 2012:**28**(15):2052–2058. https://doi.org/10.1093/bioinformatics/bts300

George J, Kahlke T, Abbriano RM, Kuzhiumparambil U, Ralph PJ, and Fabris M. Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology. Front Bioeng Biotechnol. 2020:**8**. https://doi.org/10.3389/fbioe.2020.00513

Georgiou G and Segatori L. Preparative expression of secreted proteins in bacteria: status report and future prospects. Curr Opin Biotechnol. 2005:**16**(5):538–545. https://doi.org/10.1016/j.copbio.2005.07.008

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, and Smith HO. Enzymatic assembly of DNA molecules up to several

hundred kilobases. Nat Methods. 2009:**6**(5):343–345. https://doi.org/10.1038/nmeth.1318

Giguere DJ, Bahcheli AT, Slattery SS, Patel RR, Browne TS, Flatley M, Karas BJ, Edgell DR, and Gloor GB. Telomere-to-telomere genome assembly of Phaeodactylum tricornutum. PeerJ. 2022:**10**:e13607. https://doi.org/10.7717/peerj.13607

Gill K, Negi S, Rana N, and Kumar P. Chapter 13 - RDT and genetic engineering: Basic of RDT method, PCR, and application. . In. Basic Biotechniques for Bioprocess and Bioentrepreneurship, AK Bhatt, RK Bhatia, and TC Bhalla, eds. (Academic Press), pp. 207–216. https://doi.org/10.1016/B978-0-12-816109-8.00013-1

Gnügge R and Rudolf F. Saccharomyces cerevisiae Shuttle vectors. Yeast. 2017:**34**(5):205–221. https://doi.org/10.1002/yea.3228

Gonzalez NH, Felsner G, Schramm FD, Klingl A, Maier U-G, and Bolte K. A Single Peroxisomal Targeting Signal Mediates Matrix Protein Import in Diatoms. PLoS ONE. 2011:**6**(9):e25316. https://doi.org/10.1371/journal.pone.0025316

Gordon G and Pfleger B. Regulatory tools for controlling gene expression in cyanobacteria. Adv Exp Med Biol. 2018:**1080**:281–315. https://doi.org/10.1007/978-981-13-0854-3_12

Grabenhorst E and Conradt HS. The Cytoplasmic, Transmembrane, and Stem Regions of Glycosyltransferases Specify Their *in Vivo* Functional Sublocalization and Stability in the Golgi*. J Biol Chem. 1999:**274**(51):36107–36116. https://doi.org/10.1074/jbc.274.51.36107

de Grahl I and Reumann S. Stramenopile microalgae as “green biofactories” for recombinant protein production. World J Microbiol Biotechnol. 2021:**37**(9):163. https://doi.org/10.1007/s11274-021-03126-y

Gressler M, Hortschansky P, Geib E, and Brock M. A new high-performance heterologous fungal expression system based on regulatory elements from the Aspergillus terreus terrein gene cluster. Front Microbiol. 2015:**6**. https://doi.org/10.3389/fmicb.2015.00184

Griese M, Lange C, and Soppa J. Ploidy in cyanobacteria. FEMS Microbiol Lett. 2011:**323**(2):124–131. https://doi.org/10.1111/j.1574-6968.2011.02368.x

Gu Y, Alam S, and Oliferenko S. Peroxisomal compartmentalization of amino acid biosynthesis reactions imposes an upper limit on compartment size. Nat Commun. 2023:**14**(1):5544. https://doi.org/10.1038/s41467-023-41347-x

Gu Y, Xu X, Wu Y, Niu T, Liu Y, Li J, Du G, and Liu L. Advances and prospects of *Bacillus subtilis* cellular factories: From rational design to industrial applications. Metab Eng. 2018:**50**:109–121. https://doi.org/10.1016/j.ymben.2018.05.006

Gupta SK and Shukla P. Advanced technologies for improved expression of recombinant proteins in bacteria: perspectives and applications. Crit Rev Biotechnol. 2016:**36**(6):1089–1098. https://doi.org/10.3109/07388551.2015.1084264

Hamer JE and Timberlake WE. Functional organization of the Aspergillus nidulans trpC promoter. Mol Cell Biol. 1987:**7**(7):2352–2359.

Hamilton ML, Haslam RP, Napier JA, and Sayanova O. Metabolic engineering of Phaeodactylum tricornutum for the enhanced accumulation of omega-3 long chain polyunsaturated fatty acids. Metab Eng. 2014:**22**:3–9. https://doi.org/10.1016/j.ymben.2013.12.003

Hansen EH, Møller BL, Kock GR, Bünner CM, Kristensen C, Jensen OR, Okkels FT, Olsen CE, Motawia MS, and Hansen J. De Novo Biosynthesis of Vanillin in Fission Yeast (Schizosaccharomyces pombe) and Baker's Yeast (Saccharomyces cerevisiae). Appl Environ Microbiol. 2009:**75**(9):2765–2774. https://doi.org/10.1128/AEM.02681-08

Haoujar I, Cacciola F, Abrini J, Mangraviti D, Giuffrida D, Oulad El Majdoub Y, Kounnoun A, Miceli N, Fernanda Taviano M, Mondello L, et al. The Contribution of Carotenoids, Phenolic Compounds, and Flavonoids to the Antioxidative Properties of Marine Microalgae Isolated from Mediterranean Morocco. Molecules. 2019:**24**(22):4037. https://doi.org/10.3390/molecules24224037

Harris EH. CHLAMYDOMONASAS A MODEL ORGANISM. Annu Rev Plant Biol. 2001:**52**(Volume 52, 2001):363–406. https://doi.org/10.1146/annurev.arplant.52.1.363

Harris TJ, Lowe PA, Lyons A, Thomas PG, Eaton MA, Millican TA, Patel TP, Bose CC, Carey NH, and Doel MT. Molecular cloning and nucleotide sequence of cDNA coding for calf preprochymosin. Nucleic Acids Res. 1982:**10**(7):2177–2187. https://doi.org/10.1093/nar/10.7.2177

Havkin-Frenkel D and Belanger FC. Vanillin Biosynthesis – Still Not as Simple as it Seems? . In. Handbook of Vanilla Science and Technology. (John Wiley & Sons, Ltd), pp. 435–445. https://doi.org/10.1002/9781119377320.ch24

Hayat SMG, Farahani N, Golichenari B, and Sahebkar A. Recombinant Protein Expression in Escherichia coli (E.coli): What We Need to Know. Curr Pharm Des. 2018:**24**(6):718–725. https://doi.org/10.2174/1381612824666180131121940

He D-M, Qian K-X, Shen G-F, Zhang Z-F, Li Y-N, Su Z-L, and Shao H-B. Recombination and expression of classical swine fever virus (CSFV) structural protein E2 gene in *Chlamydomonas reinhardtii* chroloplasts. Colloids Surf B Biointerfaces. 2007:**55**(1):26–30. https://doi.org/10.1016/j.colsurfb.2006.10.042

Hempel F, Lau J, Klingl A, and Maier UG. Algae as Protein Factories: Expression of a Human Antibody and the Respective Antigen in the Diatom Phaeodactylum tricornutum. PLOS ONE. 2011:**6**(12):e28424. https://doi.org/10.1371/journal.pone.0028424

Hempel F and Maier UG. An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. Microb Cell Factories. 2012:**11**:126. https://doi.org/10.1186/1475-2859-11-126

Henchion M, Hayes M, Mullen AM, Fenelon M, and Tiwari B. Future Protein Supply and Demand: Strategies and Factors Influencing a Sustainable Equilibrium. Foods. 2017:**6**(7):53. https://doi.org/10.3390/foods6070053

Hillson NJ. DNA Assembly Method Standardization for Synthetic Biomolecular Circuits and Systems. . In. Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology, H Koeppl, G Setti, M di Bernardo, and D Densmore, eds. (Springer: New York, NY), pp. 295–314. https://doi.org/10.1007/978-1-4419-6766-4_14

Hooykaas PJJ, Van Heusden GPH, Niu X, Reza Roushan M, Soltani J, Zhang X, and Van Der Zaal BJ. Agrobacterium-Mediated Transformation of Yeast and Fungi. . In. Agrobacterium Biology, SB Gelvin, ed, Current Topics in Microbiology and Immunology. (Springer International Publishing: Cham), pp. 349–374. https://doi.org/10.1007/82_2018_90

Huang W and Daboussi F. Genetic and metabolic engineering in diatoms. Philos Trans R Soc B Biol Sci. 2017:**372**(1728):20160411. https://doi.org/10.1098/rstb.2016.0411

Huemer M, Mairpady Shambat S, Brugger SD, and Zinkernagel AS. Antibiotic resistance and persistence—Implications for human health and treatment perspectives. EMBO Rep. 2020:**21**(12):e51034. https://doi.org/10.15252/embr.202051034

Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007:**9**(3):90–95. https://doi.org/10.1109/MCSE.2007.55

Ikeuchi M and Tabata S. Synechocystis sp. PCC 6803 - a useful tool in the study of the genetics of cyanobacteria. Photosynth Res. 2001:**70**(1):73–83. https://doi.org/10.1023/A:1013887908680

Ingram Z, Patkar A, Oh D, Zhang KK, Chung C, Lin-Cereghino J, and Lin-Cereghino GP. Overcoming Obstacles in Protein Expression in

the Yeast Pichia pastoris: Interviews of Leaders in the Pichia Field. Pac J Health Stockt Calif. 2021:**4**(1):2. https://doi.org/10.56031/2576-215x.1010

Javorova R, Rezuchova B, Feckova L, Novakova R, Csolleiova D, Kopacova M, Patoprsty V, Opaterny F, Sevcikova B, and Kormanec J. A new synthetic biology system for investigating the biosynthesis of antibiotics and other secondary metabolites in streptomycetes. J Biotechnol. 2024:**392**:128–138. https://doi.org/10.1016/j.jbiotec.2024.07.007

Jia H, Zhang L, Wang T, Han J, Tang H, and Zhang L. Development of a CRISPR/Cas9-mediated gene-editing tool in Streptomyces rimosus. Microbiology. 2017:**163**(8):1148–1155. https://doi.org/10.1099/mic.0.000501

Jiang Y-H, Liu Y-F, Wang K, Zhou J-Y, Guo F, Zhao Q-W, and Mao X-M. Fine-Tuning Cas9 Activity with a Cognate Inhibitor AcrIIA4 to Improve Genome Editing in *Streptomyces*. ACS Synth Biol. 2021:**10**(11):2833–2841. https://doi.org/10.1021/acssynbio.1c00141

Jørgensen MS, Skovlund DA, Johannesen PF, and Mortensen UH. A novel platform for heterologous gene expression in Trichoderma reesei (Teleomorph Hypocrea jecorina). Microb Cell Factories. 2014:**13**(1):33. https://doi.org/10.1186/1475-2859-13-33

Jung ST and Kim D. Mass Production of Full-Length IgG Monoclonal Antibodies from Mammalian, Yeast, and Bacterial Hosts. . In. Emerging Areas in Bioengineering, HN Chang, ed. (Wiley), pp. 679–695. https://doi.org/10.1002/9783527803293.ch39

Karas BJ, Diner RE, Lefebvre SC, McQuaid J, Phillips APR, Noddings CM, Brunson JK, Valas RE, Deerinck TJ, Jablanovic J, et al. Designer diatom episomes delivered by bacterial conjugation. Nat Commun. 2015:**6**(1):6925. https://doi.org/10.1038/ncomms7925

Kaspar F, Neubauer P, and Gimpel M. Bioactive Secondary Metabolites from Bacillus subtilis: A Comprehensive Review. J Nat Prod. 2019:**82**(7):2038–2053. https://doi.org/10.1021/acs.jnatprod.9b00110

Kassaw TK, Paton AJ, and Peers G. Episome-Based Gene Expression Modulation Platform in the Model Diatom Phaeodactylum tricornutum. ACS Synth Biol. 2022:**11**(1):191–204. https://doi.org/10.1021/acssynbio.1c00367

Kastberg LLB, Ard R, Jensen MK, and Workman CT. Burden Imposed by Heterologous Protein Production in Two Major Industrial Yeast Cell Factories: Identifying Sources and Mitigation Strategies. Front Fungal Biol. 2022:**3**:827704. https://doi.org/10.3389/ffunb.2022.827704

Kaur B and Chakraborty D. Biotechnological and Molecular Approaches for Vanillin Production: a Review. Appl Biochem Biotechnol. 2013:**169**(4):1353–1372. https://doi.org/10.1007/s12010-012-0066-1

Kaur J, Kumar A, and Kaur J. Strategies for optimization of heterologous protein expression in *E. coli*: Roadblocks and reinforcements. Int J Biol Macromol. 2018:**106**:803–822. https://doi.org/10.1016/j.ijbiomac.2017.08.080

Kaye Y, Grundman O, Leu S, Zarka A, Zorin B, Didi-Cohen S, Khozin-Goldberg I, and Boussiba S. Metabolic engineering toward enhanced LC-PUFA biosynthesis in *Nannochloropsis oceanica*: Overexpression of endogenous Δ12 desaturase driven by stress-inducible promoter leads to enhanced deposition of polyunsaturated fatty acids in TAG. Algal Res. 2015:**11**:387–398. https://doi.org/10.1016/j.algal.2015.05.003

Khakhar A and Voytas DF. RNA Viral Vectors for Accelerating Plant Synthetic Biology. Front Plant Sci. 2021:**12**. https://doi.org/10.3389/fpls.2021.668580

Kilian O and Kroth PG. Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. Plant J. 2005:**41**(2):175–183. https://doi.org/10.1111/j.1365-313X.2004.02294.x

Kim SR, Park Y-C, Jin Y-S, and Seo J-H. Strain engineering of Saccharomyces cerevisiae for enhanced xylose metabolism. Biotechnol Adv. 2013:**31**(6):851–861. https://doi.org/10.1016/j.biotechadv.2013.03.004

Kim WJ, Lee S-M, Um Y, Sim SJ, and Woo HM. Development of SyneBrick Vectors As a Synthetic Biology Platform for Gene Expression in Synechococcus elongatus PCC 7942. Front Plant Sci. 2017:**8**. https://doi.org/10.3389/fpls.2017.00293

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011:**2011**:bar030. https://doi.org/10.1093/database/bar030

Kocaoglan EG, Andreou A, Nirkko J, Ochoa-Villarreal M, Loake G, and Nakayama N. Mobius Assembly for Plant Systems uncovers combinatorial interactions among promoters, coding sequences, and terminators in gene regulation. 2024. https://doi.org/10.1101/2024.07.10.602858

Kuczyńska P, Jemioła-Rzemińska M, and Strzałka K. Pigments in Diatoms. . In. Diatom Photosynthesis. (John Wiley & Sons, Ltd), pp. 137–189. https://doi.org/10.1002/9781119842156.ch5

Kundu A. Vanillin biosynthetic pathways in plants. Planta. 2017:**245**(6):1069–1078. https://doi.org/10.1007/s00425-017-2684-x

Kuwata A and Jewson DH. Ecology and Evolution of Marine Diatoms and Parmales. . In. Marine Protists: Diversity and Dynamics, S Ohtsuka, T Suzaki, T Horiguchi, N Suzuki, and F Not, eds. (Springer Japan: Tokyo), pp. 251–275. https://doi.org/10.1007/978-4-431-55130-0_10

LaFave MC and Sekelsky J. Mitotic Recombination: Why? When? How? Where? PLoS Genet. 2009:**5**(3):e1000411. https://doi.org/10.1371/journal.pgen.1000411

Lauersen KJ, Baier T, Wichmann J, Wördenweber R, Mussgnug JH, Hübner W, Huser T, and Kruse O. Efficient phototrophic production of a high-value sesquiterpenoid from the eukaryotic microalga *Chlamydomonas reinhardtii*. Metab Eng. 2016:**38**:331–343. https://doi.org/10.1016/j.ymben.2016.07.013

Lauersen KJ, Wichmann J, Baier T, Kampranis SC, Pateraki I, Møller BL, and Kruse O. Phototrophic production of heterologous diterpenoids and a hydroxy-functionalized derivative from *Chlamydomonas reinhardtii*. Metab Eng. 2018:**49**:116–127. https://doi.org/10.1016/j.ymben.2018.07.005

Lee C, Kim J, Shin SG, and Hwang S. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. J Biotechnol. 2006:**123**(3):273–280. https://doi.org/10.1016/j.jbiotec.2005.11.014

Lee Y, Hwang S, Kim W, Kim JH, Palsson BO, and Cho B-K. CRISPR-aided genome engineering for secondary metabolite biosynthesis in *Streptomyces*. J Ind Microbiol Biotechnol. 2024:**51**:kuae009. https://doi.org/10.1093/jimb/kuae009

Leterme S. The oil production capacity of diatoms. Ann Aquac Res. 2015:**2**:1007.

Li C, Deans NC, and Buell CR. “Simple Tidy GeneCoEx”: A gene co-expression analysis workflow powered by tidyverse and graph-based clustering in R. Plant Genome. 2023:**16**(2):e20323. https://doi.org/10.1002/tpg2.20323

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009:**25**(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li L, Wei K, Zheng G, Liu X, Chen S, Jiang W, and Lu Y. CRISPR-Cpf1-Assisted Multiplex Genome Editing and Transcriptional Repression

in Streptomyces. Appl Environ Microbiol. 2018:**84**(18):e00827-18. https://doi.org/10.1128/AEM.00827-18

Li MZ and Elledge SJ. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods. 2007:**4**(3):251–256. https://doi.org/10.1038/nmeth1010

Linlin Y. ggvenn: Draw Venn Diagram by “ggplot2.” 2023.

Liu C-P, Tsai T-I, Cheng T, Shivatare VS, Wu C-Y, Wu C-Y, and Wong C-H. Glycoengineering of antibody (Herceptin) through yeast expression and in vitro enzymatic glycosylation. Proc Natl Acad Sci. 2018:**115**(4):720–725. https://doi.org/10.1073/pnas.1718172115

Liu J, Wang X, Dai G, Zhang Y, and Bian X. Microbial chassis engineering drives heterologous production of complex secondary metabolites. Biotechnol Adv. 2022:**59**:107966. https://doi.org/10.1016/j.biotechadv.2022.107966

Liu J, Wu X, Yao M, Xiao W, and Zha J. Chassis engineering for microbial production of chemicals: from natural microbes to synthetic organisms. Curr Opin Biotechnol. 2020:**66**:105–112. https://doi.org/10.1016/j.copbio.2020.06.013

Liu K and Wessler SR. Transposition of Mutator–like transposable elements (MULEs) resembles hAT and Transib elements and V(D)J recombination. Nucleic Acids Res. 2017:**45**(11):6644–6655. https://doi.org/10.1093/nar/gkx357

Liu X, Hempel F, Stork S, Bolte K, Moog D, Heimerl T, Maier UG, and Zauner S. Addressing various compartments of the diatom model organism Phaeodactylum tricornutum via sub-cellular marker proteins. Algal Res. 2016:**20**:249–257. https://doi.org/10.1016/j.algal.2016.10.018

Liu Y, Tsinoremas NF, Johnson CH, Lebedeva NV, Golden SS, Ishiura M, and Kondo T. Circadian orchestration of gene expression in cyanobacteria. Genes Dev. 1995:**9**(12):1469–1478. https://doi.org/10.1101/gad.9.12.1469

Liu Z, Chen O, Wall JBJ, Zheng M, Zhou Y, Wang L, Ruth Vaseghi H, Qian L, and Liu J. Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. Sci Rep. 2017:**7**(1):2193. https://doi.org/10.1038/s41598-017-02460-2

Longworth J, Wu D, Huete-Ortega M, Wright PC, and Vaidyanathan S. Proteome response of Phaeodactylum tricornutum, during lipid accumulation induced by nitrogen depletion. Algal Res. 2016:**18**:213–224. https://doi.org/10.1016/j.algal.2016.06.015

Luziatelli F, Brunetti L, Ficca AG, and Ruzzi M. Maximizing the Efficiency of Vanillin Production by Biocatalyst Enhancement and Process

Optimization. Front Bioeng Biotechnol. 2019:**7**:279. https://doi.org/10.3389/fbioe.2019.00279

Lyseng-Williamson KA. Coagulation Factor IX (Recombinant), Albumin Fusion Protein (Albutrepenonacog Alfa; Idelvion®): A Review of Its Use in Haemophilia B. Drugs. 2017:**77**(1):97–106. https://doi.org/10.1007/s40265-016-0679-8

Ma C-H, Su B-Y, Maciaszek A, Fan H-F, Guga P, and Jayaram M. A Flp-SUMO hybrid recombinase reveals multi-layered copy number control of a selfish DNA element through post-translational modification. PLOS Genet. 2019:**15**(6):e1008193. https://doi.org/10.1371/journal.pgen.1008193

Madzak C, Gaillardin C, and Beckerich J-M. Heterologous protein expression and secretion in the non-conventional yeast *Yarrowia lipolytica*: a review. J Biotechnol. 2004:**109**(1):63–81. https://doi.org/10.1016/j.jbiotec.2003.10.027

Maeda Y, Tsuru Y, Matsumoto N, Nonoyama T, Yoshino T, Matsumoto M, and Tanaka T. Prostaglandin production by the microalga with heterologous expression of cyclooxygenase. Biotechnol Bioeng. 2021:**118**(7):2734–2743. https://doi.org/10.1002/bit.27792

Martău GA, Călinoiu L-F, and Vodnar DC. Bio-vanillin: Towards a sustainable industrial production. Trends Food Sci Technol. 2021:**109**:579–592. https://doi.org/10.1016/j.tifs.2021.01.059

Matsui H, Harada H, Maeda K, Sugiyama T, Fukuchi Y, Kimura N, Nawaly H, Tsuji Y, and Matsuda Y. Coordinated phosphate uptake by extracellular alkaline phosphatase and solute carrier transporters in marine diatoms. New Phytol. 2024:**241**(3):1210–1221. https://doi.org/10.1111/nph.19410

Mellitzer A, Weis R, Glieder A, and Flicker K. Expression of lignocellulolytic enzymes in Pichia pastoris. Microb Cell Factories. 2012:**11**(1):61. https://doi.org/10.1186/1475-2859-11-61

Mertz JE and Davis RW. Cleavage of DNA by R 1 restriction endonuclease generates cohesive ends. Proc Natl Acad Sci U S A. 1972:**69**(11):3370–3374. https://doi.org/10.1073/pnas.69.11.3370

Messaabi A, Merindol N, Bohnenblust L, Fantino E, Meddeb-Mouelhi F, and Desgagné-Penix I. In vivo thrombin activity in the diatom Phaeodactylum tricornutum: biotechnological insights. Appl Microbiol Biotechnol. 2024:**108**(1):481. https://doi.org/10.1007/s00253-024-13322-z

Meyer V, Cairns T, Barthel L, King R, Kunz P, Schmideder S, Müller H, Briesen H, Dinius A, and Krull R. Understanding and controlling filamentous growth of fungal cell factories: novel tools and opportunities for targeted morphology engineering. Fungal Biol

Biotechnol. 2021:**8**(1):8. https://doi.org/10.1186/s40694-021-00115-6

Million-Weaver S and Camps M. Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? Plasmid. 2014:**0**:27–36. https://doi.org/10.1016/j.plasmid.2014.07.002

Mo Q and Yuan J. Minimal aromatic aldehyde reduction (MARE) yeast platform for engineering vanillin production. Biotechnol Biofuels Bioprod. 2024:**17**(1):4. https://doi.org/10.1186/s13068-023-02454-5

Moosburner MA, Gholami P, McCarthy JK, Tan M, Bielinski VA, and Allen AE. Multiplexed Knockouts in the Model Diatom Phaeodactylum by Episomal Delivery of a Selectable Cas9. Front Microbiol. 2020:**11**. https://doi.org/10.3389/fmicb.2020.00005

Mosmann TR, Naim I, Rebhahn J, Datta S, Cavenaugh JS, Weaver JM, and Sharma G. SWIFT—Scalable Clustering for Automated Identification of Rare Cell Populations in Large, High-Dimensional Flow Cytometry Datasets, Part 2: Biological Evaluation. Cytometry. 2014:**85**(5):422–433. https://doi.org/10.1002/cyto.a.22445

Mutalik VK, Novichkov PS, Price MN, Owens TK, Callaghan M, Carim S, Deutschbauer AM, and Arkin AP. Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. Nat Commun. 2019:**10**(1):308. https://doi.org/10.1038/s41467-018-08177-8

Muusse TW, Lee MYL, Kim H, Parat M-O, Nanson JD, Kobe B, Vajjhala PR, and Stacey KJ. Flow cytometric reporter assays provide robust functional analysis of signaling complexes. J Biol Chem. 2022:**298**(12):102666. https://doi.org/10.1016/j.jbc.2022.102666

Najah S, Saulnier C, Pernodet J-L, and Bury-Moné S. Design of a generic CRISPR-Cas9 approach using the same sgRNA to perform gene editing at distinct loci. BMC Biotechnol. 2019:**19**(1):18. https://doi.org/10.1186/s12896-019-0509-7

Nguyen TTM, Iwaki A, Ohya Y, and Izawa S. Vanillin causes the activation of Yap1 and mitochondrial fragmentation in Saccharomyces cerevisiae. J Biosci Bioeng. 2014:**117**(1):33–38. https://doi.org/10.1016/j.jbiosc.2013.06.008

Ni J, Tao F, Du H, and Xu P. Mimicking a natural pathway for *de novo* biosynthesis: natural vanillin production from accessible carbon sources. Sci Rep. 2015:**5**:13670. https://doi.org/10.1038/srep13670

Nielsen MT, Nielsen JB, Anyaogu DC, Holm DK, Nielsen KF, Larsen TO, and Mortensen UH. Heterologous Reconstitution of the Intact Geodin Gene Cluster in Aspergillus nidulans through a Simple and Versatile PCR Based Approach. PLoS ONE. 2013:**8**(8):e72871. https://doi.org/10.1371/journal.pone.0072871

Nora LC, Westmann CA, Guazzaroni M-E, Siddaiah C, Gupta VK, and Silva-Rocha R. Recent advances in plasmid-based tools for establishing novel microbial chassis. Biotechnol Adv. 2019:**37**(8):107433. https://doi.org/10.1016/j.biotechadv.2019.107433

Novikova O. Chromodomains and LTR retrotransposons in plants. Commun Integr Biol. 2009:**2**(2):158–162.

van Opijnen T, Bodi KL, and Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009:**6**(10):767–772. https://doi.org/10.1038/nmeth.1377

Ottaviani D, LeCain M, and Sheer D. The role of microhomology in genomic structural variation. Trends Genet. 2014:**30**(3):85–94. https://doi.org/10.1016/j.tig.2014.01.001

Ovide C, Kiefer-Meyer M-C, Bérard C, Vergne N, Lecroq T, Plasson C, Burel C, Bernard S, Driouich A, Lerouge P, et al. Comparative in depth RNA sequencing of P. tricornutum's morphotypes reveals specific features of the oval morphotype. Sci Rep. 2018:**8**(1):14340. https://doi.org/10.1038/s41598-018-32519-7

Pampuch M, Walker EJL, and Karas BJ. Towards synthetic diatoms: The *Phaeodactylum tricornutum* Pt-syn 1.0 project. Curr Opin Green Sustain Chem. 2022:**35**:100611. https://doi.org/10.1016/j.cogsc.2022.100611

Patel E and Kushwaha DS. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. Procedia Comput Sci. 2020:**171**:158–167. https://doi.org/10.1016/j.procs.2020.04.017

Patron NJ, Orzaez D, Marillonnet S, Warzecha H, Matthewman C, Youles M, Raitskin O, Leveau A, Farré G, Rogers C, et al. Standards for plant synthetic biology: a common syntax for exchange of DNA parts. New Phytol. 2015:**208**(1):13–19. https://doi.org/10.1111/nph.13532

Pattrick CA, Webb JP, Green J, Chaudhuri RR, Collins MO, and Kelly DJ. Proteomic Profiling, Transcription Factor Modeling, and Genomics of Evolved Tolerant Strains Elucidate Mechanisms of Vanillin Toxicity in Escherichia coli. mSystems. 2019:**4**(4):e00163-19. https://doi.org/10.1128/mSystems.00163-19

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011:**12**(85):2825–2830.

Perozeni F, Pivato M, Angelini M, Maricchiolo E, Pompa A, and Ballottari M. Towards microalga-based superfoods: heterologous

expression of zeolin in Chlamydomonas reinhardtii. Front Plant Sci. 2023:**14**. https://doi.org/10.3389/fpls.2023.1184064

Petsch D and Anspach FB. Endotoxin removal from protein solutions. J Biotechnol. 2000:**76**(2–3):97–119. https://doi.org/10.1016/s0168-1656(99)00185-6

Pouresmaeil M and Azizi-Dargahlou S. Factors involved in heterologous expression of proteins in E. coli host. Arch Microbiol. 2023:**205**(5):212. https://doi.org/10.1007/s00203-023-03541-9

Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018:**557**(7706):503–509. https://doi.org/10.1038/s41586-018-0124-0

Qiu X-B, Shao Y-M, Miao S, and Wang L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell Mol Life Sci CMLS. 2006:**63**(22):2560–2570. https://doi.org/10.1007/s00018-006-6192-6

Quan J and Tian J. Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. PLOS ONE. 2009:**4**(7):e6441. https://doi.org/10.1371/journal.pone.0006441

R Core Team. R: A Language and Environment for Statistical Computing. 2023.

Rai M and Padh H. Expression systems for production of heterologous proteins. Curr Sci. 2001:**80**(9):1121–1128.

Rajacharya GH, Sharma A, and Yazdani SS. Proteomics and metabolic burden analysis to understand the impact of recombinant protein production in E. coli. Sci Rep. 2024:**14**(1):12271. https://doi.org/10.1038/s41598-024-63148-y

Ramasamy MN, Minassian AM, Ewer KJ, Flaxman AL, Folegatti PM, Owens DR, Voysey M, Aley PK, Angus B, Babbage G, et al. Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial. Lancet Lond Engl. 2020:**396**(10267):1979–1993. https://doi.org/10.1016/S0140-6736(20)32466-1

Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, McCarthy J, Allen AE, Kersey P, Bowler C, et al. Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms. Sci Rep. 2018:**8**(1):4834. https://doi.org/10.1038/s41598-018-23106-x

Rathod JP, Vira C, Lali AM, and Prakash G. Metabolic Engineering of Chlamydomonas reinhardtii for Enhanced β-Carotene and Lutein Production. Appl Biochem Biotechnol. 2020:**190**(4):1457–1469. https://doi.org/10.1007/s12010-019-03194-9

Rico M, López A, Santana-Casiano JM, Gonzàlez AG, and Gonzàlez-Dàvila M. Variability of the phenolic profile in the diatom *Phaeodactylum tricornutum* growing under copper and iron stress. Limnol Oceanogr. 2013:**58**(1):144–152. https://doi.org/10.4319/lo.2013.58.1.0144

Rockman S, Laurie KL, Parkes S, Wheatley A, and Barr IG. New Technologies for Influenza Vaccines. Microorganisms. 2020:**8**(11):1745. https://doi.org/10.3390/microorganisms8111745

Rouches MV, Xu Y, Cortes LBG, and Lambert G. A plasmid system with tunable copy number. Nat Commun. 2022:**13**(1):3908. https://doi.org/10.1038/s41467-022-31422-0

Roulet J, Taton A, Golden JW, Arabolaza A, Burkart MD, and Gramajo H. Development of a cyanobacterial heterologous polyketide production platform. Metab Eng. 2018:**49**:94–104. https://doi.org/10.1016/j.ymben.2018.07.013

Rouzeau C, Dagkesamanskaya A, Langer K, Bibette J, Baudry J, Pompon D, and Anton-Leberre V. Adaptive response of yeast cells to triggered toxicity of phosphoribulokinase. Res Microbiol. 2018:**169**(6):335–342. https://doi.org/10.1016/j.resmic.2018.06.002

Russo MT, Aiese Cigliano R, Sanseverino W, and Ferrante MI. Assessment of genomic changes in a CRISPR/Cas9 Phaeodactylum tricornutum mutant through whole genome resequencing. PeerJ. 2018:**6**:e5507. https://doi.org/10.7717/peerj.5507

Sau S, Ghosh SK, Liu Y-T, Ma C-H, and Jayaram M. Hitchhiking on chromosomes: A persistence strategy shared by diverse selfish DNA elements. Plasmid. 2019:**102**:19–28. https://doi.org/10.1016/j.plasmid.2019.01.004

Sau S, Liu Y-T, Ma C-H, and Jayaram M. Stable persistence of the yeast plasmid by hitchhiking on chromosomes during vegetative and germ-line divisions of host cells. Mob Genet Elem. 2015:**5**(2):21–28. https://doi.org/10.1080/2159256X.2015.1031359

Schindler D. Genetic Engineering and Synthetic Genomics in Yeast to Understand Life and Boost Biotechnology. Bioengineering. 2020:**7**(4):137. https://doi.org/10.3390/bioengineering7040137

Schmitz KR, Liu J, Li S, Setty TG, Wood CS, Burd CG, and Ferguson KM. Golgi localization of glycosyltransferases requires a Vps74p oligomer. Dev Cell. 2008:**14**(4):523–534. https://doi.org/10.1016/j.devcel.2008.02.016

Schober AF, Flori S, Finazzi G, Kroth PG, and Bártulos CR. Isolation of Plastid Fractions from the Diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. Methods Mol Biol Clifton NJ. 2018:**1829**:189–203. https://doi.org/10.1007/978-1-4939-8654-5_13

Schreiber V, Dersch J, Puzik K, Bäcker O, Liu X, Stork S, Schulz J, Heimerl T, Klingl A, Zauner S, et al. The Central Vacuole of the Diatom Phaeodactylum tricornutum: Identification of New Vacuolar Membrane Proteins and of a Functional Di-leucine-based Targeting Motif. Protist. 2017:**168**(3):271–282. https://doi.org/10.1016/j.protis.2017.03.001

Sebesta J and Peebles CAM. Improving heterologous protein expression in *Synechocystis* sp. PCC 6803 for alpha-bisabolene production. Metab Eng Commun. 2020:**10**:e00117. https://doi.org/10.1016/j.mec.2019.e00117

Seo S, Jeon H, Hwang S, Jin E, and Chang KS. Development of a new constitutive expression system for the transformation of the diatom *Phaeodactylum tricornutum*. Algal Res. 2015:**11**:50–54. https://doi.org/10.1016/j.algal.2015.05.012

Seo S, Kim J, Lee J-W, Nam O, Chang KS, and Jin E. Enhanced pyruvate metabolism in plastids by overexpression of putative plastidial pyruvate transporter in Phaeodactylum tricornutum. Biotechnol Biofuels. 2020:**13**(1):120. https://doi.org/10.1186/s13068-020-01760-6

Serif M, Lepetit B, Weißert K, Kroth PG, and Rio Bartulos C. A fast and reliable strategy to generate TALEN-mediated gene knockouts in the diatom Phaeodactylum tricornutum. Algal Res. 2017:**23**:186–195. https://doi.org/10.1016/j.algal.2017.02.005

Sharma N, Simon DP, Diaz-Garza AM, Fantino E, Messaabi A, Meddeb-Mouelhi F, Germain H, and Desgagné-Penix I. Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow. Front Mar Sci. 2021:**8**:636613. https://doi.org/10.3389/fmars.2021.636613

Shen H, Stoute J, and Liu KF. Structural and catalytic roles of the human 18*S* rRNA methyltransferases DIMT1 in ribosome assembly and translation. J Biol Chem. 2020:**295**(34):12058–12070. https://doi.org/10.1074/jbc.RA120.014236

Shetty RP, Endy D, and Knight TF. Engineering BioBrick vectors from BioBrick parts. J Biol Eng. 2008:**2**(1):5. https://doi.org/10.1186/1754-1611-2-5

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, and Bowler C. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene. 2007:**406**(1):23–35. https://doi.org/10.1016/j.gene.2007.05.022

Sinha AK, Sharma UK, and Sharma N. A comprehensive review on vanilla flavor: extraction, isolation and quantification of vanillin and others constituents. Int J Food Sci Nutr. 2008:**59**(4):299–326. https://doi.org/10.1080/09687630701539350

Slattery SS, Diamond A, Wang H, Therrien JA, Lant JT, Jazey T, Lee K, Klassen Z, Desgagné-Penix I, Karas BJ, et al. An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in Phaeodactylum tricornutum. ACS Synth Biol. 2018:**7**(2):328–338. https://doi.org/10.1021/acssynbio.7b00191

Slattery SS, Giguere DJ, Stuckless EE, Shrestha A, Briere L-AK, Galbraith A, Reaume S, Boyko X, Say HH, Browne TS, et al. Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom Phaeodactylum tricornutum for pandemic diagnostics. Sci Rep. 2022:**12**(1):7010. https://doi.org/10.1038/s41598-022-11053-7

Slattery SS, Wang H, Giguere DJ, Kocsis C, Urquhart BL, Karas BJ, and Edgell DR. Plasmid-based complementation of large deletions in Phaeodactylum tricornutum biosynthetic genes generated by Cas9 editing. Sci Rep. 2020:**10**(1):13879. https://doi.org/10.1038/s41598-020-70769-6

Smith RA, Duncan MJ, and Moir DT. Heterologous Protein Secretion from Yeast. Science. 1985:**229**(4719):1219–1224. https://doi.org/10.1126/science.3939723

Snoeck S, Guidi C, and De Mey M. “Metabolic burden” explained: stress symptoms and its related responses induced by (over)expression of (heterologous) proteins in Escherichia coli. Microb Cell Factories. 2024:**23**:96. https://doi.org/10.1186/s12934-024-02370-9

Soanes DM, Alam I, Cornell M, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Oliver SG, and Talbot NJ. Comparative Genome Analysis of Filamentous Fungi Reveals Gene Family Expansions Associated with Fungal Pathogenesis. PLOS ONE. 2008:**3**(6):e2300. https://doi.org/10.1371/journal.pone.0002300

Sonnhammer EL, von Heijne G, and Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998:**6**:175–182.

Sørensen HP and Mortensen KK. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. Microb Cell Factories. 2005:**4**(1):1. https://doi.org/10.1186/1475-2859-4-1

Spangenberg SH, Zavareh RB, and Lairson LL. Protocol for high-throughput compound screening using flow cytometry in THP-1 cells. STAR Protoc. 2021:**2**(2):100400. https://doi.org/10.1016/j.xpro.2021.100400

Steiger MG, Vitikainen M, Uskonen P, Brunner K, Adam G, Pakula T, Penttilä M, Saloheimo M, Mach RL, and Mach-Aigner AR. Transformation System for Hypocrea jecorina (Trichoderma reesei) That Favors Homologous Integration and Employs Reusable Bidirectionally Selectable Markers. Appl Environ Microbiol. 2011:**77**(1):114–121. https://doi.org/10.1128/AEM.02100-10

Storms R, Zheng Y, Li H, Sillaots S, Martinez-Perez A, and Tsang A. Plasmid vectors for protein production, gene expression and molecular manipulations in *Aspergillus niger*. Plasmid. 2005:**53**(3):191–204. https://doi.org/10.1016/j.plasmid.2004.10.001

Strazza V, Rossi M, Avati A, Tiseo G, Falcone M, Cusi MG, Menichetti F, Ricciardi-Castagnoli P, Tinti C, and Pileri P. Rapid generation of human recombinant monoclonal antibodies from antibody-secreting cells using ferrofluid-based technology. Front Immunol. 2024:**15**. https://doi.org/10.3389/fimmu.2024.1341389

Stukenberg D, Zauner S, Dell'Aquila G, and Maier UG. Optimizing CRISPR/Cas9 for the Diatom Phaeodactylum tricornutum. Front Plant Sci. 2018:**9**.

Su M, Bastiaens L, Verspreet J, and Hayes M. Applications of Microalgae in Foods, Pharma and Feeds and Their Use as Fertilizers and Biostimulants: Legislation and Regulatory Aspects for Consideration. Foods. 2023:**12**(20):3878. https://doi.org/10.3390/foods12203878

Sun M, Qian K, Su N, Chang H, Liu J, and Shen G. Foot-and-mouth disease virus VP1 protein fused with cholera toxin B subunit expressed in Chlamydomonas reinhardtii chloroplast. Biotechnol Lett. 2003:**25**(13):1087–1092. https://doi.org/10.1023/A:1024140114505

Sun X, Li X, Shen X, Wang J, and Yuan Q. Recent advances in microbial production of phenolic compounds. Chin J Chem Eng. 2021:**30**:54–61. https://doi.org/10.1016/j.cjche.2020.09.001

Tan YQ, Xue B, and Yew WS. Genetically Encodable Scaffolds for Optimizing Enzyme Function. Molecules. 2021:**26**(5):1389. https://doi.org/10.3390/molecules26051389

Tanaka T, Maeda Y, Suhaimi N, Tsuneoka C, Nonoyama T, Yoshino T, Kato N, and Lauersen KJ. Intron-mediated enhancement of transgene expression in the oleaginous diatom *Fistulifera solaris* towards bisabolene production. Algal Res. 2021:**57**:102345. https://doi.org/10.1016/j.algal.2021.102345

Tang H, Wang S, Wang J, Song M, Xu M, Zhang M, Shen Y, Hou J, and Bao X. N-hypermannose glycosylation disruption enhances recombinant protein production by regulating secretory pathway and cell wall integrity in Saccharomyces cerevisiae. Sci Rep. 2016:**6**(1):25654. https://doi.org/10.1038/srep25654

Tannous A, Patel N, Tamura T, and Hebert DN. Reglucosylation by UDP-glucose:glycoprotein glucosyltransferase 1 delays glycoprotein secretion but not degradation. Mol Biol Cell. 2015:**26**(3):390–405. https://doi.org/10.1091/mbc.E14-08-1254

Taparia Y, Dolui AK, Boussiba S, and Khozin-Goldberg I. Multiplexed Genome Editing via an RNA Polymerase II Promoter-Driven sgRNA Array in the Diatom Phaeodactylum tricornutum: Insights Into the Role of StLDP. Front Plant Sci. 2022:**12**.

Taparia Y, Zarka A, Leu S, Zarivach R, Boussiba S, and Khozin-Goldberg I. A novel endogenous selection marker for the diatom Phaeodactylum tricornutum based on a unique mutation in phytoene desaturase 1. Sci Rep. 2019:**9**(1):1–12. https://doi.org/10.1038/s41598-019-44710-5

Taton A, Ecker A, Diaz B, Moss NA, Anderson B, Reher R, Leão TF, Simkovsky R, Dorrestein PC, Gerwick L, et al. Heterologous Expression of Cryptomaldamide in a Cyanobacterial Host. ACS Synth Biol. 2020:**9**(12):3364–3376. https://doi.org/10.1021/acssynbio.0c00431

Taton A, Unglaub F, Wright NE, Zeng WY, Paz-Yepes J, Brahamsha B, Palenik B, Peterson TC, Haerizadeh F, Golden SS, et al. Broad-host-range vector system for synthetic biology and biotechnology in cyanobacteria. Nucleic Acids Res. 2014:**42**(17):e136–e136. https://doi.org/10.1093/nar/gku673

Terpe K. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol. 2006:**72**(2):211–222. https://doi.org/10.1007/s00253-006-0465-8

Tiedge KJ. Piece by piece: making plant natural products accessible via heterologous biosynthesis. Synth Biol. 2022:**7**(1):ysac028. https://doi.org/10.1093/synbio/ysac028

Tirichine L, Rastogi A, and Bowler C. Recent progress in diatom genomics and epigenomics. Curr Opin Plant Biol. 2017:**36**:46–55. https://doi.org/10.1016/j.pbi.2017.02.001

Tolia NH and Joshua-Tor L. Strategies for protein coexpression in Escherichia coli. Nat Methods. 2006:**3**(1):55–64. https://doi.org/10.1038/nmeth0106-55

Tomimoto K, Fujita Y, Iwaki T, Chiba Y, Jigami Y, Nakayama K, Nakajima Y, and Abe H. Protease-deficient Saccharomyces cerevisiae strains for the synthesis of human-compatible glycoproteins. Biosci Biotechnol Biochem. 2013:**77**(12):2461–2466. https://doi.org/10.1271/bbb.130588

Tomoiaga D, Bubnell J, Herndon L, and Feinstein P. High rates of plasmid cotransformation in E. coli overturn the clonality myth and reveal colony development. Sci Rep. 2022:**12**(1):11515. https://doi.org/10.1038/s41598-022-14598-9

Tong Y, Jørgensen TS, Whitford CM, Weber T, and Lee SY. A versatile genetic engineering toolkit for E. coli based on CRISPR-prime editing. Nat Commun. 2021:**12**(1):5206. https://doi.org/10.1038/s41467-021-25541-3

Tran NT and Kaldenhoff R. Achievements and challenges of genetic engineering of the model green alga *Chlamydomonas reinhardtii*. Algal Res. 2020:**50**:101986. https://doi.org/10.1016/j.algal.2020.101986

Tschumper G and Carbon J. Copy number control by a yeast centromere. Gene. 1983:**23**(2):221–232. https://doi.org/10.1016/0378-1119(83)90054-9

United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects 2024: Summary of Results. 2024.

Valenzuela-Ortega M and French C. Joint universal modular plasmids (JUMP): a flexible vector platform for synthetic biology. Synth Biol. 2021:**6**(1):ysab003. https://doi.org/10.1093/synbio/ysab003

Vavitsas K, Rue EØ, Stefánsdóttir LK, Gnanasekaran T, Blennow A, Crocoll C, Gudmundsson S, and Jensen PE. Responses of Synechocystis sp. PCC 6803 to heterologous biosynthetic pathways. Microb Cell Factories. 2017:**16**(1):140. https://doi.org/10.1186/s12934-017-0757-y

Vojcic L, Despotovic D, Martinez R, Maurer K-H, and Schwaneberg U. An efficient transformation method for Bacillus subtilis DB104. Appl Microbiol Biotechnol. 2012:**94**:487–93. https://doi.org/10.1007/s00253-012-3987-2

Wakai S, Arazoe T, Ogino C, and Kondo A. Future insights in fungal metabolic engineering. Bioresour Technol. 2017:**245**:1314–1326. https://doi.org/10.1016/j.biortech.2017.04.095

Walker EJL, Pampuch M, Chang N, Cochrane RR, and Karas BJ. Design and assembly of the 117-kb Phaeodactylum tricornutum chloroplast genome. Plant Physiol. 2024:**194**(4):2217–2228. https://doi.org/10.1093/plphys/kiad670

Wannathong T, Waterhouse JC, Young REB, Economou CK, and Purton S. New tools for chloroplast genetic engineering allow the synthesis of human growth hormone in the green alga Chlamydomonas reinhardtii. Appl Microbiol Biotechnol.

2016:**100**(12):5467–5477. https://doi.org/10.1007/s00253-016-7354-6

Watanabe Y, Kadono T, Kira N, Suzuki K, Iwata O, Ohnishi K, Yamaguchi H, and Adachi M. Development of endogenous promoters that drive high-level expression of introduced genes in the model diatom Phaeodactylum tricornutum. Mar Genomics. 2018:**42**:41–48. https://doi.org/10.1016/j.margen.2018.06.003

Watts A, Sankaranarayanan S, Watts A, and Raipuria RK. Optimizing protein expression in heterologous system: Strategies and tools. Meta Gene. 2021:**29**:100899. https://doi.org/10.1016/j.mgene.2021.100899

Wawrosch C and Zotchev SB. Production of bioactive plant secondary metabolites through in vitro technologies—status and outlook. Appl Microbiol Biotechnol. 2021:**105**(18):6649–6668. https://doi.org/10.1007/s00253-021-11539-w

Welch LG and Munro S. A tale of short tails, through thick and thin: investigating the sorting mechanisms of Golgi enzymes. FEBS Lett. 2019:**593**(17):2452–2465. https://doi.org/10.1002/1873-3468.13553

Wentz AE and Shusta EV. A Novel High-Throughput Screen Reveals Yeast Genes That Increase Secretion of Heterologous Proteins. Appl Environ Microbiol. 2007:**73**(4):1189–1198. https://doi.org/10.1128/AEM.02427-06

Wichmann J, Baier T, Wentnagel E, Lauersen KJ, and Kruse O. Tailored carbon partitioning for phototrophic production of (*E*)-α-bisabolene from the green microalga *Chlamydomonas reinhardtii*. Metab Eng. 2018:**45**:211–222. https://doi.org/10.1016/j.ymben.2017.12.010

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. J Open Source Softw. 2019:**4**(43):1686. https://doi.org/10.21105/joss.01686

Windhagauer M, Abbriano RM, Ashworth J, Barolo L, Jaramillo-Madrid AC, Pernice M, and Doblin MA. Characterisation of novel regulatory sequences compatible with modular assembly in the diatom *Phaeodactylum tricornutum*. Algal Res. 2021:**53**:102159. https://doi.org/10.1016/j.algal.2020.102159

Windhagauer M, Doblin MA, Signal B, Kuzhiumparambil U, Fabris M, and Abbriano RM. Metabolic response to a heterologous poly-3-hydroxybutyrate (PHB) pathway in Phaeodactylum tricornutum. Appl Microbiol Biotechnol. 2024:**108**(1):104. https://doi.org/10.1007/s00253-023-12823-7

Wozniak KJ and Simmons LA. Genome editing methods for Bacillus subtilis. Methods Mol Biol Clifton NJ. 2022:**2479**:159–174. https://doi.org/10.1007/978-1-0716-2233-9_11

Wu Q, Xu C, Shi W, Li L, Zhang H, Liu T, Fan J, Cui L, and Li J. Suitable carrier protein and linker peptide significantly increase the secretory expression of human lysozyme in Aspergillus niger. Acta Biochim Biophys Sin. 2023:**55**(10):1677–1680. https://doi.org/10.3724/abbs.2023153

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021:**2**(3):100141. https://doi.org/10.1016/j.xinn.2021.100141

Xin X, Zhang R-K, Liu S-C, He Z-J, Liu R-Y, Lan H-N, Liu Z-H, Li B-Z, and Yuan Y-J. Engineering yeast to convert lignocellulose into vanillin. Chem Eng J. 2024:**485**:149815. https://doi.org/10.1016/j.cej.2024.149815

Xu P, Vansiri A, Bhan N, and Koffas MAG. ePathBrick: A Synthetic Biology Platform for Engineering Metabolic Pathways in E. coli. ACS Synth Biol. 2012:**1**(7):256–266. https://doi.org/10.1021/sb300016b

Yan L, Chen P, Zhang S, Li S, Yan X, Wang N, Liang N, and Li H. Biotransformation of ferulic acid to vanillin in the packed bed-stirred fermentors. Sci Rep. 2016:**6**:34644. https://doi.org/10.1038/srep34644

Yang G, Cozad MA, Holland DA, Zhang Y, Luesch H, and Ding Y. Photosynthetic Production of Sunscreen Shinorine Using an Engineered Cyanobacterium. ACS Synth Biol. 2018a:**7**(2):664–671. https://doi.org/10.1021/acssynbio.7b00397

Yang H, Barros-Rios J, Kourteva G, Rao X, Chen F, Shen H, Liu C, Podstolski A, Belanger F, Havkin-Frenkel D, et al. A re-evaluation of the final step of vanillin biosynthesis in the orchid Vanilla planifolia. Phytochemistry. 2017:**139**:33–46. https://doi.org/10.1016/j.phytochem.2017.04.003

Yang J-R, Chen X, and Zhang J. Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLOS Biol. 2014:**12**(7):e1001910. https://doi.org/10.1371/journal.pbio.1001910

Yang M, Lin X, Liu X, Zhang J, and Ge F. Genome Annotation of a Model Diatom Phaeodactylum tricornutum Using an Integrated Proteogenomic Pipeline. Mol Plant. 2018b:**11**(10):1292–1307. https://doi.org/10.1016/j.molp.2018.08.005

Yang Z, Li yinü, Chen F, Li D, Zhang Z, Liu Y, Zheng D, Wang Y, and Shen G. Expression of human soluble TRAIL in Chlamydomonas

reinhardtii chloroplast. Chin Sci Bull. 2006:**51**(14):1703–1709. https://doi.org/10.1007/s11434-006-2041-0

Ye X and Ho JWK. Ultrafast clustering of single-cell flow cytometry data using FlowGrid. BMC Syst Biol. 2019:**13**(2):35. https://doi.org/10.1186/s12918-019-0690-2

Yeo WL, Heng E, Tan LL, Lim YW, Lim YH, Hoon S, Zhao H, Zhang MM, and Wong FT. Characterization of Cas proteins for CRISPR-Cas editing in streptomycetes. Biotechnol Bioeng. 2019:**116**(9):2330–2338. https://doi.org/10.1002/bit.27021

Yin X, Shin H-D, Li J, Du G, Liu L, and Chen J. Pgas, a Low-pH-Induced Promoter, as a Tool for Dynamic Control of Gene Expression for Metabolic Engineering of Aspergillus niger. Appl Environ Microbiol. 2017:**83**(6):e03222-16. https://doi.org/10.1128/AEM.03222-16

Yurtsev E and Friedman J. FlowCytometryTools. 2015. https://doi.org/10.5281/ZENODO.32992

Zecchin A, Stapor PC, Goveia J, and Carmeliet P. Metabolic pathway compartmentalization: an underappreciated opportunity? Curr Opin Biotechnol. 2015:**34**:73–81. https://doi.org/10.1016/j.copbio.2014.11.022

Zeng H and Cheung Y-M. A new feature selection method for Gaussian mixture clustering. Pattern Recognit. 2009:**42**(2):243–250. https://doi.org/10.1016/j.patcog.2008.05.030

Zhang JJ, Tang X, and Moore BS. Genetic platforms for heterologous expression of microbial natural products. Nat Prod Rep. 2019:**36**(9):1313–1332. https://doi.org/10.1039/C9NP00025A

Zhang X and Zhang Y -H. P. Simple, fast and high-efficiency transformation system for directed evolution of cellulase in Bacillus subtilis. Microb Biotechnol. 2011:**4**(1):98–105. https://doi.org/10.1111/j.1751-7915.2010.00230.x

Zhang Y, Park C, Bennett C, Thornton M, and Kim D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. Genome Res. 2021:**31**(7):1290–1295. https://doi.org/10.1101/gr.275193.120

Zhang Y-HP, Sun J, and Ma Y. Biomanufacturing: history and perspective. J Ind Microbiol Biotechnol. 2017:**44**(4–5):773–784. https://doi.org/10.1007/s10295-016-1863-2

Zhao M, Ma J, Zhang L, and Qi H. Engineering strategies for enhanced heterologous protein production by Saccharomyces cerevisiae. Microb Cell Factories. 2024:**23**:32. https://doi.org/10.1186/s12934-024-02299-z

Zhao Y, Li G, Chen Y, and Lu Y. Challenges and Advances in Genome Editing Technologies in Streptomyces. Biomolecules. 2020:**10**(5):734. https://doi.org/10.3390/biom10050734

Zharikova NV, Iasakov TR, Bumazhkin BK, Patutina EO, Zhurenko EI, Korobov VV, Sagitova AI, Kuznetsov BB, and Markusheva TV. Isolation and sequence analysis of pCS36-4CPA, a small plasmid from *Citrobacter* sp. 36-4CPA. Saudi J Biol Sci. 2018:**25**(4):660–671. https://doi.org/10.1016/j.sjbs.2016.02.014

Zhu L, Qiu X, Zhu L, Wu X, Zhang Y, Zhu Q, Fan D, Zhu C, and Zhang D. Spatial organization of heterologous metabolic system in vivo based on TALE. Sci Rep. 2016:**6**(1):26065. https://doi.org/10.1038/srep26065

Zhu X, Liu X, Liu T, Wang Y, Ahmed N, Li Z, and Jiang H. Synthetic biology of plant natural products: From pathway elucidation to engineered biosynthesis in plant cells. Plant Commun. 2021:**2**(5):100229. https://doi.org/10.1016/j.xplc.2021.100229

ANNEX A: Supplementary Information from Chapter II

Table S2. 1. Primer sequences used for Gibson assembly.

Name	Sequence
FW-BiPCS-Gibson	GCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACCATGAGATCCTTTTGCATCGC
RV-BiPCS-Gibson	GTTGCG TGA CCG ATC GAA TGG GGT TAT CTT CTT CAA ATC CATC CAT GAC AAT CGT TGC TTT AC
FW-SiGol-Gibson	GCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACC ATG GCG TTT TTG CCG AAT C
EV-SiGol-Gibson	GTT GCG TGA CCG ATC GAA TGG GGT TAT CTT CTT CAA ATC CTGC ACT CAT CCA CAT GAG AAA
FW-VANCS-Gibson	CTCTGCTCATCGCACGCGTAAAGCAACGATTGTCATGG AT GGA TTT GAA GAA GAT AAC CCC
FW-VANGol-Gibson	TTTTGCCCTCAATATATGCTTTCTCATGTGGATGAGTGC A GGA TTT GAA GAA GAT AAC CCC
FW-GFPBb-Gibson	TGCCTCGTACCCGATTGTCGCCGTCATTGAAGACGGAC GA ATG GTG AGC AAG GGC GAG GAG C

RV-GFPBb-Gibson	CAGCCAAAGTCGAGGTAGTTGTTGCGGTTA GAA TTC CTT GTA CAG CTC GTC CAT GCC GAG
RV-GFPPeSi-Gibson	GAAAGTGTCCCAGCCAAAGTCGAGGTAGTTGTTGCGGT TA CAA CTT CGA GAA TTC CTT GTA CAG CTC GTC CAT GCC GAG
FW-BbGFP-Gibson	CTC GGC ATG GAC GAG CTGTACAAGGAATTC TAA CCG CAA CAA CTA CCT C
RV-BbGFP-Gibson	GCG TGA CCG ATC GAA TGG GGT TAT CTT CTT CAA ATC CCA TGGT ATT CTA TTC TCT GAT TCC TTC G
FW-BbPeSi-Gibson	GACGAGCTGTACAAGGAATTC TCGAAGTTG TAA CCG CAA CAA CTA CCT C
RV-BbCS-Gibson	ATG CCA CAG CCA AAA GGG CTG CGA TGC AAA AGG ATC TCA TGGT ATT CTA TTC TCT GAT TCC TTC G
RV-BbER-Gibson	GCA AGG CCA GTG CTG CTA CGG CAA TTC TCA TGA ACA TCA TGGT ATT CTA TTC TCT GAT TCC TTC G
RV-BbGol-Gibson	AGG CCA CGG TAT ATC TTC GTC GAT TCG GCA AAA ACG CCA TGGT ATT CTA TTC TCT GAT TCC TTC G

40SRPS8F	GCA CCA CCC CGG TGA ACA GCT CCT CGC CCT TGC TCA CCA TGG TAT TCT ATT CTC TGA TTC C
GFPR	CGT TGA TCT TGC ACC GAA GGA ATC AGA GAA TAG AAT ACC ATG GTG AGC AAG GGC GAG G
Go40SRPS8F	AGG CCA CGG TAT ATC TTC GTC GAT TCG GCA AAA ACG CCA TGG TAT TCT ATT CTC TGA TTC
GoGFPR	GCG TTG ATC TTG CAC CGA AGG AAT CAG AGA ATA GAA TAC CAT GGC GTT TTT GCC GAA TCG
CS40SRPS8F	ATG CCA CAG CCA AAA GGG CTG CGA TGC AAA AGG ATC TCA TGG TAT TCT ATT CTC TGA TTC
CSGFPR	GCG TTG ATC TTG CAC CGA AGG AAT CAG AGA ATA GAA TAC CAT GAG ATC CTT TTG CAT CGC
VANF	GTC GAG GTA GTT GTT GCG GTT AGT GGT GGT GGT GAT GGT GTC GTC CGT CTT CAA TGA CGG
PeVANF	GTT GTT GCG GTT ACA ACT TCG AGT GGT GGT GGT GAT GGT GTC GTC CGT CTT CAA TGA CGG

Table S2. 2. Sequences used in this study.

Name	Sequence
VpVAN	GGATTTGAAGAAGATAACCCCATTCGATCGGTCACG CAACGACCCGATTCCATTGAACCCGCCATTTTGGGC GTTTTGGGTTCGTGCCGCCACGCCTTTCACTTTGCC CGATTTGCCCGCCGATACGGAAAGTCGTACGGATC GGAAGAAGAAATTAAGAAGCGATTTGGAATTTTTGT CGAAAACTTGGCCTTTATTCGATCGACGAACCGTAA GGATTTGTCGTACACGTTGGGAATTAACCAATTTGC CGATTTGACGTGGGAAGAATTTCGAACGAACCGATT GGGAGCCGCCCAAAACTGTTCCGCCACGGCCCACG GAAACCACCGCTTTGTCGATGGTGTCTTGCCCGTCA CGCGTGATTGGCGCGAACAAGGAATTGTATCCCCG GTCAAAGACCAAGGTTCGTGCGGATCGTGCTGGAC GTTTTCCACCACGGGAGCCTTGGAAGCCGCCTACA CCCAATTGACGGGAAAATCGACCTCGTTGTCGGAAC AACAATTGGTTGACTGCGCTTCGGCCTTTAACAACT TTGGATGTAACGGAGGATTGCCGTCCCAGGCTTTTG AATACGTCAAATACAACGGTGGAATTGACACCGAAC AAACCTACCCCTACCTCGGAGTCAACGGAATTTGTA ACTTTAAGCAAGAAAACGTCGGAGTCAAAGTCATTG ACTCCATTAACATTACCTTGGGAGCCGAAGATGAAT TGAAACACGCCGTCGGATTGGTTCGCCCCGTCTCG GTCGCCTTTGAAGTCGTCAAGGGTTTTAACTTGTAC AAAAAGGGAGTCTACTCGTCCGACACGTGCGGACG

	AGATCCAATGGATGTCAACCACGCCGTCTTGGCCGT CGGTTACGGAGTCGAAGACGGTATTCCGTACTGGTT GATCAAAAACTCGTGGGGAACGAACTGGGGAGATA ACGGATACTTTAAGATGGAATTGGGAAAAAACATGT GCGGAGTCGCCACGTGTGCCTCGTACCCGATTGTC GCCGTCA
eGFP	ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGT GGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAA ACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAG GGCGATGCCACCTACGGCAAGCTGACCCTGAAGTT CATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGC CCACCCTCGTGACCACCCTGACCTACGGCGTGCAG TGCTTCAGCCGCTACCCCGACCACATGAAGCAGCA CGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGT CCAGGAGCGCACCATCTTCTTCAAGGACGACGGCA ACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGC GACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT CGACTTCAAGGAGGACGGCAACATCCTGGGGCACA AGCTGGAGTACAACTACAACAGCCACAACGTCTATA TCATGGCCGACAAGCAGAAGAACGGCATCAAGGTG AACTTCAAGATCCGCCACAACATCGAGGACGGCAG CGTGCAGCTCGCCGACCACTACCAGCAGAACACCC CCATCGGCGACGGCCCCGTGCTGCTGCCCGACAAC CACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGA CCCCAACGAGAAGCGCGATCACATGGTCCTGCTGG

	AGTTCGTGACCGCCGCCGGGATCACTCTCGGCATG GACGAGCTGTACAAGGAATTC
p40SRPS8	CCCTGCGATAGACCTTTTCCAAACTCACGCAGTCCA AGAAAACAAAGGGGTGAGAAGTATACGCACCTTTCG GTTTCGGCATAATTCTTAAACTCTTGTGGTCACTTTC TTGTGAAGAAGCTAGGGGCACTCGTTTTCCCTCAGA GCCTGCAAACACAAAATTCCTGCAGTCAATTGTCCC AACACTCGGCAAACCGTATGCGCAAGCAACGATGC GCAGAAGGCCGTGGATGGATGGCGACTCGCGATAT GGCTTCTTGGGTCGCCAGTGTGGTACGTCCGGCGT ATGTCAATACGCGAATTCGGACGACTGGCATCTCTA GGAGGAGGATTCCTTCTTTTATGACATGTTTATTTTA TATACATTGATGCTTTCCGACAGTCGGAAGTAATAAA TGAATTTATTTCAAGACTACCTATACTCCTTTGACTT GTTCGACTAATCTTACCGCTTACTAAAATCTCGAAAT CACGCTTGACCTCTCGCACGCAAATTTTTGCTGCTG GACGCTACGCACTCGGCCCAATTCTTCTCGGTCCTC GTCGTCGCAATTGTCGTTGCGTTGATCTTGCACCGA AGGAATCAGAGAATAGAATACC
tFcpA	CCGCAACAACTACCTCGACTTTGGCTGGGACACTTT CAGTGAGGACAAGAAGCTTCAGAAGCGTGCTATCG AACTCAACCAGGGACGTGCGGCACAAATGGGCATC CTTGCTCTCATGGTGCACGAACAGTTGGGAGTCTCT ATCCTTCCTTAAAAATTTAATTTTCATTAGTTGCAGTC ACTCCGCTTTGGTTT

ATPC^{BPS}	ATGAGATCCTTTTGCATCGCAGCCCTTTTGGCTGTG GCATCTGCCTTCACCACACAGCCAACTTCCTTCACT GTGAAGACTGCGAATGTGGGCGAACGGGCGAGTGG GGTTTTCCCTGAGCAGAGCTCTGCTCATCGCACGC GTAAAGCAACGATTGTCATGGAT
PTS1	TCGAAGTTG
XylT^{CT}	ATGGCGTTTTTGCCGAATCGACGAAGATATACCGTG GCCTGCTTGTTTATCGGTTTTGCCCTCAATATATGCT TTCTCATGTGGATGAGTGCA
Xa	TTGAAGACGGACGA

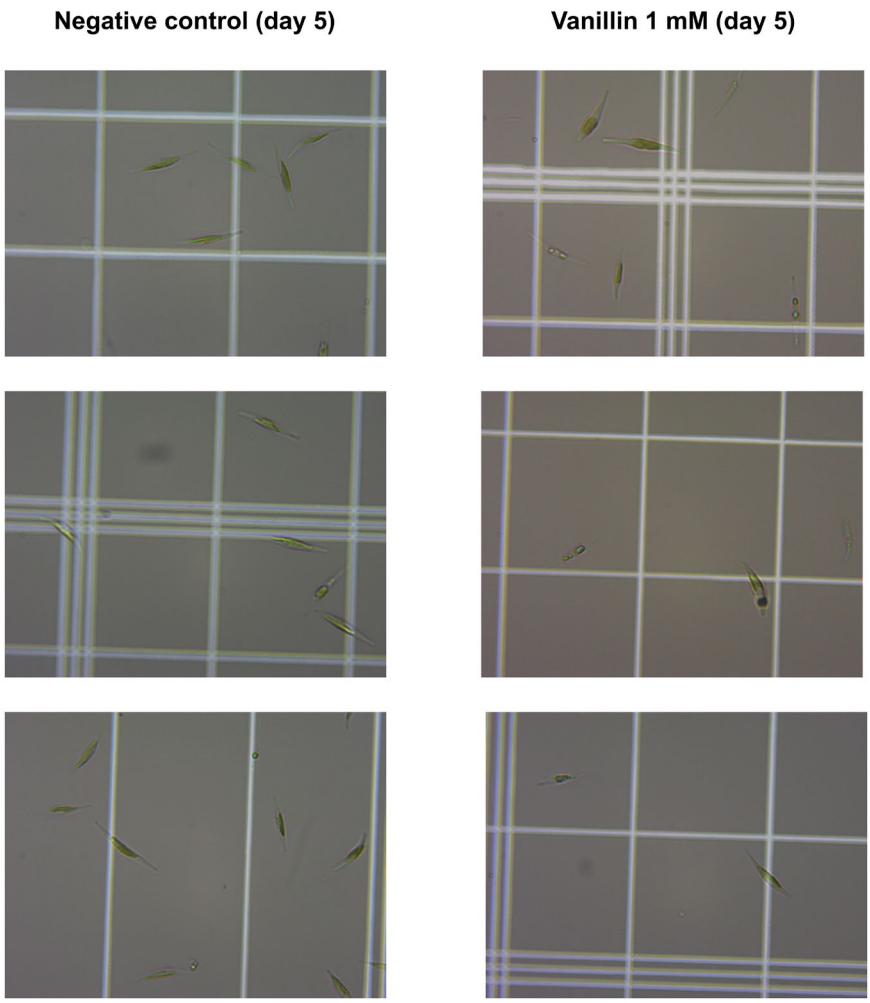


Figure S2. 1. The morphology of *P. tricornutum* cells treated with 1 mM of vanillin is similar to untreated cells at day five. Three biological replicates of the wild-type strain are shown for the non-treated (left panels) and treated cells with vanillin (right panels). Microscopy images were obtained using the 40x objective.

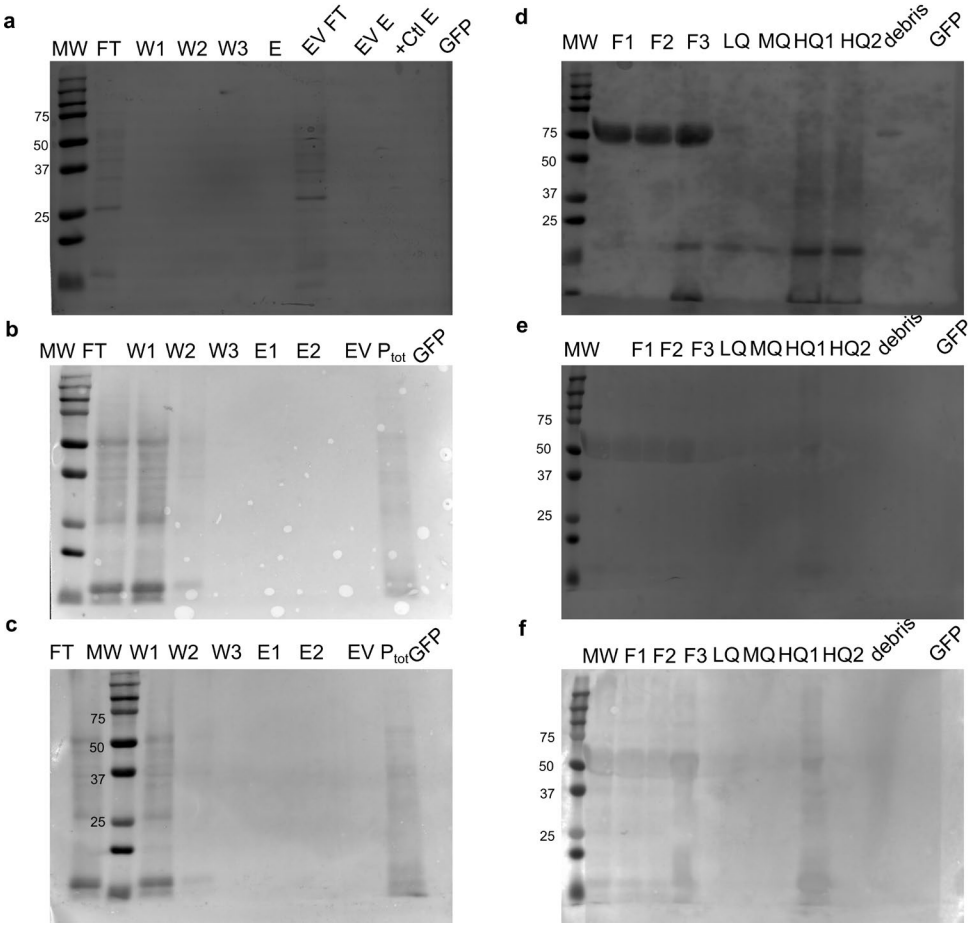


Figure S2. 2. Proteins transferred to the blotting membrane. Red ponceau staining of membranes used for western blot: (**a**-**c**) protein extracts from three different clones with cytosol constructions after purification using GFP-trap; (**d**-**f**) protein extract of plastid purification of three different clones harboring the plastid construction. MW: molecular weight marker; FT: flow through; W1-3: washes; E1-2: elution fraction; P_{tot}: total protein; +Ctl: eGFP:T2A:mCherry clone; F1: total protein extract; F2: thylakoids fragments and mitochondria; F3: plastids; LQ: low quality plastids; MQ: medium quality plastids.

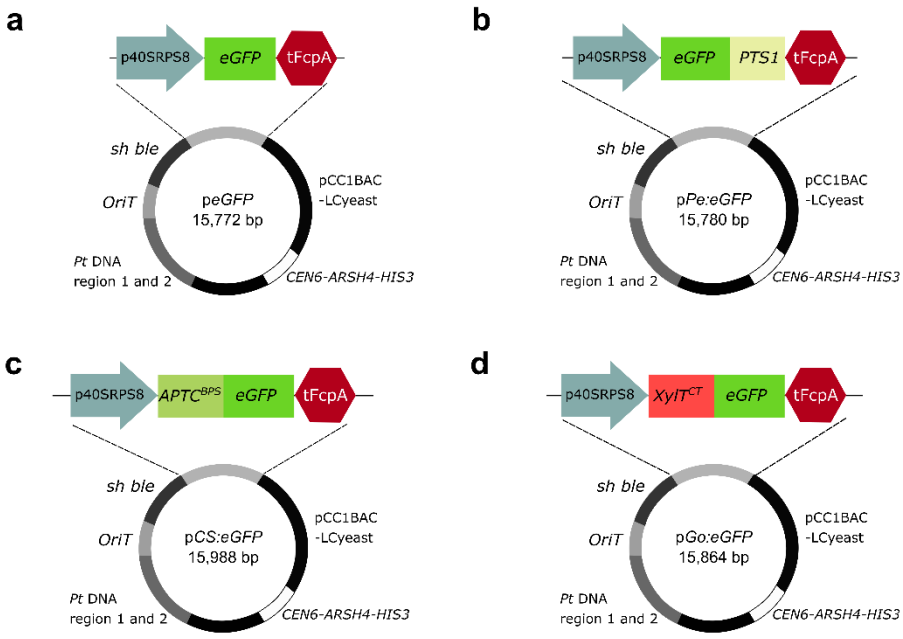


Figure S2. 3. Constructions of controls for subcellular localization. a. cytosolic construction *eGFP* without any localization signal; **b**. peroxisome construction harboring the peroxisomal targeting signal 1 (*PST1*); **c**. plastid construction with the bipartite signal of the chloroplast ATPase gamma subunit (*ATPC*^{BPS}); **d.** β1,2-xylosyltransferase cytosolic tail and transmembrane domain (CT) for medial Golgi (*XylT*^{CT}). *eGFP* expression cassette is antisense from *sh ble*.

ANNEX B: Supporting Information from Chapter III

Supplementary Figures

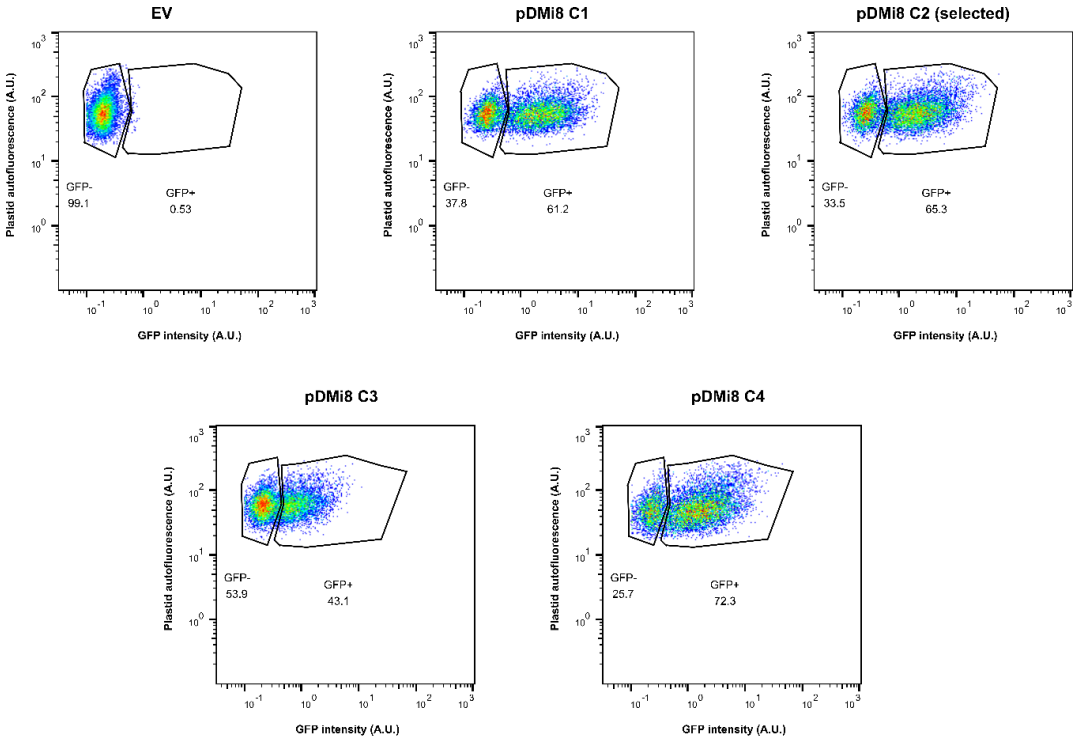


Fig. S3. 1. Initial analysis of GFP cell lines (pDMi8) do not present three stable subpopulations. Flow cytometry analysis of four cell lines containing pDMi8 plasmid and the empty vector (EV) strain.

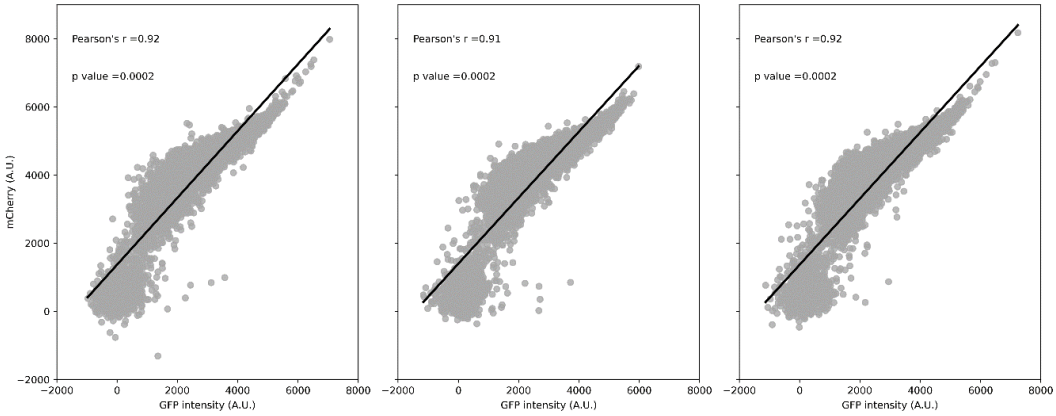


Fig. S3. 2. Fluorescence intensity of eGFP and mCherry is highly correlated. Analysis was performed in triplicates.

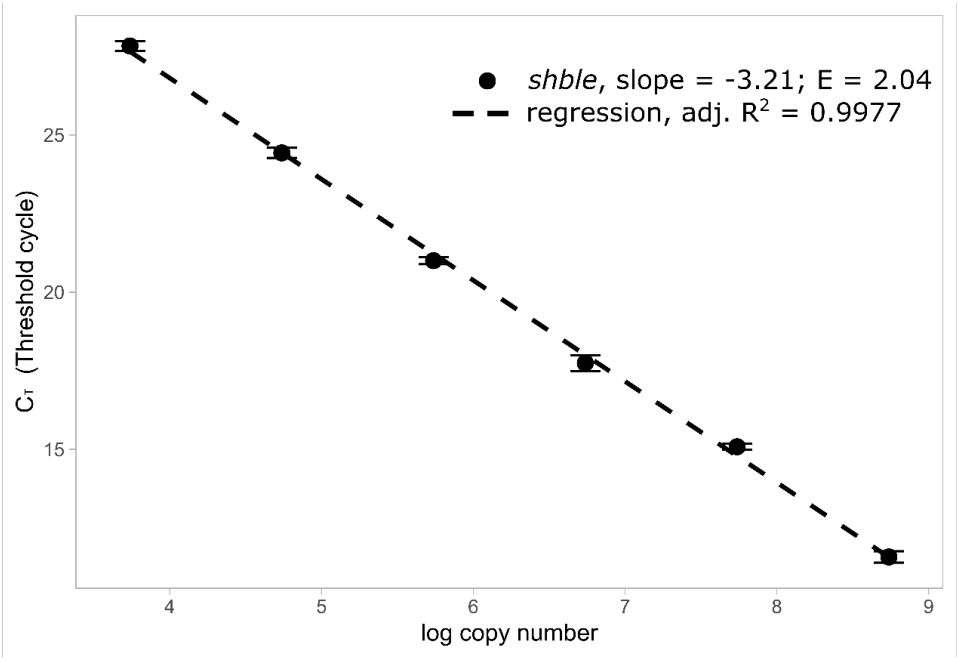


Fig. S3. 3. *Sh ble* primer pair validation for copy number quantification.

Amplification efficiency (E) was calculated by the equation $E = 10^{-1/slope} - 1$.

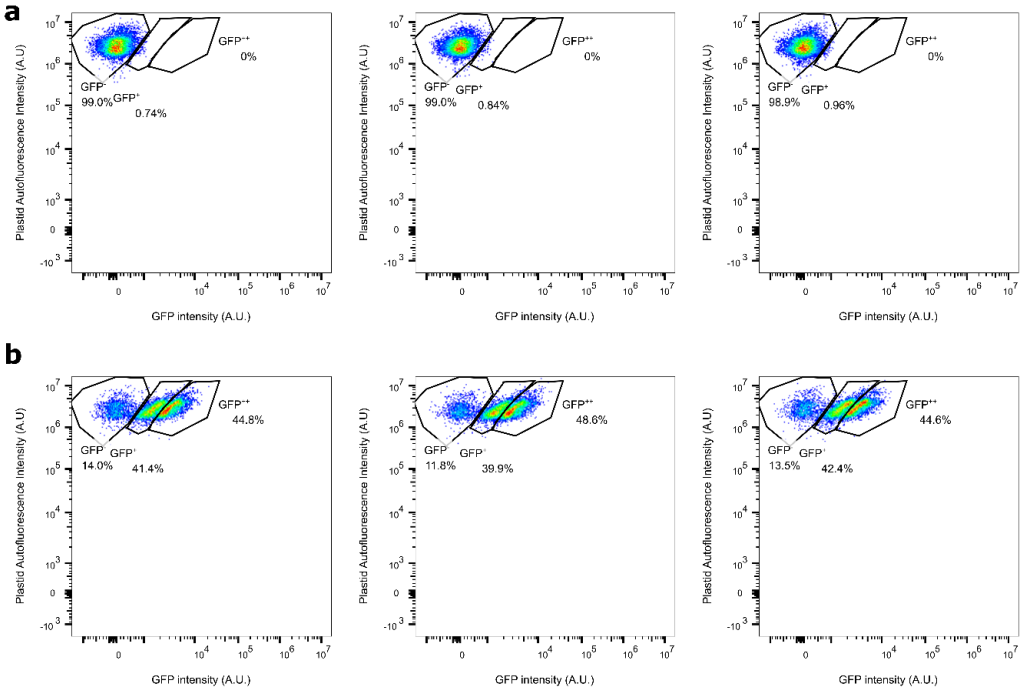


Fig. S3. 4. Subpopulations are detected in a clonally propagated *P. tricornutum* cell culture. (a) Negative control empty vector strain used for gating the GFP^- subpopulation. **(b)** GFP cell line with diverse profile of fluorescence per cell grouping in three subpopulations: GFP^-, GFP^+, and GFP^{++}. Percentages represent the number of cells in each gate.

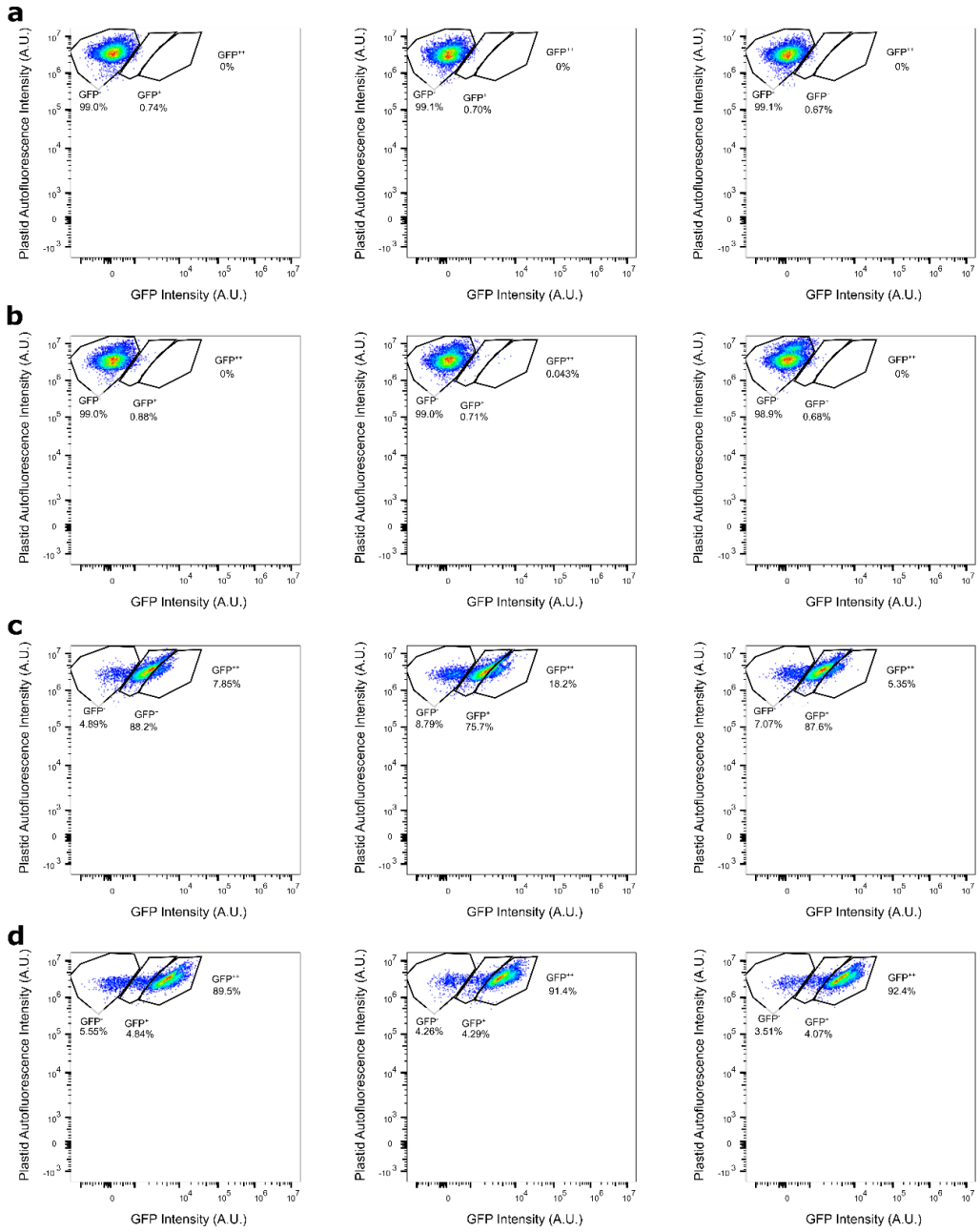


Fig. S3. 5. Enrichment of subpopulations through fluorescence activated cell sorting. (a) Sorted empty vector strain in triplicates. **(b-d)** Cultures in triplicate of enriched subpopulations from GFP cell line: **(b)** GFP^-, **(c)** GFP^+, and **(d)** GFP^{++}.

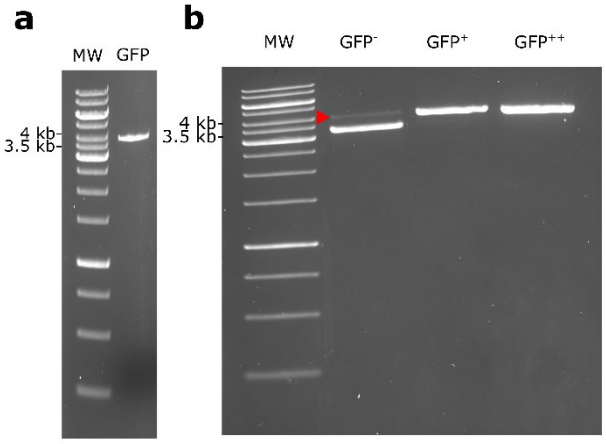


Fig. S3. 6. PCR amplification reveals two different versions of expression cassettes are present in GFP- subpopulation. Amplification of the expression cassette from plasmid extract of *P. tricornutum* cultures **(a)** before (GFP) and **(b)** after sorting (GFP^-, GFP^+, and GFP^{++}). The red arrow indicates the faint band present at the expected size of 4 kb. MW: molecular weight marker.

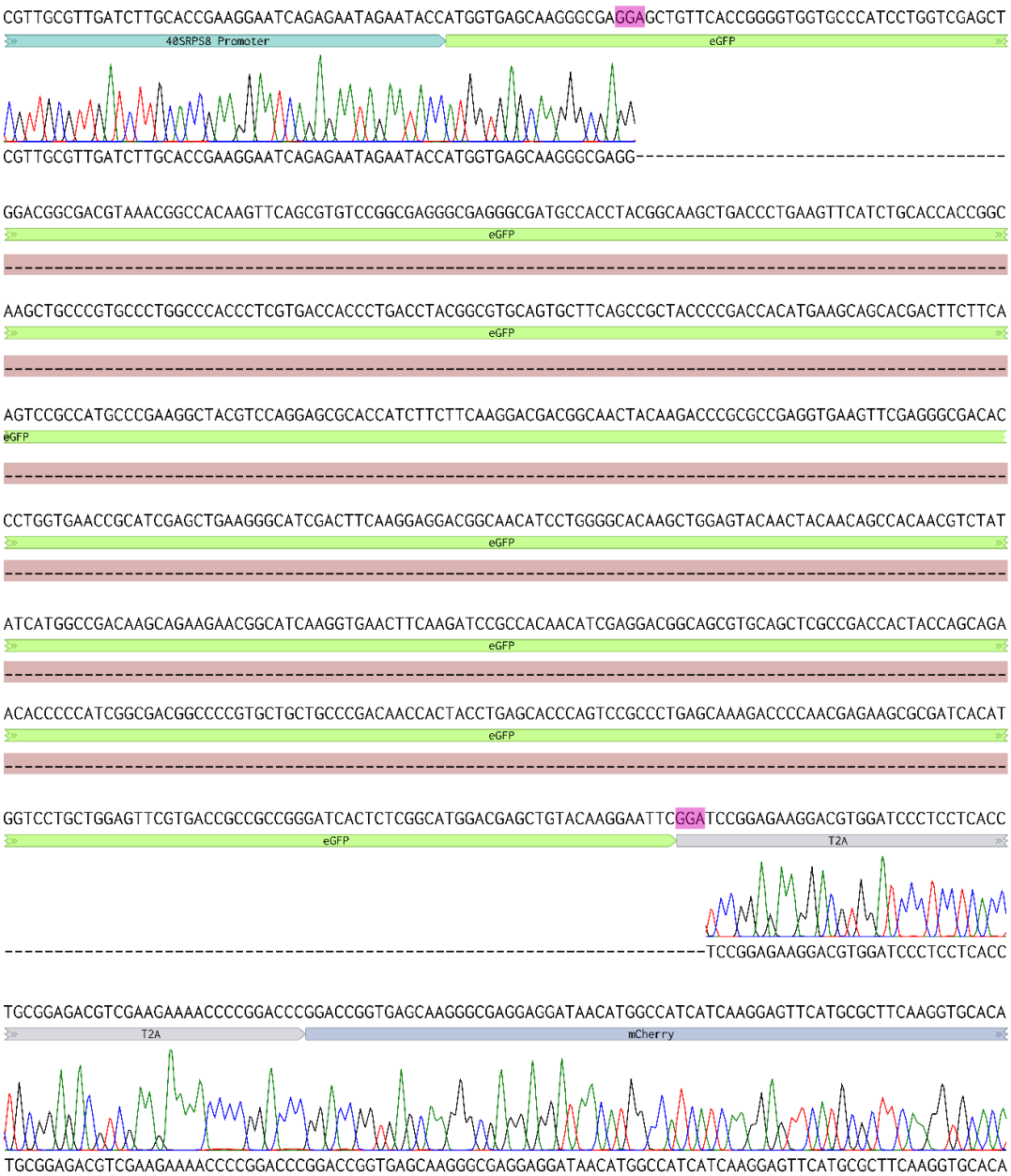


Fig. S3. 7. Deletion of 707 bp in GFP$^-$ sorted clones. Deletion map showing the sequencing results of GFP$^-$ enriched cultures, microhomologous 3 bp region are highlighted in the template sequence in magenta. Created with benchling.

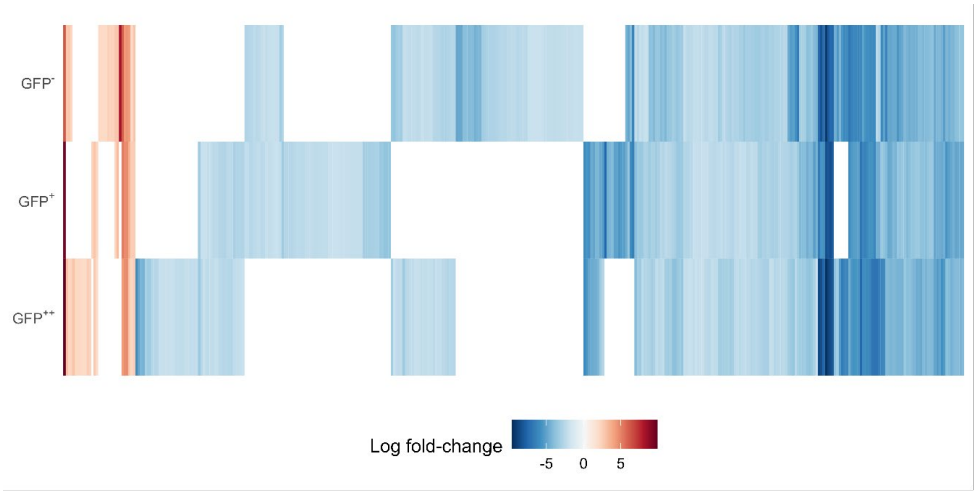


Fig. S3. 8. Commonly deregulated genes between subpopulations. Heatmap of deregulated genes. Genes down-regulated correspond to log fold-change < -2 and an adjusted p value < 0.05. Up-regulated genes correspond to log fold-change > 2 and an adjusted p value < 0.05.

Supplementary Tables

Table S3. 1. Primer sequences used for Sanger sequencing and qPCR to measure copy number.

Name	Sequence
pPtGE30 bb F	GCAATGGACAGAACAACCTAATG
pPtGE30 bb R	GGCAAAATCCTGTATATCGTGCG
p3-Sanger	GAGCCTACATCCTTCTGCAA
p4-Sanger	GAACCGCATCGAGCTGAA
p5-Sanger	TGCTTGTCGGCCATGATATAG
p6-Sanger	GACTACTTGAAGCTGTCCTTCC
qPCR-shble F	CGACGTGACCCTGTTCATC
qPCR-shble R	ACGACCTCCGACCACTC
qPCR-Ru F	CCACTCGCGGAATTCGATTA
qPCR-Ru R	GGCATCGTCCTCTTCACTTT

Table S3. 2. Highly interconnected genes from module 73 of co-expression analysis.

Gene ID	Functional domains (CLADE, DAMA)
Phatr3_EG02315	Metaviral_G glycoprotein;Mycoplasma protein of unknown function, DUF285;Putative RNA methylase family UPF0020;Leucine Rich repeats (2 copies);
Phatr3_J3183	Pyridoxal-phosphate dependent enzyme;
Phatr3_J44950	Transmembrane secretion effector;
Phatr3_J47263	RNA binding activity-knot of a chromodomain ;
Phatr3_J48068	Cytochrome c/c1 heme lyase;
Phatr3_J48362	Leucine rich repeat N-terminal domain;
Phatr3_J52285	Zn-dependent metallo-hydrolase RNA specificity domain;Beta-Casp domain;Cleavage and polyadenylation factor 2 C-terminal;Metallo-beta-lactamase superfamily;
Phatr3_J8686	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family;
Phatr3_Jdraft1387	Galactose-3-O-sulfotransferase ;
Phatr3_Jdraft144	Calcineurin-like phosphoesterase;

ANNEX C: Clustering algorithms for flow cytometry data

Flow cytometry allows us to assess protein production by *proxy* in individual cells, tagging the protein of interest with reporter proteins such as the green fluorescent protein (GFP) (Muusse et al. 2022). This advantage of flow cytometry made possible the development of protocols for high through put screening of recombinant protein production (Spangenberg et al. 2021). However, analyzing data can be time consuming and subjective by manually gating subpopulations, since it relies on visual inspection of scatter plots (Ye and Ho 2019).

The use of unsupervised algorithms to analyze flow cytometry data can speed up analysis and provide an objective manner to cluster subpopulations. In recent years, algorithms to cluster flow cytometry data have been implemented such as FlowPeaks and FLOCK, based on k-means clustering (Ge and Sealfon 2012; Dorfman et al. 2016). K-means clustering algorithm performs well with compact hyper-spherical clusters, however is less efficient with asymmetric clusters (Ahmed et al. 2020).

Another algorithm is Gaussian Mixture Models (GMM) which assigns data point to clusters using probability, with each cluster being described by a separate Gaussian distribution (Patel and Kushwaha 2020). GMM-based clustering involves the selection of the number of components (model order) and the estimation of parameters for each component, using the observed data, represented as a vector of features (Zeng and Cheung 2009).

GMM-based clustering has been explained by Zeng and Cheung, it supposes that the number of observations (N) is denoted as $X_N = \{x_1, x_2, \ldots, x_N\}$ and generated from a mixture of k* Gaussian components:

$$p(xt|\theta *) = \sum_{j=1}^{k*} \alpha *_j \, p(xt|\theta *_j)$$

Where:

$\sum_{j=1}^{k*} \alpha *_j = 1$ and $\forall 1 \leq j \leq k *, \alpha *_j > 0$

and each observation $xt(1 \leq t \leq N)$ is a vector of d-dimensional features:

$$x_{1t}, \ldots, x_{dt}$$

Also, $p(xt|\theta *_j)$ is the jth Gaussian component with the parameter $\theta *_j = \{\mu *_j, \sum_j *\}$, where $\theta *_j$ and $\sum_j *$ are, respectively, the center vector and covariance matrix of the jth component. The term $\alpha *_j$ is the mixing coefficient of the jth component in the mixture. GMM-based clustering main task is to find and estimate for $\theta * = \{\alpha *_j, \theta *_j\}_{j=1}^{k}$, denoted as from *N* observations, where k is an estimate of the true model order k*.

Gaussian mixture model-based clustering has been implemented in Scalable Weighted Iterative Flow-clustering Technique (SWIFT), to objectively identified asymmetric subpopulations (Mosmann et al. 2014). Here we used GMM-based clustering to validate the grouping of cells in the GFP cell line in three subpopulations (Figure S4. 1) presented in Chapter III. While we manually set the number of components to three, more objective analysis could be done by using the fitting scores.

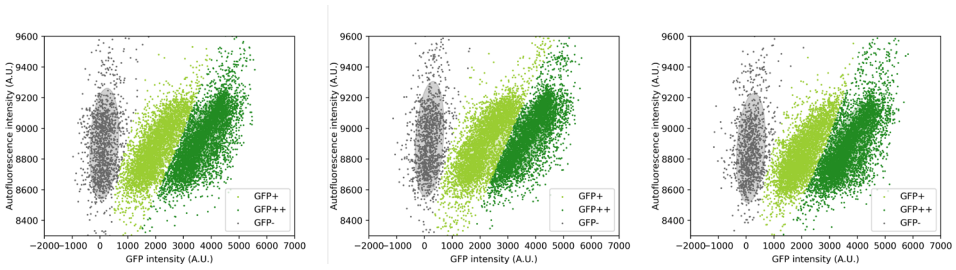


Figure S4. 1. A single GFP cell line presents three subpopulations. Analysis was done in triplicates.

ANNEX D: Collaborations and contributions

The following articles are annexed to this thesis:

Diamond A, Diaz-Garza AM, Li J, Slattery SS, Merindol N, Fantino E, Meddeb-Mouelhi F, Karas BJ, Barnabé S, and Desgagné-Penix I. Instability of extrachromosomal DNA transformed into the diatom *Phaeodactylum tricornutum*. Algal Res. 2023b:**70**:102998. https://doi.org/10.1016/j.algal.2023.102998

Sharma N, Simon DP, Diaz-Garza AM, Fantino E, Messaabi A, Meddeb-Mouelhi F, Germain H, and Desgagné-Penix I. Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow. Front Mar Sci. 2021:**8**:636613. https://doi.org/10.3389/fmars.2021.636613

Awwad F, Fantino EI, Héneault M, Diaz-Garza AM, Merindol N, Custeau A, Gélinas S-E, Meddeb-Mouelhi F, Li J, Lemay J-F, et al. Bioengineering of the Marine Diatom *Phaeodactylum tricornutum* with Cannabis Genes Enables the Production of the Cannabinoid Precursor, Olivetolic Acid. Int J Mol Sci. 2023:**24**(23):16624. https://doi.org/10.3390/ijms242316624

Fantino E, Awwad F, Merindol N, Diaz Garza AM, Gélinas S-E, Gajón Robles GC, Custeau A, Meddeb-Mouelhi F, and Desgagné-Penix I. Bioengineering *Phaeodactylum tricornutum*, a marine diatom, for cannabinoid biosynthesis. Algal Res. 2024:**77**:103379. https://doi.org/10.1016/j.algal.2023.103379

frontiers in Marine Science

REVIEW
published: 23 February 2021
doi: 10.3389/fmars.2021.636613

Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow

*Nikunj Sharma[1], Daris Pazhukkunnel Simon[1], Aracely Maribel Diaz-Garza[1], Elisa Fantino[1], Anis Messaabi[1], Fatma Meddeb-Mouelhi[1,2], Hugo Germain[1,2] and Isabel Desgagné-Penix[1,2]**

[1] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada, [2] Groupe de Recherche en Biologie Végétale, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

The benefits of the complex microscopic and industrially important group of microalgae such as diatoms is not hidden and have lately surprised the scientific community with their industrial potential. The ability to survive in harsh conditions and the presence of different pore structures and defined cell walls have made diatoms ideal cell machinery to produce a variety of industrial products. The prospect of using a diatom cell for industrial application has increased significantly in synch with the advances in microscopy, metabarcoding, analytical and genetic tools. Furthermore, it is well noted that the approach of industry and academia to the use of genetic tools has changed significantly, resulting in a well-defined characterization of various molecular components of diatoms. It is possible to conduct the primary culturing, harvesting, and further downstream processing of diatom culture in a cost-effective manner. Diatoms hold all the qualities to become the alternative raw material for pharmaceutical, nanotechnology, and energy sources leading to a sustainable economy. In this review, an attempt has been made to gather important progress in the different industrial applications of diatoms such as biotechnology, biomedical, nanotechnology, and environmental technologies.

Keywords: diatoms, microalgae biotechnology, metabolic engineering, metabarcoding, sustainable economy, biofuel, lipids

OPEN ACCESS

Edited by:
Marlen Ines Vasquez, Cyprus University of Technology, Cyprus

Reviewed by:
Gianluca Maria Farinola, University of Bari Aldo Moro, Italy
Muriel Bardor, Université de Rouen, France

****Correspondence:***
Isabel Desgagné-Penix
Isabel.Desgagne-Penix@uqtr.ca

Specialty section:
This article was submitted to Marine Biotechnology, a section of the journal Frontiers in Marine Science

Received: *01 December 2020*
Accepted: *26 January 2021*
Published: *23 February 2021*

Citation:
Sharma N, Simon DP, Diaz-Garza AM, Fantino E, Messaabi A, Meddeb-Mouelhi F, Germain H and Desgagné-Penix I (2021) Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow. Front. Mar. Sci. 8:636613. doi: 10.3389/fmars.2021.636613

INTRODUCTION

The global trend of economy and society is shifting toward building a greener and more sustainable society to combat climate and health issues. This is a critical issue, which is being approached with various interdisciplinary strategies to produce a wide range of sustainable products. For instance, biotechnology research has invested a significant number of resources, time, and money in studying microorganisms to exploit them for human consumption in multiple ways. Furthermore, the decades of research and improvisation in cultivation strategies, extraction, and harvesting protocols strongly support a good return on investment in industrial applications of microbes. A pinch of soil and a drop of water contain a diversity of microbes that controls major biogeochemical cycles and subsequently have the potential of producing an abundance of sustainable products. Since the beginning of this century, a high amount of research work has been published on industrial applications of microbes such as bacteria, yeast, and microalgae (**Figure 1**). But, limited attention

has been paid to diatoms which have the potential of becoming a robust sustainable industry because diatoms can continuously grow with an average annual yield of 132 MT dry diatoms ha^{-1} over almost 5 years (Wang and Seibert, 2017).

Diatoms are dynamic microorganisms with rich diversity and detailed membrane design. They are the most dominating phytoplankton with an overall number of around 200,000 species having complex variability in dimensions and shapes (Round et al., 1990; Smetacek, 1999; Mann and Vanormelingen, 2013). Diatoms' distinctive characteristic compared to the phytoplankton community is their silica cell wall, known as a frustule. This innate ability to uptake silicon from the environment has made them an interesting community of microbes since the 19th century. Few studies have stated the role of frustule biosilicate as pH buffering material which facilitates shifting of bicarbonate to CO_2 dissolved in cell fluids (the latter is readily metabolized by diatoms) (Milligan and Morel, 2002).

The access to advanced microscopes and modern genetic tools enabled us to study the detailed frustule structure and validate metabolic pathways involved in absorption, transportation, and polymerization of silicon and other biomolecules like lipids (Knight et al., 2016; Zulu et al., 2018). Furthermore, this advanced knowledge of metabolic pathways and validation of diatom structure can be applied to produce a wide range of renewable products such as optoelectronics, biofuels, nutritional supplements, ecology tools, etc. (Marella et al., 2020).

Other common factors that have shaped the evolution of diatoms are their ability to adapt and grow in various natural resources; fresh and marine water, wastewater, rivers, and oceans. Their abundance and adaptability in a wide range of climate and geographical areas make them suitable for different applications (Jin and Agustí, 2018). It was reported that diatoms are responsible to produce yearly, 40% of the organic carbon and 20% of oxygen (Tréguer et al., 1995; Falkowski et al., 1998; Afgan et al., 2016). Besides, these photoautotrophic organisms are involved in biogeochemical cycles, which play a significant role in global carbon fixation, carbon sequestration, and silicon cycle. They are also suitable candidates to capture nitrogen and carbon from various sources, which can be exploited by waste management and the biofuel industry to create carbon-neutral fuels (Singh et al., 2017). Furthermore, these algae are used to produce nutraceutical compounds, such as vegetarian proteins, omega, and other essential fatty acids for pharmaceutical industries (Wen and Chen, 2001a,b).

Multiple epidemiological, clinical, and pre-clinical studies have shown that omega fatty acids such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) are useful in slowing down age-related diseases such as cardiovascular diseases and cancer (Cole et al., 2010; Dyall, 2015; Thomas et al., 2015; Wang and Daggy, 2017). The development of diatoms strains rich in omega fatty acids can replace the dependence on fish as a source of omega oils and reduce the problems associated with seasonal variations and ocean pollution which might affect the biochemical composition of fish oil (Alves Martins et al., 2013). Also, various marine diatoms are considered for the commercial production of antioxidant pigments such as fucoxanthin and other carotenoids. It has been reported that these pigments exhibit various protective effects such as strong antioxidant activities (Xia et al., 2013).

Thus, the flexible and complex nature of diatoms offers immense possibilities to develop a wide range of sustainable products and contributes to carbon neutrality. Because of its dimensions, pore distributions, and geometries, it is studied to develop tools for nanotechnology and biomedical industry such as nanofabrication techniques, chemo and biosensing, particle sorting, and control of particles in micro- and nano-fluidics (Mishra et al., 2017). Silica and biosilica can be used to develop advanced nanomaterial for electronic and optical technologies which can be employed for ultra-sensitive detection of biological compounds (Dolatabadi et al., 2011).

Recent accomplishment in diatoms metabarcoding, a reference database of the global population of diatoms, has advanced its use extensively in studying ecological problems such as climate change, acidification, and eutrophication (Nanjappa et al., 2014). Because of its robust nature and potential to inhabit different photic regions, from the equator to the poles, diatoms offer the potential to develop tools and products for all geographical regions (Medlin, 2016). The technological and infrastructure advancements of diatoms-based applications are at a new level. Besides, it requires different kinds of optimization either in laboratory or large-scale research such as energy utilization for different steps, financial modeling, and collaborating with different industries to make diatom-based products commercially successful. However, the standardization at various levels such as optimization of culture conditions, genetic tools, genome and transcriptome sequencing make diatoms based products commercially viable.

Therefore, this review aims to provide a better understanding of the potential of diatoms research at a laboratory scale. We have tried to provide comprehensive information on a variety of diatoms applications such as energy, biomedical products, and environment monitoring which are being investigated at different levels. All these applications have the potential to contribute toward a greener tomorrow. The purpose of the research is to increase the sustainable economy while reducing the dependence on non-renewable resources. Therefore, recovering and producing various sustainable products like biofuels, feed, bioactive molecules, and services like environment monitoring embedded in diatoms is a promising opportunity to be seized as shown in **Figure 2**.

BIOFUEL INDUSTRY

Fast globalization and industrialization have impacted the ecosystem widely but shutting or slowing down the globalization is not the solution. At the moment, almost 95% of all the transportation industry is based on a non-renewable source of energy (Rodrigue and Notteboom, 2013). Therefore, developing sustainable and carbon-neutral fuels could reduce the existing dependence on fossil fuels and contribute to bringing back harmony in nature without disrupting the existing economic development. Few economic aspects of biofuel production from microalgae such as biodiesel productivity, land use, and oil

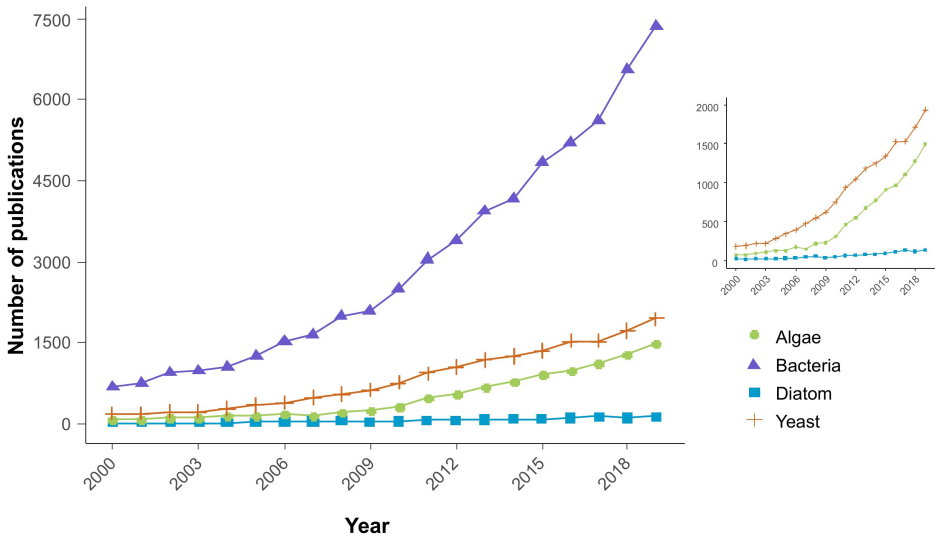


FIGURE 1 | Approximate number of research articles indexed in Scopus database (September 14, 2020) in the area of industrial application of different microbes (bacteria, yeast, algae, and diatoms).

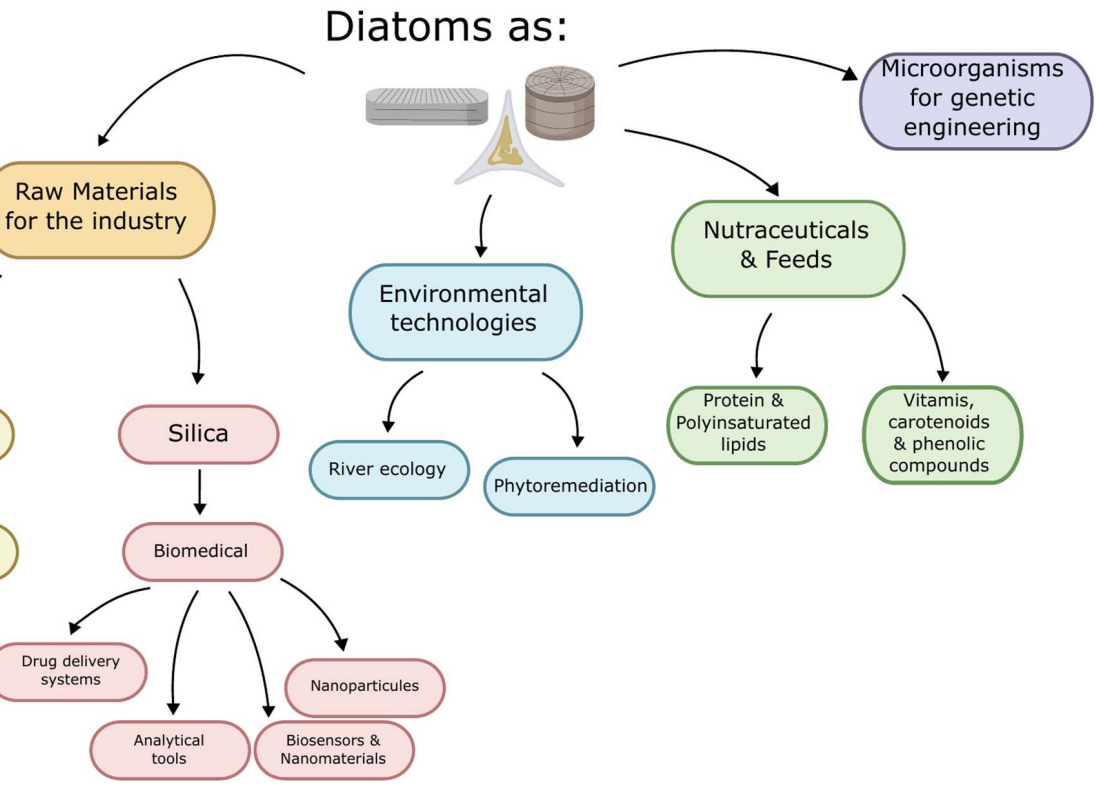


FIGURE 2 | Scheme of the different uses of diatoms for green industry.

yield support the use of microalgae for commercial production as compared to corn and other food crops. The oil yield for microalgae with high oil content is almost 15-fold more as compared to corn. Whereas, the land use for corn and maize is 66-fold more as compared to microalgae (Brocks et al., 2003).

The microalgae such as diatoms are the promising feedstock to replace non-renewable sources of energy. It has been proven by geochemists that algal lipids are the major feedstocks of petroleum and these lipids act as the biomarker remaining stable for millions of years (Brocks et al., 2003). The main biomarker

for the diatoms is the ratio of C28 and C29 steranes and highly branched isoprenoid alkenes which are found in high-quality oil fields around the globe (Katz et al., 2004).

Moreover, targeting the diatom lipids by manipulating and optimizing the growth and culture conditions such as light, stress, and nutrients can provide an interesting alternative to help meet the existing demands of commercial production of biofuel. Knowing the potential of diatoms to accumulate high lipids and varied compositions of fatty acids, diatoms are an underexploited area of the biofuel industry. The most predominant saturated and unsaturated fatty acids in diatom species are 14:0; 16:0, 16:1, 16:2, 16:3, 18:1, 18:2, 18:3, 20:4, and 20:5 (Dunstan et al., 1993; Sharma et al., 2020). Various reports have been published on different species of diatoms regarding the lipid yield and triacylglycerol accumulation (TAG) under different treatments as shown in **Table 1**.

It is possible to improve the quality of biodiesel by optimizing the content of different fatty acids that impacts biodiesel properties; cetane number, level of emissions, cold flow, oxidative stability, viscosity, and lubricity (Knothe, 2005). Fatty acids with chain lengths from C16 to C18 should contribute the maximum amount in the final product (Knothe, 2009). Some researchers have reported that a high percentage of mono-unsaturation is also desirable for biodiesel (Knothe, 2012). Thus, optimizing the fatty acid profile along with increased biomass will significantly enhance their economic value.

Statistical analyses predicted that 100 mt/ha/year biomass of diatoms is required for commercial biofuel production (Gallagher, 2011). Over 10 years, productivity range was observed to be between 29 and 142 mt/ha/year (Sheehan et al., 1998; Huesemann and Benemann, 2009), these values motivate the researchers and industry experts to study diatom cell in-depth for the biofuel industry in both lab-scale and large scale.

Furthermore, the availability of advanced genetic tools can help to achieve the missing targets in developing diatoms cells as biofuel machinery (Radakovits et al., 2010; Tibocha-Bonilla et al., 2018). Based on theoretical calculations about the land area, lipid production, and photosynthetic energy conversion, the biofuel demand of the complete United States population could be met using only 5% of United States land (Levitan et al., 2014). Although various other factors that define the efficacy of biodiesel such as engine performance, that is based on (cylinder pressure, brake mean effective pressure, frictional mean effective pressure, power, torque, brake specific fuel combustion, brake thermal efficiency). The statistical data supports the use of microalgae-based biofuel but there are various limitations at a technological level for large-scale implementation of this project. Therefore, one of the alternatives is to use the blended form of

TABLE 1 | Lipid content and productivities of different microalgae diatom species (-: no data).

Microalgae	Culture condition	Lipid (% dry weight)	Lipid productivity mg L^{-1} day^{-1}	TAG productivity μ mol L^{-1} day^{-1}	% of TAG	References
Thalassiosira weissflogii P09	–	29.94 ± 1.17	7.27 ± 0.28	–	51.0 ± 3.2	d'Ippolito et al. (2015)
	Nitrogen limitation	–	–	19 (+20%)	–	d'Ippolito et al. (2015)
Thalassiosira weissflogii CCMP 1010	–	38.84 ± 0.78	4.87 ± 0.10	–	53.0 ± 1.9	d'Ippolito et al. (2015)
Thalassiosira pseudonana CCMP 1335	–	29.33 ± 1.17	1.72 ± 0.07	–	19.0 ± 0.9	d'Ippolito et al. (2015)
	High CO_2 20,000 ppm	–	–	45.5 ± 26 (exponential) (+285%)	–	Jensen et al. (2020)
Cyclotella cryptica CCMP 331	–	41.97 ± 1.26	2.98 ± 0.09	–	55.0 ± 2.1	d'Ippolito et al. (2015)
	Nitrogen limitation	–	–	45 (+20%)		d'Ippolito et al. (2015)
Phaeodactylum tricornutum CCMP 632	–	9.32 ± 0.28	2.09 ± 0.06	–	19.0 ± 0.6	d'Ippolito et al. (2015)
	Tn19745_1 strain + nitrogen limitation	–	–	–	45-fold increase	Daboussi et al. (2014)
	Dark	+2.3-fold	–	–	–	Bai et al. (2016)
	High CO_2	–	–	75.7 ± 9 (stationary) (+50%)	–	Jensen et al. (2020)
textitNavicula pelliculosa (marine)	High CO_2	–	–	158.4 ± 29 (stationary) (+35%)	–	Jensen et al. (2020)

biodiesel. It would be more efficient to make a blended version of petro-diesel and microalgae/diatoms based fuel for large-scale operation. The comparative studies of blended (20% microalgae fuel plus 80% petrodiesel) and 100% petrodiesel have no major performance variations. Furthermore, it was reported that there was a reduction in the CO, unburnt HC, and smoke emissions in blended form as compared to pure diesel (Soni et al., 2020).

BIOMEDICAL INDUSTRY

Drug Delivery Systems

The cost required to bring a new drug to the market has been estimated by the Tufts Centre for the Study of Drug Development at approximately 2.6 billion dollars (DiMasi et al., 2016). In addition, the current drug delivery systems have limited solubility, poor bio-distribution, lack of selectivity, premature degradation, and unfavorable pharmacokinetics (Aw et al., 2011a,b). Therefore, these limitations have motivated the research and development of alternative drug delivery systems to improve the performance of existing drugs (i.e., increasing bioavailability), while reducing undesirable effects. There is no doubt that existing biomedical technologies have increased the life span but the human society wants to improvise the quality of life further by adopting environment friendly methods. Therefore, we should speed up the process and conduct in-depth research on using diatom frustules, even other bio-inspired alternatives for biomedical applications.

Among the available drug delivery tools (liposomes, nanogels, carbon nanotubes), the intricate frustule characteristics of diatoms such as specific surface area, thermal stability, biocompatibility, and alterable surface chemistry, have attracted attention for its use in drug and gene delivery. It took million years of evolution for diatoms to manufacture this level of complex and delicate structure to protect from the unwanted conditions like high temperature and variable light fluctuations. 3-D section analyses of diatom frustules have shown the availability of multiple pore patterns that range from nanometer to micrometer (Chandrasekaran et al., 2014; Cicco et al., 2015; Ragni et al., 2017). These characteristics are sufficient to explore alternative and low-priced silica-based materials for the biomedical industry (Mishra et al., 2017; Terracciano et al., 2018). Diatoms' frustule structure changes its homogenous nature, space, and intricate nature according to various environmental factors and silicon uptake efficiency (Knight et al., 2016). This ability can be used to change the frustule shape and pore size, which has multiple applications in the biomedical and nanotechnology industry. The process of biosilicification in diatoms is quite complex, it includes the role of silicic acid transporters, transportation of silica, and polymerization of silica monomers among other processes that have been extensively explained (Martin-Jézéquel et al., 2000; Knight et al., 2016). Moreover, a detailed investigation is being conducted to make the natural 3D porous structure an efficient substitute for delivery systems attributed to its chemical and mechanical features. For instance, some diatom species such as *Coscinodiscus concinnus* sp. (Gnanamoorthy et al., 2014), *Thalassiosira weissflogii* sp. (Aw et al., 2011a) are potential drug carriers candidates due to their amorphous nature and morphology. Additionally, various studies have shown that diatoms microcapsules are effective carriers for poorly soluble and water-soluble drugs, which can be applied in both oral and implant applications (Aw et al., 2011a; Ragni et al., 2017).

The defined structural architecture of diatoms, such as pore volume and controllable particle size, allows the synthesis of biomolecules at the micro- to nano-scale (Losic et al., 2005, 2010; Slowing et al., 2008). The growth of fibroblast and osteoblast has been observed on functionalized frustules supporting the idea of using biosilica from diatoms as smart support for cell growth (Ragni et al., 2017). Regarding modified diatoms, Losic et al. (2010) have designed the magnetically guided drug carrier via a functional surface of diatoms with dopamine-modified iron oxide. This modification has shown the capability of sustained release of poorly soluble drugs for 2 weeks, presenting an enhanced performance for drug delivery (Losic et al., 2010). Moreover, genetically modified biosilica has been used to selectively deliver anticancer drugs to tumor sites (Delalat et al., 2015). Overall, these findings have opened the doors to novel drug delivery systems using renewable material. Therefore, all properties of diatoms such as uniform pore structure, chemically inert and biocompatible, non-toxic, easy to transport, filtration efficiency, and specific drug delivery make it a potential model for drug delivery tools (Curnow et al., 2012; Milović et al., 2014; Rea et al., 2014; Vasani et al., 2015).

Analytical Tools

The controlled production of nanostructured silica is possible through chemical and mechanical treatment for a wide range of applications. This nanopore structure has a huge potential to attach the desired biomolecule (enzymes, DNA, antibodies) and develop label-free analytical tools or enhance the catalytic properties. It has also been shown that enzymes and DNA (oligonucleotides) can be conjugated to silica (Losic et al., 2005; Zamora et al., 2009). The encapsulation of enzymes in diatom biosilica exhibits improved enzymatic properties as compared to other immobilization technologies (Kato et al., 2020).

Additionally, luminescent nano- and micro-particles have gained the attention of the interdisciplinary scientific community (biology, chemistry, and physics). Current available fluorescent labeling agents are quantum dots, lanthanide-doped compounds, and organic fluorophore-tagged nanobeads, which offer good optical properties and a broad excitation spectrum. However, these agents have limitations in properties such as photobleaching and biocompatibility. For instance, De Stefano et al. (2009) studied diatoms' potential to incorporate fluorophores with increased stability used to study the molecular event of antibody-antigen identification. Moreover, molecular recognition between antibody and antigen was observed in relation to the change in the photoluminescence spectrum of diatoms. Concluding that diatom's frustules, due to their high sensitivity, low-cost, and availability are ideal alternative candidates for lab-on-particle applications (De Stefano et al., 2008, 2009).

There is no concrete evidence of diatoms' presence in land animal bodies. Although, various studies showed the presence of diatoms in the internal organs and circulatory system of alive or dead animals in an aquatic environment (Ludes et al., 1996; Lunetta et al., 1998; Hürlimann et al., 2000; Lunetta and Modell, 2005; Horton et al., 2006; Levkov et al., 2017). The siliceous cell wall of this organism is resistant to degradation even under high acidic conditions for a long period (Lunetta and Modell, 2005). The investigation on the occurrence of these organisms inside dead bodies of aquatic environment that died from different causalities opened up a new possibility of forensic analysis through the examination of diatoms called 'diatom axiom' or 'diatom test'(Lunetta et al., 1998). The diatom test is based on the hypothesis that the microalgae will not enter the systemic circulation and reach other internal organs and tissues such as bone marrow unless the circulation is functional. A forensic examiner can determine whether the individual was alive when it was entering the water by checking the presence of diatoms in various organs and tissues (Levkov et al., 2017). In addition, since diatoms are highly sensitive to environmental conditions, different water bodies have different diatom species abundance which allows forensics to identify the drowning site (Zhou et al., 2020).

Despite being a distinguishable method, the diatom test has limitations also. One of the major issues is the occurrence of diatoms in a drowning medium. The absence or low presence of diatoms in a water body can lead to a false positive or negative result. The presence of diatoms in different layers (water base, deeper, and surface) of the water body also can be varied (Levkov et al., 2017). Rapid death is another situation where the diatom test can be wrong. Instant death when an animal or human enters the water body for various reasons such as cold shock and cardiac diseases will give a negative result in the diatom test (Smol and Stoermer, 2010). The use of alcohol or drugs is another factor that can mislead in the diatom test (Ago et al., 2011). Recent advances in DNA Barcoding and pyrosequencing opened the possibility of increasing the accuracy of the diatom test by checking the presence of plankton specific genes (e.g., Rubisco gene) in animal tissue (Fang et al., 2019).

Biosensors and Nanomaterials

The advances in biotechnological tools have made it effective to characterize the frustules of diatoms for the fabrication of optoelectronics. The uptake of various elements such as zinc and germanium by diatom like *Stephanodiscus hantzschii*, *Thalassiosira pseudonana, etc.* to change the pore size, shape, and other characteristics which are being studied for a variety of functions such as paleolimnological indicator and photonic device application (Qin et al., 2008; Jaccard et al., 2009). It has been reported a relationship between the amount of Zn/Si (zinc/silicon) and free zinc ions which can be used as a proxy of paleolimnological indicators (Jaccard et al., 2009). The studies have raised intriguing questions about the uptake and the process of various elements which need detailed validations. Although, they have reported that they could only detect Zn and Fe as chemical elements. The analysis of various trace elements could be used as an environmental indicator which indeed will reduce the total workload needed to monitor large water bodies (Ellwood and Hunter, 2000).

The complex nanobiochemical machinery of diatoms can be exploited to fabricate a wide range of nanostructures with diverse optical and electronic properties (Rorrer et al., 2007). The ability to manufacture different pore size nanostructure molecules has inspired many research groups and industries to use diatoms in biosensing (De Tommasi, 2016). The incorporation of chemical elements such as germanium significantly affects the structure and size of frustule pores. A study tested the possibility of using Si-Germanium composite material in living diatoms in a two-stage photobioreactor cultivation process which reduced the pore size without disturbing the morphology (Rorrer et al., 2007). Another study reported that insertion of germanium in *Nitzschia frustulum* induces the nanocomb structure with blue photoluminescence (Qin et al., 2008). These nanostructure materials exhibit optical properties suitable for use in semiconductors and optoelectronics. Manufacturing of these materials combined with the silica frustule will improve the overall durability and range of applications in nanotechnology industries. These lab-scale scientific discoveries have shown that it is possible to create advanced nanomaterials in living diatoms.

Nanoparticles

The development of well-defined, advanced, and eco-friendly nanoparticles has attracted the attention of many researchers in the area of nanotechnology and its applications. Nanoparticles can be applied to study antimicrobial activity, catalyst, and filtering waste and chemical compounds. Biosynthesis of metallic nanoparticles in photoautotrophic organisms has gained the attention of nanotechnology researchers. Various approaches such as the sol-gel process, atomic layer deposition, chemical bath deposition, and inkjet printing process, have been used to modify the chemical composition of frustules. In this regard, an inexpensive chemical deposition technique was tested to deposit cadmium sulfide (CdS) on the surface on *Pinnularia* sp. without changing its morphology, since CdS has a wide range of applications in photodetectors and solar cells (Gutu et al., 2009).

Recently, it has been reported that diatoms can biosynthesize the nanoparticles such as gold and silver which has shown strong cytotoxicity against harmful microorganisms. Additionally, a highly ductile and malleable metal platinum (Pt) has been introduced in presence of dihydrogen hexachloroplatinate (IV) hexahydrate (DHH) in the living diatom *Melosira nummuloides*, without interfering the native morphology (Yamazaki et al., 2010). This is due to the platinum's excellent resistance to corrosion and stability at high temperatures, hence having application in a broad spectrum of industries, besides biomedicine. Other various examples of the on-going investigation of diatoms silica-based materials and their applications in biomedicine are shown in **Table 2**.

We have discussed the major application of diatoms for established industries such as biofuels, nanomaterials, and biomedicine. However, diatoms also have other fascinating applications in environment monitoring, animal feed, and aquaculture, which indeed have a huge potential considering

climate change and devastating impacts of globalization on ecology and environment.

ENVIRONMENTAL TECHNOLOGIES

River Ecology

Environment monitoring is an important aspect that is considered a necessity to deal with irregular changes or disturbances in our ecosystem. Therefore, researchers are developing tools using biotechnology and informatics to monitor the environment cost-effectively. Water resources are always under the influence of damaging anthropogenic pressures such as plastic waste and industrial sewage, which ultimately change or disturb the biogeochemical cycles and biodiversity. Besides, water is a universal solvent that holds the industries and economies together.

It is a well-established fact that diatoms hold the primary role in maintaining the aquatic ecosystem. Therefore, biodiversity assessment of diatom species in an environmental sample is one of the well-known strategies for biomonitoring. Presently, morphological assessment of the diatoms using microscopy is largely used which is time-consuming and requires special expertise (Larras et al., 2014). However, environmental metabarcoding has opened a quick way of analyzing the microbial DNA diversity in a natural environment such as flora and fauna (Bik et al., 2012; Taberlet et al., 2012). The metabarcoding approach is based on DNA sequencing a specific region (barcode) of the whole DNA extracted from an environmental sample (eDNA). For example, the sequencing data obtained from diatom metabarcoding are then used to assign precise taxonomic identification of the diatoms present in the eDNA sample, which are further compared with the conventional morphological database to confirm the efficacy of metabarcoding results. Diatoms metabarcoding tool has been optimized significantly to quantify the diversity of diatoms at the genus and species level (Vasselon et al., 2017; Kelly et al., 2018).

Currently, this approach is still in development, since various questions have been raised especially when deciding which are the most suitable barcodes. The barcodes that had been used are the ribosomal small subunit, cytochrome c, and the internal transcribed spacer region combined with the 5.8S rRNA gene (Zimmermann et al., 2011; Luddington et al., 2012).

Another main issue is processing the sequencing output data through computing. This method must be consistent with government policies for environmental regulation. For instance, MOTHUR is a comprehensive and efficient platform to study microbial diversity, but there are other bioinformatics software such as R, QIIME2 (Caporaso et al., 2010), LotuS (Hildebrand et al., 2014), and PIPITS (Gweon et al., 2015) that can be used to process a larger amount of data.

Additionally, various other research studies have supported the use of the diatoms metabarcoding approach as an alternative strategy to monitor river ecology on a timely basis. The results provide an estimated number of abundant and scarce species in samples obtained from different locations. Also, they give great insights into the fundamental status of the aquatic ecosystem (Larras et al., 2014). For instance, detailed evidence has been published by the Environmental Agency of the United Kingdom using diatoms indexes for river classification (Kelly et al., 2018). A similar study on detailed information on diatom biodiversity using metabarcoding has been conducted using environmental samples from Mayotte Island, France (Vasselon et al., 2017). Moreover, a recently published work studied the impact of treated effluents on benthic diatom communities that showed a systematic change in diatom community composition (Chonova et al., 2019). Concluding that detailed information about diatom diversity will give in-depth insights into climate change, micropollutants, and other organic pollutants, to study the disturbing effects of anthropogenic pressure on rivers. The use of metabarcoding for analyzing biodiversity is rapidly increasing and has been adopted by academic institutes and various companies/industries like Spygen (Canada), Naturemetrics (United Kingdom), IGAtech (Italy), Sinsoma (Austria), to name a few. This particular strategy has been adopted by public authorities as well and has shown the potential to be used as an additional screening tool to replace the existing methods, which require excessive

TABLE 2 | Biomedical applications of diatom silica-based materials using different diatom species.

Application	Organism	References
Specific nanoporous biosilica delivery system of chemotherapeutic drug, consisting in the attachment of antibodies and hydrophobic drug molecules, without using cross-linking, to the diatoms biosilica.	*T. pseudonana*	Delalat et al. (2015)
Modified frustule with self-assembled antibacterial aromatic amino acid conjugates Tyr−Zn^{II} as a zinc carrier for its controlled release to bacteria and inhibiting the bacterial growth.	*N. palea*	Singh et al. (2020)
Genetically modified frustule with chimeric fusion proteins: diatom-derived silica targeting peptide Sil3T8 and a small synthetic antibody derivative to detect *Bacillus anthracis*	*T. pseudonana*	Ford et al. (2020)
Rapid and selective detection of typhoid using cross-linked amine-functionalized diatom photoluminescent biosensor.	*Amphora* sp.	Selvaraj et al. (2018)
Nano composite of nanoporous diatom-ZrO_2 selective and highly sensitive sensor for non-enzymatic detection of methyl parathion.	*P. tricornutum*	Gannavarapu et al. (2019)
Biomaterial for negative electrode composed by a 3D-structured diatom biosilica for lithium-ion batteries, showing increased charge capacity compared to graphite.	*P. trainorii*	Nowak et al. (2019)
Improved capacitor performance of *in situ* coating of FeOx on live diatoms as a potential material for super capacitor electrodes.	*P. tricornutum*	Karaman et al. (2019)

infrastructure and human resources. It is indeed possible to make it a primary and permanent tool for river monitoring with advancements in sequencing, big data science, and artificial intelligence tools.

Phytoremediation

Besides the monitoring of river quality, water treatment is one of the major concerns for many countries around the world. In fact, human consumption has undoubtedly increased in the last few decades, subsequently, incrementing waste products presence in aquatic communities (Walker, 1983). Globally, almost 80% of the wastewater generated worldwide is discharged on rivers creating health and environmental hazards. The rise of nutrient accumulation in the aquatic system needs to be neutralized to maintain the balance in the environment. Increasing of pollution is disturbing the basic biogeochemical cycles, killing fish, depleting the dissolved oxygen, and producing different toxins, i.e., neurotoxins (Boyd, 1990). Hence, there is an urgent need to explore new ways and upscale the existing systems to test reports and mitigate pollution from rivers and lakes worldwide.

The use of microalgae for wastewater treatment has been a subject of research for a long period which could be applied in collaboration with small- and large-scale industries. The excess of industrial waste discharged in the aquatic system can be used as nutrient supply by diatoms. Different kinds of wastewater such as brewery (Choi, 2016), aquaculture (Tossavainen et al., 2019), and textile (El-Kassas and Mohamed, 2014) have been studied for phytoremediation capability and have shown interesting results. The published studies have established that diatoms and microalgae can treat the wastewater to an extent, therefore, it would be less damaging to treat the wastewater with microalgae/diatoms before discharging in water bodies. In addition, use the harvested biomass for different industrial products such as biofuel. It is safe to assume that it is possible to develop small scale business in collaboration with restaurants, breweries, textile industries, to name a few, to treat wastewater, and use the biomass for the production of valuable products such as fertilizers (Suleiman et al., 2020).

Heavy metal pollution is one of the major challenges which comes from the industries working with chemicals and dyes. Diatoms species are desirable organisms to study heavy metal pollution because of the simplicity of metal exposure, absorption, and detoxification of metal ions by single cells. This is a unique detoxification process of diatoms and microalgae due to metal-binding peptides known as phytochelatins (PCs) that protect photosynthetic organisms from heavy metals (Grill et al., 1985). Some intracellular PCs have been characterized in cultures of *P. tricornutum* exposed to different metals such as Cd, Pb, or Zn. Besides, they are used widely in waste degradation considering the unique structure of diatoms and their ability to respond to the changing environment (Glazer and Nikaido, 2007).

A study published in 2015 have reported a novel diatom *Bacillariophyta* sp. (BD1IITG) from petroleum biorefinery wastewater that can degrade phenol in a concentration range of 50−250 mg/L in Fog's media (Das et al., 2016). Another example of the degradation of toxic molecules like phenylalanine hydroxylase into less toxic compounds using simple enzymatic oxidation has been identified in diatoms during the metabolism of phenanthrene and pyrene (Wang and Zhao, 2007). These results are relevant considering that around seven billion kg of phenol is produced for oil refining, pesticide production, and to use in the pharmaceutical industry. Traditional phenol removal techniques involve several steps including the generation of by-products, which increments the cost of the treatment (Senthilvelan et al., 2014). However, there are very few reports available on exploiting the potential of diatoms in biodegrading waste materials. It is interesting to note that the studies have shown interesting results but the field of algae biotechnology requires more entrepreneurs to join the pieces of industrial and academic research to build a successful circular economy. Furthermore, there are some upcoming and growing ventures and companies in microalgae working in diverse applications and producing valuable products such as healthcare, animal feed, water management, chocolates, etc. (**Table 3**).

TABLE 3 | Different industries producing variety of products from microalgae and diatoms around the world.

Company	Products/services	Country	Website
Algae Biotechnologia	Wastewater treatment, animal nutrition, carbon dioxide fixation, biofuels, human health	Brazil	http://www.algae.com.br/site/pt/
Algae Farm	Omega3, diatom, water treatment and reuse, nutraceuticals, cosmeceuticals, algae based solar fuels cell, die sensitized solar panel, bioplastics	Canada	https://www.algaefarm.us/
Algorigin	Nutritional supplements	Switzerland	https://algorigin.com/en/
Algaetoomega	Omega 3, astaxanthin, animal feed	United States	https://algae2omega.com/
Algae Control Canada	Pond and lake water management	Canada	https://www.algaecontrol.ca/
The Algae Factory	Chocolate	Netherlands	http://thealgaefactory.com/
Algae Health	Antioxidants	United States	https://www.algaehealthsciences.com/
Swedish Algae Factory	Personal care products	Sweden	https://swedishalgaefactory.com/
Sabrtech	Recombinant proteins, fuel, nutraceuticals, aquaculture, etc.	Canada	https://www.sabrtech.ca/
Pondtech	Astaxanthin, aquaculture	Canada	https://www.pondtech.com/

DIATOMS AS NUTRACEUTICALS AND FEEDS

Multiple epidemiological and clinical trials have shown the health benefits of omega fatty acids from fish oils and algae extracts (Cole et al., 2010; Cottin et al., 2011; Thomas et al., 2015; Wang and Daggy, 2017). Besides, there are few publications on cardio-protective and cognitive performance of omega fatty acids which have led to the commercial production of infant foods, infant formula, fortified snack bars, and other dairy products supplemented with omega fatty acids (Arterburn et al., 2007; Cottin et al., 2011).

Diatoms have an immense nutritional value that can be used to produce novel compounds such as antioxidants, vitamins, animal feed, and vegetarian protein supplements. Several photosynthetic pigments have been identified in diatoms including carotenoids such as fucoxanthin (Kuczynska et al., 2015). Additionally, *Nitzschia laevis*, *Nitzschia inconspicua*, *Navicula saprophila,* and *Phaeodactylum tricornutum* extracts have a noticeable amount of EPA and DHA that can be used as a nutritional feed in human diet and animal feed (Kitano et al., 1997; Wen and Chen, 2001a,b; Wah et al., 2015; Tocher et al., 2019).

Moreover, diatoms are known to have diverse defense mechanisms in form of chemical substances for them to be protected against pathogens. For instance, *P. tricornutum* has a high amount of omega-7 monounsaturated fatty acids such as palmitoleic acid (C16:1) and other bioactive compounds that are active against gram-positive pathogens (Desbois et al., 2009). Furthermore, the EPA-rich marine diatom, *Odontella aurita*, used as a dietary supplement has shown antioxidant effects in rats (Haimeur et al., 2012). *O. aurita* has been approved to be commercialized as food in France by following EC regulation 258/97 in 2002 (Pulz and Gross, 2004; Buono et al., 2014).

Increasing the content of these bioactive molecules in diatoms has attracted a large amount of research. Some studies have managed to enhance the production of flavonoid and polyphenol content by culture modifications, for instance, cultivation temperature and nutrient supplementation in *Amphora* sp. (Chtourou et al., 2015). The general tendency when changing the culture temperature is an increase in lipid content in most species, while the chemical composition varied between species (Renaud et al., 2002). For example, the total amount of saturated and monounsaturated fatty acids increases with temperature in *Rhodomonas* sp. (NT15) and *Cryptomonas* sp. (CRFI01). Whereas, there was a comparative decrease in polyunsaturated fatty acids in both *Rhodomonas* sp. (NT15) and *Cryptomonas* sp. (CRFI01) (Renaud et al., 2002).

GENETIC ENGINEERING OF DIATOMS

The debate on using genetically modified microalgae and diatoms is on-going. However, it is a more controlled alternative for the production of recombinant proteins or any precursor molecules, considering the use of bioreactors for their production. The employment of genetic engineering tools in diatoms, to produce or increase the yield of compounds, allows the companies to optimize their use in the applications mentioned above. Therefore, genetic engineering is a promising method and an important branch to be used in the diatoms industry to further enhance the economic value of diatoms. However, it comes with two big challenges, firstly, to redesign the natural metabolic pathways in order to increase the production of desired endogenous compounds, and secondly, producing new heterologous compounds.

In the last 20 years, several projects have shown that these challenges can be solved at lab scale, by optimization of transformation methods, utilization of different gene promoters, expression of recombinant proteins, gene silencing, and genome editing methods; such as targeted mutagenesis techniques using meganucleases, gene knockouts, TALENS, and CRISPR/Cas9. Marketable bioproducts like lipids, pigments, nanomaterials, food supplements, fuel, syntheses of chemicals, drugs, and metabolites have been produced in *P. tricornutum, T. pseudonana*, and other diatoms species. While most of these analyses are related to lipid production for biofuel or bioenergy purposes, other studies showed that diatoms are biological factories that can generate a wide range of products from food to pharmaceutics biomaterial industry (Lauritano et al., 2016; Mishra et al., 2017; Slattery et al., 2018; Dhaouadi et al., 2020; Sharma et al., 2020). In addition, there are few companies such as Algenol Biofuels, Synthetic Genomics, which have reported the use of genetically modified microalgae for the production of biofuels.

TABLE 4 | Sequence Database of different diatoms species.

Species	Genome database
Phaeodactylum tricornutum CCAP 1055/1	http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index
Thalassiosira pseudonana CCMP 1335	https://genome.jgi.doe.gov/Thaps3/Thaps3.home.html
Thalassiosira oceanica CCMP 1005	https://genome.jgi.doe.gov/Thaoce1/Thaoce1.info.html
Thalassiosira weissflogii CCMP1030	https://genome.jgi.doe.gov/portal/
Fragilariopsis cylindrus CCMP 1102	https://genome.jgi.doe.gov/Fracy1/Fracy1.info.html
Pseudo-nitzschia multiseries CLN-47	https://genome.jgi.doe.gov/Psemu1/Psemu1.home.html
Pseudo-nitzschia multistriata B856	http://apollo.tgac.ac.uk/Pseudo-nitzschia_multistriata_V1_4_browser/sequences
Seminavis robusta D6	https://genome.jgi.doe.gov/portal/Semrobnscriptome/Semrobnscriptome.info.html
Fistulifera solaris JPCC DA058	https://trace.ddbj.nig.ac.jp/DRASearch/submission?acc5DRA002403
Cyclotella cryptica CCMP332	http://genomes.mcdb.ucla.edu/Cyclotella/download.html

TABLE 5 | Diatoms genetic engineering.

Species/strain	Genetic and molecular tools				
	Transformation methods and target compartment	**Promoters: (S) strong, (I) inducible and (H) heterologous**	**Reporters (R) and resistance (Re) genes**	**Expression of recombinant proteins**	**Genome editing methods and gene silencing**
Phaeodactylum tricornutum CCAP 1055/1	Biolistic (Cho et al., 2015) Electroporation (Niu et al., 2012) Conjugation (Zaboikin et al., 2017) Nuclear and chloroplast transformation (Xie et al., 2014)	**(S):** Lhcf (Fcp), light responsive (Karas et al., 2015), EF-1a, 40SRPS8, g-tubulin, RBCMT (Erdene-Ochir et al., 2019) and EF2 (Nymark et al., 2013), h4 (Fabris et al., 2020), HASP1 (De Riso et al., 2009). **(I):** rbcL (Xie et al., 2014), NR, low NO3 induce (Schellenberger Costa et al., 2012), V-ATPase C, AP1 low P induce (Lin et al., 2017) Fbp1, Fld, Isi1 iron-responsive (Yoshinaga et al., 2014) ca1, ca2 CO_2-responsive (Harada et al., 2005; Tanaka et al., 2016), U6, RNA polymerase III transcribed (Nymark et al., 2016) **(H):** CdP1, ClP1, ClP2, TnP1, TnP2 (Erdene-Ochir et al., 2016), CMV, RSV-LTR, PCMV, CaMV35S (Sakaue et al., 2008)	**(R):** GUS, GFP (Zhang and Hu, 2014), YFP, CFP (Zaboikin et al., 2017) cat (Karas et al., 2015), LUC (Cho et al., 2015), Aequorin (Falciatore et al., 2000) **(Re):** Zeocin and Phleomycin/sh ble, Nourseothricin/nat, Blasticidin-S/bsr, Streptothricin/sat, Neomycin/nptII (Karas et al., 2015)	Expression of Acyl-ACP thioesterases, increased accumulation of shorter chain (Radakovits et al., 2011). Malic enzyme (Trentacoste et al., 2013). G6PD (Wu et al., 2019), enhanced lipid productivity. Heterologous biosynthesis of the MIAs by CrGES expression under phototrophic conditions (Slattery et al., 2018), Vanillin production (Erdene-Ochir et al., 2019). PHBs for Bioplastics production (Hempel et al., 2011a). Human IgGαHBsAg:(Hempel et al., 2011b) and IgG1/kappa Ab CL4mAb: antibody to hepatitis B virus surface protein against the nucleoprotein of Marburg virus (Hempel and Maier, 2012). Over expression of DXS increased fucoxanthin synthesis (Eilers et al., 2016).	Targeted mutagenesis methods: meganucleases, gene knockouts, TALENS, and CRISPR/Cas9 (Poulsen and Kröger, 2005). Development of auxotrophic strains of *P. tricornutum* by CRISPR/Cas9 (Sakaguchi et al., 2011). A lipid producing strain through the disruption of the UDP-glucose pyrophosphorylase gene (Daboussi et al., 2014).
Thalassiosira pseudonana CCMP 1335	Biolistic (Poulsen et al., 2006) Electroporation (Buggé, 2015) Conjugation (Zaboikin et al., 2017)	**(S):** Lhcf9 **(I):** nr (161) SIT1, Si-starvation inducible (Davis et al., 2017), Thaps3_9619, Si-starvation inducible (Shrestha and Hildebrand, 2017), U6, RNA polymerase III transcribed (Weyman et al., 2015)	**(R):** YFP (Zaboikin et al., 2017) **(Re):** sh ble, nat (Poulsen et al., 2006)	Overexpression a multiple plasmids can be cotransformed; cloning multiple genes of interest Secretion of recombinant proteins has been shown. Localization of SiMat1-GFP (Kotzsch et al., 2016). Expression of the protective HslbpA DR2 antigen for the production of a vaccine against bovine respiratory disease (Davis et al., 2017), scFvTNT scFv and sdAbEA1 to detected *Bacillus anthracis* (Ford et al., 2016).	Targeted mutagenesis methods: meganucleases, TALENS, and CRISPR/Cas9 (Weyman et al., 2015). Gene silencing and gene knockouts are well established (Shrestha and Hildebrand, 2015).
Thalassiosira weissflogii (CCMP1030)	Biolistics (Cho et al., 2015)	**(S):** Lhcf2 (Cho et al., 2015)	**(R):** GUS (Cho et al., 2015)		

(Continued)

TABLE 5 | Continued

Species/strain	Genetic and molecular tools				
	Transformation methods and target compartment	**Promoters: (S) strong, (I) inducible and (H) heterologous**	**Reporters (R) and resistance (Re) genes**	**Expression of recombinant proteins**	**Genome editing methods and gene silencing**
Pseudo-nitzschia multistriata B856	Biolistics (Sabatino et al., 2015)	**(S):** h4 (Sabatino et al., 2015)	**(Re):** sh ble (Sabatino et al., 2015)		
Pseudo-nitzschia arenysensis B858	Biolistics (Sabatino et al., 2015)		**(R):** GUS, GFP (Sabatino et al., 2015)		
Fistulifera solaris JPCC DA058	Biolistics (Muto et al., 2015)	**(S):** Lhcf2 and h4 **(H):** RSV and CaMV35S (Muto et al., 2015)	**(R):** GFP **(Re):** nptII (Muto et al., 2015)	Overexpression of the endogenous *GK* improve lipid productivity (Muto et al., 2015)	
Cylindrotheca fusiformis CCAP 1017/2 –CYL	Biolistics (Kong et al., 2019)	**(I):** nr (Kong et al., 2019)	**(R):** GFP **(Re):** sh ble (Kong et al., 2019)		
Navicula saprophila NAVICI	Biolistics (Dunahay et al., 1995)	**(S):** ACCase (Dunahay et al., 1995)	**(Re):** nptII (Dunahay et al., 1995)		
Chaetoceros gracilis UTEX LB2658	Biolistics (Ifuku et al., 2015)	**(S):** Lhcr5 **(I):** nr (Ifuku et al., 2015)	**(R):** GFP, LUC **(Re):** nat (Ifuku et al., 2015)		

DXS, 1-deoxy-D-xylulose 5-phosphate synthase; 40SRPS8, 40S ribosomal protein S8; ACCase, acetyl-CoA carboxylase; Acyl-ACP thioesterases, acyl–acyl carrier protein thioesterases; AP1, alkaline phosphatase 1; bsr, blasticidin-S resistance gene; Ca1, carbonic anhydrase 1; CaMV35S, cauliflower mosaic virus 35S; CdP, Chaetoceros debilis-infecting DNA virus; ClP, Chaetoceros lorenzianus-infecting DNA virus; cat, chloramphenicol acetyl transferase conferring resistance to chloramphenicol; CRISPR, clustered regularly interspaced short palindromic repeats; CFP, cyan fluorescent protein gene; CMV, cytomegalovirus; Fcp, diatom light-regulated promoters of the fucoxanthin chlorophyll a/c-binding protein genes Lhcf; EF-1α, elongation factor 1 alpha; EF2, elongation factor 2; Fbp1, ferrichrome binding protein1; Fld, flavodoxin; CrGES, Catharanthus roseus geraniol synthase; G6PD, glucose-6-phosphate dehydrogenase; GK, glycerol kinase; GFP, green fluorescent protein gene; HASP1, highly abundant secreted protein 1; h4, histone H4; human IgGαHBsAg, antibody against hepatitis B virus surface IgG1/kappa Ab CL4mAb; HslbpA DR2, IbpA DR2 antigen from Histophilus somni; Isi1, iron-starvation-induced gene 1; MIAs, monoterpenoid indole alkaloids; nptII, neomycin phosphotransferase II; NR, nitrate reductase; nat, nourseothricin acetyl transferase; P, phosphate; PHBs, polyhydroxybutyrate; PCMV, promoter sequences of the cytomegalovirus; psba, PSII reaction center core 2 quinones are associated with D1; Lhcr5, red algal-like LHCRs; RBCMT, ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit N-methyltransferase I; RSV-LTR, Rous sarcoma virus long terminal repeat; rbcL, Rubisco large subunit; SiMat1, silica matrix protein; SIT1, silicon transporter; scFvTNT, single chain antibodies; sdAbEA1, single domain antibodies; U6, small nuclear RNA of the U6 complex; sh ble, Streptoalloteichus hindustanus bleomycin resistance gene; TnP, Thalassionema nitzschioides-infecting DNA virus; TALENs, transcription activator-like effector nucleases; sat, treptothricin acetyl transferase; TAG, triacylglycerol; g-tubulin, tubulin gamma chain; GUS, uidA b-glucuronidase-encoding gene; V-ATPase C, vacuolar H+-ATPase; YFP, yellow fluorescent protein gene.

Moreover, the approach of synthetic biology along with high throughput sequencing technologies open the doors to understanding the whole genome, the proteins that it encodes, and the regulatory elements of the cell during cellular growth and division (Hildebrand and Lerch, 2015; Huang and Daboussi, 2017). Several sequencing projects have been performed in *P. tricornutum* and *T. pseudonana* strains (Armbrust et al., 2004; Bowler et al., 2008; Koester et al., 2018; Rastogi et al., 2018), generating the transcriptomic and proteomic data sets that make possible precise reconstructions of metabolic networks (Fabris et al., 2012; Levering et al., 2016). Recently, the Synthetic Diatoms Project website has been launched as a platform to provide information to grow, transform, edit, and analyze *P. tricornutum* and *T. pseudonana*[1] These projects have been used as a springboard to facilitate genome annotation for other diatoms species: *T. oceanica, T. weissflogii, Fragilariopsis cylindrus*, *Pseudo-nitzschia multiseries*, *Pseudo-nitzschia multistriata*, *Seminavis robusta, Fistulifera solaris*, *Cyclotella cryptica* (**Table 4**).

Diatoms are a robust model for genome editing and cell transformation. Optimized methods of DNA delivery have been developed using biolistic or via electroporation. In both techniques, the transgenes are randomly integrated into the genome, with multiple integration events, variable transgene copy numbers, and chromosomal positions. The biolistic gene transfer method affects genome integrity due to the break and repair of the DNA double-strand by non-homologous end joining (NHEJ) (Zaboikin et al., 2017). However, this method is needed if the aim is to transform the chloroplast genome. An alternative transformation technique is the extrachromosomal-based expression approach that depends on vectors containing a yeast-derived sequence, which can be delivered through bacterial conjugation using *E. coli* (Karas et al., 2015).

An important element for genetic engineering is the promoter. The most commonly used are the light-regulated promoters of the fucoxanthin chlorophyll a/c-binding protein genes fcpA/B/C/D (LHCF) (Zaslavskaia et al., 2000; Nymark et al., 2013). Alternatively, the elongation factor 2 (EF2) promoter sequence is a constitutive promoter (Seo et al., 2015). Recently, the most abundant secreted protein in *P. tricornutum* was identified, named "highly abundant secreted protein 1" (HASP1), and the activities of its promoter and the signal peptide were characterized using green fluorescent protein (GFP) as a reporter (Erdene-Ochir et al., 2019). A couple of inducible promoters have been reported: like nitrate reductase (NR) and alkaline phosphatase gene promoters in *P. tricornutum*, which are induced under nitrogen or phosphate starvation respectively (Slattery et al., 2018; Fabris et al., 2020) and glutamine synthetase gene promoter, induced by a blue light pulse (De Riso et al., 2009; Erdene-Ochir et al., 2016). In addition, promoter regions containing diatom-infecting viruses (DIVs) mediated a significantly higher level expression of the reporter gene in cells in the stationary phase compared to the exponential phase of growth (Kadono et al., 2015). Other elements needed for genetic engineering are reporter genes and selection markers. Among reporter genes, beta-glucuronidase *uidA* (GUS), fluorescent proteins like GFP/YFP/CFP, chloramphenicol acetyltransferase conferring resistance to chloramphenicol (CAT) and luciferase (LUC) are the most employed, other reporter proteins are listed in **Table 5**. The classic selection markers in diatoms are genes that confer resistance to zeocin, phleomycin, and nourseothricin, as shown in **Table 5** are the most used. An alternative to using selective markers is the use of auxotrophic strains, such as uracil, histidine, and tryptophan auxotrophs (Sakaguchi et al., 2011; Slattery et al., 2020). Moreover, it is considered that the urease gene, either in an inactive or edited form, is an interesting tool for the selection of *P. tricornutum* and *T. pseudonana* strains (Weyman et al., 2015; Hopes et al., 2016; Slattery et al., 2018). An endogenous selectable marker in diatoms was generated by point mutations at a conserved residue Gly290 to Ser/Arg in the phytoene desaturase (PDS1) gene, which confers resistance to the herbicide norflurazon (Taparia et al., 2019).

Concerning heterologous recombinant protein expression, diatom gene codon optimization is required for optimal expression; to avoid silencing expression and better protein translation. Although it has not been reported in diatoms, different projects which were done in green algae, have shown that including introns in the expression cassette can increase transcript abundance (Baier et al., 2018, 2020; Kong et al., 2019). In addition, 5′-UTR and 3′-UTR of nitrate reductase (NR) allow the control of timing and level of transgene expression in *C. fusiformis* (Poulsen and Kröger, 2005). Down-regulation of gene expression can be achieved through silencing by expressing antisense repeat sequences of target genes (**Table 5**).

Industrial processes using diatoms are cost-effective and have performed well in large-scale cultures (Benedetti et al., 2018). This is supported by the plasticity to adapt to extreme environmental conditions of diatoms, making them great candidates for sustainable biofactories (Kung et al., 2012; Cho et al., 2015; d'Ippolito et al., 2015). Altogether, these developments in metabolic pathways and synthesis of heterologous compounds represent promising insights for the improvement of yield, quality of products, and sustainability in the use of diatoms as cell factories.

CONCLUSION AND FUTURE PERSPECTIVES

The documented studies stated the astounding nature and possible all-round use of diatoms. This is one of the approaches to increase human consumption of renewable products and contributes toward reducing carbon emissions. Although the commercial application of diatoms still needs improvements, it is indeed a crucial research area for human wellbeing. For example, developments in diatoms research can lead to innovative products in domains of drug delivery, sensing, and detection parts to build complex biomedical devices and nanoparticles for waste degradation. Moreover, recent advancements in sequencing technology and processing large biological datasets

[1] https://www.syntheticdiatoms.org/

have made it possible to label and store the global biodiversity of diatoms in all geographical locations.

One of the major challenges in diatom-based industries is scaling up the process for large-scale manufacturing which is dependent on many micro and macro factors such as cultivation, harvesting, drying, genetic modification, lack of genomic, proteomic, and metabolic information, etc. However, it is possible to overcome these challenges in near future with advancements in genetic tools, bioreactors, and other infrastructure changes. In general, there are many challenges in bio-based industries at different levels; academic/industrial research, infrastructure, policies, education, and information gaps. The advancements in academic research and discoveries are consistent considering the publications but it requires support from other domains such as the development of infrastructure, reducing the knowledge gaps between scientific researcher and entrepreneurs, changes in the policies at both national and international level. And to conclude, the recent research phenomenon blasted in the last decade, which is diatoms' industrial potential, still leaves many unsolved questions. Major questions will involve studying the extent of genetic or artificial manipulation without compromising its intact structure and delicate silica pattern. The unfolding of various missing links in genetic engineering, cultivation, and harvesting will make it possible to replicate complex plant pathways in diatoms. These tools have opened the door to study diatoms for eco-friendly processes.

Although the use of silica for food and agriculture has been approved by the FDA and is also labeled/classified as non-carcinogenic by the International Agency for Research on Cancer, this could be a big step toward accelerating its use at the biomedical level. It is not yet approved for biomedicine as it requires long-term evidence (Terracciano et al., 2018). All the biomedical inventions are scrutinized by multiple stakeholders like research leaders, public authorities such as provincial and federal government, before they reach the stage of commercial distribution. It is understandable considering that it will be used directly in the human body. Therefore, an innovative and different approach is required to bring in the academic researchers and bio-entrepreneurs to speed up the innovation rate in biomedical industry without harming the screening process set by public health authorities. The collaboration between entrepreneurs and researchers will allow thorough evaluation of the market for new inventions, manufacturing, investment, and globalization of the product. It seems plausible considering the rapid advancements in the biomedical infrastructure around the world. This has been demonstrated by the quick inventions in response to COVID-19 and should be adopted to be applied in other biotech based industries (Harris et al., 2020). The simultaneous advancements in the use of silica-based support system for drug delivery along with the change in infrastructure in pharmaceutical industries and hospitals to deliver these technologies to the users is possible in the near future. The other requirement is to join the gap of vast and complex scientific information and knowledge between entrepreneurs and academic researchers.

AUTHOR CONTRIBUTIONS

NS and ID-P conceived, designed, and led the study. NS, DS, AD-G, EF, and AM collected and analyzed the data, and prepared the figures and tables. NS, DS, AD-G, EF, AM, FM-M, HG, and ID-P authored and reviewed the drafts of the manuscript, and approved the final manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We acknowledge that financial support for this review was funded by the Canada Research Chair on plant specialized metabolism Award No. 950-232164 to ID-P. Thanks are extended to the Canadian taxpayers and to the Canadian Government for supporting the Canada Research Chairs Program. Additional support in the form of scholarships to NS, DS, AD-G, EF, and AM from Mitacs—Acceleration program grants nos. IT12310 and IT16463 to ID-P is also acknowledged.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Èech, M., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10.

Ago, K., Hayashi, T., Ago, M., and Ogata, M. (2011). The number of diatoms recovered from the lungs and other organs in drowning deaths in bathwater. *Leg. Med. (Tokyo)* 13, 186–190. doi: 10.1016/j.legalmed.2011.04.002

Alves Martins, D., Rocha, F., Castanheira, F., Mendes, A., Pousão-Ferreira, P., Bandarra, N., et al. (2013). Effects of dietary arachidonic acid on cortisol production and gene expression in stress response in *Senegalese sole* (*Solea senegalensis*) post-larvae. *Fish Physiol. Biochem.* 39, 1223–1238. doi: 10.1007/s10695-013-9778-6

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., et al. (2004). The genome of the diatom thalassiosira pseudonana : ecology, evolution, and metabolism. *Science* 306, 79–86. doi: 10.1126/science.1101156

Arterburn, L. M., Oken, H. A., Hoffman, J. P., Bailey-Hall, E., Chung, G., Rom, D., et al. (2007). Bioequivalence of docosahexaenoic acid from different algal oils in capsules and in a DHA-Fortified food. *Lipids* 42:1011. doi: 10.1007/s11745-007-3098-5

Aw, M. S., Simovic, S., Addai-Mensah, J., and Losic, D. (2011a). Polymeric micelles in porous and nanotubular implants as a new system for extended delivery of poorly soluble drugs. *J. Mater. Chem.* 21, 7082–7089. doi: 10.1039/c0jm04307a

Aw, M. S., Simovic, S., Addai-Mensah, J., and Losic, D. (2011b). Silica microcapsules from diatoms as new carrier for delivery of therapeutics. *Nanomedicine (Lond)* 6, 1159–1173. doi: 10.2217/nnm.11.29

Bai, X., Song, H., Lavoie, M., Zhu, K., Su, Y., Ye, H., et al. (2016). Proteomic analyses bring new insights into the effect of a dark stress on lipid biosynthesis in *Phaeodactylum tricornutum*. *Sci. Rep.* 6:25494.

Baier, T., Jacobebbinghaus, N., Einhaus, A., Lauersen, K. J., and Kruse, O. (2020). Introns mediate post-transcriptional enhancement of nuclear gene expression in the green microalga *Chlamydomonas reinhardtii*. *PLoS Genet.* 16:e1008944.doi: 10.1371/journal.pgen.1008944

Baier, T., Wichmann, J., Kruse, O., and Lauersen, K. J. (2018). Intron-containing algal transgenes mediate efficient recombinant gene expression in the green microalga *Chlamydomonas reinhardtii*. *Nucleic Acids Res.* 46, 6909–6919. doi: 10.1093/nar/gky532

Benedetti, M., Vecchi, V., Barera, S., and Dall'Osto, L. (2018). Biomass from microalgae: the potential of domestication towards sustainable biofactories. *Microbial Cell Fact.* 17:173.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* 27, 233–243. doi: 10.1016/j.tree.2011.11.010

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., et al. (2008). The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456, 239–244.

Boyd, C. E. (1990). *Water Quality in Ponds for Aquaculture*. Auburn StateAL: Alabama Agricultural Experiment Station.

Brocks, J. J., Buick, R., Logan, G. A., and Summons, R. E. (2003). Composition and syngeneity of molecular fossils from the 2.78 to 2.45 billion-year-old mount bruce supergroup, pilbara craton, Western Australia. *Geochimica et Cosmochimica Acta* 67, 4289–4319. doi: 10.1016/s0016-7037(03)00208-4

Buggé, J. A. (2015). *Electroporation-Mediated Transformation and Post-Transcriptional Gene Regulation of Nitrate Reductase in the Marine Diatom Thalassiosira Pseudonana*. Worcester, MA: Clark University.

Buono, S., Langellotti, A. L., Martello, A., Rinna, F., and Fogliano, V. (2014). Functional ingredients from microalgae. *Food Funct.* 5, 1669–1685. doi: 10.1039/c4fo00125g

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336.

Chandrasekaran, S., Sweetman, M. J., Kant, K., Skinner, W., Losic, D., Nann, T., et al. (2014). Silicon diatom frustules as nanostructured photoelectrodes. *Chem. Commun.* 50, 10441–10444. doi: 10.1039/c4cc04470c

Cho, C., Choi, S. Y., Luo, Z. W., and Lee, S. Y. (2015). Recent advances in microbial production of fuels and chemicals using tools and strategies of systems metabolic engineering. *Biotechnol. Adv.* 33, 1455–1466. doi: 10.1016/j.biotechadv.2014.11.006

Choi, H.-J. (2016). Parametric study of brewery wastewater effluent treatment using *Chlorella vulgaris* microalgae. *Environ. Eng. Res.* 21, 401–408. doi: 10.4491/eer.2016.024

Chonova, T., Kurmayer, R., Rimet, F., Labanowski, J., Vasselon, V., Keck, F., et al. (2019). Benthic diatom communities in an alpine river impacted by waste water treatment effluents as revealed using DNA metabarcoding. *Front. Microbiol.* 10:653. doi: 10.3389/fmicb.2019.00653

Chtourou, H., Dahmen, I., Jebali, A., Karray, F., Hassairi, I., Abdelkafi, S., et al. (2015). Characterization of *Amphora sp*., a newly isolated diatom wild strain, potentially usable for biodiesel production. *Bioprocess Biosystems Eng.* 38, 1381–1392. doi: 10.1007/s00449-015-1379-6

Cicco, S. R., Vona, D., De Giglio, E., Cometa, S., Mattioli-Belmonte, M., Palumbo, F., et al. (2015). Chemically modified diatoms biosilica for bone cell growth with combined drug-delivery and antioxidant properties. *ChemPlusChem* 80, 1104–1112. doi: 10.1002/cplu.201402398

Cole, G. M., Ma, Q.-L., and Frautschy, S. A. (2010). Dietary fatty acids and the aging brain. *Nutr. Rev.* 68 (Suppl 2), S102–S111.

Cottin, S. C., Sanders, T. A., and Hall, W. L. (2011). The differential effects of EPA and DHA on cardiovascular risk factors. *Proc. Nutrition Soc.* 70, 215–231. doi: 10.1017/s0029665111000061

Curnow, P., Senior, L., Knight, M. J., Thamatrakoln, K., Hildebrand, M., and Booth, P. J. (2012). Expression, purification, and reconstitution of a diatom silicon transporter. *Biochemistry* 51, 3776–3785. doi: 10.1021/bi3000484

d'Ippolito, G., Sardo, A., Paris, D., Vella, F. M., Adelfi, M. G., Botte, P., et al. (2015). Potential of lipid metabolism in marine diatoms for biofuel production. *Biotechnol. Biofuels* 8:28. doi: 10.1186/s13068-015-0212-4

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., et al. (2014). Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. *Nat. Commun.* 5:3831.

Das, B., Mandal, T. K., and Patra, S. (2016). Biodegradation of phenol by a novel diatom BD1IITG-kinetics and biochemical studies. *Int. J. Environ. Sci. Technol.* 13, 529–542. doi: 10.1007/s13762-015-0857-3

Davis, A., Crum, L. T., Corbeil, L. B., and Hildebrand, M. (2017). Expression of *Histophilus somni* IbpA DR2 protective antigen in the diatom *Thalassiosira pseudonana*. *Appl. Microbiol. Biotechnol.* 101, 5313–5324. doi: 10.1007/s00253-017-8267-8

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., and Falciatore, A. (2009). Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.* 37:e96. doi: 10.1093/nar/gkp448

De Stefano, L., Lamberti, A., Rotiroti, L., and De Stefano, M. (2008). Interfacing the nanostructured biosilica microshells of the marine diatom *Coscinodiscus wailesii* with biological matter. *Acta Biomaterialia* 4, 126–130. doi: 10.1016/j.actbio.2007.09.003

De Stefano, L., Rotiroti, L., De Stefano, M., Lamberti, A., Lettieri, S., Setaro, A., et al. (2009). Marine diatoms as optical biosensors. *Biosens. Bioelectron.* 24, 1580–1584. doi: 10.1016/j.bios.2008.08.016

De Tommasi, E. (2016). Light manipulation by single cells: the case of diatoms. *J. Spectroscopy* 2016:2490128.

Delalat, B., Sheppard, V. C., Rasi Ghaemi, S., Rao, S., Prestidge, C. A., McPhee, G., et al. (2015). Targeted drug delivery using genetically engineered diatom biosilica. *Nat. Commun.* 6:8791.

Desbois, A. P., Mearns-Spragg, A., and Smith, V. J. (2009). A fatty acid from the diatom *Phaeodactylum tricornutum* is antibacterial against diverse bacteria including multi-resistant *Staphylococcus aureus* (MRSA). *Mar. Biotechnol.* 11, 45–52. doi: 10.1007/s10126-008-9118-5

Dhaouadi, F., Awwad, F., Diamond, A., and Desgagne-Penix, I. (2020). Diatoms' breakthroughs in biotechnology: *Phaeodactylum tricornutum* as a model for producing high-added value molecules. *Am. J. Plant Sci.* 11, 1632–1670. doi: 10.4236/ajps.2020.1110118

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econom.* 47, 20–33. doi: 10.1016/j.jhealeco.2016.01.012

Dolatabadi, J. E. N., Mashinchian, O., Ayoubi, B., Jamali, A. A., Mobed, A., Losic, D., et al. (2011). Optical and electrochemical DNA nanobiosensors. *TrAC Trends Anal. Chem.* 30, 459–472. doi: 10.1016/j.trac.2010.11.010

Dunahay, T. G., Jarvis, E. E., and Roessler, P. G. (1995). Genetic transformation of the diatoms cyclotella cryptica and navicula saprophila. *J. Phycol.* 31, 1004–1012. doi: 10.1111/j.0022-3646.1995.01004.x

Dunstan, G. A., Volkman, J. K., Barrett, S. M., Leroi, J.-M., and Jeffrey, S. W. (1993). Essential polyunsaturated fatty acids from 14 species of diatom (Bacillariophyceae). *Phytochemistry* 35, 155–161. doi: 10.1016/s0031-9422(00)90525-9

Dyall, S. C. (2015). Long-chain omega-3 fatty acids and the brain: a review of the independent and shared effects of EPA, DPA and DHA. *Front. Aging Neurosci.* 7:52. doi: 10.3389/fnagi.2015.00052

Eilers, U., Bikoulis, A., Breitenbach, J., Büchel, C., and Sandmann, G. (2016). Limitations in the biosynthesis of fucoxanthin as targets for genetic engineering in *Phaeodactylum tricornutum*. *J. Appl. Phycol.* 28, 123–129. doi: 10.1007/s10811-015-0583-8

El-Kassas, H. Y., and Mohamed, L. A. (2014). Bioremediation of the textile waste effluent by *Chlorella vulgaris*. *Egyptian J. Aquatic Res.* 40, 301–308. doi: 10.1016/j.ejar.2014.08.003

Ellwood, M. J., and Hunter, K. A. (2000). The incorporation of zinc and iron into the frustule of the marine diatom *Thalassiosira pseudonana*. *Limnol. Oceanography* 45, 1517–1524. doi: 10.4319/lo.2000.45.7.1517

Erdene-Ochir, E., Shin, B.-K., Huda, M. N., Kim, D. H., Lee, E. H., Song, D.-G., et al. (2016). Cloning of a novel endogenous promoter for foreign gene expression in *Phaeodactylum tricornutum*. *Appl. Biol. Chem.* 59, 861–867. doi: 10.1007/s13765-016-0235-y

Erdene-Ochir, E., Shin, B.-K., Kwon, B., Jung, C., and Pan, C.-H. (2019). Identification and characterisation of the novel endogenous promoter HASP1 and its signal peptide from *Phaeodactylum tricornutum*. *Sci. Rep.* 9:9941.

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo-Madrid, A. C., Abbriano, R. M., et al. (2020). Extrachromosomal genetic engineering of the marine diatom phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biol.* 9, 598–612. doi: 10.1021/acssynbio.9b00455

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., and Baart, G. J. E. (2012). The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway. *Plant J.* 70, 1004–1014. doi: 10.1111/j.1365-313x.2012.04941.x

Falciatore, A., d'Alcalà, M. R., Croot, P., and Bowler, C. (2000). Perception of environmental signals by a marine diatom. *Science* 288, 2363–2366. doi: 10.1126/science.288.5475.2363

Falkowski, P. G., Barber, R. T., and Smetacek, V. V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200–207. doi: 10.1126/science.281.5374.200

Fang, T., Liao, S., Chen, X., Zhao, Y., Zhu, Q., Cao, Y., et al. (2019). Forensic drowning site inference employing mixed pyrosequencing profile of DNA barcode gene (rbcL). *Int. J. Legal Med.* 133, 1351–1360. doi: 10.1007/s00414-019-02075-4

Ford, N. R., Hecht, K. A., Hu, D., Orr, G., Xiong, Y., Squier, T. C., et al. (2016). Antigen binding and site-directed labeling of biosilica-immobilized fusion proteins expressed in diatoms. *ACS Synth Biol.* 5, 193–199. doi: 10.1021/acssynbio.5b00191

Ford, N. R., Xiong, Y., Hecht, K. A., Squier, T. C., Rorrer, G. L., and Roesijadi, G. (2020). Optimizing the design of diatom biosilica-targeted fusion proteins in biosensor construction for bacillus anthracis detection. *Biology* 9:14. doi: 10.3390/biology9010014

Gallagher, B. J. (2011). The economics of producing biodiesel from algae. *Renewable Energy* 36, 158–162. doi: 10.1016/j.renene.2010.06.016

Gannavarapu, K. P., Ganesh, V., Thakkar, M., Mitra, S., and Dandamudi, R. B. (2019). Nanostructured Diatom-ZrO2 composite as a selective and highly sensitive enzyme free electrochemical sensor for detection of methyl parathion. *Sensors Actuators B: Chem.* 288, 611–617. doi: 10.1016/j.snb.2019.03.036

Glazer, A. N., and Nikaido, H. (2007). *Microbial Biotechnology: Fundamentals of Applied Microbiology*. Cambridge: Cambridge University Press.

Gnanamoorthy, P., Anandhan, S., and Prabu, V. A. (2014). Natural nanoporous silica frustules from marine diatom as a biocarrier for drug delivery. *J. Porous Mater.* 21, 789–796. doi: 10.1007/s10934-014-9827-2

Grill, E., Winnacker, E. L., and Zenk, M. H. (1985). Phytochelatins: the principal heavy-metal complexing peptides of higher plants. *Science* 230, 674–676. doi: 10.1126/science.230.4726.674

Gutu, T., Gale, D. K., Jeffryes, C., Wang, W., Chang, C.-H., Rorrer, G. L., et al. (2009). Electron microscopy and optical characterization of cadmium sulphide nanocrystals deposited on the patterned surface of diatom biosilica. *J. Nanomater.* 2009:860536.

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., et al. (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the illumina sequencing platform. *Methods Ecol. Evol.* 6, 973–980. doi: 10.1111/2041-210x.12399

Haimeur, A., Ulmann, L., Mimouni, V., Guéno, F., Pineau-Vincent, F., Meskini, N., et al. (2012). The role of *Odontella aurita*, a marine diatom rich in EPA, as a dietary supplement in dyslipidemia, platelet function and oxidative stress in high-fat fed rats. *Lipids Health Dis.* 11, 147–147. doi: 10.1186/1476-511x-11-147

Harada, H., Nakatsuma, D., Ishida, M., and Matsuda, Y. (2005). Regulation of the expression of intracellular β -Carbonic anhydrase in response to CO 2 and light in the marine diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 139, 1041–1050. doi: 10.1104/pp.105.065185

Harris, M., Bhatti, Y., Buckley, J., and Sharma, D. (2020). Fast and frugal innovations in response to the COVID-19 pandemic. *Nat. Med.* 26, 814–817. doi: 10.1038/s41591-020-0889-1

Hempel, F., and Maier, U. G. (2012). An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. *Microbial Cell Factories* 11:126. doi: 10.1186/1475-2859-11-126

Hempel, F., Bozarth, A. S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., et al. (2011a). Microalgae as bioreactors for bioplastic production. *Microbial Cell Factories* 10:81. doi: 10.1186/1475-2859-10-81

Hempel, F., Lau, J., Klingl, A., and Maier, U. G. (2011b). Algae as protein factories: expression of a human antibody and the respective antigen in the diatom *Phaeodactylum tricornutum*. *PLoS One* 6:e28424.doi: 10.1371/journal.pone.0028424

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., and Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* 2:30. doi: 10.1186/2049-2618-2-30

Hildebrand, M., and Lerch, S. J. L. (2015). Diatom silica biomineralization: parallel development of approaches and understanding. *Sem. Cell Dev. Biol.* 46, 27–35. doi: 10.1016/j.semcdb.2015.06.007

Hopes, A., Nekrasov, V., Kamoun, S., and Mock, T. (2016). Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods* 12:49.

Horton, B. P., Boreham, S., and Hillier, C. (2006). The development and application of a diatom-based quantitative reconstruction technique in forensic science. *J. Forensic Sci.* 51, 643–650. doi: 10.1111/j.1556-4029.2006.00120.x

Huang, W., and Daboussi, F. (2017). Genetic and metabolic engineering in diatoms. *Philos. Trans. R. Soc. B: Biol. Sci.* 372:20160411. doi: 10.1098/rstb.2016.0411

Huesemann, M. H., and Benemann, J. R. (2009). "Biofuels from microalgae: review of products, processes and potential, with special focus on *Dunaliella sp*," in *The Alga Dunaliella: Biodiversity, Physiology, Genomics and Biotechnology*, eds A. Ben-Amotz, J. E. Polle, and D. V. W. Subba Rao (New Hampshire, NH: Science Publishers).

Hürlimann, J., Feer, P., Elber, F., Niederberger, K., Dirnhofer, R., and Wyler, D. (2000). Diatom detection in the diagnosis of death by drowning. *Int. J. Legal Med.* 114, 6–14. doi: 10.1007/s004149900122

Ifuku, K., Yan, D., Miyahara, M., Inoue-Kashino, N., Yamamoto, Y. Y., and Kashino, Y. (2015). A stable and efficient nuclear transformation system for the diatom *Chaetoceros gracilis*. *Photosynth Res.* 123, 203–211. doi: 10.1007/s11120-014-0048-y

Jaccard, T., Ariztegui, D., and Wilkinson, K. J. (2009). Incorporation of zinc into the frustule of the freshwater diatom *Stephanodiscus hantzschii*. *Chem. Geol.* 265, 381–386. doi: 10.1016/j.chemgeo.2009.04.016

Jensen, E. L., Yangüez, K., Carrière, F., and Gontero, B. (2020). Storage compound accumulation in diatoms as response to elevated CO2 concentration. *Biology* 9:5. doi: 10.3390/biology9010005

Jin, P., and Agustí, S. (2018). Fast adaptation of tropical diatoms to increased warming with trade-offs. *Sci. Rep.* 8:17771.

Kadono, T., Miyagawa-Yamaguchi, A., Kira, N., Tomaru, Y., Okami, T., Yoshimatsu, T., et al. (2015). Characterization of marine diatom-infecting virus promoters in the model diatom *Phaeodactylum tricornutum*. *Sci. Rep.* 5:18708.

Karaman, E. S., Wang, Z., Di Benedetto, G., Zunino, J. L., Meng, X., and Mitra, S. (2019). Fabrication of supercapacitors and flexible electrodes using biosilica from cultured diatoms. *Mater. Today Energy* 11, 166–173. doi: 10.1016/j.mtener.2018.11.004

Karas, B. J., Diner, R. E., Lefebvre, S. C., McQuaid, J., Phillips, A. P. R., Noddings, C. M., et al. (2015). Designer diatom episomes delivered by bacterial conjugation. *Nat. Commun.* 6:6925.

Kato, K., Lee, S., and Nagata, F. (2020). Efficient enzyme encapsulation inside sol-gel silica sheets prepared by poly-L-lysine as a catalyst. *J. Asian Ceramic Soc.* 8, 396–406. doi: 10.1080/21870764.2020.1747167

Katz, M. E., Finkel, Z. V., Grzebyk, D., Knoll, A. H., and Falkowski, P. G. (2004). Evolutionary trajectories and biogeochemical impacts of marine eukaryotic phytoplankton. *Annual Rev. Ecol. Evol. Systemat.* 35, 523–556. doi: 10.1146/annurev.ecolsys.35.112202.130137

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D., Pass, D., et al. (2018). *A DNA Based Diatom Metabarcoding Approach for Water Framework Directive Classification of Rivers*. Bristol: Environnment Agency.

Kitano, M., Matsukawa, R., and Karube, I. (1997). Changes in eicosapentaenoic acid content of *Navicula saprophila*, *Rhodomonas salina* and *Nitzschia sp*. under mixotrophic conditions. *J. Appl. Phycol.* 9, 559–563.

Knight, M. J., Senior, L., Nancolas, B., Ratcliffe, S., and Curnow, P. (2016). Direct evidence of the molecular basis for biological silicon transport. *Nat. Commun.* 7, 11926–11926.

Knothe, G. (2005). Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters. *Fuel Process. Technol.* 86, 1059–1070. doi: 10.1016/j.fuproc.2004.11.002

Knothe, G. (2009). Improving biodiesel fuel properties by modifying fatty ester composition. *Energy Environ. Sci.* 2, 759–766. doi: 10.1039/b903941d

Knothe, G. (2012). Fuel properties of highly polyunsaturated fatty acid methyl esters. prediction of fuel properties of algal biodiesel. *Energy Fuels* 26, 5265–5273. doi: 10.1021/ef300700v

Koester, J. A., Berthiaume, C. T., Hiranuma, N., Parker, M. S., Iverson, V., Morales, R., et al. (2018). Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species *Thalassiosira pseudonana*. *Sci. Rep.* 8:10492.

Kong, F., Yamaoka, Y., Ohama, T., Lee, Y., and Li-Beisson, Y. (2019). Molecular genetic tools and emerging synthetic biology strategies to increase cellular

oil content in *Chlamydomonas reinhardtii*. *Plant Cell Physiol.* 60, 1184–1196. doi: 10.1093/pcp/pcz022

Kotzsch, A., Pawolski, D., Milentyev, A., Shevchenko, A., Scheffel, A., Poulsen, N., et al. (2016). Biochemical composition and assembly of biosilica-associated insoluble organic matrices from the diatom *Thalassiosira pseudonana*. *J. Biol. Chem.* 291, 4982–4997. doi: 10.1074/jbc.m115.706440

Kuczynska, P., Jemiola-Rzeminska, M., and Strzalka, K. (2015). Photosynthetic pigments in diatoms. *Mar. Drugs* 13, 5847–5881. doi: 10.3390/md13095847

Kung, Y., Runguphan, W., and Keasling, J. D. (2012). From fields to fuels: recent advances in the microbial production of biofuels. *ACS Synthetic Biol.* 1, 498–513. doi: 10.1021/sb300074k

Larras, F., Keck, F., Montuelle, B., Rimet, F., and Bouchez, A. (2014). Linking diatom sensitivity to herbicides to phylogeny: a step forward for biomonitoring? *Environ. Sci. Technol.* 48, 1921–1930. doi: 10.1021/es4045105

Lauritano, C., Andersen, J. H., Hansen, E., Albrigtsen, M., Escalera, L., Esposito, F., et al. (2016). Bioactivity screening of microalgae for antioxidant, anti-inflammatory, anticancer, anti-diabetes, and antibacterial activities. *Front. Mar. Sci.* 3:68. doi: 10.3389/fmars.2016.00068

Levering, J., Broddrick, J., Dupont, C. L., Peers, G., Beeri, K., Mayers, J., et al. (2016). Genome-Scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One* 11:e0155038.doi: 10.1371/journal.pone.0155038

Levitan, O., Dinamarca, J., Hochman, G., and Falkowski, P. G. (2014). Diatoms: a fossil fuel of the future. *Trends Biotechnol.* 32, 117–124. doi: 10.1016/j.tibtech.2014.01.004

Levkov, Z., Williams, D. M., Nikolovska, D., Tofilovska, S., Èakar, Z., Williams, M., et al. (2017). *The Use of Diatoms in Forensic Science: Advantages and limitations of the diatom test in cases of drowning. The Archaeological and Forensic Applications of Microfossils: A Deeper Understanding of Human History*. London: Geological Society of London,

Lin, G., Wang, Y., Guo, L., Ding, H., Hu, Y., Liang, S., et al. (2017). Verification of mutagen function of Zeocin in *Nannochloropsis oceanica* through transcriptome analysis. *J. Ocean University country-regionChina* 16, 501–508. doi: 10.1007/s11802-017-3231-x

Losic, D., Mitchell, J. G., and Voelcker, N. H. (2005). Complex gold nanostructures derived by templating from diatom frustules. *Chem. Commun.* 39, 4905–4907. doi: 10.1039/b508733c

Losic, D., Yu, Y., Aw, M. S., Simovic, S., Thierry, B., and Addai-Mensah, J. (2010). Surface functionalisation of diatoms with dopamine modified iron-oxide nanoparticles: toward magnetically guided drug microcarriers with biologically derived morphologies. *Chem. Commun.* 46, 6323–6325. doi: 10.1039/c0cc01305f

Luddington, I. A., Kaczmarska, I., and Lovejoy, C. (2012). Distance and character-based evaluation of the V4 region of the 18S rRNA Gene for the identification of diatoms (*Bacillariophyceae*). *PLoS One* 7:e45664.doi: 10.1371/journal.pone.0045664

Ludes, B., Coste, M., Tracqui, A., and Mangin, P. (1996). Continuous river monitoring of the diatoms in the diagnosis of drowning. *J. Forensic Sci.* 41, 425–428.

Lunetta, P., and Modell, J. H. (2005). "Macroscopical, microscopical, and laboratory findings in drowning victims," in *Forensic Pathology Reviews. Forensic Pathology Reviews*, ed. M. Tsokos (Totowa, NJ: Humana Press).

Lunetta, P., Penttilä, A., and Hällfors, G. (1998). Scanning and transmission electron microscopical evidence of the capacity of diatoms to penetrate the alveolo-capillary barrier in drowning. *Int. J. Legal Med.* 111, 229–237. doi: 10.1007/s004140050159

Mann, D. G., and Vanormelingen, P. (2013). An inordinate fondness? the number, distributions, and origins of diatom species. *J. Eukaryotic Microbiol.* 60, 414–420. doi: 10.1111/jeu.12047

Marella, T. K., López-Pacheco, I. Y., Parra-Saldívar, R., Dixit, S., and Tiwari, A. (2020). Wealth from waste: diatoms as tools for phycoremediation of wastewater and for obtaining value from the biomass. *Sci. Total Environ.* 724:137960. doi: 10.1016/j.scitotenv.2020.137960

Martin-Jézéquel, V., Hildebrand, M., and Brzezinski, M. A. (2000). Silicon metabolism in diatoms: implications for growth. *J. Phycol.* 36, 821–840. doi: 10.1046/j.1529-8817.2000.00019.x

Medlin, L. K. (2016). Evolution of the diatoms: major steps in their evolution and a review of the supporting molecular and morphological evidence. *Phycologia* 55, 79–103. doi: 10.2216/15-105.1

Milligan, A. J., and Morel, F. M. M. (2002). A proton buffering role for silica in diatoms. *Science* 297, 1848–1850. doi: 10.1126/science.1074958

Milović, M., Simović, S., Lošić, D., Dashevskiy, A., and Ibrić, S. (2014). Solid self-emulsifying phospholipid suspension (SSEPS) with diatom as a drug carrier. *Eur. J. Pharm. Sci.* 63, 226–232. doi: 10.1016/j.ejps.2014.07.010

Mishra, M., Arukha, A. P., Bashir, T., Yadav, D., and Prasad, G. B. K. S. (2017). All new faces of diatoms: potential source of nanomaterials and beyond. *Front. Microbiol.* 8:1239. doi: 10.3389/fmicb.2017.01239

Muto, M., Tanaka, M., Liang, Y., Yoshino, T., Matsumoto, M., and Tanaka, T. (2015). Enhancement of glycerol metabolism in the oleaginous marine diatom *Fistulifera solaris* JPCC DA0580 to improve triacylglycerol productivity. *Biotechnol. Biofuels* 8:4. doi: 10.1186/s13068-014-0184-9

Nanjappa, D., Audic, S., Romac, S., Kooistra, W. H. C. F., and Zingone, A. (2014). Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. *PLoS One* 9:e103810.doi: 10.1371/journal.pone.0103810

Niu, Y. F., Yang, Z. K., Zhang, M. H., Zhu, C. C., Yang, W. D., Liu, J. S., et al. (2012). Transformation of diatom Phaeodactylum tricornutum by electroporation and establishment of inducible selection marker. *Biotechniques* 52, 1–3.

Nowak, A. P., Sprynskyy, M., Brzozowska, W., and Lisowska-Oleksiak, A. (2019). Electrochemical behavior of a composite material containing 3D-structured diatom biosilica. *Algal Res.* 41:101538. doi: 10.1016/j.algal.2019.101538

Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M., and Winge, P. (2016). A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* 6:24951.

Nymark, M., Valle, K. C., Hancke, K., Winge, P., Andresen, K., Johnsen, G., et al. (2013). Molecular and photosynthetic responses to prolonged darkness and subsequent acclimation to re-illumination in the diatom *Phaeodactylum tricornutum*. *PLoS One* 8:e58722.doi: 10.1371/journal.pone.0058722

Poulsen, N., and Kröger, N. (2005). A new molecular tool for transgenic diatoms. *FEBS J.* 272, 3413–3423. doi: 10.1111/j.1742-4658.2005.04760.x

Poulsen, N., Chesley, P. M., and Kröger, N. (2006). Molecular genetic manipulation of the diatom thalassiosira *Pseudonana* (*Bacillariophyceae*)1. *J. Phycol.* 42, 1059–1065. doi: 10.1111/j.1529-8817.2006.00269.x

Pulz, O., and Gross, W. (2004). Valuable products from biotechnology of microalgae. *Appl. Microbiol. Biotechnol.* 65, 635–648. doi: 10.1007/s00253-004-1647-x

Qin, T., Gutu, T., Jiao, J., Chang, C.-H., and Rorrer, G. L. (2008). Biological fabrication of photoluminescent nanocomb structures by metabolic incorporation of germanium into the biosilica of the diatom *Nitzschia frustulum*. *ACS Nano* 2, 1296–1304. doi: 10.1021/nn800114q

Radakovits, R., Eduafo, P. M., and Posewitz, M. C. (2011). Genetic engineering of fatty acid chain length in *Phaeodactylum tricornutum*. *Metab. Eng.* 13, 89–95. doi: 10.1016/j.ymben.2010.10.003

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

Ragni, R., Cicco, S., Vona, D., Leone, G., and Farinola, G. M. (2017). Biosilica from diatoms microalgae: smart materials from bio-medicine to photonics. *J. Mater. Res.* 32, 279–291. doi: 10.1557/jmr.2016.459

Rastogi, A., Maheswari, U., Dorrell, R. G., Vieira, F. R. J., Maumus, F., Kustka, A., et al. (2018). Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci. Rep.* 8:4834.

Rea, I., Martucci, N. M., De Stefano, L., Ruggiero, I., Terracciano, M., Dardano, P., et al. (2014). Diatomite biosilica nanocarriers for siRNA transport inside cancer cells. *Biochim. Biophys. Acta* 1840, 3393–3403. doi: 10.1016/j.bbagen.2014.09.009

Renaud, S. M., Thinh, L.-V., Lambrinidis, G., and Parry, D. L. (2002). Effect of temperature on growth, chemical composition and fatty acid composition of tropical Australian microalgae grown in batch cultures. *Aquaculture* 211, 195–214. doi: 10.1016/s0044-8486(01)00875-4

Rodrigue, J.-P., and Notteboom, T. (2013). The geography of cruises: itineraries, not destinations. *Appl. Geography* 38, 31–42. doi: 10.1016/j.apgeog.2012.11.011

Rorrer, G., Jeffryes, C., Chang, C.-H., Lee, D.-H., Gutu, T., Jiao, J., et al. (2007). *Biological Fabrication of Nanostructured Silicon-germanium Photonic Crystals Possessing Unique Photoluminescent and Electroluminescent Properties*. Bellingham, DC: SPIE.

Round, F. E., Crawford, R. M., Mann, D. G., and Press, C. U. (1990). *Diatoms: Biology and Morphology of the Genera*. Cambridge: Cambridge University Press.

Sabatino, V., Russo, M. T., Patil, S., d'Ippolito, G., Fontana, A., and Ferrante, M. I. (2015). Establishment of genetic transformation in the sexually reproducing diatoms pseudo-nitzschia multistriata and pseudo-nitzschia arenysensis and inheritance of the transgene. *Mar. Biotechnol. (NY)* 17, 452–462. doi: 10.1007/s10126-015-9633-0

Sakaguchi, T., Nakajima, K., and Matsuda, Y. (2011). Identification of the UMP synthase gene by establishment of uracil auxotrophic mutants and the phenotypic complementation system in the marine diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 156, 78–89. doi: 10.1104/pp.110.169631

Sakaue, K., Harada, H., and Matsuda, Y. (2008). Development of gene expression system in a marine diatom using viral promoters of a wide variety of origin. *Physiol. Plant.* 133, 59–67. doi: 10.1111/j.1399-3054.2008.01089.x

Schellenberger Costa, B., Jungandreas, A., Jakob, T., Weisheit, W., Mittag, M., and Wilhelm, C. (2012). Blue light is essential for high light acclimation and photoprotection in the diatom *Phaeodactylum tricornutum*. *J. Exp. Bot.* 64, 483–493. doi: 10.1093/jxb/ers340

Selvaraj, V., Muthukumar, A., Nagamony, P., and Chinnuswamy, V. (2018). Detection of typhoid fever by diatom-based optical biosensor. *Environ. Sci. Poll. Res.* 25, 20385–20390. doi: 10.1007/s11356-017-9362-1

Senthilvelan, T., Kanagaraj, J., Panda, R. C., and Mandal, A. B. (2014). Biodegradation of phenol by mixed microbial culture: an eco-friendly approach for the pollution reduction. *Clean Technol. Environ. Pol.* 16, 113–126. doi: 10.1007/s10098-013-0598-2

Seo, S., Jeon, H., Hwang, S., Jin, E., and Chang, K. S. (2015). Development of a new constitutive expression system for the transformation of the diatom *Phaeodactylum tricornutum*. *Algal Res.* 11, 50–54. doi: 10.1016/j.algal.2015.05.012

Sharma, N., Fleurent, G., Awwad, F., Cheng, M., Meddeb-Mouelhi, F., Budge, S., et al. (2020). Red light shift an effective alternative to boost biomass and lipids in *Phaeodactylum tricornutum*. *Appl. Sci.* 10:2531. doi: 10.3390/app10072531

Sheehan, J., Dunahay, T., Benemann, J., and Roessler, P. (1998). *Look Back at the U.S. Department of Energy's Aquatic Species Program: Biodiesel from Algae*. Golden, CO: National Renewable Energy Laboratory.

Shrestha, R. P., and Hildebrand, M. (2015). Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. *Eukaryot. Cell* 14, 29–40. doi: 10.1128/ec.00209-14

Shrestha, R. P., and Hildebrand, M. (2017). Development of a silicon limitation inducible expression system for recombinant protein production in the centric diatoms *Thalassiosira pseudonana* and *Cyclotella cryptica*. *Microbial Cell Factories* 16:145.

Singh, A. K., Sharma, N., Farooqi, H., Abdin, M. Z., Mock, T., and Kumar, S. (2017). Phycoremediation of municipal wastewater by microalgae to produce biofuel. *Int. J. Phytoremediation* 19, 805–812. doi: 10.1080/15226514.2017.1284758

Singh, R., Khan, M. J., Rane, J., Gajbhiye, A., Vinayak, V., and Joshi, K. B. (2020). Biofabrication of diatom surface by tyrosine-metal complexes:smart microcontainers to inhibit bacterial growth. *ChemistrySelect* 5, 3091–3097. doi: 10.1002/slct.201904248

Slattery, S. S., Diamond, A., Wang, H., Therrien, J. A., Lant, J. T., Jazey, T., et al. (2018). An expanded plasmid-based genetic toolbox enables cas9 genome editing and stable maintenance of synthetic pathways in *Phaeodactylum tricornutum*. *ACS Synthetic Biol.* 7, 328–338. doi: 10.1021/acssynbio.7b00191

Slattery, S. S., Wang, H., Giguere, D. J., Kocsis, C., Urquhart, B. L., Karas, B. J., et al. (2020). Plasmid-based complementation of large deletions in *Phaeodactylum tricornutum* biosynthetic genes generated by Cas9 editing. *Sci. Rep.* 10:13879.

Slowing, I. I., Vivero-Escoto, J. L., Wu, C.-W., and Lin, V. S. Y. (2008). Mesoporous silica nanoparticles as controlled release drug delivery and gene transfection carriers. *Adv. Drug Deliv. Rev.* 60, 1278–1288. doi: 10.1016/j.addr.2008.03.012

Smetacek, V. (1999). Diatoms and the ocean carbon cycle. *Protist* 150, 25–32. doi: 10.1016/s1434-4610(99)70006-4

Smol, J. P., and Stoermer, E. F. (2010). *The Diatoms: Applications for the Environmental and Earth Sciences*. Cambridge: Cambridge Press.

Soni, A. K., Kumar, S., and Pandey, M. (2020). Performance comparison of microalgae biodiesel blends with petro–diesel on variable compression ratio engine. *J. Instit. Eng. (India): Series E.* https://doi.org/10.1007/s40034-020-00183-0

Suleiman, A. K. A., Lourenço, K. S., Clark, C., Luz, R. L., da Silva, G. H. R., Vet, L., et al. (2020). From toilet to agriculture: fertilization with microalgal biomass from wastewater impacts the soil and rhizosphere active microbiomes, greenhouse gas emissions and plant growth. *Resources Conserv. Recycl.* 161:104924. doi: 10.1016/j.resconrec.2020.104924

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. Environmental DNA. *Mol. Ecol.* 21, 1789–1793.

Tanaka, A., Ohno, N., Nakajima, K., and Matsuda, Y. (2016). Light and CO 2 /cAMP signal cross talk on the promoter elements of chloroplastic β -carbonic anhydrase genes in the marine diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 170, 1105–1116. doi: 10.1104/pp.15.01738

Taparia, Y., Zarka, A., Leu, S., Zarivach, R., Boussiba, S., and Khozin-Goldberg, I. (2019). A novel endogenous selection marker for the diatom *Phaeodactylum tricornutum* based on a unique mutation in phytoene desaturase 1. *Sci. Rep.* 9:8217.

Terracciano, M., De Stefano, L., and Rea, I. (2018). Diatoms green nanotechnology for biosilica-based drug delivery systems. *Pharmaceutics* 10:242. doi: 10.3390/pharmaceutics10040242

Thomas, J., Thomas, C. J., Radcliffe, J., and Itsiopoulos, C. (2015). Omega-3 fatty acids in early prevention of inflammatory neurodegenerative disease: a focus on Alzheimer's Disease. *BioMed. Res. Int.* 2015:172801.

Tibocha-Bonilla, J. D., Zuñiga, C., Godoy-Silva, R. D., and Zengler, K. (2018). Advances in metabolic modeling of oleaginous microalgae. *Biotechnol. Biofuels* 11:241.

Tocher, D. R., Betancor, M. B., Sprague, M., Olsen, R. E., and Napier, J. A. (2019). Omega-3 long-chain polyunsaturated fatty acids, EPA and DHA: bridging the gap between supply and demand. *Nutrients* 11:89. doi: 10.3390/nu11010089

Tossavainen, M., Lahti, K., Edelmann, M., Eskola, R., Lampi, A.-M., Piironen, V., et al. (2019). Integrated utilization of microalgae cultured in aquaculture wastewater: wastewater treatment and production of valuable fatty acids and tocopherols. *J. Appl. Phycol.* 31, 1753–1763. doi: 10.1007/s10811-018-1689-6

Tréguer, P., Nelson, D. M., Van Bennekom, A. J., DeMaster, D. J., Leynaert, A., and Quéguiner, B. (1995). The silica balance in the world ocean: a reestimate. *Science* 268, 375–379. doi: 10.1126/science.268.5209.375

Trentacoste, E. M., Shrestha, R. P., Smith, S. R., Glé, C., Hartmann, A. C., Hildebrand, M., et al. (2013). Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. *Proc. Natl. Acad. Sci. U S A.* 110, 19748–19753. doi: 10.1073/pnas.1309299110

Vasani, R. B., Losic, D., Cavallaro, A., and Voelcker, N. H. (2015). Fabrication of stimulus-responsive diatom biosilica microcapsules for antibiotic drug delivery. *J. Mater. Chem. B* 3, 4325–4329. doi: 10.1039/c5tb00648a

Vasselon, V., Rimet, F., Tapolczai, K., and Bouchez, A. (2017). Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol. Indicat.* 82, 1–12. doi: 10.1016/j.ecolind.2017.06.024

Wah, N. B., Ahmad, A. L. B., Chieh, D. C. J., and Hwai, A. T. S. (2015). Changes in lipid profiles of a tropical benthic diatom in different cultivation temperature. *Asian J. Appl. Sci. Eng.* 4, 91–101.

Walker, W. W. Jr. (1983). Significance of eutrophication in water supply reservoirs. *J. AWWA* 75, 38–42. doi: 10.1002/j.1551-8833.1983.tb05056.x

Wang, H., and Daggy, B. P. (2017). The role of fish oil in inflammatory eye diseases. *Biomed. Hub.* 2, 1–12. doi: 10.1159/000455818

Wang, J.-K., and Seibert, M. (2017). Prospects for commercial production of diatoms. *Biotechnol. Biofuels* 10:16.

Wang, X. C., and Zhao, H. M. (2007). Uptake and biodegradation of polycyclic aromatic hydrocarbons by marine seaweed. *J. Coastal Res.* SI50, 1056–1061.

Wen, Z. Y., and Chen, F. (2001a). A perfusion–cell bleeding culture strategy for enhancing the productivity of eicosapentaenoic acid by *Nitzschia laevis*. *Appl. Microbiol. Biotechnol.* 57, 316–322. doi: 10.1007/s002530100786

Wen, Z.-Y., and Chen, F. (2001b). Application of statistically-based experimental designs for the optimization of eicosapentaenoic acid production by the diatom *Nitzschia laevis*. *Biotechnol. Bioeng.* 75, 159–169. doi: 10.1002/bit.1175

Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., McCarthy, J. K., Heuberger, A. L., et al. (2015). Inactivation of *Phaeodactylum tricornutum* urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnol. J.* 13, 460–470. doi: 10.1111/pbi.12254

Wu, S., Gu, W., Huang, A., Li, Y., Kumar, M., Lim, P. E., et al. (2019). Elevated CO2 improves both lipid accumulation and growth rate in the glucose-6-phosphate dehydrogenase engineered *Phaeodactylum tricornutum*. *Microbial Cell Factories* 18:161.

Xia, S., Wang, K., Wan, L., Li, A., Hu, Q., and Zhang, C. (2013). Production, characterization, and antioxidant activity of fucoxanthin from the marine diatom *Odontella aurita*. *Mar. Drugs* 11, 2667–2681. doi: 10.3390/md11072667

Xie, W. H., Zhu, C. C., Zhang, N. S., Li, D. W., Yang, W. D., Liu, J. S., et al. (2014). Construction of novel chloroplast expression vector and development of an efficient transformation system for the diatom *Phaeodactylum tricornutum*. *Mar. Biotechnol. (NY)* 16, 538–546. doi: 10.1007/s10126-014-9570-3

Yamazaki, T., Sasanuma, H., Mayama, S., and Umemura, K. (2010). Cultivation of melosira nummuloides cells in the presence of platinum: preparation of metal-containing frustules. *Phys. Status Solidi c* 7, 2759–2762. doi: 10.1002/pssc.200983808

Yoshinaga, R., Niwa-Kubota, M., Matsui, H., and Matsuda, Y. (2014). Characterization of iron-responsive promoters in the marine diatom *Phaeodactylum tricornutum*. *Mar. Genom.* 16, 55–62. doi: 10.1016/j.margen.2014.01.005

Zaboikin, M., Zaboikina, T., Freter, C., and Srinivasakumar, N. (2017). Non-Homologous end joining and homology directed DNA repair frequency of double-stranded breaks introduced by genome editing reagents. *PLoS One* 12:e0169931.doi: 10.1371/journal.pone.0169931

Zamora, P., Narváez, A., and Domínguez, E. (2009). Enzyme-modified nanoparticles using biomimetically synthesized silica. *Bioelectrochemistry* 76, 100–106. doi: 10.1016/j.bioelechem.2009.05.006

Zaslavskaia, L. A., Lippmeier, J. C., Kroth, P. G., Grossman, A. R., and Apt, K. E. (2000). Transformation of the diatom *Phaeodactylum tricornutum* (*Bacillariophyceae*) with a variety of selectable marker and reporter genes. *J. Phycol.* 36, 379–386. doi: 10.1046/j.1529-8817.2000.99164.x

Zhang, C., and Hu, H. (2014). High-efficiency nuclear transformation of the diatom *Phaeodactylum tricornutum* by electroporation. *Mar. Genomics* 16, 63–66. doi: 10.1016/j.margen.2013.10.003

Zhou, Y., Cao, Y., Huang, J., Deng, K., Ma, K., Zhang, T., et al. (2020). Research advances in forensic diatom testing. *Forensic Sci. Res.* 5, 98–105. doi: 10.1080/20961790.2020.1718901

Zimmermann, J., Jahn, R., and Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organ. Divers. Evol.* 11:173. doi: 10.1007/s13127-011-0050-6

Zulu, N. N., Zienkiewicz, K., Vollheyde, K., and Feussner, I. (2018). Current trends to comprehend lipid metabolism in diatoms. *Progress Lipid Res.* 70, 1–16. doi: 10.1016/j.plipres.2018.03.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Algal Research 70 (2023) 102998

Contents lists available at ScienceDirect

Algal Research

journal homepage: www.elsevier.com/locate/algal

Instability of extrachromosomal DNA transformed into the diatom *Phaeodactylum tricornutum*

Andrew Diamond [a,1], Aracely Maribel Diaz-Garza [a,1], Jessica Li [c], Samuel S. Slattery [c], Natacha Merindol [a], Elisa Fantino [a], Fatma Meddeb-Mouelhi [a,b], Bogumil J. Karas [c], Simon Barnabé [a,b], Isabel Desgagné-Penix [a,b,*]

[a] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada
[b] Plant Biology Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada
[c] Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

ARTICLE INFO

Keywords:
Metabolic engineering
Bacterial conjugation
Episome rearrangements
Flow cytometry
Screening through fluorescence
2A self-cleaving peptide
Gene toxicity

ABSTRACT

Phaeodactylum tricornutum has been highly studied for its potential as a platform for metabolic engineering. While the possible applications of extrachromosomal expression via an episome have been investigated, there is still a lack of information concerning its efficacy and limitations. Therefore, we studied the episome expression system in *P. tricornutum,* aiming to elucidate its limitations regarding heterologous protein production and episome rearrangement events. Our objectives were to screen positive transconjugants by fluorescent signal indicating as a proxy for the production of the proteins of interest that could be used for vanillin synthesis, and to characterize the transconjugants by flow cytometry and whole plasmid sequencing. We designed an episome harboring an expression cassette that consisted of the enhanced green-fluorescent-protein (eGFP) linked by *Thosea asigna* virus 2A self-cleaving peptide (T2A) to a fusion protein of enoyl-CoA hydratase/aldolase (ech) and feruloyl-CoA synthetase (fcs), both from *Streptomyces* sp. strain V-1. This construction resulted in a percentage of fluorescent transconjugants lower than 10 % and it presented rearranged episomes in the fluorescent and the non-fluorescent transconjugants. The replacement of the fusion protein ech-fcs in the expression cassette with the fluorescent protein mCherry increased the percentage of eGFP fluorescent transconjugants over 80 % suggesting a toxicity of the *ech-fcs* gene expression and in turn forcing selection for rearranged episomes. A comparison of flow cytometry results and sequencing analysis demonstrated that a successful transformation with an unaltered expression cassette could lead to diatoms that do not produce the protein. On the other hand, transconjugants with mutations or rearrangements in the genes encoding the fusion ech-fcs protein led to fluorescent signal detection. Here, we show that using fluorescent reporters can mislead the selection of positive transconjugants by not being able to identify rearrangements in the genes of interest, and intact cassettes can lack fluorescent signal due to lack of heterologous protein production.

1. Introduction

Phaeodactylum tricornutum is the model organism for pennate diatoms and a suitable platform for metabolic engineering. It is not only the best-characterized diatom so far known to accumulate high-value products but also is a viable organism for large-scale culture [1]. *P. tricornutum* has been successfully used for recombinant protein production attributed to its high growth rates and high efficiency to express complex eukaryotic genes [2–5]. In addition, the sequencing and annotation of the diatom's genome done in 2008 [6], and revised in 2021 [7] and 2022 [8], combined with the development of a variety of genetic tools, have enabled its use in biotechnology [9,10].

Biolistic transformation of *P. tricornutum* led to the successful accumulation of docosahexaenoic acid (DHA) [11] and the production of betulin and its precursors [12]. This transformation method leads to random integration into the genome that could interrupt non-target genes by partial or multiple integrations of the expression cassette requiring high throughput screening methods of positive clones [13,14].

* Corresponding author at: Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada.
E-mail address: Isabel.Desgagne-Penix@uqtr.ca (I. Desgagné-Penix).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.algal.2023.102998
Received 8 April 2022; Received in revised form 24 December 2022; Accepted 29 January 2023
Available online 3 February 2023

To overcome these undesirable effects, the episomal expression method was recently added to the molecular tools for diatoms, allowing the design of extrachromosomal consistent, complex, and predictable platforms for protein production [13,14]. Besides, when compared to extrachromosomal episomal expression systems, random integration into the genome produced a higher, but variable, fluorescent signal detected from the reporter protein mVenus whereas the fluorescence intensity was more stable between the transconjugants harboring the linearized episome [15]. This expression variability of the heterologous genes following nuclear transformation can be due to the random integration impacting genes responsible for growth and development or to silencing due to the integration site.

Although episomes have recently been successfully used for protein synthesis and production of metabolites [15,16], there is still a lack of knowledge on the success of episomal expression and the screening particularly when referring to sequence stability and segregation patterns, impacting the production of heterologous proteins or metabolites. In a previous study, Slattery et al. (2018) have demonstrated the possibility to introduce eight genes encoding enzymes of the vanillin biosynthesis pathway including *Vanilla planifolia* vanillin synthase (*Vp*VAN) [17] into *P. tricornutum* in a single episome. Despite the absence of mutations in the genes, enzyme production and activity in the selected transconjugants was not detected (data not shown), justifying the need to a more in-depth study of alternative pathways such as episomal expression systems.

Therefore, our aim was to study the episomal expression system in *P. tricornutum* focusing on the production of heterologous proteins and episome rearrangement events. To the best of our knowledge, this kind of molecular events in extrachromosomal expression system propagated by *P. tricornutum* has not been reported previously. The objectives were (1) to screen positive colonies using a fluorescent reporter linked to a fusion protein that could potentially be used for vanillin production; (2) to characterize the transconjugants by flow cytometry and whole plasmid sequencing. Briefly, we used a construction containing the enhanced green fluorescent protein (eGFP) linked to the fusion protein composed of feruloyl-CoA synthetase (fcs) and enoyl-CoA hydratase/aldolase (ech) soluble enzymes from *Streptomyces* sp. strain V-1 (Fig. 1a). Since these two enzymes have already been characterized for the conversion of ferulic acid to vanillin in *E. coli* [18,19], they were chosen over *Vp*VAN whose catalytical activity remains controversial [20]. We demonstrated that the screening of *P. tricornutum* transconjugants by fluorescence detection using eGFP as a reporter protein can be misleading. On the one hand, no fluorescence could be detected using microscopy and flow cytometry on zeocin resistant transconjugants,

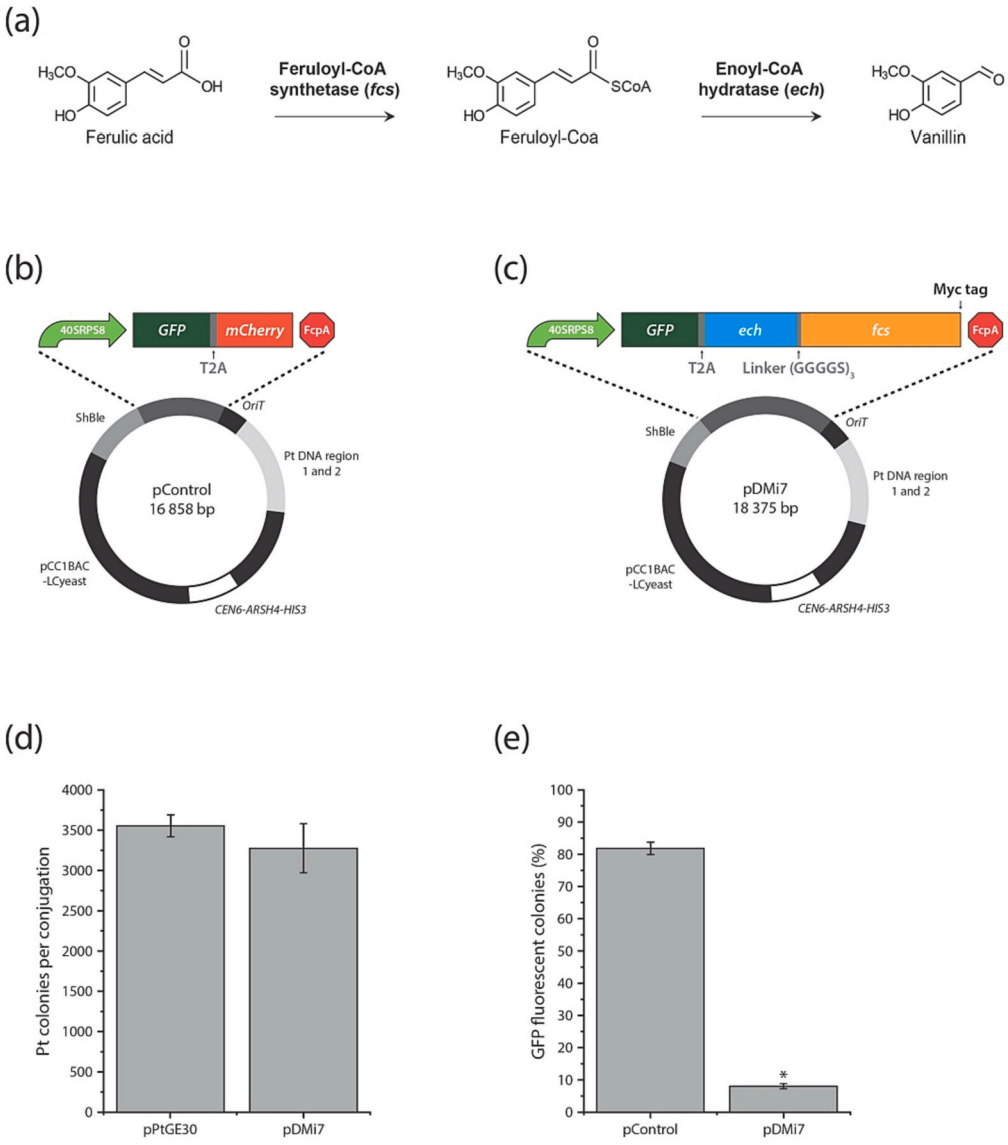


Fig. 1. Transformation of *P. tricornutum* cells with a vanillin biosynthetic pathway. (a) Biosynthetic pathway of two bacterial enzymes from *Streptomyces* sp. that convert ferulic acid into vanillin. (b) Plasmid scheme of pControl. The scheme is on scale, with the exceptions of the size of the promoter and terminator. (c) Plasmid scheme of pDMi7 and the expression cassette containing the vanillin biosynthetic pathway. The scheme is on scale, with the exceptions of the size of the promoter and terminator. (d) Total count of colonies per bacterial conjugation done in parallel with an empty vector (pPtGE30) and pDMi7. (e) Percentage of GFP fluorescent colonies obtained from bacterial conjugations with pDMi7 and pControl observed by fluorescence microscopy. Data in (d) and (e) are means of three biological replicates except for the percentage of GFP fluorescent colonies from pDMi7 that was calculated from two biological replicates. The asterisk annotation indicates a significant difference from pControl as determined by Welch's t-test (p = 8.1E-06). 40SRPS8, 40SRPS8 promoter; eGFP, enhanced green fluorescent protein; T2A, *Thosea asigna* virus 2A self-cleaving peptide; ech, enoyl-CoA hydratase/aldolase; fcs, feruloyl-CoA synthetase; FcpA, FcpA terminator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

despite confirming a genetically intact whole plasmid by sequencing. On the other hand, in the zeocin resistant eGFP fluorescent transconjugants, the fcs protein was altered with deletions, shifts in the open reading frame (ORF), or nonsynonymous substitution mutations. Thus, the screening based on the fluorescence of eGFP led us to the selection of transconjugants that did not contain an intact episome nor an unaltered expression cassette. While, previous studies proved the efficiency of the episome for the genetic engineering of *P. tricornutum* [15,16,21–24], our work suggests that the episomes can be rearranged at an unknown step during the conjugation to diatoms or subsequent episome propagation.

2. Material and methods

2.1. Microbial strains and growth conditions

Escherichia coli (NEB® 10-beta, New England Biolabs, Canada) was grown in Luria Broth (LB) supplemented with appropriate antibiotics (chloramphenicol (15 mg L^{-1})). *Escherichia coli* (Epi300, Epicenter) was grown in Luria Broth (LB) supplemented with appropriate antibiotics (gentamicin (20 mg L^{-1}) or chloramphenicol (15 mg L^{-1}) and gentamicin (20 mg L^{-1})). *Phaeodactylum tricornutum* (CCAP 1055/1, Culture Collection of Algae and Protozoa) was grown in modified L1 medium without silica at 18 °C under cool white fluorescent lights (75 μE m^{-2} s^{-1}) and a photoperiod of 16 h light:8 h dark with an agitation of 130 rpm for liquid cultures.

2.2. P. tricornutum growth medium

L1 media preparation was adapted from [10]. It consisted of 917 mL of autoclaved distilled water, 50 mL of 20× stock of NaCl and Na_2SO_4, 10 mL of 100× stock of anhydrous salt, 20 mL of 50× stock solution of hydrous salt, 2 mL of sodium phosphate (NP) stock, 1 mL L1 trace metals stock, 0.5 mL f/2 vitamin solution.

The 20× stock of NaCl and Na_2SO_4 sterilized by autoclave consisted of 245 g L^{-1} NaCl and 40.9 g L^{-1} Na_2SO_4. The 100× stock of anhydrous salt sterilized by autoclave consisted of 35 g L^{-1} KCl, 10 g L^{-1} $NaHCO_3$, 5 g L^{-1} KBr, 1.5 g L^{-1} H_3BO_3, and 0.15 g L^{-1} NaF. The 50× stock of hydrous salt sterilized by autoclave consisted of 277.5 g L^{-1} $MgCl_2{\cdot}6H_2O$ and 38.5 g L^{-1} $CaCl_2{\cdot}2H_2O$. The sodium phosphate (NP) stock was made in 100 mL H_2O and consisted of 37.5 g $NaNO_3$ and 2.5 g $NaH_2PO_4{\cdot}H_2O$ and was sterilized through a 0.2 μm filter. The L1 trace metal stock solution was made by mixing 3.15 g $FeCl_3{\cdot}6H_2O$, 4.36 g $Na_2EDTA{\cdot}2H_2O$, 0.25 mL (9.8 g L^{-1} dH_2O) $CuSO_4{\cdot}5H_2O$, 3.0 mL (6.3 g L^{-1} dH_2O) $Na_2MoO_4{\cdot}2H_2O$, 1.0 mL (22.0 g L^{-1} dH_2O) $ZnSO_4{\cdot}7H_2O$, 1.0 mL (10.0 g L^{-1} dH_2O) $CoCl_2{\cdot}6H_2O$, 1.0 mL (180.0 g L^{-1} dH_2O) $MnCl_2{\cdot}4H_2O$, 1.0 mL (1.3 g L^{-1} dH_2O) H_2SeO_3, 1.0 mL (2.7 g L^{-1} dH_2O) $NiSO_4{\cdot}6H_2O$, 0.1 mL (100 mM, Cat. P0758S, NEB) Na_3VO_4, and 1.0 mL (1.94 g L^{-1} dH_2O) K_2CrO_4 in 1 L H_2O and was sterilized through a 0.2-μm filter. The F/2 vitamin stock solution was made by mixing 200 mg thiamine-HCl, 10 mL of a 0.1 g L^{-1} biotin stock, and 1 mL of a cyanocobalamin 1 g L^{-1} stock in 1 L H_2O and was sterilized through a 0.2-μm filter.

For agar plates, equal parts sterilized liquid L1 medium and autoclaved 2 % agar were combined and poured into Petri dishes.

2.3. Plasmid construction

All plasmid constructs were done by Gibson assembly using the NEBuilder® HiFi DNA Assembly Bundle for Large Fragments (New England Biolabs, Canada). Amplicons used to do the assemblies were amplified by PCR with PrimeSTAR GXL DNA Polymerase (Takara Bio, Japon) following the manufacturer's protocol. Episome pDMi7 was made by replacing the URA3 element in pPtGE30 with an expression cassette containing 40SRPS8 promoter and FcpA terminator driving *eGFP* linked with a T2A self-cleaving peptide linker to the enzyme ech fused with the enzyme fcs by a flexible linker $(GGGGS)_3$ (Fig. 1a). Expression vector pDMi8 was derived from pDMi7 by inserting *mCherry* ORF next to the T2A peptide instead of *ech*, flexible linker $(GGGGS)_3$, and *fcs* genes.

The *ech*, and *fcs* containing a Myc tag in 3′ genes were codon optimized and were synthesized by Bio Basic (Markham, Ontario, Canada) using the codon usage list extracted from the High-performance Integrated Virtual Environment-Codon Usage Tables (HIVE-CUTs) platform [25] with the refseq database on September 21st in 2018 (Supplementary Fig. S9). The *eGFP* and *T2A* DNA were amplified from pPtGE33, and the 40SRPS8 promoter was amplified from pPtGE19 [10].

All forward and reverse primers used are listed in Supplementary Table S4. Plasmids list used and constructed for this study is available in Supplementary Table S5. The list of all DNA sequences used in this study is available in Supplementary Table S6.

2.4. Transformation of P. tricornutum by bacterial conjugation from E. coli cells

Conjugation protocol was adapted from Karas et al. (2015) and Slattery et al. (2018).

2.4.1. Preparation of P. tricornutum cells

A liquid culture of *P. tricornutum* of 4 to 8 days old was used as the starter culture. The cell concentration of 250 μL culture was adjusted to 1.0×10^8 cells ml^{-1} based on the optical density of a 1/5 diluted sample and calibration curve (Supplementary Fig. S10). To adjust the concentration, the cells were centrifuged at 3500 ×*g* at room temperature. The 250 μL of concentrated culture is plated on 1/2 × L1 1% agar plates and grown for 4 days. The cells were then scrapped three times with 400 μL of L1 media and the concentration was adjusted to 5.0×10^8 cells mL^{-1}.

2.4.2. Preparation of E. coli cells

The *E. coli* transconjugants used for bacterial conjugation contain the conjugative plasmid pTA-Mob [26] and the cargo plasmid with the expression cassette of interest. A 50 mL of LB was inoculated with 1 mL of an overnight 5 mL culture of *E. coli* and incubated at 37 °C to A_{600} of 0.8–1.0, centrifuged for 10 min at 3000 ×*g* and resuspended in 500 μL of SOC media.

2.4.3. Bacterial conjugation

Then, 200 μL of *P. tricornutum* and 200 μL of *E. coli* cells were mixed to initiate the conjugation. The cell mixture was plated on 1/2 × L1 5% LB 1 % agar plates, incubated for 2 h at 30 °C in the dark, and then moved to grow for two days at 18 °C in the light. After 2 days, the cells were scrapped twice with 650 μL of L1 media. The scrapped cells were plated with 200 μL three times and the remaining volume (<200 μL) on the fourth plates of 1/2 × L1 1% agar plates supplemented with Zeocin 50 μg/mL (or nourseothricin 200 μg/mL for pPtGE33 transconjugants only). Colonies appeared after 10–14 days of incubation at 18 °C with a photoperiod of 16 h light: 8 h dark.

Counts of *P. tricornutum* (Pt) colonies per conjugation were compared statistically by Welch's *t*-test ($\alpha = 0.05$) in Microsoft Excel with the Data AnalysisToolPak.

The *P. tricornutum* transconjugants were named as follows in our study: (Name of the episome used for conjugation) - (Round of conjugation)(Number of the *E. coli* colony obtained from the Gibson assembly of pDMi7) - (Number of *P. tricornutum* colony after the bacterial conjugation)→As an example: DMi7-21-1

2.5. Fluorescence microscopy

Colonies were analysed 14 days after conjugation under a Fluorescent Stereo Microscope Leica M165 FC with GFP filter for eGFP fluorescence and RFP filter for mCherry fluorescence detection. Colonies were observed with a magnification of 80 to 120×.

The percentage of eGFP fluorescent colonies were compared

statistically by Welch's *t*-test ($\alpha = 0.05$) in Microsoft Excel with the Data AnalysisToolPak.

2.6. Episome DNA isolation from P. tricornutum and episome rescue

The recovery of episomes from *P. tricornutum* was adapted from Karas et al. (2015) and Slattery et al. (2018), and the manufacturer's protocol of the Large Plasmid Mini Kit (Geneaid Biotech Ltd., Taiwan). Briefly, 5 mL of a 7 days old culture of *P. tricornutum* was centrifuged for 10 min at 3500 ×*g*. The pellet was resuspended in 235 μL of PDL1 buffer (Geneaid) supplemented with 5 μL of hemicellulase (100 mg/mL), 5 μL of lysozyme (25 mg/mL) and 5 μL of 20 T zymolyase solution (10 mg/mL). The mixed solution was incubated for 30 min at 37 °C. To initiate the lysis, 250 μL of PDL2 buffer was added to the solution and mixed by inversions 5 to 10 times and incubated 2 min at room temperature. The lysis was neutralized by the addition of 375 μL of PDL3 buffer and mixed by inversions 5 to 10 times and incubated 2 min at room temperature. The mixture was centrifuged for 3 min on a microcentrifuge at maximal speed at room temperature. The manufacturer's protocol from the Large Plasmid Mini Kit (Geneaid) was then followed for the steps of DNA binding, washing, and elution.

To complete the episome rescue, 2 μL of the miniprep were transformed by heat shock in NEB 10-beta chemically competent *E. coli* cells following the manufacturer's protocol up to the spreading of transformed cells on LB plates. At this point, 100 μL of the cell mixture was spread on a LB plate with chloramphenicol. The remaining volume was centrifuged 2000 ×*g* at room temperature for 5 min. The supernatant was partially removed (700 μL) and the remaining volume is used to resuspend the pellet. The total volume of the cell mixture is then plated on a LB plate with chloramphenicol. After an overnight incubation at 37 °C, an isolated colony was used to inoculate 5 mL of LB culture to proceed for a miniprep following the manufacturer's protocol from the Large Plasmid Mini Kit (Geneaid).

2.7. Whole plasmid sequencing

Episome pDMi7 was sequenced following the Gibson assembly and served as the reference sequence for further alignment made with CLC Main Workbench 7.7 (QIAGEN, Germany) with the "very accurate" alignment parameter. pDMi7 episomes from Gibson assembly and episome rescue were completely sequenced by CCIB DNA Core (Massachusetts General Hospital, United States of America) through their next Next-Generation sequencing Illumina MiSeq platform.

2.8. Promoter region prediction

To analyze the sequence of a putative promoter in DMi7-21-3 clone, a region of 1000 bp before the first ATG of the longest predicted open reading frame from the sequence result of episome rearrangements was analysed using PlantCARE [27] software for predicting transcription factor binding sites. Besides, to determine if it had the potential consensus transcription initiation sequence from *P. tricornutum* we searched for "TCAH$_{+1}$W" in the selected region [28].

2.9. Protein extraction

One-week-old cultures were centrifuged at 1500 ×*g* for 20 min at 4 °C. Pellets were weighed and resuspended with a ratio 1.3 g FW/mL in an extraction buffer (50 mM Tris pH 7.4, 500 mM NaCl, 0.1 % Tween20, 1× protease inhibitor cocktail). Sonication was performed 6 times at 35 % amplitude, 30 s on, 30 s off for 3 min total. Protein extracts were centrifuged at 20,000 ×*g* for 30 min at 4 °C. Supernatant containing the total soluble protein fractions were kept at −20 °C to be used for western blot. The proteins were quantified with the RC DC™ Protein Assay Kit I (Bio-Rad cat # 5000121).

2.10. Western blot

For protein detection, 50 μg of total proteins were loaded in 10 % SDS-PAGE. The proteins were then transferred to the 0.2 μm PVDF membrane and transfer settings were; 100 V constant and 400 mA for 2 h. Primary antibodies were incubated overnight at 4 °C. Primary antibody for eGFP was purchased from Cedarlane (Ontario L7L 5R2 Canada, cat. #CLH106AP) and for Myc Tag from ThermoFisher Scientific (Illinois 61101 USA, cat. #MA1-21316). Both were used at a 1:1000 dilution in 3 % BSA. After three washes with Tris-buffered saline, 0.1 % Tween 20 (TBST) solution, the blots were incubated for 1 h in a 1:20,000 dilution, in 5 % milk, of Immun-Star Goat Anti-Mouse (GAM)-HRP Conjugate from Bio-Rad (Ontario L5T 1C9 Canada, cat. #1705047). *P. tricornutum* clone containing pPtGE33 was used as a positive control for eGFP expression. A quantity of 10 ng of Multiple Tag from GenScript (cat. # M0101) was used as a positive control for Myc Tag detection. After three washes of the membrane with TBST solution, protein detection was realized by using Clarity Max Western ECL Substrate-Luminol solution from Bio-Rad (cat # 1705062S). Chemiluminescence detection and Ponceau S stained (Glacial Acetic Acid 5 % v/v, Ponceau Red dye 0.1 % m/v) of the blots were visualized using ChemiDoc Imaging System with Image Lab Touch Software (Bio-Rad cat # 12003153) and Image Lab™ Software (Bio-Rad cat # 1709690). The molecular weight of the protein corresponding to the detected band was calculated with Image Lab™ Software and the method point-to-point (semi-log).

2.11. Flow cytometry and fluorescence-activated cell sorting (FACS)

The BD FACSMelody (BD Biosciences, La Jolla, CA, USA) equipped with blue (488 nm), red (640 nm) and violet (405 nm) lasers was used to sort *P. tricornutum* transformed transconjugants according to eGFP production. Prior to the first sort, selected transconjugants were grown in L1 liquid medium supplemented with corresponding antibiotic and grown for 7 days. *P. tricornutum* cultures were washed in L1 medium, filtered on a 100 μm Nylon Net filter (Merck Millipore, Ireland) and diluted to an $OD_{730} = 0.1$ in L1 media prior to sorting.

Events were acquired at a fixed flow rate of 1 and at least 10,000 events were analysed. Cells were gated according to FSC-A (forward scatter area) and SSC-A (side scatter area) parameters and doublons were excluded according to FSC-H (height) vs. FSC-W (width) and SSC-H vs SSC-W plots. Chloroplast autofluorescence was measured on the PerCP channel (700/54 nm). Cells with high levels of PerCP fluorescence were further gated whereas cells with non-specific high autofluorescence were excluded based on their emission in the 448/45 nm channel. eGFP was further analysed on the 527/32 nm band-pass filter channel. Sorting parameters were set on purity parameter. Sorted cells were collected in 1.5 mL Eppendorf containing 500 μL of L1 media without antibiotics. Sorted cells were centrifuged 10 min at 3500 ×*g* and 90–95 % of the supernatant was removed and replaced by 500 μL of L1 media supplemented with zeocin 50 μg/mL and chloramphenicol 25 μg/mL.

For the second round of sorting, 1st round-sorted cultures were incubated for 7 days and diluted in 1 mL to an OD_{730} of 0.1. The cultures were then grown for another 7 days before being sorted a second time, according to the same procedure. Figures and statistics were analysed using BD FlowJo version 10 software (BD Biosciences, La Jolla, CA, USA, 2020). At least 10,000 events were acquired for each sample.

2.12. Total RNA extraction and RT-PCR

Total RNA from *P. tricornutum* from flash frozen biomass (10 mL of a 7 days old culture) was extracted using the RNeasy Plant Mini Kit according to the manufacturer's protocol (QIAGEN, Germany). Followed by DNase I (NEB, Canada) treatment at 37 °C for 15 min according to the manufacturer's protocol to remove episome contamination and purification using RB columns from Plant Total RNA Mini Kit (GENEAID,

Taiwan). The RNA quality was confirmed by migration at 100 V on a 1 % agarose gel for 35 min. RT-PCR was carried out using High-Capacity cDNA Reverse Transcription Kit according to the manufacturer's protocol (Applied Biosystems, USA). For PCR reaction, all the transconjugants were tested using a forward primer binding at *ech* (DMI7_F) and a reverse primer that binds at the beginning of *fcs* (DMI7_R). The reaction conditions were: initial denaturation at 95 °C for 30s; 30 cycles of 95 °C (30s), 50 °C (40s) and 68 °C (16 s), with a final extension at 68 °C for 5 min. PCR products were visualized in 1 % agarose gel and the length of 259 bp was expected for the positive amplification. The forward and reverse primers sequence used for PCR are listed in Supplementary Table S4.

3. Results and discussion

3.1. Fluorescence microscopy analysis of transconjugants demonstrated different fluorescent patterns in a single transformation event

In a previous study, we constructed an episome encoding eight heterologous genes involved in vanillin biosynthesis that was stable and propagated in *P. tricornutum* over four months with no evidence of rearrangements [10]. However, we could not detect vanillin, *Vp*VAN enzyme activity or accumulation in the transconjugants. Since there is a controversy surrounding the function of *Vp*VAN in vanillin biosynthesis, we engineered a new construction to study protein production linked to vanillin biosynthesis. It is based on a simpler pathway using two genes encoding enzymes already characterized for the conversion of ferulic acid to vanillin [18,19]. Specifically, the fcs enzyme can convert the ferulic acid into feruloyl-CoA which will be converted into vanillin by the catalytic activity of ech (Fig. 1a). The new episomal construction was built into the previously characterized plasmid pPtGE33 containing the *Thosea asigna* virus 2A self-cleaving peptide (T2A) linker cloned between eGFP and mCherry sequences, and the selection marker *ShBle* as resistance cassette obtained from pPtGE30 plasmid [10]. There, the expression of the bi-cistronic gene construction of *eGFP-T2A-mCherry* is under the control of the 40SRPS8 promoter and its native terminator which is known for increased heterologous gene expression [10]. Furthermore, the fluorescence from the reporter genes, eGFP and mCherry, will allow for high-throughput screening of transconjugants from liquid culture by flow cytometry or directly from plate using fluorescence microscopy (Supplementary fig. S1). Using this system, we constructed two episomes; pControl and pDMi7 (Fig. 1b, c). In pDMi7, the expression cassette *eGFP-T2A-mCherry* of pControl was modified by replacing the *mCherry* sequence with the *ech* and *fcs* coding sequence linked by a flexible peptide sequence $(GGGGS)_3$ (Fig. 1c).

First, we evaluated the transformation efficiency on zeocin resistant transconjugants 14 days after the transformation of *P. tricornutum* with pPtGE30 (empty vector) and pDMi7 by *E. coli* conjugation. The number of colonies obtained from pControl and pDMi7 conjugations was not significantly different (p-value = 0.25) (Fig. 1d and Supplementary Table S2), indicating that the transformation was successful.

To evaluate the potential of screening directly on L1 agar plate media, the eGFP fluorescence was observed using fluorescence microscopy on *P. tricornutum* zeocin resistant transconjugants. For pControl, the fluorescence of eGFP and mCherry was observed respectively on 82.1 ± 1.8 % and 76.0 ± 1.0 % of the analysed colonies (Fig. 1e and Supplementary Table S1). Whereas, the absence of eGFP and mCherry fluorescence was observed in 17.9 % and 24.0 % of pControl colonies, respectively. The non fluorescent clones could be due to an absence of fluorescent proteins accumulation that might be caused by a partial expression cassette resulting from an incomplete episome transfer during conjugation, or by mutations in the expression cassette sequence.

Unexpectedly, a percentage of 8.2 ± 0.8 % of pDMi7 colonies displayed eGFP fluorescence (Fig. 1e) which is significantly different (p-value < 0.00001) from the pControl results as determined by Welch's t-test. The only difference between pControl and pDMi7 is located in the expression cassettes downstream the T2A sequence which is coding respectively for mCherry or ech-fcs (Fig. 1b-c). Regarding the protein of interest eGFP-T2A-ech-fcs from pDMi7, the eGFP fluorescence should be produced by the cleaved protein eGFP-T2A, or the uncleaved product eGFP-T2A-ech-fcs. Compared to pControl, the low percentage of pDMi7 colonies suggests an instability or a potential toxicity related to the *ech-fcs* at gene level that could lead to low or non-fluorescent transconjugants. It is possible that the fusion protein ech-fcs promotes the instability of the uncleaved protein eGFP-T2A-ech-fcs resulting in its degradation or in its incapacity to produce fluorescence. However, it cannot explain the absence of fluorescence from the cleaved form eGFP-T2A since this protein is cleaved and released from the ribosomes before the translation of *ech-fcs* downstream the same mRNA. In the case of the non-fluorescent pDMi7 colonies, there is then also lack of accumulation of the protein eGFP-T2A. It then means that the gene *eGFP-T2A-ech-fcs* is not translated, or that the T2A peptide is not cleaved resulting only in eGFP-T2A-ech-fcs proteins that are not fluorescent or not degraded by *P. tricornutum*.

Regarding the possible toxicity of ech-fcs, the number of zeocin resistant colonies obtained from pPtGE30 (empty vector) is not significantly different than conjugations with pDMi7 (Fig. 1d and Supplementary Table S2). It indicates that if the *ech-fcs* gene was lethal for *P. tricornutum*, the diatom was able to counter its toxic property through translation inhibition or degradation of the protein of interest. It would allow its survival and would explain the diminution of fluorescence transconjugants. It is possible that the toxicity is an effect of the catalytic activity of the ech-fcs fusion protein in *P. tricornutum*'s metabolism, either by producing vanillin or catalyzing another reaction, thus consuming an important metabolite or producing something toxic. A literature search did not yield any information on the toxicity from ech or fcs enzymes.

It is interesting to note that the fluorescent areas inside the pDMi7 transconjugant colonies were not always uniform. Some colonies were almost completely fluorescent (Supplementary fig. S2a), and others were partially fluorescent (Supplementary fig. S2b-f). We noticed that some colonies were weakly fluorescent as the GFP signal was observed only from small dots (Supplementary fig. S2b). In other cases, the colonies were fluorescent, but they had regions where the fluorescence signal was undetectable (Supplementary fig. S2c-f). An overlap between a fluorescent colony and a non-fluorescent one could explain this pattern. The irregular shapes of some colonies and their fluorescence pattern support this hypothesis (Supplementary fig. S2c-e). However, for some round colonies, it was more ambiguous to determine if two colonies had merged or if the heterogeneous fluorescence pattern was originating from a unique colony (Supplementary fig. S2f). Theses observations were also present among pControl transconjugants colonies (Supplementary fig. S3).

3.2. The sequences of the episomes recovered from P. tricornutum pDMi7 transconjugants showed rearrangements and mutations

We selected ten positive pDMI7 transconjugants displaying antibiotic resistance to investigate eGFP fluorescence from the reporter gene. Six out of the ten transconjugants were eGFP positives (GFP+) whereas four were not (GFP-). To further investigate this, we assessed the integrity of the pDMI7 with an episomes rescue experiment. For that, the recovered episomes from *P. tricornutum* cultures were transformed and replicated in *E. coli* from our six GFP+ transconjugants (namely DMi7-21-1, DMi7-21-3, DMi7-31-1, DMi7-31-3, DMi7-31-4, and DMi7-31-8) and four non-fluorescent (GFP-) ones (DMi7-21-14, DMi7-21-15, DMi7-21-16, and DMi7-31-10), extracted and digested (Fig. 2a). The double digestion of the recovered episomes showed variable profiles compared to the control plasmid which is the initial pDMi7 following Gibson assembly. Only DMi7-21-16 exhibited an expected restriction enzyme digestion profile corresponding to the plasmid control (Fig. 2a). This suggests that DNA rearrangements occurred in nine out of the ten

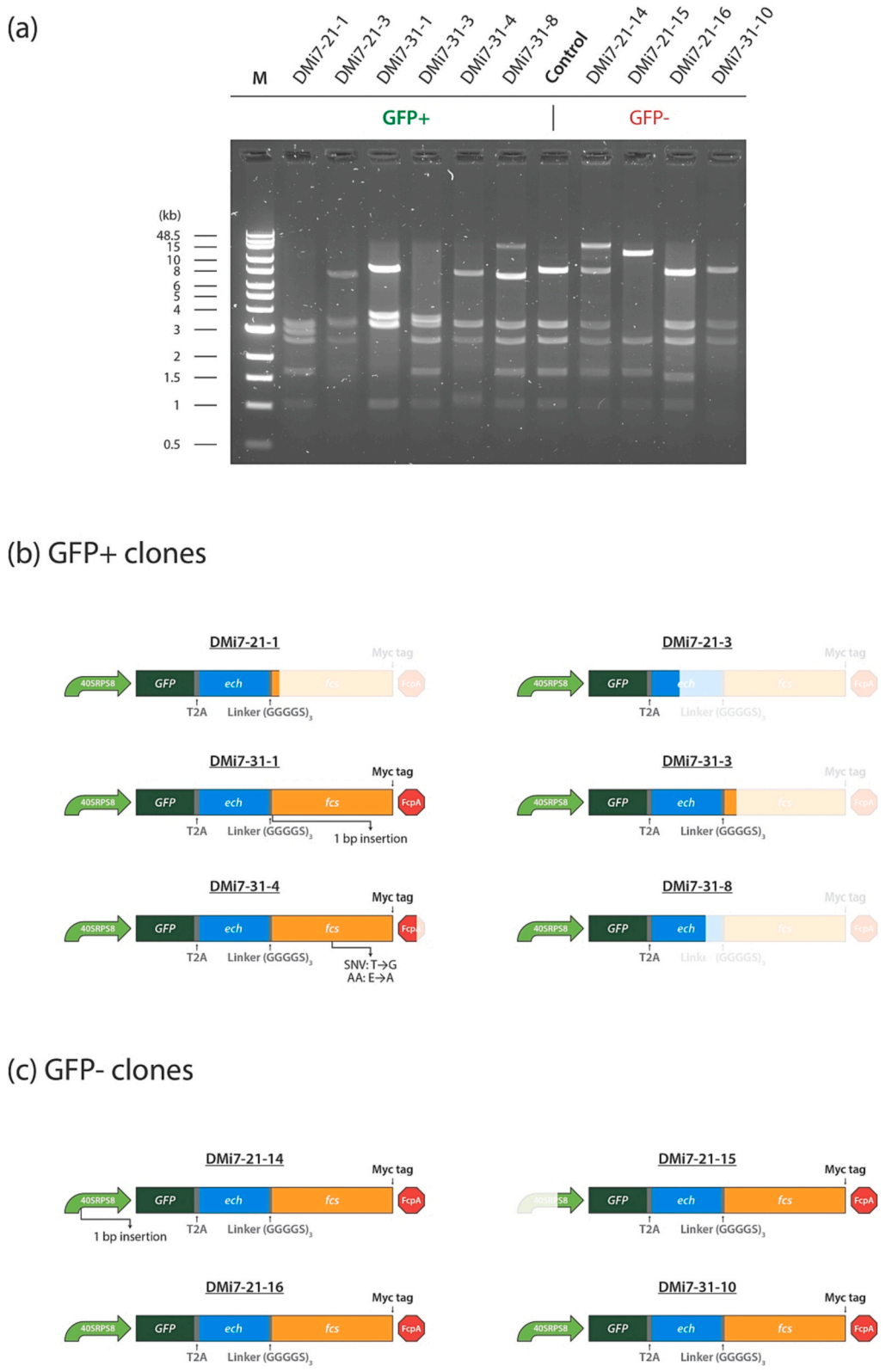


Fig. 2. DNA digestion profiles and representation of the rearrangements of the recovered episomes from pDMi7 *P. tricornutum* transconjugants
(a) Double digestion by *Pst*I and *Spe*I of the episomes recovered from *P. tricornutum* transconjugants showing rearrangement in their sequence. The control is the digestion of the plasmid pDMi7 used for the conjugation. (b and c) Schematic representation of the rearrangements in the expression cassette from the episomes recovered from *P. tricornutum* transconjugants containing pDMi7. The GFP+ transconjugants are the one where GFP fluorescent cells have been detected by fluorescent microscopy and flow cytometry. Shaded parts represent deletions in the expression cassettes and is on scale. As a reference the schemes of DMi7–21-16 and DMi7–31-10 are identical to the designed expression cassette from pDMi7. 40SRPS8, 40SRPS8 promoter; eGFP, enhanced green fluorescent protein; T2A, *Thosea asigna* virus 2A self-cleaving peptide; ech, enoyl-CoA hydratase/aldolase; fcs, feruloyl-CoA synthetase; FcpA, FcpA terminator; SNV, single nucleotide variation (T to G); AA, amino acid (E to A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

episomes recovered from pDMi7 transconjugants whether they produced eGFP or not (Fig. 2a).

To further analyze the profile of rearrangements, the rescued episomes were sequenced by whole plasmid next-generation sequencing with the Illumina MiSeq platform. Interestingly, the episome sequence recovered from the six GFP+ clones showed the *fcs* coding sequence altered by mutations, or from complete or partial deletions (Fig. 2b, c). The complete analysed changes in the DNA sequences are listed in supplementary table S3. The sequence of the expression cassette was truncated with total or partial deletion of the *fcs* gene and complete deletion of the FcpA terminator for DMi7-21-1, DMi7-21-3, DMi7-31-3, and DMi7-31-8, (Fig. 2b). The total length of the deletions starting in the expression cassette for DMi7-21-1 and DMi7-31-3 were of 5523 bp and 4873 bp respectively. Regarding DMi7-31-8, the deletion of 2552 bp was replaced by a partial and inverted duplicate from the *ech* sequence of 155 bp. In the case of DMi7-21-3 there was also a partial deletion of 5004 bp starting in the *ech* gene. Moreover, in this transconjugant, the expression cassette was also present in two extra partial duplicates. The first one carries the partial sequence of *ech-fcs*. The second one consists of a partial *eGFP* sequence with the complete gene encoding the fusion protein (ech-fcs), however, lacking the promoter sequence of 40SRPS8. Regarding the clone DMi7-31-1, there was a change in the open reading frame of the *fcs* gene with a single insertion after the first twelve nucleotides as represent in Fig. 2b. A nonsynonymous substitution was detected in the sequence of the *fcs* gene of the clone DMi7-31-4 causing a substitution of negatively charged amino acid glutamic acid (E) with a non-polar aliphatic alanine (A). Further work needs to be done to evaluate the impact of this mutation on the enzymatic activity of *fcs*. Altogether, the six GFP+ transconjugants demonstrated instability in their episomal sequences that was linked to the *fcs* sequence. This type of extrachromosomal DNA instability which influences the production of an heterologous protein has never been reported from diatoms before. It is possible that the lack of data in the literature concerning this matter could be due to the fact that the episome in *P. tricornutum* has been discovered fairly recently [13]. However, it is conceivable that this instability should not be specific to the *fcs* coding sequence used in this study. If so, it would then be possible that the episome instability could be problematic in some metabolic engineering experiments as it was reported for development of Cyanobacteria strains [29].

Intriguingly, no mutations were observed in the *eGFP-T2A-ech-fcs* coding sequence of the expression cassette in any of the non-fluorescent GFP- transconjugants (Fig. 2c). Moreover, DMi7-21-16 and DMi7-31-10 had no modification in the 40SRPS8 promoter and in the FcpA terminator sequences. Regarding the two other eGFP negative transconjugants, the expression cassette from the clone DMi7-21-14 contained a single nucleotide insertion in the promoter sequence and DMi7-21-15 was harboring a deletion of the first half of the promoter sequence. Thus, the absence of eGFP fluorescence of those transconjugants could not be explained by rearrangements and/or mutations in the expression cassette. Besides, the absence of fluorescence in the recombinant *P. tricornutum* transconjugants is not related to the DNA sequence, which suggests that regulation at other levels such as RNA silencing, protein degradation or low protein production below the sensitivity levels, could be responsible for this phenomenon.

Interestingly, a similar observation was reported recently. Defrel et al. (2021) transformed *P. tricornutum* by biolistic with the *uidA* gene, which encodes the β-glucuronidase (GUS) enzyme, linked by the 2A peptide to the *nourseothricin N-acetyl transferase* gene (NAT). No GUS activity was detected in two transconjugants with a colorimetric assay despite the absence of sequence modification of their transgene. With the development of a fluorometric assay, GUS activity was detected in both transconjugants and was 30 and 40 times lower compared to the clone with the strongest activity. The reason for the low activity from these two transconjugants was not elucidated. In their case, the integration of the transgenes in poorly expressed genomic regions might be the cause of their low activity. This hypothesis cannot be applied to our work since the episomes are extrachromosomal expression systems. However, it is possible that the heterologous expression systems, either by random DNA integration or by episome, would be affected by defense mechanisms of *P. tricornutum* like RNA silencing. Regarding RNA silencing, De Riso et al. showed that the GUS reporter gene expression can be successfully silenced using constructs with sense and antisense guide RNAs in *P. tricornutum* [30]. Besides, they were able to modulate the expression of two endogenous genes encoding for phytochrome (Dph1) and cryptochrome/photolyase family 1 (CPF1), proving that RNA silencing can occur in *P. tricornutum* [30]. Screening of potential miRNAs would allow to determine whether this mechanism could be the cause of the absence of fluorescent signal in transconjugants with no mutations. In this regard, Huang et al. identify by sequencing and bioinformatic analysis novel miRNAs which may play an important role in the regulation of *P. tricornutum* metabolism [31].

3.3. Fluorescent and non-fluorescent transconjugants tested positive for gene expression of the ech-fcs fusion protein

To investigate the heterologous protein production from the pDMi7 transconjugants, we first tested the gene expression of the cassette by an endpoint RT-PCR, using primers that amplified a 209 bp long fragment that covered the end of *ech* and the beginning of the *fcs* genes. For this purpose, six transconjugants were analysed including GFP+ and GFP- ones. To remove traces of episomal DNA that co-purified with the total RNA and could give false positives from the PCR analysis, the samples were treated with DNase I. Total RNA samples without reverse transcription were used as negative control for the PCR. All the transconjugants tested were positive indicating that they express the gene encoding the ech-fcs fusion protein (Fig. 3). Based on the construction design, the eGFP should be produced from the same mRNA as the ech-fcs fusion protein, and then separated by the self-cleavage T2A linker.

The clone DMi7-21-3 gave a positive result even though the partial cassette where the 40SRPS8 promoter and the *eGFP* sequences are complete does not have the reverse primer binding site (Fig. 2b and Supplementary table 3). In the episome recovered from this transconjugants, there were three partial copies of the cassette where one of them contains the sequence that would give correct RT-PCR amplification. The first partial copy is presented in Fig. 2b, which is composed of 40SRPS8 promoter, *eGFP*, and *ech* partial sequence lacking the binding site for the reverse primer (DMi7_R). The second partial copy of 2400 bp is inserted after the reference position 13,888 bp. However, it is composed of *ech* partial sequence (3′-end) lacking the promoter, eGFP, and binding site for the forward primer (DMi7_F). Regarding the other copy, it is an insertion of 3561 bp before the reference position 14,879. In this case, the 40SRPS8 promoter sequence and the first nucleotide of *eGFP* are lacking, but the *ech-fcs* sequence is complete and contains the

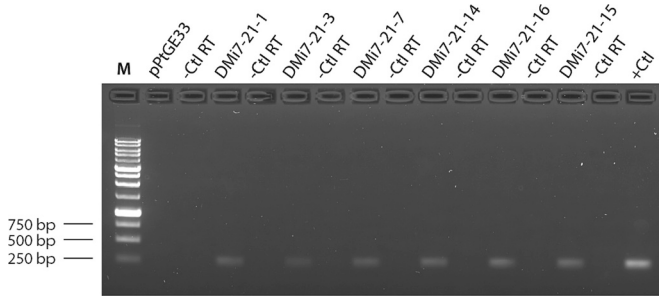


Fig. 3. Expression of the cassette of interest from pDMi7 *P. tricornutum* transconjugants
Endpoint RT-PCR of *P. tricornutum* transconjugants that have been observed by fluorescence microscopy as fluorescent GFP+ (DMi7-21-1, -3, and -7) or non-fluorescent GFP- (DMi7-21-14, -15, -16). The plasmid pDMi7 has been used as a positive control for PCR (+Ctl) and the respective RNA samples (without reverse transcription) were used as negative controls (-Ctl RT).

RT-PCR primers binding sites. This could explain the cDNA amplification from DMi7–21-3. We hypothesized that the sequence upstream the duplicate, product of the episome rearrangements, could drive the expression of the *eGFP:c.1delA-T2A-ech-fcs* cassette. For this, we analysed the upstream sequence of the forward promoter binding site and found transcription factor binding sites predicted by PLANTCARE software [27] and the presence of the transcription initiation (Inr) like consensus motif "TCAH$_{+1}$W" (Supplementary Fig. S4) which has been characterized before in this diatom [28].

3.4. Detection of heterologous proteins in P. tricornutum transconjugants depends on the percentage of producing cells in the total population and on the DNA sequence of the expression cassette

To confirm the presence of eGFP fluorescence, transconjugants that were assessed by endpoint RT-PCR were also analysed by flow cytometry. We confirmed that fluorescent *P. tricornutum* transconjugants observed under fluorescence microscopy contained both eGFP+ and eGFP- cells, with a majority of the cells that were non-fluorescent (Fig. 4 a-b, initial cultures). As such, the percentages of fluorescent cells obtained by flow cytometry were 33.8 %, 5.65 %, and 10.2 % for DMi7-21-1, DMi7-21-3, and DMi7-21-7 respectively (Fig. 4b; left panels). Regarding the non-fluorescent transconjugants (DMi7-21-14; DMi7-21-15 and DMi7-21-16), no eGFP fluorescent cells were detected by flow cytometry (Supplementary fig. S5; right panel).

Furthermore, we used cell sorting with flow cytometry to enrich the proportion of fluorescent cells in DMI7-21-1, DMI7-21-3, and DMI7-21-7 cultures (Fig. 4). A first sorting was performed on the initial transconjugants DMi7-21-1, DMi7-21-3, and DMi7-21-7 (Fig. 4b and supplementary fig. S6; left panels). Both fluorescent (named sGFP+ cultures) and non-fluorescent cells (named sGFP- cultures) were individually sorted. The sGFP+ and the sGFP- cell cultures were grown for 7

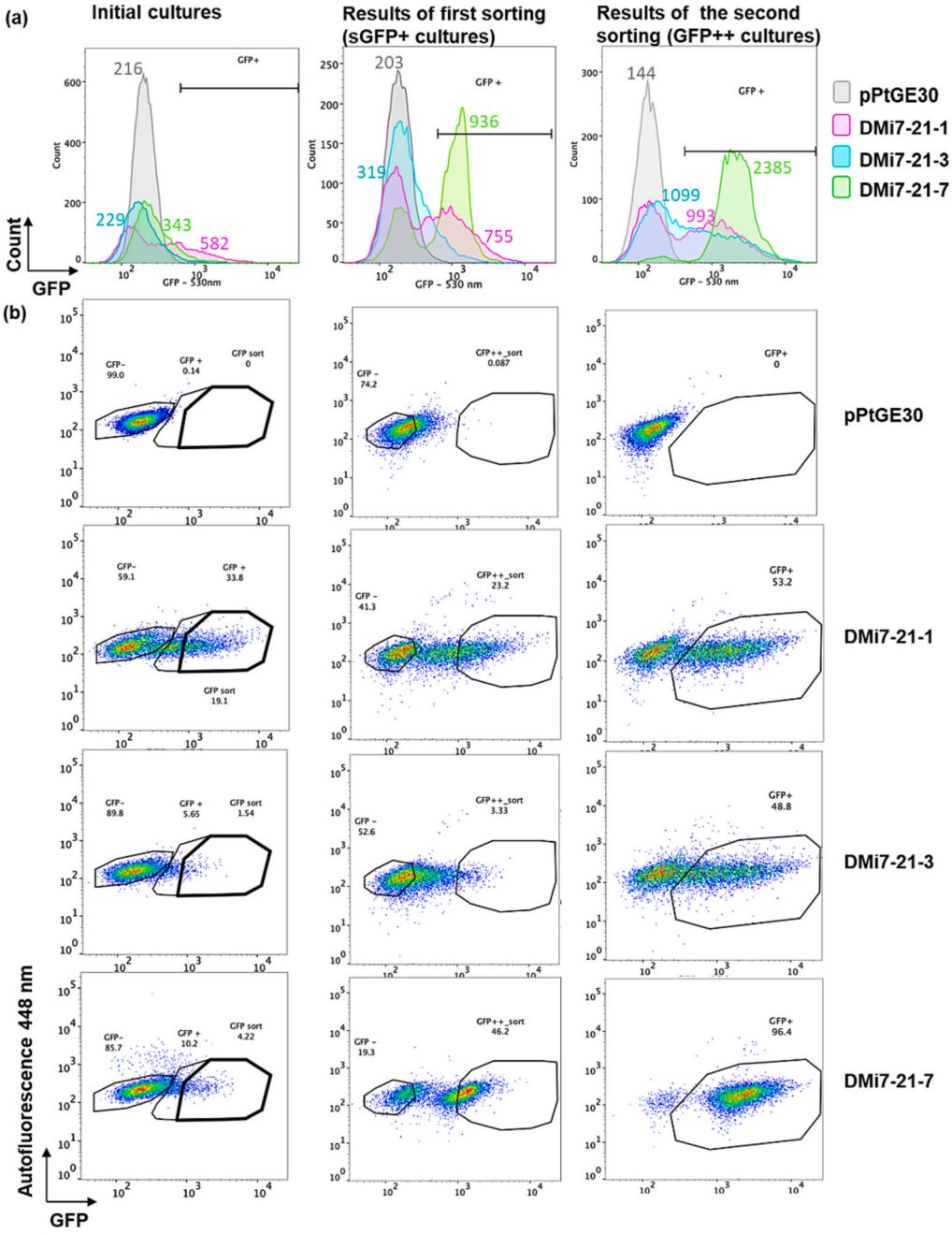


Fig. 4. Enrichment of *P. tricornutum* GFP+ through fluorescence activated cell sorting.
(a). Histogram plots of GFP intensity profile in cultures of DMi7-21-1 (pink), DMi7-21-3 (turquoise) and DMi7-21-7 (green) compared to control pPtGE30 (grey). Background autofluorescence across days following two sequences of enrichment (initial cultures and results from first sorting). The total GFP+ population gate is shown as reference. Numbers indicate the mean fluorescence intensity of GFP in the total chlorophyll+ population. (b). Pseudocolor dot plots of control pPtGE30 and transconjugant DMi7-21-1, DMi7-21-3, and DMi7-21-7 cells at 530 nm (GFP) in the x axis, and autofluorescence at 448 nm in the y axis from initial culture used for sorting (left panels). The cultures grown for 7 days after first sorting (sGFP+ cultures in center panels) were used for second sorting. The enriched cultures following the second sorting were then subcultured after 2 days and grown for 5 days (GFP++, right panels). Gates and frequencies of total GFP+, GFP+ sorted (sGFP+ sort and GFP++ sort) and GFP- populations were designed according to the autofluorescence of the negative control pPtGE30 at each day of experiment and are shown as reference. GFP, green fluorescent protein. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

days and then sorted for a second time (Fig. 4b and supplementary fig. S6; center panels). For this last round of sorting, only the fluorescent cells were sorted from the sGFP+ cultures and the resulting cultures from this enrichment were named GFP++ (e.g., DMi7-21-1^{GFP++}) (Fig. 4b; right panels). From the sGFP- enrichment, only the non-fluorescent cells were sorted and the resulting cell cultures were named GFP– (Supplementary fig. S6; right panel). With the enrichment in eGFP+ cells, the percentage of fluorescent *P. tricornutum* cells increased from 2 to 9 fold as shown in Fig. 4 from 33.8 % to 53.2 % for DMi7-21-1^{GFP++}, from 5.65% to 48.8% for DMi7-21-3^{GFP++}, and from 10.2% to 96.4% for DMi7-21-7^{GFP++}.

Among the transconjugants tested, only DMi7-21-7^{GFP++} reached almost 100% of fluorescent cells. Even after three subculturing of DMi7-21-7^{GFP++}, the enriched proportion did not decrease and showed over 99% of fluorescent cells (Fig. 5). The stability of the fluorescent cells population in the enriched DMi7-21-7^{GFP++} culture and the failure to recover the episomes from the initial culture cells of DMi7-21-7 could suggest that the DNA episome was integrated into the diatom genome. Intriguingly, the GFP- sorted cells such as DMi7-21-1$^{GFP–}$ and DMi7-21-3$^{GFP–}$ come from transconjugants with an intact eGFP sequence (Fig. 2b). This suggests that protein production such as GFP can be stably turned off in *P. tricornutum* episome (Supplementary Fig. S6). Although, DMi7-21-1^{GFP++} and DMi7-21-3^{GFP++} display similar eGFP frequency and mean fluorescence intensity (MFI), a distinct subpopulation of eGFP+ cells was only observed in DMi7-21-1^{GFP++}, compared to a continuous range of eGFP intensity (from negative to positive cells) was observed in DMi7-21-3^{GFP++} (Fig. 4a; right panel). Based on this observation, it is possible that some colonies of pDMi7 transconjugants are originating from a mix or heterogenous transconjugants as shown in (Supplementary Figs. S2 and S3) with a subpopulation of cells producing eGFP and another not producing eGFP. In all, the results demonstrated that using cell sorting by flow cytometry to enrich a cell population can increase the production of a heterologous protein from a culture of *P. tricornutum* transconjugants without optimization of the media or the culture conditions. This can also be achieved by the droplet-based microfluidic techniques that can be used to accelerate the screening of microalgal transconjugants and to increase the cell population producing the eGFP [32].

Next, we assessed the production of eGFP using western blot analysis (Fig. 6a). Proteins were extracted from *P. tricornutum* transconjugants including empty vector controls, and initial unsorted (DMi7-21-1, DMi7-21-3, and DMi7-21-7), and sorted (GFP++ and GFP–) cell cultures. As expected, the negative control (empty vector pPTGE30) showed no GFP specific bands (Fig. 6a). Positive control (pPtGE33) showed three bands (28. 36, and 58 kDa) where the lower (28 kDa) and the upper (58 kDa) bands corresponded to the cleaved and uncleaved form of the protein eGFP-T2A-mCherry that have a theoretical molecular weight of 29 and 56 kDa respectively (Fig. 6a). The results are consistent with previously reported ones [10]. The 36 kDa band appears to be a byproduct of the 2A. Indeed, green and red fluorescent proteins linked by a 2A peptide have been used to study the 2A cleavage efficiency. In some cases, it was reported that byproducts were detected in immunoblotting experiments with anti-GFP antibodies [33–35]. It raised the question of whether the assigned bands reported in these investigations as cleaved and uncleaved proteins and the byproducts detected by western blot could be eGFP cleaved and/or uncleaved variants with post-translational modifications.

Regarding the unsorted cultures, eGFP was detected only in DMi7-21-1 (Fig. 6a) and this transconjugant exhibited 33.4 % of eGFP+ cells (Fig. 6b, c). The analysis of the DMi7-21-1 episome sequence revealed the deletion of the initial stop codon and a partial loss of the *fcs* gene sequence (Fig. 2b and Supplementary table S3). Based on the protein size of 59 kDa detected in the western blot, it suggests that the eGFP was not cleaved by the self-cleaving T2A peptide. From the sequence of DMi7-21-1 episome, the expected protein sizes were 29 kDa and 65 kDa for the cleaved and uncleaved eGFP, respectively. If the *eGFP-T2A-ech-fcs* cassette would have been unaltered, the cleaved (29 kDa) and uncleaved protein (115 kDa) could have been expected. However, it appears that a deletion in the DNA led to an unexpected 59 kDa eGFP protein. Regarding the unsorted initial cultures of DMi7-21-3 and DMI7-21-7 transconjugants, no band was detected, and only 8.0 % and 5.9 % of cells were fluorescent, respectively (Fig. 6a-c). For the group of transconjugants enriched in non-fluorescent cells (GFP–), no proteins were detected. The percentages of eGFP fluorescent cells for the cultures of DMi7–21-1$^{GFP–}$ and DMi7–21-3$^{GFP–}$ are close to 0 %. The cultures of DMi7–21-7$^{GFP–}$ reached back to 7.82 % which is close to the level of the

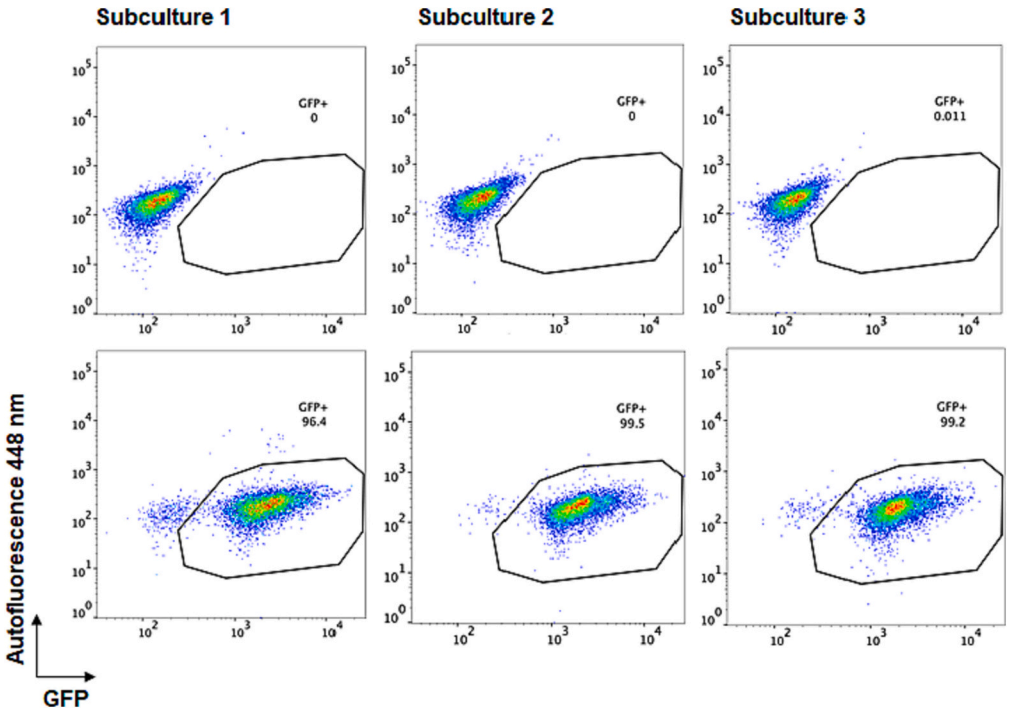
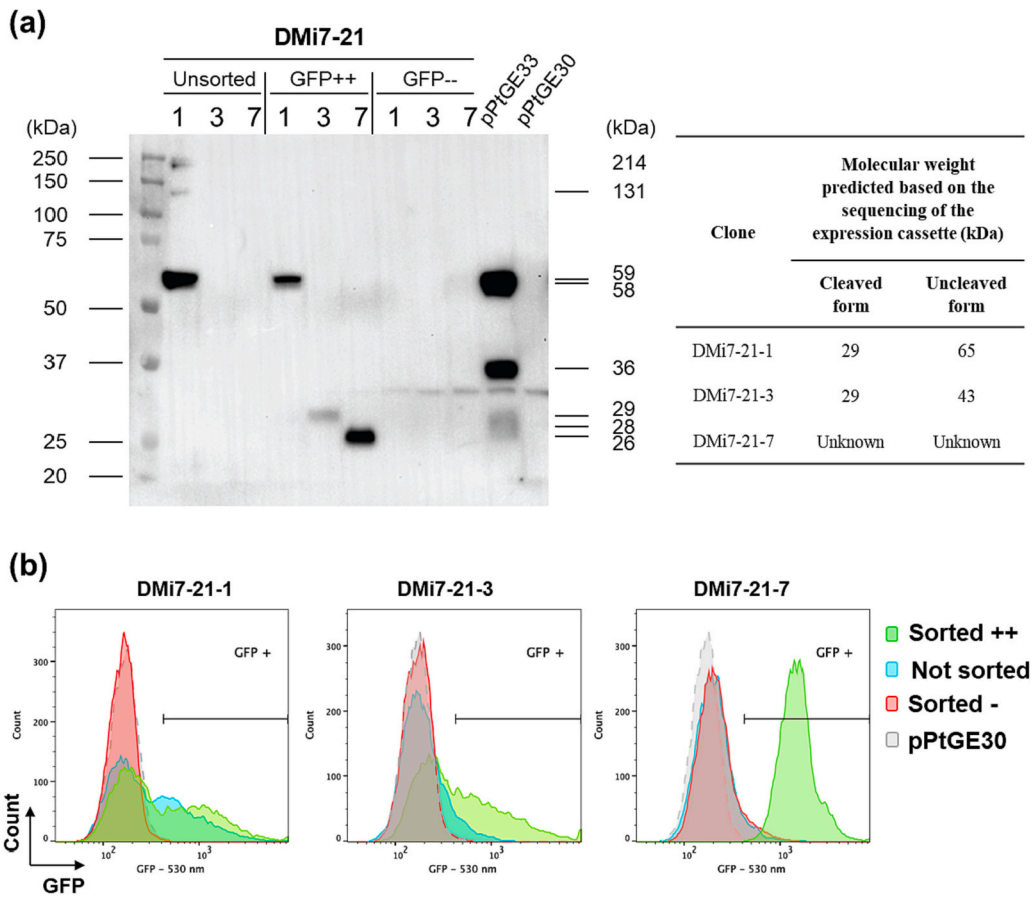


Fig. 5. Stable and high level of GFP expression in DMi7-21-7 across subculturing.
Pseudocolor dot plots of pPtGE30 (upper lane) and DMi7-21-7GFP++ (lower lane) cultures with 530 nm (GFP) in the *x* axis and autofluorescence at 448 nm in the *y* axis of the first three subculturing after the second sorting of GFP+ cells. Gates and frequencies of total GFP were designed according to the negative control pPtGE30 autofluorescence and are shown as reference. GFP, green fluorescent protein. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Clone	Molecular weight predicted based on the sequencing of the expression cassette (kDa)	
	Cleaved form	Uncleaved form
DMi7-21-1	29	65
DMi7-21-3	29	43
DMi7-21-7	Unknown	Unknown

(c)

	pPtGE30		DMi7-21-1		DMi7-21-3		DMi7-21-7	
	MFI	% GFP	MFI	% GFP	MFI	% GFP	MFI	% GFP
Not sorted	175	**0.6**	504	**33.4**	228	**8.03**	233	**5.93**
++	na	**na**	720	**42.2**	790	**45.3**	1628	**99.6**
-	na	**na**	165	**0.3**	179	**0.71**	243	**7.82**

Fig. 6. Comparison of GFP detected by western blot and by flow cytometry. (a) Western blot anti-GFP of unsorted, GFP+ sorted (GFP++), GFP- sorted (GFP–) cultures of DMi7-21-1, DMi7-21-3 and DMi7-21-7 compared to pPtGE30 (empty vector) and pPtGE33 (positive control). The size of the proteins from the ladder are identified at the left of the blot. The sizes measured for the detected proteins are indicated at the right of the blot. The table at the right of the blot contains the expected size of the proteins based on their sequencing result. (b) Histogram plots from the same culture of GFP++ sorted (green), GFP- sorted (red) and unsorted (blue) cultures of DMi7-21-1, DMi7-21-3 and DMi7-21-7 compared to pPtGE30 (grey) autofluorescence. The GFP+ population gate is shown. (c) Table of mean fluorescence intensity (MFI) at 530 nm (GFP) and frequency of GFP+ population (% GFP in bold) in sorted and unsorted DMi7-21-1, -3 and -7 cultures, as compared to pPtGE30 (autofluorescence). Sorted ++, GFP positive population was sorted twice; Sorted –, GFP negative population was sorted twice; na, not applicable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

unsorted culture (Fig. 6c).

As for enriched sorted GFP++ cells, different sizes of eGFP were detected in each of the three transconjugants (Fig. 6a). For example, the previously observed 59 kDa band was detected in *P. tricornutum* DMi7-21-1^{GFP++} but not in the other enriched transconjugants. The clones DMi7-21-1^{GFP++}and DMi7-21-3^{GFP++} showed a band at 29 kDa which matches the theoretical molecular weight of eGFP associated with the cleaved form of the T2A peptide. Thus, DMi7-21-1^{GFP++} seems then to produce a protein with an uncleaved T2A peptide, while DMi7-21-3^{GFP++} is producing a cleaved protein. In HeLa cells and in the TnT Quick Coupled Transcription/Translation System, it was demonstrated that the sequence upstream of the 2A peptide influences its cleavage efficiency [34]. This investigation demonstrated that an addition of a spacer composed of three amino acids (Gly-Ser-Gly) between the N-terminal protein and the 2A peptide can increase the efficiency of the cleavage. In the expectation that this spacer would also increase the cleavage efficiency of the 2A peptides in *P. tricornutum*, we added this spacer upstream of the T2A peptide in all the episomes designed in this study. No difference in the sequence upstream the T2A region that can explain the difference between DMi7-21-1 and DMi7-21-3 related to the cleavage of the heterologous proteins produced. This was also confirmed by sequencing analysis following the episome rescue (Fig. 2 and Supplementary table S3). In the case of the enriched culture of DMi7-21-7^{GFP++}, composed of almost 100% of eGFP fluorescent cells, the molecular weight of the detected band is 26 kDa which is lower compared to the band obtained from DMi7-21-3^{GFP++} (Fig. 6a). In the impossibility to recover the episome from this clone and sequence it, we cannot analyze the sequence of this protein, but it suggests a truncated form of eGFP. Interestingly, DMi7-21-3^{GFP++} production of only cleaved protein suggests that either the length of the C-terminal protein or the possible modifications of the mRNA secondary structure could affect the ribosome skipping producing mostly cleaved protein (which was detected by western blot) with a shorter sequence in the case of DMi7-21-3^{GFP++} and an uncleaved form in the case were the sequence of the complete ech protein is present (DMi7-21-1^{GFP++}). This could lead to think that when ech protein is complete it is less stable by itself, and it is not detected by western blot, either because of its enzymatic activity or the non-codon optimized sequence, since both rearranged episomes have lost the stop codon. In this regard, studies have shown that protein synthesis levels as well as a functional polycistronic 2A construct could be affected either by the N- or C- terminal fusion of partial 2A sequences and/or the identity and order of adjacent genes [36].

Based on the western blot results (Fig. 6a), there is knowledge that needs to be acquired regarding the cleavage efficiency of the T2A

peptide in *P. tricornutum*. Multiple 2A peptides have been compared for their cleavage efficiency in *Drosophila* [37], in CHO cells [38], in yeast [35], in the silkworm *Bombyx mori* [39], in the zebrafish embryos [33], in human cell lines [33,36], in adult mice [33], in mouse cell lines [36], and recently in *P. tricornutum* [40]. In all, the 2A peptides with the best cleavage efficiency were the T2A, the porcine teschovirus-1 2A (P2A), and the ERBV-1 peptide. In the case of *P. tricornutum*, Defrel et al. (2021) compared the activity of the GUS enzyme linked to NAT by the P2A or T2A peptides. A higher β-glucuronidase activity was observed when the P2A was used between the two enzymes and when the GUS enzyme was in C-terminal position. Based on a relative comparison with their reference strain named Gus5, the relative β-glucuronidase intensity with GUS in C-terminal was higher than 50 % for 10 of 24 and for 23 of 31 transconjugants, respectively for T2A and P2A groups. When *GUS* gene was upstream of the 2A peptide sequence, the transconjugants with a relative β-glucuronidase intensity that was higher than 50 % were 2 of 18 for the group with the T2A peptide and 5 of 20 for the ones with the P2A peptide. The authors proposed that when the GUS enzyme is cloned in N-terminal of the construct, the residual 20 amino acids resulting from the cleavage of the 2A peptide could negatively affect its β-glucuronidase activity. The cleavage efficiency of the 2A peptides was not measured by Defrel et al. (2021) and it was then not correlated with the β-glucuronidase activity. While we also analysed the production of the ech-fcs fusion protein by western blot detecting an anti-myc tag antibody. None of the fusion proteins tagged with a Myc tag in C-terminal were detected on the blot (Supplementary Fig. S7). More studies are required for better understanding 2A peptides for *P. tricornutum* engineering. For its potential as a molecular tool, it will be necessary to investigate the correlation between the sequence of the 2A peptides, their cleavage efficiency, the position effect of the proteins in the expression cassette, and the activities of the proteins of interest.

3.5. The sequence of the expression cassette has an impact on the percentage of fluorescent transconjugant obtained by bacterial conjugation

Based on the sequencing results from the rescued episomes and flow cytometry, we observed three contexts in which the reporter protein eGFP was detected in the DMi7 transconjugants. First, there was a partial or complete deletion of the *fcs* sequence in DMi7-21-1, DMi7-21-3, DMi7-31-3, and DMi7-31-8. Second, there was a change in the ORF affecting the *fcs* sequence in DMi7-31-1. Third, there was a non-synonymous substitution in the *fcs* sequence of DMi7-31-4 that could have an impact on the structure of the protein. In all these scenarios, the eGFP-T2A-ech-fcs protein could not be produced unaltered (Fig. 2b and Supplementary table 3).

There still could be a possibility that those modifications in the *fcs* sequence and the rearrangement in the episomes could have removed toxic elements or parts that inhibited *P. tricornutum* translation of the unaltered *eGFP-T2A-ech-fcs* gene sequence. These modifications could also be consequences of a random phenomenon. To determine what could be the impact of rearrangements, episomal DNA was extracted and recovered from *P. tricornutum* clones and then transformed into *E. coli* for a new round of bacterial conjugation into wild type strain of *P. tricornutum*. The episomes rescued from non-fluorescent DMi7-21-14, DMi7-21-15, and DMi7-31-10 transconjugants were successfully transformed in *P. tricornutum* and led to frequencies of 13.1 %, 12.6 %, and 15.0 % of fluorescent colonies, respectively, which are similar to the results obtained with the conjugation of the original pDMi7 (Fig. 1e and Table 1). In these rescued episomes, the coding sequence of the expression cassette did not contain any mutation or rearrangement (Fig. 2c and Supplementary Table S3). From this point of view, there is no difference between transforming *P. tricornutum* with those three episomes or with the original episome pDMi7. The eGFP fluorescence of the transconjugants from these rescued episomes is then possibly originating from mutated or rearranged episomes, like the transconjugants from the conjugation with the episome pDMi7. Different results were obtained with the episomes rescued from the two remaining fluorescent transconjugants. Episome originating from the fluorescent clone DMi7-31-1 contained a frameshift in the ORF of the *fcs* sequence (Fig. 2b and Supplementary Table S3). Once it was reintroduced into *P. tricornutum*, 84.1 % of the colonies were detected as fluorescent using fluorescence microscopy (Table 1). This percentage of fluorescent colonies is close to what was obtained from the conjugation with the episome pControl where the *ech-fcs* sequence was replaced by *mCherry* (Fig. 1b, d). Regarding the episome from the clone DMi7-21-3, the number of colonies was lower than what was obtained with the others rescued episomes (Supplementary Fig. S8) with 46 % of fluorescent colonies (Table 1). Based on the sequencing of the episome recovered from the clone DMi7-21-3, the bacterial conjugation should not have been possible as this episome contains only two partial duplicates of the OriT cassette which is necessary to activate the transfer of the episome from *E. coli* to *P. tricornutum* [26]. The first partial copy of the OriT cassette contains the first 131 bp and the second one has the first 158 bp on a total length of 771 bp. It seems that those two partial sequences could be sufficient to activate the conjugation, but at a lower efficiency. As expected, no colonies were obtained from the conjugation of the episomes recovered from the transconjugants DMi7-21-1, DMi7-21-16, DMi7-31-3, DMi7-31-4, and DMi7-31-8. Following the sequencing analysis of these episomes (Supplementary Table S3), it revealed the complete or partial deletion of the OriT sequence.

Table 1
Count of fluorescent colonies using fluorescence microscopy 14 days after the bacterial conjugation with the episomes recovered from the *P. tricornutum* cells initially transformed with pDMi7.

Subclones of	GFP fluorescent colonies (%)
DMi7-21-3 Replicate #1	46.4 (13/28)
DMi7-21-14	13.1 (47/358)
DMi7-21-15	12.6 (42/333)
DMi7-31-1	84.1 (334/397)
DMi7-31-2	14.3 (36/252)
DMi7-31-10	15.0 (53/354)

4. Conclusion

To engineer *P. tricornutum* to produce enzymes linked to vanillin biosynthesis, we designed a bicistronic expression cassette containing a T2A peptide into an episomal system. We observed rearrangement of the episomal cassette in *P. tricornutum*. Based on western blot results and the sequencing of the episomes, it appears that the sequence downstream of the T2A peptide could have an impact on its cleavage efficiency. Future investigations focusing on 2A peptides in *P. tricornutum* should elucidate the links between the selected 2A peptides, their cleavage efficiency, the genes order in the expression cassette, and their impacts on the proteins produced downstream or upstream the 2A peptide.

The results presented here introduce the possibilities of mechanisms used by *P. tricornutum* to prevent the production of a heterologous protein of interest from extrachromosomal expression systems. Indeed, the bacterial conjugation of *P. tricornutum* with an episome harboring the expression cassette for the polyprotein eGFP-T2A-mCherry resulted in 82.1 % of colonies producing the eGFP fluorescence. A significant reduction to 8.1 % of fluorescent colonies was observed by replacing the *mCherry* sequence by the coding sequences of the enzymes ech and fcs from *Streptomyces* sp. strain V-1.

We herein demonstrated that a screening that is not specific to the production of the proteins or the metabolites of interest can lead to the selection of transconjugants that contain a rearranged or mutated episome at the level of the expression cassette. A screening could have been based on the sequence of the episome. However, it would most likely have selected transconjugants where the reporter protein eGFP and the proteins of interest ech and fcs are not produced.

We successfully enriched cultures of *P. tricornutum* by fluorescence-

activated cell sorting by flow cytometry. Following the enrichment, it was possible to increase the protein production of a clone without optimizing the culture conditions. Further investigation will be necessary to understand the dynamic of the cell population of the *P. tricornutum* transconjugants. The observations by fluorescence microscopy showed that some colonies are almost completely fluorescent, and some others are only partially fluorescent. We could not confirm the reason for these differences in fluorescence patterns and what could be the impact on the cultures made from these colonies.

It is the first time that rearrangement of episomes in *P. tricornutum* is investigated. We demonstrated that the episome can be unstable as a molecular in in this microalgae species. Nevertheless, recent publications by other groups presented successful uses of this extrachromosomal expression system. This article should therefore be interpreted as a warning about the limits of the episome in *P. tricornutum* and the importance of the construction strategy. It can be concluded that the screening strategy used in this paper led to the selection of transconjugants that did not contain the designed episome.

Accession numbers

fcs (GenBank accession # KC847405.1)
ech (GenBank accession # KC847406.1)

CRediT authorship contribution statement

Andrew Diamond: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Visualization, Writing- Original draft preparation and Writing- Reviewing and Editing. **Aracely Maribel Diaz-Garza:** Conceptualization, Methodology, Software, Visualization and Writing- Reviewing. **Jessica Li:** Methodology, Validation and Writing- Reviewing. **Samuel S. Slattery:** Methodology, Validation and Writing- Reviewing. **Natacha Merindol:** Methodology, Investigation, Validation, Visualization and Writing- Reviewing. **Elisa Fantino:** Methodology, Investigation, Validation and Writing- Reviewing. **Fatma Meddeb-Mouelhi:** Resources, Project administration and Writing-Reviewing. **Bogumil J. Karas:** Conceptualization, Resources, Supervision and Writing- Reviewing. **Simon Barnabé:** Resources, Supervision and Writing- Reviewing. **Isabel Desgagné-Penix:** Conceptualization, Visualization, Writing- Original draft, Project administration, Funding acquisition, Supervision and Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgements

We acknowledge that financial support for this study was funded by the Canada Research Chair on plant specialized metabolism Award No 950-232164 to I.D-P. Thanks are extended to the Canadian taxpayers and to the Canadian government for supporting the Canada Research Chairs Program. Additional support in the form of scholarship to A.D. was provided by the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program from the Natural Sciences and Engineering Research Council of Canada. A.M.D-G., and E.F. were supported by Mitacs—Acceleration program grants no IT12310 and IT16463 to I. D-P.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.algal.2023.102998.

References

[1] T. Butler, R.V. Kapoore, S. Vaidyanathan, Phaeodactylum tricornutum: a diatom cell factory, Trends Biotechnol. (2020), https://doi.org/10.1016/j.tibtech.2019.12.023.
[2] F. Hempel, U.G. Maier, An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency, Microb. Cell Factories 11 (2012) 126, https://doi.org/10.1186/1475-2859-11-126.
[3] Y. Taparia, et al., A novel endogenous selection marker for the diatom Phaeodactylum tricornutum based on a unique mutation in phytoene desaturase 1, Sci. Rep. 9 (1) (2019) 8217, https://doi.org/10.1038/s41598-019-44710-5.
[4] F. Hempel, et al., From hybridomas to a robust microalgal-based production platform: molecular design of a diatom secreting monoclonal antibodies directed against the Marburg virus nucleoprotein, Microb. Cell Factories 16 (1) (2017) 131, https://doi.org/10.1186/s12934-017-0745-2.
[5] A. Pudney, et al., Multifunctionalizing the marine diatom Phaeodactylum tricornutum for sustainable co-production of omega-3 long chain polyunsaturated fatty acids and recombinant phytase, Sci. Rep. 9 (1) (2019) 11444, https://doi.org/10.1038/s41598-019-47875-1.
[6] C. Bowler, et al., The phaeodactylum genome reveals the evolutionary history of diatom genomes, Nature 456 (7219) (2008) 239–244, https://doi.org/10.1038/nature07410.
[7] G.V. Filloramo, et al., Re-examination of two diatom reference genomes using long-read sequencing, BMC Genomics 22 (1) (2021) 379, https://doi.org/10.1186/s12864-021-07666-3.
[8] D.J. Giguere, et al., Telomere-to-telomere genome assembly of Phaeodactylum tricornutum, PeerJ 10 (2022), e13607, https://doi.org/10.7717/peerj.13607.
[9] F. Daboussi, et al., Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology, Nat. Commun. 5 (2014) 3831, https://doi.org/10.1038/ncomms4831.
[10] S.S. Slattery, et al., An expanded plasmid-based genetic toolbox enables Cas9 genome editing and stable maintenance of synthetic pathways in Phaeodactylum tricornutum, ACS Synth. Biol. 7 (2) (2018) 328–338, https://doi.org/10.1021/acssynbio.7b00191.
[11] M.L. Hamilton, et al., Metabolic engineering of Phaeodactylum tricornutum for the enhanced accumulation of omega-3 long chain polyunsaturated fatty acids, Metab. Eng. 22 (100) (2014) 3–9, https://doi.org/10.1016/j.ymben.2013.12.003.
[12] S. D'Adamo, et al., Engineering the unicellular alga Phaeodactylum tricornutum for high-value plant triterpenoid production, Plant Biotechnol. J. 17 (1) (2019) 75–87, https://doi.org/10.1111/pbi.12948.
[13] B.J. Karas, et al., Designer diatom episomes delivered by bacterial conjugation, Nat. Commun. 6 (2015) 6925, https://doi.org/10.1038/ncomms7925.
[14] R.E. Diner, et al., Refinement of the diatom episome maintenance sequence and improvement of conjugation-based DNA delivery methods, Front. Bioeng. Biotechnol. 4 (2016) 65, https://doi.org/10.3389/fbioe.2016.00065.
[15] J. George, et al., Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology, Front. Bioeng. Biotechnol. 8 (2020) 513, https://doi.org/10.3389/fbioe.2020.00513.
[16] M. Fabris, et al., Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids, ACS Synth. Biol. 9 (3) (2020) 598–612, https://doi.org/10.1021/acssynbio.9b00455.
[17] N.J. Gallage, et al., Vanillin formation from ferulic acid in Vanilla planifolia is catalysed by a single enzyme, Nat. Commun. 5 (2014) 4037, https://doi.org/10.1038/ncomms5037.
[18] W. Yang, et al., Characterization of two streptomyces enzymes that convert ferulic acid to vanillin, PLoS ONE 8 (6) (2013), e67339, https://doi.org/10.1371/journal.pone.0067339.
[19] J. Ni, et al., Mimicking a natural pathway for de novo biosynthesis: natural vanillin production from accessible carbon sources, Sci. Rep. 5 (2015) 13670, https://doi.org/10.1038/srep13670.
[20] H. Yang, et al., A re-evaluation of the final step of vanillin biosynthesis in the orchid Vanilla planifolia, Phytochemistry 139 (2017) 33–46, https://doi.org/10.1016/j.phytochem.2017.04.003.
[21] M.A. Moosburner, et al., Multiplexed knockouts in the model diatom phaeodactylum by episomal delivery of a selectable Cas9, Front. Microbiol. (2020) 11, https://doi.org/10.3389/fmicb.2020.00005.
[22] M.T. Russo, et al., Assessment of genomic changes in a CRISPR/Cas9 Phaeodactylum tricornutum mutant through whole genome resequencing, PeerJ 6 (2018), e5507, https://doi.org/10.7717/peerj.5507.
[23] S.S. Slattery, et al., Plasmid-based complementation of large deletions in Phaeodactylum tricornutum biosynthetic genes generated by Cas9 editing, Sci. Rep. 10 (1) (2020) 13879, https://doi.org/10.1038/s41598-020-70769-6.
[24] A.K. Sharma, et al., Transgene-free genome editing in marine algae by bacterial conjugation - comparison with biolistic CRISPR/Cas9 transformation, Sci. Rep. 8 (1) (2018) 14401, https://doi.org/10.1038/s41598-018-32342-0.
[25] J. Athey, et al., A new and updated resource for codon usage tables, BMC Bioinformatics 18 (1) (2017) 391, https://doi.org/10.1186/s12859-017-1793-7.
[26] T.A. Strand, et al., A new and improved host-independent plasmid system for RK2-based conjugal transfer, PLoS One 9 (3) (2014), e90372, https://doi.org/10.1371/journal.pone.0090372.

[27] M. Lescot, et al., PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences, Nucleic Acids Res. 30 (1) (2002) 325–327, https://doi.org/10.1093/nar/30.1.325.
[28] T. Kadono, et al., Characterization of marine diatom-infecting virus promoters in the model diatom Phaeodactylum tricornutum, Sci. Rep. 5 (2015) 18708, https://doi.org/10.1038/srep18708.
[29] P.R. Jones, Genetic instability in cyanobacteria – an elephant in the room? Front. Bioeng. Biotechnol. (2014) 2, https://doi.org/10.3389/fbioe.2014.00012.
[30] V. De Riso, et al., Gene silencing in the marine diatom Phaeodactylum tricornutum, Nucleic Acids Res. 37 (14) (2009), e96, https://doi.org/10.1093/nar/gkp448.
[31] A. Huang, L. He, G. Wang, Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis, BMC Genomics 12 (2011) 337, https://doi.org/10.1186/1471-2164-12-337.
[32] Z. Yu, et al., Droplet-based microfluidic screening and sorting of microalgal populations for strain engineering applications, Algal Res 56 (2021), https://doi.org/10.1016/j.algal.2021.102293. None.
[33] J.H. Kim, et al., High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice, PLoS One 6 (4) (2011), e18556, https://doi.org/10.1371/journal.pone.0018556.
[34] E. Minskaia, M.D. Ryan, Protein coexpression using FMDV 2A: effect of "linker" residues, Biomed. Res. Int. 2013 (2013), 291730, https://doi.org/10.1155/2013/291730.
[35] T.M. Souza-Moreira, et al., Screening of 2A peptides for polycistronic gene expression in yeast, FEMS Yeast Res. 18 (5) (2018), https://doi.org/10.1093/femsyr/foy036.
[36] Z. Liu, et al., Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector, Sci. Rep. 7 (1) (2017) 2193, https://doi.org/10.1038/s41598-017-02460-2.
[37] R.W. Daniels, et al., Expression of multiple transgenes from a single construct using viral 2A peptides in drosophila, PLoS One 9 (6) (2014), e100637, https://doi.org/10.1371/journal.pone.0100637.
[38] J. Chng, et al., Cleavage efficient 2A peptides for high level monoclonal antibody expression in CHO cells, MAbs 7 (2) (2015) 403–412, https://doi.org/10.1080/19420862.2015.1008351.
[39] Y. Wang, et al., 2A self-cleaving peptide-based multi-gene expression system in the silkworm Bombyx mori, Sci. Rep. 5 (2015) 16273, https://doi.org/10.1038/srep16273.
[40] G. Defrel, et al., Identification of loci enabling stable and high-level heterologous gene expression, Front. Bioeng. Biotechnol. 9 (2021), 734902, https://doi.org/10.3389/fbioe.2021.734902.

International Journal of
Molecular Sciences

MDPI

Article

Bioengineering of the Marine Diatom *Phaeodactylum tricornutum* with Cannabis Genes Enables the Production of the Cannabinoid Precursor, Olivetolic Acid

Fatima Awwad [1,†], Elisa Ines Fantino [1,†], Marianne Héneault [1], Aracely Maribel Diaz-Garza [1], Natacha Merindol [1,2], Alexandre Custeau [1], Sarah-Eve Gélinas [1], Fatma Meddeb-Mouelhi [1,2], Jessica Li [3], Jean-François Lemay [4], Bogumil J. Karas [3] and Isabel Desgagne-Penix [1,2,*]

[1] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, 3351 Boulevard des Forges, Trois-Riviere, QC G9A 5H7, Canada
[2] Groupe de Recherche en Biologie Végétale, Université du Québec à Trois-Rivières, Trois-Riviere, QC G9A 5H7, Canada
[3] Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, ON N6A 5C1, Canada
[4] Centre National en Électrochimie et en Technologies Environnementales Inc., 2263 Avenue du Collège, Shawinigan, QC G9N 6V8, Canada
* Correspondence: isabel.desgagne-penix@uqtr.ca
† These authors contributed equally to this work.

Abstract: The increasing demand for novel natural compounds has prompted the exploration of innovative approaches in bioengineering. This study investigates the bioengineering potential of the marine diatom *Phaeodactylum tricornutum* through the introduction of cannabis genes, specifically, tetraketide synthase (TKS), and olivetolic acid cyclase (OAC), for the production of the cannabinoid precursor, olivetolic acid (OA). *P. tricornutum* is a promising biotechnological platform due to its fast growth rate, amenability to genetic manipulation, and ability to produce valuable compounds. Through genetic engineering techniques, we successfully integrated the cannabis genes *TKS* and *OAC* into the diatom. *P. tricornutum* transconjugants expressing these genes showed the production of the recombinant TKS and OAC enzymes, detected via Western blot analysis, and the production of cannabinoids precursor (OA) detected using the HPLC/UV spectrum when compared to the wild-type strain. Quantitative analysis revealed significant olivetolic acid accumulation (0.6–2.6 mg/L), demonstrating the successful integration and functionality of the heterologous genes. Furthermore, the introduction of TKS and OAC genes led to the synthesis of novel molecules, potentially expanding the repertoire of bioactive compounds accessible through diatom-based biotechnology. This study demonstrates the successful bioengineering of *P. tricornutum* with cannabis genes, enabling the production of OA as a precursor for cannabinoid production and the synthesis of novel molecules with potential pharmaceutical applications.

Keywords: diatom; metabolic engineering; olivetolic acid cyclase; tetraketide synthase; synthetic biology; cannabinoids; microalgae

Citation: Awwad, F.; Fantino, E.I.; Héneault, M.; Diaz-Garza, A.M.; Merindol, N.; Custeau, A.; Gélinas, S.-E.; Meddeb-Mouelhi, F.; Li, J.; Lemay, J.-F.; et al. Bioengineering of the Marine Diatom *Phaeodactylum tricornutum* with Cannabis Genes Enables the Production of the Cannabinoid Precursor, Olivetolic Acid. *Int. J. Mol. Sci.* **2023**, *24*, 16624. https://doi.org/10.3390/ijms242316624

Academic Editors: Chiara Lauritano and Assunta Saide

Received: 2 September 2023
Revised: 19 November 2023
Accepted: 20 November 2023
Published: 22 November 2023

1. Introduction

Over a thousand of the molecules produced in *Cannabis sativa* plants have been reported in the literature, including more than a hundred cannabinoids (CBs) [1]. The antinociceptive and appetite-stimulating properties of CBs have been studied thoroughly, as well as their association with relieving the symptoms associated with different diseases such as multiple sclerosis and the relief of chemotherapy side effects [2–5]. The most studied CBs are Δ9-tetrahydrocannabinol (THC), which has psychoactive potential, and cannabidiol (CBD), which acts on human endocannabinoid receptors to modulate pain and is already contained in many pharmaceutical products available on the market, such

as Sativex and Bediol [6–8]. Other therapeutic properties of specific CBs that could be administrated in a purified form are still under investigation and offer a promising alternative for different human pathology applications [5]. In planta, phytocannabinoids are produced in glandular trichomes via fatty acid and terpenoid precursors [9]. The genes and enzymes implicated in the cannabinoid pathway were characterized by several plant biologists shortly after the sequencing of the *C. sativa* genome [9–12]. The first step in the CB biosynthetic pathway is the formation of olivetolic acid (OA) from malonyl-CoA and hexanoyl-CoA, through a two-step reaction performed by a type-III polyketide synthase named tetraketide synthase (*Cs*TKS) followed by a cyclase named olivetolic acid cyclase (*Cs*OAC), both of which are localized in the cytosol or cells from the glandular trichomes (Figure 1). The enzymes belonging to the type-III polyketide synthase (PKS) superfamily are responsible for producing a wide array of central frameworks that are found in specialized plant metabolites of significant medicinal value. These metabolites encompass various categories, such as flavonoids, stilbenes, chromones, pyrones, phloroglucinols, resorcinols, xanthones, acridones, and quinolones [13,14]. Examples of PKS include chalcone synthase, which accepts *p*-coumaroyl-CoA as the starter substrate to catalyze three successive condensations with malonyl-CoA to generate naringenin chalcone [15,16] and stilbene synthase, which generates resveratrol [17]. The polyketide synthase from *C. sativa* (also named tetraketide synthase or olivetol synthase) was characterized and has been shown to be dependent on OAC for the formation of OA (Figure 1) [9]. The 3D structure of *Cs*OAC is most accurately described as a dimeric α + β barrel (DABB) protein, with similarities to other plant DABBs, especially those implicated in stress responses and to the DABB proteins found in *Streptomyces* bacteria [9]. The aromatic precursor OA is a resorcinol carrying an alkyl chain (named alkylresorcinolic acid), which is condensed with a monoterpenoid moiety (geranyl diphosphate, GPP) to yield the general structure of the most well-known CBs, as reported in feeding studies of 13C-labeled glucose in the early 1960s [18,19]. Although not all CBs are derived from the OA precursor, it remains the most widely characterized pathway in *C. sativa* [20]. The prenyl moiety of GPP is transferred to OA by an aromatic prenyltransferase (*Cs*APT) to form cannabigerolic acid (CBGA) (Figure 1). This core intermediate is the backbone of the remaining reaction, which diverges to form cannabidiolic acid (CBDA) and tetrahydrocannabinolic acid (THCA) through CBDA-synthase (CBDAS) and THCA-synthase (THCAS), respectively (Figure 1), as well as other CBs such as cannabichromenic acid (CBCA) via cannabichromenic acid cyclase (CBCAS) [9,11,21,22]. The neutral phytocannabinoids THC and CBD are generated from cannabinolic acids by non-enzymatic decarboxylation and are often produced upon exposure to light or heat (Figure 1) [21]. Aside from CBs, the cannabis plants also produce other specialized metabolites from shared precursors with cannabinoids, such as fatty acids, competing with olivetolic acid synthesis, and terpenoids such as limonene [23], competing for the GPP pool [24,25].

The yield of CB in plants is limited, both in quality and in quantity; the maximum production is due to a mixture of these metabolites accumulating in the flowers (20 mg/g fresh weight) [6,26]. In addition, the lack of scientific literature on cannabis physiology leads to the absence of standardized culturing protocols, resulting in a variation in CB profiles, quality, and quantity. In addition, cannabis pathogens such as the gray mold causal agent (*Botrytis cinerea*) affect the overall plant yield and metabolite composition [27]. As well as multiple unknown facets of CB production in planta, this suggests that the implementation of a heterologous expression system with fewer components and fewer competing enzymes might be a solution for achieving a higher production yield of single CB components, such as CBD or THC. Therefore, synthetic biology offers an important alternative method to produce CBs for the pharmaceutical field with less risk and fewer crop management challenges.

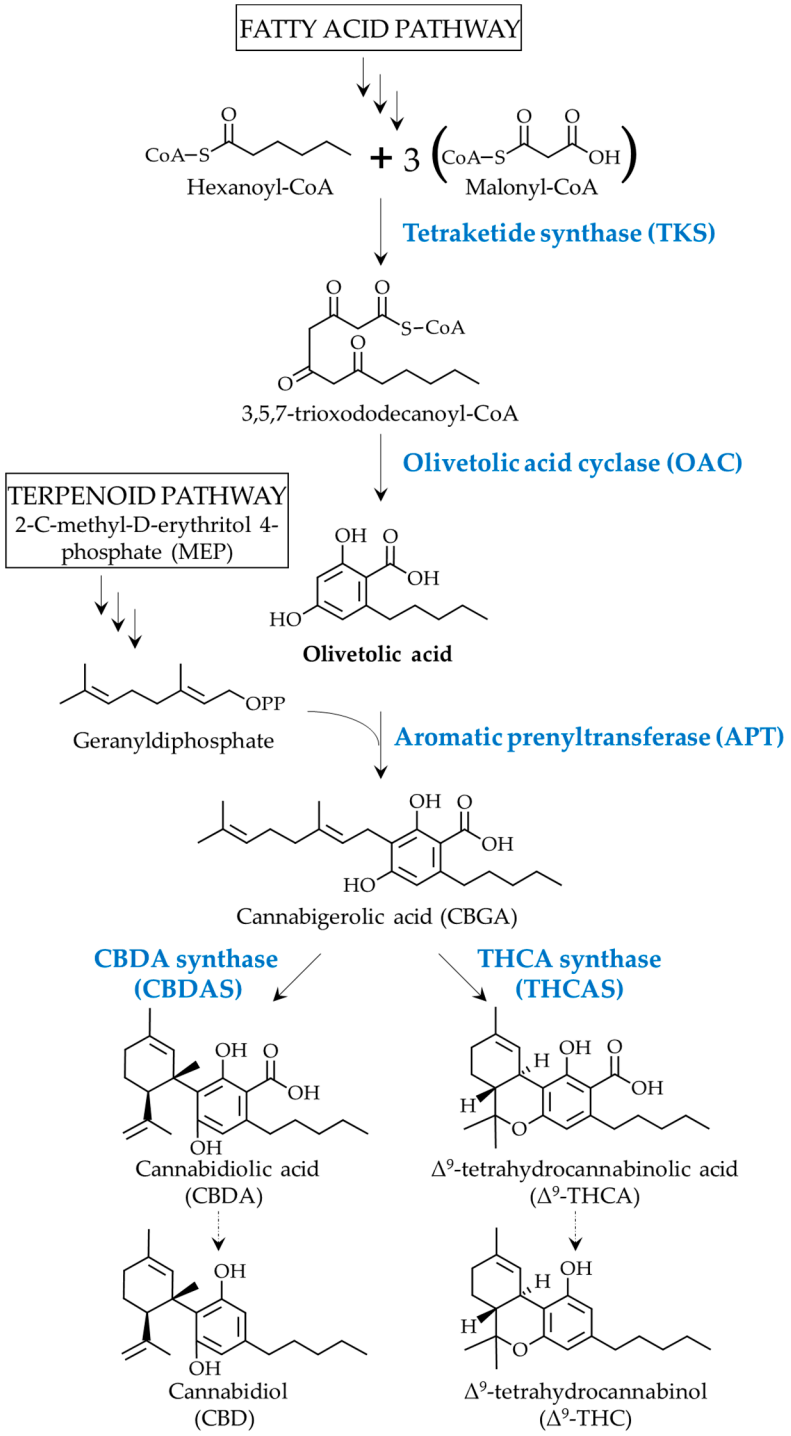


Figure 1. The proposed cannabinoid biosynthetic pathway, leading to the two major phytocannabinoids, THC and CBD. Enzymes are shown in blue. Bolded arrows represent an enzymatic reaction, whereas the dotted arrows represent non-enzymatic (heat or light) decarboxylation.

Several research groups aimed to design de novo or improve pre-existing cannabinoid production in plant tissue cultures, but apart from the transformation rates, in vitro culture viability and yield are still to be optimized [6]. For instance, few groups succeeded in transforming yeast (*Saccharomyces cerevisiae*) and bacteria (*Escherichia coli*) to produce the different metabolites of the CB pathway, such as OA or CBGA [28,29]. The production of these metabolites in yeast and bacteria often involves the addition of costly precursors such as acetyl-CoA, hexanoyl-CoA, or olivetolic acid [28,30], or involves supplementing the

media with sugars [31], such as galactose, to provide a more affordable alternative with a long series of gene stacking. For instance, the production of OA in *Dyctiostelium discoideum* on a 300-L scale led to 4.8 µg/L [32]. OA production in *E. coli* reached 80 mg/L after introducing a series of modifications to increase the CoAs pool [33]. However, the bacterial and yeast translation and post-translational mechanisms, as well as their metabolic chassis, are distinct from the mechanisms found in higher plants. This resulted in low yields, such as picomoles per unit of optical density, from these microorganisms when compared to the bioengineering and supplementation costs [31,34]. By contrast, diatoms are photosynthetic organisms that share the main metabolic paths with plants, making them promising candidates for the production of heterologous compounds [9]. All the precursor pathways, such as acetyl-CoA, hexanoyl-CoA, GPP, etc., are present in diatom cells at different times of the cell cycle, which reduces the need to supplement the culture media with costly precursors or sugars [35]. Also, the extrachromosomal transformation of diatoms via bacterial conjugation allows the fairly stable expression of heterologous genes [36,37]. Similar advances in the bioengineering of diatoms have been shown recently, allowing the production of monoterpenoids such as geraniol, an insect repellant, in the model diatom *Phaeodactylum tricornutum* [36].

Thus, the aim of this study is to produce the CB precursor OA in the marine microalga, *P. tricornutum*. Here, we inserted cannabis *Cs*TKS and *Cs*OAC genes into a stable optimized episome that was assembled in *S. cerevisiae* [38,39]. The recombinant episome was then transformed via a trans-kingdom conjugation between *E. coli* and *P. tricornutum* [38]. The transconjugant strains were screened for their ability to produce molecules of interest, using a high-performance liquid chromatograph coupled to an ultra-violet detector (HPLC-UV). We detected OA (0.6–2.6 mg/L) in *P. tricornutum* transconjugants that issued from two different expression cassettes. Although the production of the OA compounds of interest was temporary, this work sheds light on a powerful and suitable system for plant metabolite heterologous production in a cost-effective phototropic system employing *P. tricornutum*.

2. Results

2.1. Heterologous Protein Expression and Localization in the Diatom P. tricornutum

2.1.1. Transformation Validation and Transconjugant Characterization

To inform and design the metabolic engineering strategy, we first performed computational analyses to assess the availability of the lipid-derived precursors required to produce OA and cannabinoids in *P. tricornutum* with in silico tools and showed that upstream pathways existed in the host organism. According to the Biocyc database [40], hexanoyl-CoA can be formed in *P. tricornutum* in the peroxisome via the fatty acid β-oxidation pathway, mainly by long-chain acyl-CoA synthase (*Pt*ACS3 and *Pt*ACS4), while malonyl-CoA is formed via the activity of an ACYL-CoA carboxylase (ACCase). Moreover, the geranyl pyrophosphate (GPP) pool is present at detectable levels in *P. tricornutum* [36].

To produce OA in *P. tricornutum*, three different expression cassettes were designed using p*Pt*GE30 as a backbone plasmid DNA (Figure 2A), namely, *Pt*OA1, *Pt*OA2, and *Pt*OA3, with various combinations of *C. sativa* genes encoding for *Cs*TKS and *Cs*OAC (Figure 2B). Each expression cassette was codon-optimized for *P. tricornutum* and assembled (Appendix A, Figure A1, Tables A1–A3) according to the HIVE lab codon usage table (CUT) [41].

The transconjugant strains *Pt*OA1, *Pt*OA2, and *Pt*OA3 did not present significant differences in growth and shape compared to *P. tricornutum* when transformed with the p*Pt*GE30 empty vector or wild-type strains (Figure 3A–C). In addition, morphological properties such as morphotype, size, granulosity, and chlorophyll levels did not significantly differ between the studied strains, as evaluated by microscopy (Figure 3A) and flow cytometry (Figure 4).

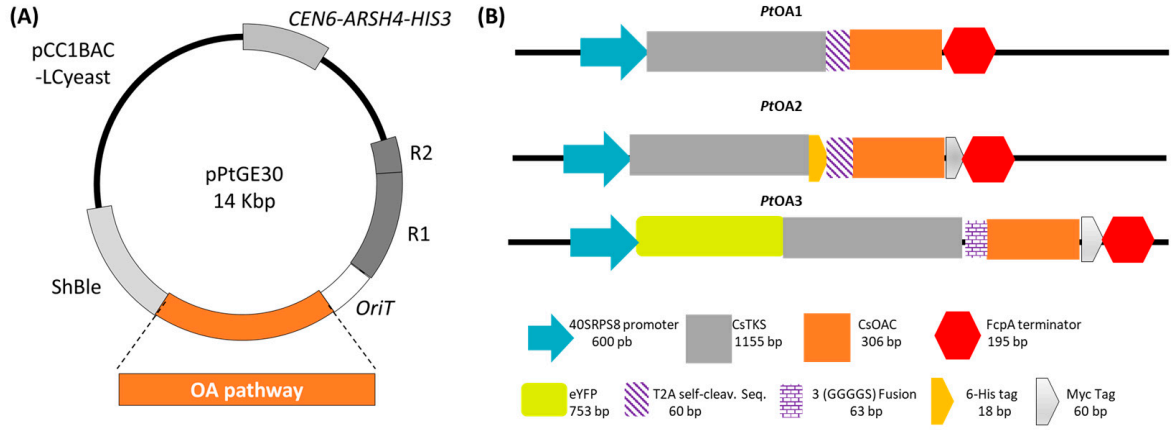


Figure 2. Design of the expression vectors *Pt*OA1, *Pt*OA2, and *Pt*OA3. (**A**) Schematic representation of the p*Pt*GE30 expression vector. (**B**) Description of the recombinant cassettes used to express *C. sativa* polyketide synthase (*Cs*TKS) and olivetolic acid cyclase (*Cs*OAC) under *P. tricornutum*'s constitutive promoter 40SRPS8, the fucoxanthin–chlorophyll binding protein A (FcpA) terminator, with or without tags (His = histidine, eYFP = enhanced yellow fluorescent protein, Seq = sequence, T2A = the *Thosea asigna* virus 2A cleavable sequence), and reporter genes of different sizes (base pair = bp).

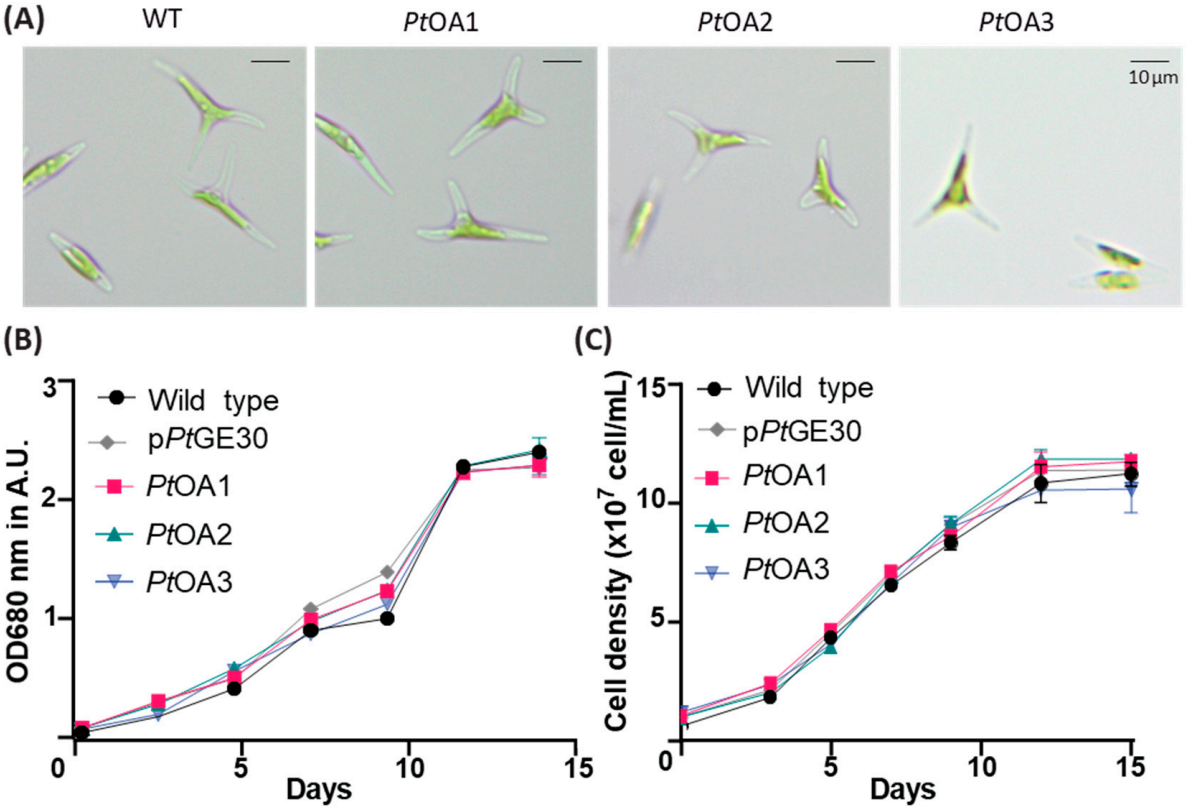


Figure 3. Cell shapes and growth curves of the transformed clones in comparison with wild-type *P. tricornutum*. (**A**) Optical microscope images of wild-type (WT) *P. tricornutum* and the transformed cells *Pt*OA1, *Pt*OA2, and *Pt*OA3. (**B**,**C**) *P. tricornutum* WT, transformed with pPtGE30, *Pt*OA1, *Pt*OA2, and *Pt*OA3 cell culture growth curves, followed by absorbance at 680 nm (**B**) or via cell count (**C**) during the 15 days of the cycle. The plotted values represent the means of three replicates and error bars represent the standard deviation.

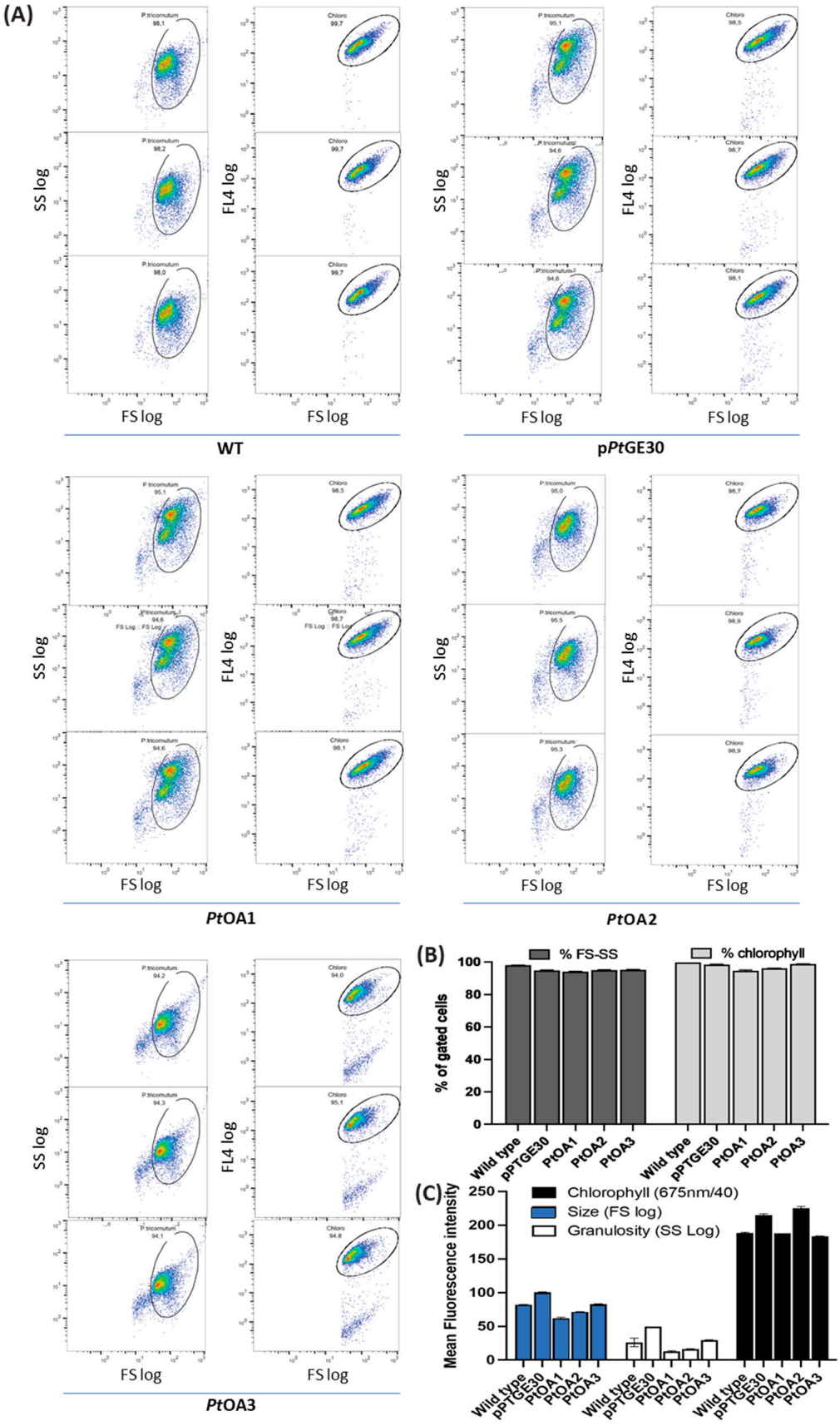


Figure 4. Flow cytometry observation for cell size and chlorophyll auto-fluorescence among *P. tricornutum* transconjugants and wild-type cultures. (**A**) Scatter plots for each strain and (**B**) a graphical

representation of chlorophyll and size percentages among gated cells of wild-type *P. tricornutum* (WT), and *P. tricornutum*, harboring the empty vector p*Pt*GE30, *Pt*OA1, *Pt*OA2, and *Pt*OA3. (**C**) Histogram representing the mean chlorophyll fluorescence intensity, mean granulosity, and mean size scatter.

2.1.2. Heterologous Protein Detection and Localization

The successful expression of cannabis genes and the accumulation of the corresponding enzymes was observed in *P. tricornutum* transconjugants. Indeed, the accumulation of the *Cs*TKS-T2A (45 kDa band) protein in *Pt*OA1C1 (Figure 5A), a 68 kDa uncleaved protein corresponding to CsTKS-His-T2A-CsOAC-cMyc in *Pt*OA2C2 (Figure 5B), and an 85 kDa band corresponding to YFP-CsTKS-3(GGGGS)-CsOAC-cMyc in *Pt*OA3C1 (Figure 5C) was detected via Western blot analysis in PCR-positive *P. tricornutum* transconjugants (Appendix A, Figure A1B). However, the construction with the self-cleavable sequence did not yield detectable successful single proteins in *Pt*OA2. This could be caused by the tag (6xHis) used in the OA2 construct but not in the OA1 construct, changing the efficiency of cleavage (Figure 5B). Then, we studied the localization of *Cs*TKS in *PtOA*3, taking advantage of the presence of the fluorescent YFP fusion protein. We observed under confocal microscopy that the heterologous recombinant protein (YFP-*Cs*TKS-3(GGGGS)-CsOAC-cMyc) produced in *PtOA*3 accumulated and was localized to the cytosol (Figure 5D).

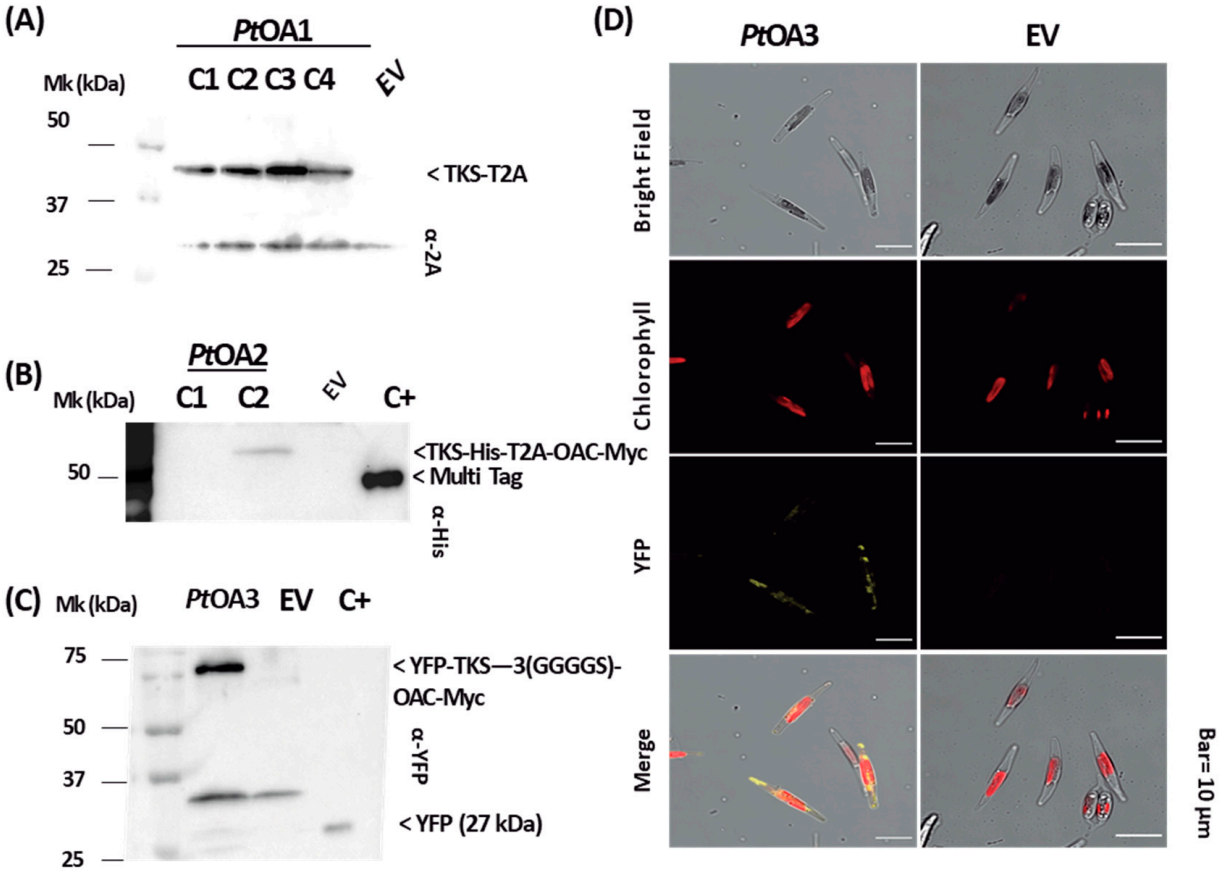


Figure 5. Expression and localization of heterologous proteins in transconjugant *P. tricornutum* strains. (**A**) Whole-cell extraction of four *Pt*OA1 (TKS-T2A-OAC) strains 1–4; TKS protein detection (45 kDa) was performed using anti-T2A antibodies. (**B**) Whole-cell extraction of two *Pt*OA2 (TKS-His-T2A-OAC-cMyc) clones 1 and 2; TKS:T2A:OAC protein detection (68 kDa) was performed using anti-His antibodies. (**C**) Whole-cell extraction of *Pt*OA3 (YFP-TKS-3(GGGGS)-OAC-cMyc) strains; YFP:TKS:OAC protein detection (85 kDa) was performed using anti-YFP antibodies. Mk: protein ladder; whole-cell extracts of the pPtGE30 culture were used as negative control (empty vector, EV);

C+: Multi Tag was used as a positive control. All Western blots were obtained from 12% SDS-PAGE gel. For each construct, four clones were characterized and the ones showing protein production are presented in this figure. (**D**) Confocal microscopic images of *Pt*OA3 and pPtGE30 cells in a bright field, with autofluorescence of the chloroplast, YFP:TKS:OAC fluorescence, and the merging of three fields are shown.

2.2. Temporal Metabolite Production

2.2.1. Metabolite Detection and Identification

Selected positive transconjugant strains, displaying protein accumulation via Western blot analysis, were cultivated, extracted, and used for targeted metabolite analysis. OA was detected at 2.6 mg/kg in *Pt*OA1 and 0.54 mg/kg in *Pt*OA2, using HPLC-UV, whereas no peaks at the OA retention time were detected in the *Pt*OA3 extracts (Figure 6 and Appendix A, Figure A3).

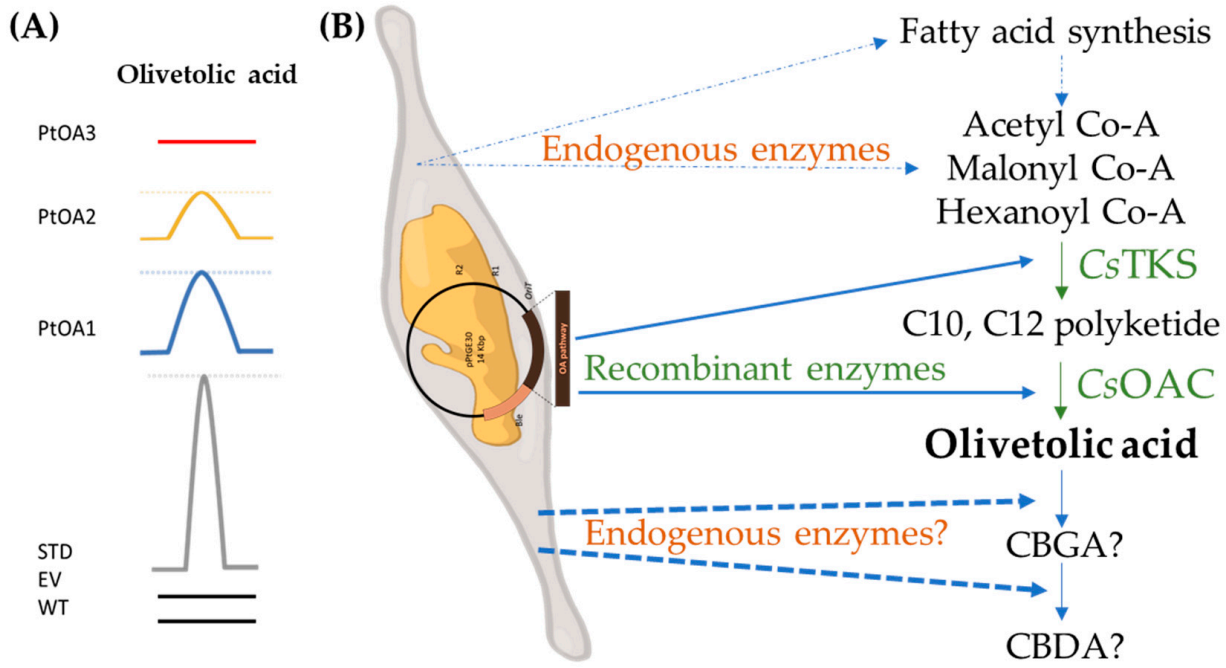


Figure 6. Olivetolic acid production, obtained via HPLC-UV analysis at 280 nm. (**A**) Curves represent the peak at the OA retention time of chromatograms, obtained following the injection of 10 µL of each extracted sample. At the bottom, a concentration of 10 µg/mL of commercial standards (STD) is shown: metabolite extraction from the *P. tricornutum* wild type (WT); *P. tricornutum,* transformed with an empty vector (EV = p*Pt*GE30), *Pt*OA1, *Pt*OA2, and *Pt*OA3. The peaks represent olivetolic acid. (**B**) Schematic representation of the bioengineered *P. tricornutum* for the production of cannabinoid precursors.

Interestingly, additional peaks with similar retention times to the cannabinoid standards of CBN, CBD, and THC, and their corresponding acid forms were detected in the extracts of OA-producing transconjugants (*Pt*OA1C1 and *Pt*OA2C2) (Appendix A, Figure A4). Liquid chromatography–mass spectrometry (LC-MS) analyses using electrospray ionization in the positive mode (ESI+) were performed to validate if the mass-to-charge ratio (m/z) of these peaks corresponded to protonated CBs $(M + H)^+$. The compounds were first scanned for specific masses corresponding to the metabolites of interest, then, in the case of a positive scan, another LC-MS analysis was performed to make sure that the fragments corresponded to the target compound. Unexpectedly, the daughter m/z obtained after employing collision-induced energy (CID) at 40 V of the selected RT and m/z corresponded to the cannabinoid standards (Appendix A, Figure A4). This suggests that *P. tricornutum* may contain endogenous enzymes that are capable of converting OA into CB-like metabo-

lites such as CBGA, CBDA, and THCA. Such enzymes could be mimicking the activity of cannabis APT (using the endogenous pool of GPP) and CBDAS/THCAS (Figure 1). Candidate enzymes from the diatom database were searched and low-similarity sequences for APT and THCAS were identified (Appendix B, Figure A6). However, further characterization is needed and might help find closer candidates among microalgae. Alternatively, candidates from a completely different type of enzyme could be involved.

To test for the presence of such endogenous enzymes, we performed a supplementation assay with *P. tricornutum* transconjugants harboring the p*Pt*GE30, which were devoid of cannabis genes (Table 1). We first added OA and GPP for 16 h and could subsequently detect small amounts (circa 3–5 mg/kg) of CBGA and other CB-like metabolites using HPLC analysis. This suggests the presence of an APT-like enzyme in *P. tricornutum* (Table 1). Next, we supplemented the sample with CBGA and detected metabolites that behaved in a similar way to CB on an HPLC chromatogram, but we were unable to confirm their identity (not all were appropriate m/z fragments) via LC-MS.

Table 1. Results of a supplementation assay of *P. tricornutum* with olivetolic acid (OA), geranyl diphosphate (GPP), and cannabigerolic acid (CBGA). The results show the mean detection levels that were found in three different strains of *P. tricornutum* transconjugants harboring p*Pt*GE30.

Supplemented mM			Detected			
OA	**GPP**	**CBGA**	**HPLC**	**MS (*m/z*) Confirmation**	**HPLC**	**MS (*m/z*) Confirmation**
0.45	1	0	CBGA	yes	CBDA, CBNA	No
0	0	0.27	CBDA	yes	THCA, CBNA	No

2.2.2. Metabolite Production Timeline

Olivetolic acid and cannabinoid-like metabolites were detected up to three months after obtaining and subculturing each transconjugant in optimal conditions with no supplementation of any kind. However, a gradual decrease in the concentration of the OA and CB-like metabolites was observed. After six months, the transconjugants completely lost the ability to produce these desired metabolites. Along with production loss, we noticed that the *P. tricornutum* transconjugant morphotype also changed from triradiate (Figure 3A) to fusiform (Figure 5D), and that the initial triradiate morphotype of the transconjugants was never restored.

3. Discussion

The bioengineering of *P. tricornutum* with cannabis genes represents a promising approach to addressing the challenges associated with traditional cannabinoid production. By harnessing the unique capabilities of marine diatoms, this study paves the way for sustainable and controlled cannabinoid biosynthesis. Furthermore, the successful expression of cannabis genes in a non-native host highlights the versatility of synthetic biology in the context of metabolic engineering. This work represents the first report of cannabinoid engineering in brown microalgae and provides a proof-of-concept evaluation of *P. tricornutum* for the heterologous production of OA and cannabinoids. We provide evidence that the natural metabolism of the marine diatom *P. tricornutum* is well-suited for use in metabolic engineering to produce the cannabinoid precursor, OA, and CB-like metabolites.

The episomal vectors, *Pt*OA1 and *Pt*OA2, contained both *CsTKS* and *CsOAC* genes, separated by a self-cleaving peptide sequence and under the control of the same promoter, to ensure similar levels of expression of both genes and, subsequently, comparable levels of accumulated enzymes for performing the two-step reactions. The transconjugants *Pt*OA1 and *Pt*OA2 only differed in terms of the addition of a 6xHis tag on *Cs*TKS for easier detection of the protein. In *Pt*OA3, both genes were linked together with a linker and tagged with YFP for visualization and localization studies.

Although a previous study demonstrated the occurrence of episomal rearrangements [42], in this study, the *P. tricornutum* transconjugant strains obtained for each episomal vector did not show significant DNA sequence rearrangement after 1 year (Appendix A, Figure A1 and Table A4). This suggests that episomal rearrangements could be sequence- or DNA-motif-specific. Also, the *Pt*OA1-3 transconjugant strains grew similarly to the WT and EV strains (Figures 3 and 4). This suggests that the presence and expression of the transgenes and the produced metabolites did not significantly affect the growth and division of *P. tricornutum*. Genome analysis of green algae, diatoms, and higher plants revealed similarities in lipid metabolism [43,44]. Therefore, the fact that there is little to no effect of *Cs*TKS and *Cs*OAC expression on *P. tricornutum* growth and fitness could be a result of the nature of the pathway, as an extension of fatty acid metabolism that is fairly close to the diatom chassis, and a preference for energy conversion into lipids.

Heterologous protein detection was not consistent among the different transconjugant strains (Figure 5 and Appendix A, Figure A2). For example, the four analyzed transconjugant strains of *Pt*OA1 show a similar pattern of accumulation, whereas, in the *Pt*OA2 strains, the T2A sequence did not appear to be cleaved. This could be explained either by the possible presence of negatively acting *cis* elements in the sequence of *Pt*OA2 transgenes [43] or by the nature of the amino acid environment that surrounds the T2A sequence cleavage site [45]. It was also reported that the gene following the T2A sequence would be affected on the expression level [46] and, thus, would yield fewer *Cs*OAC transcripts, but the effect of this phenomenon would be validated if the C10–C12 polyketide product was available as a standard for HPLC detection, which is not the case.

When compared to previous studies, the production of OA in *Yarrowia lypolytica* slightly changed when *Cs*TKS and *Cs*OAC were expressed under different promoters, compared to when the two genes were fused [42]; in *P. tricornutum*, the *Pt*OA1 showed a higher quantity of OA (Figure A3) compared to there being none in *Pt*OA3 when the two genes were fused by a glycine-serine linker. This is not applicable in *Pt*OA2 since the Western blot did not show cleaved proteins, which could indicate that the cleaved proteins were below detection levels.

To our knowledge, this is the first time that a brown microalgae host has demonstrated the capacity to accumulate OA and use it for the synthesis of CB-like metabolites. The heterologous production of OA was observed in the different transconjugants of *P. tricornutum*, transformed with episomal constructions and harboring various combinations of *C. sativa TKS* and *OAC* genes (Figure 2; Appendix A, Figure A1 and Table A2). The cassette designs in this study were fairly simple and direct, with only two main cannabis genes being introduced. When compared to the literature, the first approaches regarding yeast achieved the production of only 0.48 mg/L of OA by expressing *Cs*TKS and *Cs*OAC, and by feeding the culture with sodium hexanoate [9]. Here, we obtained higher yields (0.56–2.6 mg/kg) compared to yeast with and without hexanoate supplementation [9]. In *Y. lipolytica*, it was bioengineered with *Pseudomonas* sp. *LvaE*, encoding a short-chain acyl-CoA synthetase, acetyl-CoA carboxylase, pyruvate dehydrogenase bypass, NADPH-generating malic enzyme, as well as the activation of the peroxisomal β-oxidation pathway and the ATP export pathway to redirect the carbon flux toward OA synthesis [47]; the yield was also less than 0.5 mg/L until batch culture optimization.

Further optimization of the *P. tricornutum* transconjugant strains by boosting the pool of precursors could increase the production of OA. In *E. coli*, besides introducing *Cs*TKS and *Cs*OAC, metabolic engineering consisted of the overexpression of the acetyl-CoA carboxylase subunits A, B, C, and D from the same host to ensure the high production of malonyl-CoA, and a reversal of beta-oxidation that allowed the accumulation of hexanoyl-CoA [33]. This led to a higher quantity of OA (46.3 mg/L) after the optimization of culture conditions and induction parameters [33]. In *P. tricornutum*, hexanoyl-CoA and malonyl-CoA occur naturally, due to the long-chain acyl-CoA synthase (*Pt*ACS3 and *Pt*ACS4) for hexanoyl-CoA and to the acetyl-CoA ATP-dependent carboxylase (*Pt*ACCase) for malonyl-CoA [43]. *P. tricornutum* was also shown to accumulate intracellular amounts

of GPP, along with neryl diphosphate (NPP) [36,48]; both precursors are donors of the prenyl moiety of several cannabinoids, such as CBGA and cannabinerolic acid (CBNA) in cannabis [46]. However, overexpression of these upstream genes could eliminate the need for supplementation while maintaining the necessary pools of precursors for more stable production of cannabinoids, which is similar to the process used with yeast [31,34].

Along with precursor bioavailability, endogenous enzymes are required to enable the conversion of OA into CB or CB-like compounds, as observed in this study, either by direct production (Figure A4) or by supplementation assays (Table 1). Bioinformatic analyses led to the identification of two putative enzymes (Appendix B, Figure A6) that are encoded by Phatr3_J37858 and Phatr3_J2738.t1 and are capable of prenyltransferase activity, thereby allowing the production of CBGA. Both enzymes are UbiA prenyltransferase family members that could be used originally by the diatom to produce terpenes or flavonoids [49,50]. The in silico identification of THCAS and CBDAS-like endogenous candidate genes is more intricate, given the specificity of these enzymes. Our approach can also be used to further characterize endogenous diatom enzymes with multiple functions by testing the reactivity of diatoms to drug precursors in higher-level plants.

The episomal sequences were well-conserved in the timeframe of this study in all three constructions of interest (in DNA and proteins). However, after such a timeframe, the production of OA was not detectable, perhaps due to silencing via an epigenetic mechanism [42], mRNA or protein stability [37], and, less probably, episomal rearrangements, as seen in other cases [42]. It would be useful to characterize a larger number of transformed *P. tricornutum* using a method similar to the process used with yeast [31,34], or to try randomly integrated chromosomal expression (RICE). Interestingly, a loss of production was observed in parallel with a change in cell morphotype from a triradiate to a fusiform population, which could also be associated with changes in the intracellular levels of certain precursors related to general metabolism [51]. A loss of heterologous production of specialized metabolites was previously observed for *E. coli*, where, after the 31st generation, lycopene production decreased from 3.3% of its initial yield [52] due to segregational plasmid instability. Similarly, the use of cyanobacteria as a heterologous production platform faces many challenges, such as a loss of productivity due to the burden or loss of the metabolic pathway and genetic instability [52,53]. In yeast, low expression-cassette instability in large fermentation batches led to a decrease in recombinant protein production [54]. Taken together, these studies show that a loss of production seems to occur elsewhere in other systems, as a result of different underlying problems. In addition, these further demonstrate that like other systems, diatoms need further optimization of the metabolic pathway-encoding constructs discussed below to ensure a suitable platform for the heterologous production of specialized metabolites. Recently, cloning the *Metarhizium anisopliae ARSEF23* highly reducing PKS enzyme (*MaOvaA*), in addition to a non-reducing PKS (*MaOvaB)* and a domain PKS (*MaOvaC)* consisting of an acyl carrier protein and a thioesterase, in the fungus *Aspergillus nidulans* yielded high titers of OA (80 mg/L) [55]. Thus, it would be interesting to compare the expression and behavior of cannabis enzymes regarding fungal enzymes such as highly reducing and non-reducing PKS, along with thioesterase and acyl carrier enzymes [50]. Another approach to increase OA titers could be the construction of alternative pathways using enzymes found in other OA-producing plants than cannabis, in order to achieve OA and other CB precursors while using fungal enzymes or mutation-enhanced enzymes. It is possible, from an evolutionary point of view, that other sources of transgenes are more suitable for heterologous expression in diatoms.

Moreover, in this study, low OA yield could be explained by the detection of cannabinoid-like compounds in higher titers (Figure A4) similar to those observed in other studies, where flavonoid precursors were consumed to form flavonoids in bioengineered yeast [56]. Further steps of compartmentalization of the products would be an interesting strategy, given the possible cytotoxicity of the cannabinoids in many systems [12,57].

The application of *P. tricornutum* as a production platform for OA raises intriguing questions about its metabolic compatibility with the broader CB biosynthesis pathway. Further research is warranted to understand how the engineered diatom handles the downstream conversion of OA into various cannabinoids. Additionally, optimizing the growth conditions, such as light intensity, nutrient availability, and cultivation scale, will be crucial for maximizing production yields [51,58,59]. The successful synthesis of olivetolic acid in *P. tricornutum* opens the door to a range of future possibilities. Further pathway engineering could enable the production of specific compounds, offering a renewable and controllable source of cannabinoids for medical and industrial applications. The engineered diatom strains could be integrated into larger bioprocessing systems, harnessing their natural growth advantages for cost-effective production.

4. Materials and Methods

4.1. Microbial Strains and Growth Conditions

Saccharomyces cerevisiae VL6-48 (ATCC MYA-3666: MATα his3-Δ200 trp1-Δ1 ura3-52 lys2 ade2-1 met14 cir0) was used for the yeast assembly, as described previously [38]. Positive yeast strains containing the His selection were grown on minimal yeast media without histidine. A pool of the grown yeasts was harvested 5 days after assembly and the total DNA was extracted, as described previously [60]. The assembled plasmids were then transformed via electroporation and amplified in *Escherichia coli* (Epi300, Epicenter, Biosearch Technologies, Guelph, ON, Canada), which was grown on Luria Broth (LB) media supplemented with the appropriate antibiotic chloramphenicol (25 mg·L^{-1}) overnight at 37 °C. The plasmids were then extracted from the chloramphenicol-*E. coli* colonies, which were tested by cPCR using a miniprep kit, allowing the extraction of large vectors (EZ10 DNA miniprep kit, Bio Basic Inc., Markham, ON, Canada).

4.2. Diatom Transformation, Selection, and Subculturing

The plasmids were verified by next-generation sequencing and were then amplified in an *E. coli* Epi 300 strain containing the pTA-MOB plasmid to enable conjugation with wild-type diatoms, as described in the literature [38]. *P. tricornutum* (Culture Collection of Algae and Protozoa, CCAP 1055/1), was kindly provided by Prof. Bogumil Karas. The transfer of plasmid DNA to *P. tricornutum* via conjugation from *E. coli* was performed as described by Karas et al. [38]. For this process, 250 µL of wild-type *P. tricornutum* culture was adjusted to a density of 10^8 cells/mL, while the *P. tricornutum* cell density was obtained by plating 1 mL of wild-type *P. tricornutum* on ½ × L1 1% agar plates and growing them at 18 °C under cool fluorescent lights (75 µE $m^{-2}s^{-1}$) on a light/dark cycle of 16/8 h for 4 days. Prior to transformation, 1 mL of L1 media was added to each agar plate, and the cells were first scraped and then harvested via pipetting in a sterile tube. The cells were then diluted and mounted in an improved Neubauer hemacytometer (BLAUBRAND® counting chamber, MilliporeSigma Canada Ltd., Oakville, ON, Canada) to be counted, then the cell concentration was adjusted to 5.0 × 10^8 cells mL^{-1}. A volume of 50 mL of *E. coli* culture containing the assembled and pTA-MOB plasmids was grown at 37 °C under agitation to reach an OD_{600} of 0.9, then the volume was centrifuged at 3000× *g* for 10 min and resuspended in 500 µL of SOC media. Conjugation was initiated by adding 200 µL of *P. tricornutum* to 200 µL of *E. coli* cells. The cell mixture was then plated on ½ L1, 5% LB, 1% agar plates, incubated at 30 °C for 90 min in the dark, and then transferred to 18 °C in the light and grown for 48 h. Two days later, 1 mL of L1 media was added to the plates and the cells were collected by scraping, then a volume of 200 µL of cells was plated on ½ L1, 1% agar plates supplemented with zeocin at 50 µg mL^{-1} for selection and were incubated at 18 °C under light (75 µE $m^{-2}s^{-1}$). Two weeks later, positive colonies appeared and were streaked again on ½ L1, 1% agar plates supplemented with zeocin at 50 µg mL^{-1} for the verification of plasmid stability. The recombinant colonies of *P. tricornutum* were checked via cPCR. Four positive colonies were retained from each construct, from which the plasmids were extracted for further confirmation. The extracted vectors were then

amplified in *E. coli* Epi300; these were extracted using the EZ10 kit and sequenced using next-generation sequencing via Illumina MiSeq technology at the Massachusetts General Hospital's Center of Computational and Integrative Biology (MGH CCIB DNA Core, Boston, MA, USA). In parallel, transformed *P. tricornutum* culture was launched from a single colony until obtaining an OD_{680} nm of 0.06–0.1. The bioengineered *P. tricornutum* strains were then sub-cultured every 10 days in a 1/3 volume ratio of fresh L1 Si^{-} medium containing zeocin (50 mg·L^{-1}).

4.3. Transgene Sequences Validation

C. sativa tetraketide synthase (*Cs*TKS) and olivetolic acid cyclase (*Cs*OAC) amino acid sequences were obtained from *C. sativa* through Genbank, under the respective accession numbers of ACD76855.1 and JN679224.1 (Appendix A, Table A1). The amino acid sequence was then converted into a DNA sequence with respect to *P. tricornutum*'s codon usage table, which is accessible through NCBI Genbank, as well as the HIVE laboratory's codon usage table (CUT) [41]. Several cassettes containing the genes of interest were designed; in this article, we describe three of them. (1) *Pt*OA1 was a construction where TKS and OAC were linked by a self-cleavable sequence from the *Thosea asigna* virus (T2A) [45] without tags, to ensure a similar enzymatic site exposition as that seen in *C. sativa*. (2) *Pt*OA2 is a construction where TKS and OAC were tagged, respectively, in the C-terminal with 6xHis and c-Myc tags for possible protein detection or purification and were linked by a T2A sequence. (3) *Pt*OA3 is a construction designed to present TKS coupled with a reporter gene fluorescent protein eYFP on the N-terminal fused by $(GlyGlyGlyGlySer)_3$ to OAC, which is tagged with c-Myc on the C-terminal. Each inserted construction was designed to be expressed under the strong constitutive promoter 40SRPS8 and the FcpA terminator in a p*Pt*GE30 backbone vector [61] that contained the zeocin resistance gene *ShBle* as a selection marker. The plasmid p*Pt*GE30 contains a centromeric yeast fragment, allowing it to remain extrachromosomal [38]. All gene sequences were codon-optimized for *P. tricornutum*'s optimal codon usage, with a GC content ranging from 48 to 55%. A representation of each plasmid design is detailed in Figure 1. The primers used for the gene amplification and DNA fragment assembly are listed in Appendix A, Table A1. The fragments and transgene sequences that have been optimized for higher expression in *P. tricornutum* are detailed in Appendix A, Table A2. The sequencing results of the plasmid DNA obtained from *P. tricornutum* clones described in this study are detailed in Appendix A, Table A3.

4.4. Heterologous Protein Detection

Protein detection in the positively transformed strains of *P. tricornutum* and the negative control (wild type, transformed with p*Pt*GE30) was performed on 50 mL (OD_{680} nm 0.1 a.u.) cultures that were 6 days old with OD_{680} nm values between 0.9 and 1.7 a.u. The cells were centrifuged at 4000 rpm for 20 min at 4 °C. The resulting pellets were weighed and resuspended in loading buffer 1× (0.8% sodium dodecyl sulfate, 0.05 M Tris-HCl pH 6.8, 6% glycerol, bromophenol blue, and 3.2% β-mercaptoethanol) to a final concentration of 500 mg FW·mL^{-1}. The cell lysates were boiled at 95 °C for 5 min and were then centrifuged at maximum speed for 5 min at room temperature. A volume of 35 µL of the supernatant was loaded into 12% SDS-PAGE gel and migrated at 80 volts until the proteins passed through to the stacking gel, then the voltage was raised to 120 volts. The gel was then transferred using the BioRad Trans-Blot Turbo Transfer system at 2.5 A for 15 min. The blot was equilibrated in 1× TBS solution and blocked with 5% milk for 1 h at room temperature, before washing it three times with TBST solution and adding the primary antibody overnight at 4 °C. In this study, the various primary antibodies used for the T2A sequence were purchased from Sigma Aldrich (cat. #ABS31, Boston, MA, USA), the 6X-His Tag Monoclonal antibody from Thermofisher (cat. #MA1-21315, Waltham, MA, USA), and anti-GFP/CFP/YFP from Cedarlane (cat. #CLH106AP, Burlington, ON, Canada), all at a 1:1000 dilution in 3% BSA. After three washes with TBST solution, the *Pt*OA1 sample blot was incubated for 1 h in a 1:20,000 dilution in 5% milk of the Immun-Star Goat Anti-Rabbit

(GAR)-HRP conjugate from Bio-Rad (Mississauga, ON, Canada, cat. #1705046). Meanwhile, the *Pt*OA2 and *Pt*OA3 blots were incubated under the same conditions with Immun-Star Goat Anti-Mouse (GAM)-HRP conjugate from BioRad (cat. #1705047, Hercules, CA, USA). The multiple Tag protein (GenScript, cat. # M0101, Piscataway, NJ, USA) and the recombinant YFP purified protein (10 ng) from *E. coli* were used as the positive controls. After three washes with TBST solution, protein detection was performed using the Clarity Max Western ECL Substrate-Luminol solution from Bio-Rad (cat #1705062S, CA, USA). The blots and Coomassie-stained gels were visualized using a ChemiDoc Imaging System with Image Lab Touch software 2.4 (Bio-Rad, cat #12003153, Hercules, CA, USA) and Image Lab™ 5.2 software (Bio-Rad, cat # 1709690, Hercules, CA, USA).

4.5. Subcellular Localization of YFP

Live cell images of the four-day-old culture were captured using a Leica SP8 confocal laser microscope (Leica, Wetzlar, Germany) with an HCX PL APO 60×/1.25–0.75 Oil CS objective. The excitation of YFP and chlorophyll fluorescence occurred at 488 nm with a 65 mW argon laser. YFP fluorescence emission was detected between 552 and 560 nm, whereas chlorophyll fluorescence was detected at a bandwidth of 625–720 nm. Bright-field light microscopy images were also taken. The images were analyzed using the Fiji software (https://imagej.net/software/fiji/) for Windows (64-bit) [62].

4.6. Flow Cytometry Analysis

The BD FACS Melody device (BD Biosciences, La Jolla, CA, USA), equipped with blue (488 nm), red (640 nm), and violet (405 nm) lasers, and a Beckn Cytoflex FC500 equipped with Argon (488 nm) and HeNe (633 nm) lasers were used to measure YFP fluorescence on the FITC channel (527/32 and 525/15 nm) and chloroplast autofluorescence on the APC channel (660/10 nm and 675/15 nm, respectively). The *P. tricornutum* samples were diluted to an OD_{680} of 0.1 and were then analyzed at a fixed flow rate of 1 for at least 10,000 events per sample, conducted with 3 replicates. The diatoms were first gated using side-scattered light (SSC) versus forward scatter plots to determine the targeted population and were then gated using the chlorophyll levels (650 nm) (Figure 4). All data acquisition and analysis were carried out with the BD FlowJo version 10 software (BD Biosciences, La Jolla, CA, USA, 2020). All FACS experiments were conducted in triplicate.

4.7. Metabolite Extraction Method

The wild-type and positive transconjugant strains were pre-cultured in L1 medium, as mentioned previously. Approximately 100 mg of wet biomass from each culture was harvested by centrifuging 50 mL of culture at 1500× *g* and at 4 °C for 10 min. The pellets were resuspended in 5 mL of ethanol 95% (cat.# P016EA95), vortexed 5 times for 1 min each time, and stored at −20 °C overnight. The metabolite extracts were separated from the cells by centrifuging at 1500× g and at 4 °C for 10 min. The supernatants were transferred to clean tubes and ethanol fractions were evaporated via a speedvac vacuum concentrator (Thermo Savant SPD 2010) over 3 h at ramp level 5, without heat. The dried extracts were reconstituted in 250 µL of mobile phase solution, consisting of formic acid at 0.1% *v*/*v* in methanol and water (85:15), for the purposes of HPLC analysis. After homogenization, 200 µL of the extract was transferred to an HPLC vial with a 300 µL adaptor.

4.8. HPLC Detection Method for Cannabinoid Precursors

Olivetolic acid (CAS 491-72-5) and olivetol (CAS 500-66-3) standards were purchased from Santa Cruz Biotechnologies (Dallas, TX, USA). Extracts from *P. tricornutum* transconjugant strains and the wild type were then analyzed via HPLC (Agilent 1260 Infinity II) with a DAD detector. The column used was a ZORBAX Eclipse XDB-C18, 80Å, 4.6 × 250 mm, and 5 µm (Agilent; PN: 990967-902, Missisauga, ON, Canada). The sampler harvested and injected 10 µL of the sample into the system; the analysis conditions were the following: the temperature applied to the column was 30 °C, while the isocratic mobile phase comprised

methanol at 85% and formic acid at 0.1%, mixed with 15% pure water at a flow rate of 0.4 mL per minute. The total run time of the analysis was 70 min per sample. The analyzed wavelengths were chosen based on the maximal peak of each standard (220, 230, and 280 nm).

4.9. Mass Spectrum Validation

All the LC-MS analyses in this study were performed at the Centre National en Électrochimie et en Technologies Environnementales, Inc. (CNETE, Shawinigan, QC, Canada). To validate the presence of the compounds of interest, the samples were analyzed under the same LC conditions but used the MS detector, achieving this by diluting each sample 10 times in methanol at LC-MS grade containing 0.1% formic acid and injecting the samples into a Thermo Scientific UPLC Dionex Ultimate 3000 MS LTQ XL UPLC-MS system equipped with a ZORBAX Eclipse XDB-C18 at 80 Å, 4.6 × 250 mm, and 5 µm (Agilent; PN.: 990967-902). The MS detector was set to analyze five scan events during the same period of analysis, taking one SIM event for each compound and one for large-scale TIC analysis (the event parameters are detailed in Appendix A, Figure A5). The analysis conditions used were the following: the temperature applied to the column was 30 °C and the isocratic mobile phase used comprised methanol 85% and formic acid 0.1%, mixed with 15% pure water at a flow rate of 0.4 mL per minute. The total run time of an analysis was 70 min per sample. The sample injection volume was 2.5 µL. The events observed at 220 nm were analyzed for 5 main events via the MS spectra corresponding to each of the main cannabinoids, while fragmentation was compared to the corresponding standard and theoretical fragmentation using the software program Mass Frontier 7.0.

4.10. Precursor Supplementation Assay

P. tricornutum's empty vector (p*Pt*GE30) was grown for a full cycle from an established liquid culture until the OD_{680} was greater than 2 a.u. Afterward, six new subcultures were made with a starting inoculum of 0.1 OD_{680} in L1 media and then supplemented with 50 mg/L of zeocin for 9 days. The final volume of all subcultures was around 360 mL. Then, this volume was distributed as nine 30 mL cultures and placed into 50 mL Falcon tubes. The first set of triplicates was supplemented with ethanol as a negative control, the second set was supplemented with 1 mM of geranyl diphosphate ammonium salt (Sigma, CAS no 763-10-0) and 0.45 mM of olivetolic acid dissolved in L1 media, and the third set was supplemented with 0.27 mM of CBGA diluted in ethanol. Along with the p*Pt*GE30 subcultures, this assay was performed on at least two other recombinant strains containing at least one of the *Pt*OA cassettes. The Falcon tubes were returned to normal growth conditions for 16 h; the next day, each culture was centrifuged to separate the supernatant, then the pellet was washed and the washing liquid was kept for analysis. The metabolites were extracted from the cell pellets, as described in Section 4.7. The supernatant, washes, and cell extracts were analyzed via HPLC, and the signals of interest were confirmed later via MS analysis.

5. Conclusions

In conclusion, the bioengineering of *P. tricornutum* with cannabis genes represents a significant advancement in CB production technology. This innovative approach not only provides a sustainable alternative to traditional cannabis cultivation but also highlights the potential of marine microorganisms as versatile bioreactors for valuable natural products. As synthetic biology and metabolic engineering continue to evolve, the intersection of marine biology and cannabis biosynthesis could shape the future of CB production.

Author Contributions: F.A., E.I.F., N.M., F.M.-M. and I.D.-P. contributed to the study's conception and design. Methodology, material preparation, data collection, and analysis were performed by F.A., E.I.F., M.H., A.M.D.-G., N.M., A.C., S.-E.G. and F.M.-M. Episomal constructions, diatom conjugation, and molecular biology experiments were planned and conducted by F.A., E.I.F., A.M.D.-G., J.L., N.M., F.M.-M., B.J.K. and I.D.-P. HPLC analyses were performed by M.H., A.C. and S.-E.G., whereas all LC-MS analyses were performed by J.-F.L. The first draft of the manuscript was written by F.A., E.I.F., A.M.D.-G. and N.M. F.M.-M. and I.D.-P. edited the previous versions of the manuscript. I.D.-P. supervised the project and secured the funding. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada through the Alliance Program (award nos. ALLRP 554429-20 and ALLRP 570476-2021) to I.D.-P. Additional support in the form of scholarships was awarded to F.A. from MITACS-Elevation IT15971, to E.I.F. and A.M.D.-G. from the MITACS-Acceleration program (grant nos. IT12310, IT16463, and IT19432), and to I.D.-P. is also acknowledged. B.J.K. is supported by Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-06172).

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors would like to acknowledge Patrik Quessy from the Centre National en Électrochimie et en Technologies Environnementales Inc. for LC-MS analyses, and Mélodie B. Plourde, Pamela Lavoie, and Hugo Germain from the University du Québec à Trois-Rivières for their guidance, mentoring, and assistance.

Conflicts of Interest: The authors declare that this study received funding from Algae-C. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Abbreviations

Au	Absorbance units
CB	Cannabinoid
CBD	Cannabidiol
CBDA	Cannabidiolic acid
CBGA	Cannabigerolic acid
cPCR	Colony polymerase chain reaction
FACS	Fluorescence-activated cell sorting
Gly	Glycine
GPP	Geranyl pyrophosphate
His	Histidine
OA	Olivetolic acid
OAC	Olivetolic acid cyclase
OD	Optical density
Ser	Serine
T2A	*Thosea asigna* virus self-cleavable sequence
THC	Δ-9-tetrahydrocannabinol
THCA	Delta (9)-tetrahydrocannabinolic acid
TKS	Tetraketide synthase
YFP	Yellow fluorescent protein

Appendix A

Our study showed the construction of three episomes in the p*Pt*GE30 backbone containing *C. sativa* genes for the production of OA (Appendix A, Figure A1 and Tables A1–A3). Each episome was inserted into *P. tricornutum* via bacterial conjugation and the corresponding enzymes were successfully accumulated (Appendix A, Figure A2 and Table A4).

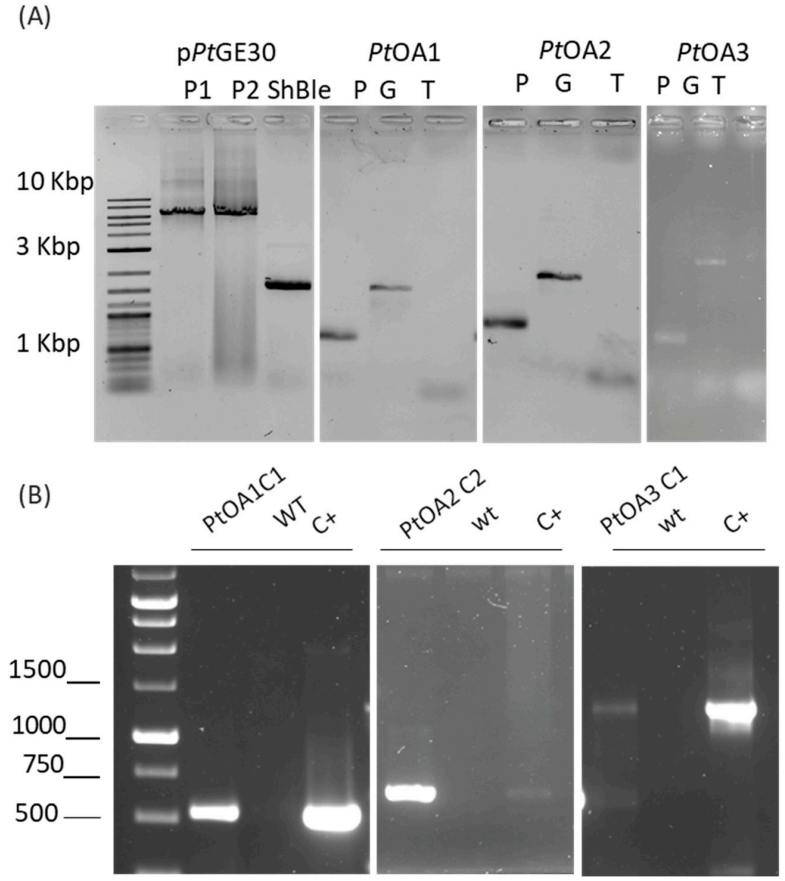


Figure A1. *P. tricornutum* plasmid fragments for yeast assembly and the colony PCR of recombinant *P. tricornutum* after bacterial conjugation. (**A**) Agarose gel electrophoresis of the amplified fragments (the backbone moieties (P1 and P2), promoter (P), gene (G), terminator (T), and selection marker *ShBle* encoding for zeocin resistance) for the plasmid constructs *Pt*OA1 and *Pt*OA2. (**B**) Agarose gel (2%) electrophoresis of the colonies. PCR was performed on the *P. tricornutum* colonies for each construct, with the cPCR primers from Appendix A, Table A4.

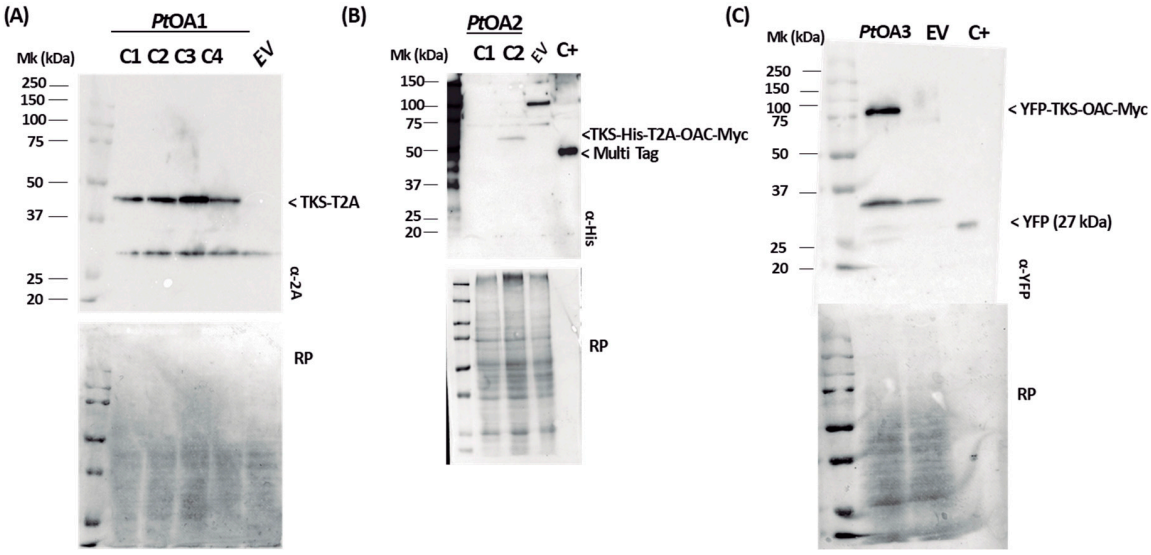


Figure A2. Western blot membranes and the Ponceau red coloration of gels of (**A**) *Pt*OA1, (**B**) *Pt*OA2, and (**C**) *Pt*OA3. The membranes were obtained from different 12% SDS gels, and the antibodies used are described in the main text.

Table A1. The amino acid sequences of *Cs*TKS and *Cs*OAC used in this study to build the *Pt*OA1, *Pt*OA2, and *Pt*OA3 cassettes. CDS: coding sequence.

Sequence Name	Amino Acid Sequence	Length (a.a.)	Accession Number (CDS)	Accession Number AA Seq
C. sativa type 3 tetraketide synthase (TKS)	MNHLRAEGPASVLAIGTANPENILLQDEFPDYYFRVTKSEHMTQLKEKFRKICDK SMIRKRNCFLNEEHLKQNPRLVEHEMQTLDARQDMLVVEVPKLGKDACAKAIKEWGQPKS KITHLIFTSASTTDMPGADYHCAKLLGLSPSVKRVMMYQLGCYGGGTVLRIAKDIAENN KGARVLAVCCDIMACLFRGPSESDLELLVGQAIFGDGAAAVIVGAEPDESVGERPIFELVSTG QTILPNSEGTIGGHIREAGLIFDLHKDVPMLISNNIEKCLIEAFTPIGISDWNSIFWITH PGGKAILDKVEEKLHLKSDKFVDSRHVLSEHGNMSSSTVLFVMDELRKRSLEEGKSTTGDGFE WGVLFGFGPGLTVERVVVRSVPIKY	385	EU551165.1	ACD76855.1
C. sativa olivetolic acid cyclase (OAC)	MAVKHLIVLKFKDEITEAQKEEFFKTYVNLVNIIPAMKDVYWGKDVTQKNKEEGYTHIVEV TFESVETIQDYIIHPAHVGFGDVYRSFWEKLLIFDYTPRK	101	NM_001397939.1	NP_001384868.1

Table A2. List of the DNA sequence parts for the plasmid constructs, with the corresponding references and accession numbers. The OA1, OA2, and OA3 parts are color-coded as follows: *Cannabis sativa tetraketide synthase* CsTKS (AB164375), *Cannabis sativa olivetolic acid cyclase* CsOAC (JN679224), self-cleavable sequence from the *Thosea asigna* virus T2A, c-Myc and 6His tags, 3(GGGGS) fusion, and the yellow fluorescent protein, YFP. All the mentioned sequences are codon-optimized for the best *Phaeodactylum tricornutum* usage.

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 1	CAGGGTAATATAGATCTTCCGCTGCATAACCCTGCTTCGGGGTCATTATAGCGATTTTTTCGGTATATCCATCCTTTTTCGCA CGATATACAGGATTTTGCCAAAGGGTTCGTGTAGACTTTCCTTGGTGTATCCAACGGCGTCAGCCGGGCAGGATAGGTGAAGTAG GCCCACCCGCGAGCGGGTGTTCCTTCTTCACTGTCCCTTATTCGCACCTGGCGGTGCTCAACGGGAATCCTGCTCTGCGAGGCTG GCCGGCTACCGCCGGCGTAACAGATGAGGGCAAGCGGATGGCTGATGAAACCAAGCCAACCAGGAAGGGCAGCCCACCTATCAAG GTGTACTGCCTTCCAGACGAACGAAGAGCGATTGAGGAAAAGGCGGCGGCGGCCGGCATGAGCCTGTCGGCCTACCTGCTGGCCG TCGGCCAGGGCTACAAAATCACGGGCGTCGTGGACTATGAGCACGTCCGCGAGCTGGCCCGCATCAATGGCGACCTGGGCCGCCT GGGCGGCCTGCTGAAACTCTGGCTCACCGACGACCCGCGCACGGCGCGGTTCGGTGATGCCACGATCCTCGCCCTGCTGGCGAAG ATCGAAGAGAAGCAGGACGAGCTTGGCAAGGTCATGATGGGCGTGGTCCGCCCGAGGGCAGAGCCATGACTTTTTTAGCCGCTAA AACGGCCGGGGGGTGCGCGTGATTGCCAAGCACGTCCCCATGCGCTCCATCAAGAAGAGGCACTTCGAGCTGTAAGTACATCACC GACGAGCAAGGCAAGACGATCCGCGCCTGTTTTATTGAGAACGTTGTTCGTGTTGGCCTCAATGGTAGCGATGCGTCATTCAGCG AAGTTCTGGTGCTGATGATGTGGTTCGCTTTGCCACTGGTCAATGTGGTAAGCCCGTGTAATGTCAGTAACCTTTTTACTGATCT CAGCTTGAGCACGGTCGCTGATGAGCTTATCCATGGCCCCACGGTAACGGATATGATCCTCTAGGGCGTTGACAAATTCTTTGTC GGTTTTCATGGGGTAGACGTCAGTGACCAGGGAACGTCTCCCTACAAAAAGTTGCGCATACTTTGCTCCGTTGTCGACGGCAGGG GTGTCGGACCAGATTGTATCTGTGGCAACGGCCTCATTGCGTCAATGGACCTTTAAAGCCGGGAAACGAGACTTGAAATGTTTAC GTAGAGGTGCATTATAGACCTCACGCGCATATTGAGTGGTTGCAAAAATGGTTTTGCGAACGGTATCACTGGAGACCCATGCCAG GCAAGGACGGAGGGCATCATAATCATGTTCATTGCGTTGGACAGCATGCTTGTTACACGTAAGTATACGGTCGAGAGCATCGTGA CGTTGAGAATTACAGATGGCAACGTGTCGGTGCATGTATTGGCCGAGATTGTCAAATCGTGGCTCAATATATGGCATGCCAGGTA AGTCATGGATGTCGGTATGCCATTCAGCCTTCATGTCGATTTCGTTGTCGAGAACTGATGGGTCCCAGTCAACGTCAGAAGTAAG AACAACATGGGGAAGGGAATGAAGTTCATGACTGGTGGGTGCGCGCATATCCATGTACGGAAGACCCTGTCGTATGTTTAGTGGG ATGATGTAATCATCCAGGGTAACAATACGTTGGTTGCCGCCCACGGTACGAGACCGATCTTGAACACAATTGTGGCAATGTTCAA	Slattery et al., 2018 [61]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 1	GTTGGGCGCTGGAATGGATGGTTTTACCTTTACCAAGATGCGCGTATTGATGCATGATAACAACAATGGGCCCTCGTTGCGATTC GACAAGACCGGCAGCCGTGACAATGTCTAGATTGGACAAGGTATGTTCATTTATACCGGTTATGTTGGCTGAACGTCCAGTTTTG TGCAGTACGGTAACATCACTCCCCGCAAGTCCACCGTTGGCACCGCGATCGACAAGTGCAGATGTGGTATTTTGTACTGTGTGTCG GGAGACTTGGTACTGTAACACATTTGACTGAAGTGACTTGGTGGTACCATGAGGACGGAGGGAACCCGTTGATACGGTGTTGCGT GACGCTGCAAGTACCTTACGTATGTCGCCTGGATCCATACGTCCAACCCGATCAGAAAGATGTGCTAGGAGTTCCGTTTCTGGTG CACAATCATGAAAGGTATCCACGGGAGTTCCATTGTTGTCGTCCGGGGCTGCGTCGCGACTGCCAGTTGCATGTACATGCGCCAT CAAGGGACGTGCCAATCCATTACTAACCGGGAGAACCTTAGCTGGGGCAGCCAATCCCTGAAGTATAGCCTTAGCATCATCTGAG AGCTGGTTCCACATATCTCTAGGGATATAAGGTCGTTCACGGTTAGTGGGCGTACTATGATGTCGGTTATTAGTACAATTCTCCC TGCGGGCATGCGCATTGGCTTCATAGAGTACGGAAGGTGACAAGTCAATATTGTAGTTAACATCCGGATCAGTGTCAAAGTCTGT AGGATGGTAAGAAAGATCAGTAGAATGAATACTACGCTTTTCCTTGGGACTACGGGAATTAGAGAAGTTGTTTCCTTTATTGTAG AGTGACGCTGAAGCAAGTAGAAGACTAAGGTAACTCTCGTAGCTAATAGGATTACCTCCTTTGGCTAGGTCAAGAGTGGCTGTGA TCTTCACTTGACAAAGTTCCGGTACATTGTGGACAGCATTCTCCAAAAGACTAAGACACAGTTGCTTTGGGAGTTGCTCAGCCAT TGGTACAGTATTGTGGTAGATACAAAGGTGGTTCTTCCAATGAAGGATAAATCCTTCTGCTGTGCCTGTCCATGAGGATCCATAT TTCGCCGTAGTTAGGTAACCAAGCGTGGTGGCTGAACTGATCTTCGCACTTGCTGATTCCGTATAGTGTTTGACAACTTTACAAA ACACTTCTTGCGCAGTTCGCTCTAGCCTGGACAGAGAAGGGACAATCTTCGTCGTTAAGACTCGTGCGAATAGCAAAAGATCACA AAATAGCACATCGGCACCGACCAACGATTATTTCCAAGGAAAAAAAGAATGCTTCACTACAAGAAATTGTGTCATCCCTATACAG AGTCTTGTTACTGTGACAGAAAATTGATGGAAGATGTGGCGGATTGCCTTTACACTAGCCAACTTGTTCGACTAATTGCAGCTTCT TCTGAGAGGCTTCACCGAGTAACGCGAAGAACACCGGTGTCTCGTACATGCTCGTCGGTGAACGCTCGTCCAATGACACCCCCCA CTTTGTATCAATATCCCAACTTGGTAGTGAACTGGAATGATACATGCAATTTCGCGCCGCATCAACAGCCACGGGCACCATCGACG AATAGACTCGGTCGAGCTGGTTGCCCTCGCCGCTGGGCTGGCGGCCGTCTATGGCCCTGCAAACGCGCCAGAAACGCCGTCGAAGC CGTGTGCGAGACACCGCGGCCGGCCGCCGGCGTTGTGGATACCTCGCGGAAAACTTGGCCCTCACTGACAGATGAGGGGCGGACGT TGACACTTGAGGGGCCGACTCACCCGGCGCGGCGTTGACAGATGAGGGGCAGGCTCGATTTCGGCCGGCGACGTGGAGCTGGCCAG CCTCGCAAATCGGCGAAAACGCCTGATTTTACGCGAGTTTCCCACAGATGATGTGGACAAGCCTGGGGATAAGTGCCCTGCGGTAT TGACACTTGAGGGGCGCGACTACTGACAGATGAGGGGCGCGATCCTTGACACTTGAGGGGCAGAGTGCTGACAGATGAGGGGCGCA CCTATTGACATTTGAGGGGCTGTCCACAGGCAGAAAATCCAGCATTTGCAAGGGTTTCCGCCCGTTTTTCGGCCACCGCTAACCTG TCTTTTAACCTGCTTTTAAACCAATATTTATAAACCTTGTTTTTTAACCAGGGCTGCGCCCTGTGCGCGTGACCGCGCACGCCGAAG GGGGGTGCCCCCCCTTCTCGAACCCTCCCGGTCGAGTGAGCGAGGAAGCACCAGGGAACAGCACTTATATATTCTGCTTACACACG ATGCCTGAAAAAACTTCCCTTGGGGTTATCCACTTATCCACGGGGATATTTTTATAATTATTTTTTTTATAGTTTTTAGATCTTCT TTTTTAGAGCGCCTTGTAGGCCTTTATCCATGCTGGTTCTAGAGAAGGTGTTGTGACAAATTGCCCTTTCAGTGTGACAAATCACC CTCAAATGACAGTCCTGTCTGTGACAAATTGCCCTTAACCCTGTGACAAATTGCCCTCAGAAGAAGCTGTTTTTTCACAAAGTTAT CCCTGCTTATTGACTCTTTTTTATTTAGTGTGACAATCTAAAAACTTGTCACACTTCACATGGATCTGTCATGGCGGAAACAGCGG TTATCAATCACAAGAAACGTAAAAATAGCCCGCGAATCGTCCAGTCAAACGACCTCACTGAGGCGGCATATAGTCTCTCCCGGGAT CAAAAACGTATGCTGTATCTGTTCGTTGACCAGATCAGAAAATCTGATGGCACCCTACAGGAACATGACGGTATCTGCGAGATCCA TGTTGCTAAATATGCTGAAATATTCGGATTGACCTCTGCGGAAGCCAGTAAGGATATACGGCAGGCATTGAAGAGTTTCGCGGGGA AGGAAGTGGTTTTTTATCGCCCTGAAGAGGATGCCGGCGATGAAAAAGGCTATGAATCTTTTCCTTGGTTTATCAAACGTGCGCAC AGTCCATCCAGAGGGCTTTACAGTGTACATATCAACCCATATCTCATTCCCTTCTTTATCGGGTTACAGAACCGGTTTACGCAGTTT CGGCTTAGTGAAACAAAAGAAATCACCAATCCGTATGCCATGCGTTTATACGAATCCCTGTGTCAGTATCGTAAGCCGGATGGCTCA GGCATCGTCTCTCTGAAAATCGACTGGATCATAGAGCGTTACCAGCTGCCTCAAAGTTACCAGCGTATGCCTGACTTCCGCCGCCG CTTCCTGCAGGTCTGTGTTAATGAGATCAACAGCAGAACTCCAATGCGCCTCTCATACATTGAGAAAAAGAAAGGCCGCCAGACGA CTCATATCGTATTTTCCTTCCGCGATATCACTTCCATGACGACAGGATAGTCTGAGGGTTATCTGTCACAGATTTGAGGGTGGTTCG	Slattery et al., 2018 [61]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 1	TCACATTTGTTCTGACCTACTGAGGGTAATTTGTCACAGTTTTGCTGTTTCCTTCAGCCTGCATGGATTTTCTCATACTTTTTGAA CTGTAATTTTTAAGGAAGCCAAATTTGAGGGCAGTTTGTCACAGTTGATTTCCTTCTCTTTCCCTTCGTCATGTGACCTGATATC GGGGGTTAGTTCGTCATCATTGATGAGGGTTGATTATCACAGTTTATTACTCTGAATTGGCTATCCGCGTGTGTACCTCTACCTGG AGTTTTTCCCACGGTGGATATTTCTTCTTGCGCTGAGCGTAAGAGCTATCTGACAGAACAGTTCTTCTTTGCTTCCTCGCCAGTT CGCTCGCTATGCTCGGTTACACGGCTGCGGCGAGCATCACGTGCTATAAAAATAATTATAATTTAAATTTTTTAATATAAATATA TAAATTAAAAATAGAAAGTAAAAAAAAGAAATTAAAGAAAAAATAGTTTTTGTTTTCCGAAGATGTAAAAGACTCTAGGGGGATCG CCAACAAATACTACCTTTTACCTTGCTCTTCCTGCTCTCAGGTATTAATGCCGAATTGTTTCATCTTGTCTGTGTAGAAGACCAC ACACGAAAATCCTGTGATTTTACATTTTACTTATCGTTAATCGAATGTATATCTATTTAATCTGCTTTTCTTGTCTAATAAATATA TATGTAAAGTACGCTTTTTGTTGAAATTTTTTAAACCTTTGTTTATTTTTTTTTCTTCATTCCGTAACTCTTCTACCTTCTTTATT TACTTTCTAAAATCCAAATACAAAACATAAAAATAAATAAACACAGAGTAAATTCCCAAATTATTCCATCATTAAAAGATACGAGG CGCGTGTAAGTTACAGGCAAGCGATCCTAGTACACTCTATATTTTTTTATGCCTCGGTAATGATTTTCATTTTTTTTTTCCACCTA GCGGATGACTCTTTTTTTTTCTTAGCGATTGGCATTATCACATAATGAATTATACATTATATAAAGTAATGTGATTTCTTCGAA GAATATACTAAAAAATGAGCAGGCAAGATAAACGAAGGCAAAGATGACAGAGCAGAAAGCCCTAGTAAAGCGTATTACAAATGA AACCAAGATTCAGATTGCGATCTCTTTAAAGGGTGGTCCCCTAGCGATAGAGCACTCGATCTTCCCAGAAAAAGAGGCAGAAGCA GTAGCAGAACAGGCCACACAATCGCAAGTGATTAACGTCCACACAGGTATAGGGTTTCTGGACCATATGATACATGCTCTGGCCA AGCATTCCGGCTGGTCGCTAATCGTTGAGTGCATTGGTGACTTACACATAGACGACCATCACACCACTGAAGACTGCGGGATTGC TCTCGGTCAAGCTTTTAAAGAGGCCCTAG	Slattery et al., 2018 [61]
p*Pt*GE30-part 2	TCGAGCTGGTTGCCCTCGCCGCTGGGCTGGCGGCCGTCTATGGCCCTGCAAACGCGCCAGAAACGCCGTCGAAGCCGTGTGCGA GACACCGCGGCCGGCCGCCGGCGTTGTGGATACCTCGCGGAAAACTTGGCCCTCACTGACAGATGAGGGGCGGACGTTGACACTT GAGGGGCCGACTCACCCGGCGCGGCGTTGACAGATGAGGGGCAGGCTCGATTTCGGCCGGCGACGTGGAGCTGGCCAGCCTCGCA AATCGGCGAAAACGCCTGATTTTACGCGAGTTTCCCACAGATGATGTGGACAAGCCTGGGGATAAGTGCCCTGCGGTATTGACAC TTGAGGGGCGCGACTACTGACAGATGAGGGGCGCGATCCTTGACACTTGAGGGGCAGAGTGCTGACAGATGAGGGGCGCACCTAT TGACATTTGAGGGGCTGTCCACAGGCAGAAAATCCAGCATTTGCAAGGGTTTCCGCCCGTTTTTCGGCCACCGCTAACCTGTCTT TTAACCTGCTTTTAAACCAATATTTATAAACCTTGTTTTTAACCAGGGCTGCGCCCTGTGCGCGTGACCGCGCACGCCGAAGGGG GGTGCCCCCCCTTCTCGAACCCTCCCGGTCGAGTGAGCGAGGAAGCACCAGGGAACAGCACTTATATATTCTGCTTACACACGAT GCCTGAAAAAACTTCCCTTGGGGTTATCCACTTATCCACGGGGATATTTTTATAATTATTTTTTTTATAGTTTTTAGATCTTCTT TTTTAGAGCGCCTTGTAGGCCTTTATCCATGCTGGTTCTAGAGAAGGTGTTGTGACAAATTGCCCTTTCAGTGTGACAAATCACC CTCAAATGACAGTCCTGTCTGTGACAAATTGCCCTTAACCCTGTGACAAATTGCCCTCAGAAGAAGCTGTTTTTTCACAAAGTTA TCCCTGCTTATTGACTCTTTTTTATTTAGTGTGACAATCTAAAAACTTGTCACACTTCACATGGATCTGTCATGGCGGAAACAGC GGTTATCAATCACAAGAAACGTAAAAATAGCCCGCGAATCGTCCAGTCAAACGACCTCACTGAGGCGGCATATAGTCTCTCCCGG GATCAAAAACGTATGCTGTATCTGTTCGTTGACCAGATCAGAAAATCTGATGGCACCCTACAGGAACATGACGGTATCTGCGAGA TCCATGTTGCTAAATATGCTGAAATATTCGGATTGACCTCTGCGGAAGCCAGTAAGGATATACGGCAGGCATTGAAGAGTTTCGC GGGGAAGGAAGTGGTTTTTTATCGCCCTGAAGAGGATGCCGGCGATGAAAAAGGCTATGAATCTTTTCCTTGGTTTATCAAACGT GCGCACAGTCCATCCAGAGGGCTTTACAGTGTACATATCAACCCATATCTCATTCCCTTCTTTATCGGGTTACAGAACCGGTTTA CGCAGTTTCGGCTTAGTGAAACAAAAGAAATCACCAATCCGTATGCCATGCGTTTATACGAATCCCTGTGTCAGTATCGTAAGCC GGATGGCTCAGGCATCGTCTCTCTGAAAATCGACTGGATCATAGAGCGTTACCAGCTGCCTCAAAGTTACCAGCGTATGCCTGAC TTCCGCCGCCGCTTCCTGCAGGTCTGTGTTAATGAGATCAACAGCAGAACTCCAATGCGCCTCTCATACATTGAGAAAAAGAAAG GCCGCCAGACGACTCATATCGTATTTTCCTTCCGCGATATCACTTCCATGACGACAGGATAGTCTGAGGGTTATCTGTCACAGAT TTGAGGGTGGTTCGTCACATTTGTTCTGACCTACTGAGGGTAATTTGTCACAGTTTTGCTGTTTCCTTCAGCCTGCATGGATTTT CTCATACTTTTTGAACTGTAATTTTTAAGGAAGCCAAATTTGAGGGCAGTTTGTCACAGTTGATTTCCTTCTCTTTCCCTTCGTC	Slattery et al., 2018 [61]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 2	ATGTGACCTGATATCGGGGGTTAGTTCGTCATCATTGATGAGGGTTGATTATCACAGTTTATTACTCTGAATTGGCTATCCGCGT GTGTACCTCTACCTGGAGTTTTTCCCACGGTGGATATTTCTTCTTGCGCTGAGCGTAAGAGCTATCTGACAGAACAGTTCTTCTT TGCTTCCTCGCCAGTTCGCTCGCTATGCTCGGTTACACGGCTGCGGCGAGCATCACGTGCTATAAAAATAATTATAATTTAAATT TTTTAATATAAATATATAAATTAAAAATAGAAAGTAAAAAAAAGAAATTAAAGAAAAAATAGTTTTTGTTTTCCGAAGATGTAAAA GACTCTAGGGGGATCGCCAACAAATACTACCTTTTACCTTGCTCTTCCTGCTCTCAGGTATTAATGCCGAATTGTTTCATCTTGT CTGTGTAGAAGACCACACACGAAAATCCTGTGATTTTACATTTTACTTATCGTTAATCGAATGTATATCTATTTAATCTGCTTTT CTTGTCTAATAAATATATATGTAAAGTACGCTTTTTGTTGAAATTTTTTAAACCTTTGTTTATTTTTTTTTCTTCATTCCGTAACT CTTCTACCTTCTTTATTTACTTTCTAAAATCCAAATACAAAACATAAAAATAAATAAACACAGAGTAAATTCCCAAATTATTCCA TCATTAAAAGATACGAGGCGCGTGTAAGTTACAGGCAAGCGATCCTAGTACACTCTATATTTTTTTATGCCTCGGTAATGATTTT CATTTTTTTTTTCCACCTAGCGGATGACTCTTTTTTTTTCTTAGCGATTGGCATTATCACATAATGAATTATACATTATATAAAG TAATGTGATTTCTTCGAAGAATATACTAAAAAATGAGCAGGCAAGATAAACGAAGGCAAAGATGACAGAGCAGAAAGCCCTAGTA AAGCGTATTACAAATGAAACCAAGATTCAGATTGCGATCTCTTTAAAGGGTGGTCCCCTAGCGATAGAGCACTCGATCTTCCCAG AAAAAGAGGCAGAAGCAGTAGCAGAACAGGCCACACAATCGCAAGTGATTAACGTCCACACAGGTATAGGGTTTCTGGACCATAT GATACATGCTCTGGCCAAGCATTCCGGCTGGTCGCTAATCGTTGAGTGCATTGGTGACTTACACATAGACGACCATCACACCACTG AAGACTGCGGGATTGCTCTCGGTCAAGCTTTTAAAGAGGCCCTAGGGGCCGTGCGTGGAGTAAAAAGGTTTGGATCAGGATTTGC GCCTTTGGATGAGGCACTTTCCAGAGCGGTGGTAGATCTTTCGAACAGGCCGTACGCAGTTGTCGAACTTGGTTTGCAAAGGGAG AAAGTAGGAGATCTCTCTTGCGAGATGATCCCGCATTTTCTTGAAAGCTTTGCAGAGGCTAGCAGAATTACCCTCCACGTTGATT GTCTGCGAGGCAAGAATGATCATCACCGTAGTGAGAGTGCGTTCAAGGCTCTTGCGGTTGCCATAAGAGAAGCCACCTCGCCCAA TGGTACCAACGATGTTCCCTCCACCAAAGGTGTTCTTATGTAGTTTTACACAGGAGTCTGGACTTGACGCTAGTGATAATAAGTG ACTGAGGTATGTGCTCTTCTTATCTCCTTTTGTAGTGTTGCTCTTATTTTAAACAACTTTGCGGTTTTTTGATGACTTTGCGATTT TGTTGTTGCTTTGCAGTAAATTGCAAGATTTAATAAAAAAACGCAAAGCAATGATTAAAGGATGTTCAGAATGAAACTCATGGAA ACACTTAACCAGTGCATAAACGCTGGTCATGAAATGACGAAGGCTATCGCCATTGCACAGTTTAATGATGACAGCCCGGAAGCGA GGAAAATAACCCGGCGCTGGAGAATAGGTGAAGCAGCGGATTTAGTTGGGGTTTCTTCTCAGGCTATCAGAGATGCCGAGAAAGCA GGGCGACTACCGCACCCGGATATGGAAATTCGAGGACGGGTTGAGCAACGTGTTGGTTATACAATTGAACAAATTAATCATATGCG TGATGTGTTTGGTACGCGATTGCGACGTGCTGAAGACGTATTTCCACCGGTGATCGGGGTTGCTGCCCATAAAGGTGGCGTTTACA AAACCTCAGTTTCTGTTCATCTTGCTCAGGATCTGGCTCTGAAGGGGCTACGTGTTTTGCTCGTGGAAGGTAACGACCCCCAGGGA ACAGCCTCAATGTATCACGGATGGGTACCAGATCTTCATATTCATGCAGAAGACACTCTCCTGCCTTTCTATCTTGGGGAAAAGG ACGATGTCACTTATGCAATAAAGCCCACTTGCTGGCCGGGGCTTGACATTATTCCTTCCTGTCTGGCTCTGCACCGTATTGAAACT GAGTTAATGGGCAAATTTGATGAAGGTAAACTGCCCACCGATCCACACCTGATGCTCCGACTGGCCATTGAAACTGTTGCTCATG ACTATGATGTCATAGTTATTGACAGCGCGCCTAACCTGGGTATCGGCACGATTAATGTCGTATGTGCTGCTGATGTGCTGATTGT TCCCACGCCTGCTGAGTTGTTTGACTACACCTCCGCACTGCAGTTTTTCGATATGCTTCGTGATCTGCTCAAGAACGTTGATCTTA AAGGGTTCGAGCCTGATGTACGTATTTTGCTTACCAAATACAGCAATAGCAATGGCTCTCAGTCCCCGTGGATGGAGGAGCAAAT TCGGGATGCCTGGGGAAGCATGGTTCTAAAAAATGTTGTACGTGAAACGGATGAAGTTGGTAAAGGTCAGATCCGGATGAGAACTG TTTTTGAACAGGCCATTGATCAACGCTCTTCAACTGGTGCCTGGAGAAATGCTCTTTCTATTTGGGAACCTGTCTGCAATGAAATT TTCGATCGTCTGATTAAACCACGCTGGGAGATTAGATAATGAAGCGTGCGCCTGTTATTCCAAAACATACGCTCAATACTCAACCG GTTGAAGATACTTCGTTATCGACACCAGCTGCCCCGATGGTGGATTCGTTAATTGCGCGCGTAGGAGTAATGGCTCGCGGTAATGC CATTACTTTGCCTGTATGTGGTCGGGATGTGAAGTTTACTCTTGAAGTGCTCCGGGGTGATAGTGTTGAGAAGACCTCTCGGGTAT GGTCAGGTAATGAACGTGACCAGGAGCTGCTTACTGAGGACGCACTGGATGATCTCATCCCTTCTTTTCTACTGACTGGTCAACAG ACACCGGCGTTCGGTCGAAGAGTATCTGGTGTCATAGAAATTGCCGATGGGAGTCGCCGTCGTAAAGCTGCTGCACTTACCGAAAG TGATTATCGTGTTCTGGTTGGCGAGCTGGATGATGAGCAGATGGCTGCATTATCCAGATTGGGTAACGATTATCGCCCAACAAGTG	Slattery et al., 2018 [61]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 2	CTTATGAACGTGGTCAGCGTTATGCAAGCCGATTGCAGAATGAATTTGCTGGAAATATTTCTGCGCTGGCTGATGCGGAAAATATT TCACGTAAGATTATTACCCGCTGTATCAACACCGCCAAATTGCCTAAATCAGTTGTTGCTCTTTTTTCTCACCCCGGTGAACTATC TGCCCGGTCAGGTGATGCACTTCAAAAAGCCTTTACAGATAAAGAGGAATTACTTAAGCAGCAGGCATCTAACCTTCATGAGCAGA AAAAAGCTGGGGTGATATTTGAAGCTGAAGAAGTTATCACTCTTTTAACTTCTGTGCTTAAAACGTCATCTGCATCAAGAACTAGT TTAAGCTCACGACATCAGTTTGCTCCTGGAGCGACAGTATTGTATAAGGGCGATAAAATGGTGCTTAACCTGGACAGGTCTCGTGTT CCAACTGAGTGTATAGAGAAAATTGAGGCCATTCTTAAGGAACTTGAAAAGCCAGCACCCTGATGCGACCTCGTTTTAGTCTACGT TTATCTGTCTTTACTTAATGTCCTTTGTTACAGGCCAGAAAGCATAACTGGCCTGAATATTCTCTCTGGGCCCACTGTTCCACTTG TATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCG TATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCG TATCGTCGGTCTGATTATTAGTCTGGGACCATGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCG TATCGTCGGTCTGATTATTAGTCTGGAACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCG TATCGTCGGTCTGATTATTAGTCTGGGACCACGATCCCACTCGTGTTGTCGGTCTGATTATCGGTCTGGGACCACGGTCCCACTTG TATTGTCGATCAGACTATCAGCGTGAGACTACGATTCCATCAATGCCTGTCAAGGGCAAGTATTGACATGTCGTCGTAACCTGTAG AACGGAGTAACCTCGGTGTGCGGTTGTATGCCTGCTGTGGATTGCTGCTGTGTCCTGCTTATCCACAACATTTTGCGCACGGTTAT GTGGACAAAATACCTGGTTACCCAGGCCGTGCCGGCACGTTAACCGGGCTGCATCCGATGCAAGTGTGTCGCTGTCGACGAGCTCG CGAGCTCGGACATGAGGTTGCCCCGTATTCAGTGTCGCTGATTTGTATTGTCTGAAGTTGTTTTTACGTTAAGTTGATGCAGATCA ATTAATACGATACCTGCGTCATAATTGATTATTTGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGATGATAATCATT ATCACTTTACGGGTCCTTTCCGGTGATCCGACAGGTTACGGGGCGGCGACCTCGCGGGTTTTCGCTATTTATGAAAATTTTCCGGTT TAAGGCGTTTCCGTTCTTCTTCGTCATAACTTAATGTTTTTATTTAAAATACCCTCTGAAAAGAAAGGAAACGACAGGTGCTGAAA GCGAGCTTTTTGGCCTCTGTCGTTTCCTTTCTCTGTTTTTGTCCGTGGAATGAACAATGGAAGTCCGAGCTCATCGCTAATAACTT CGTATAGCATACATTATACGAAGTTATATTCGATGCGGCCGCAAGGGGTTCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCT TAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCACACAGATGCGTAAGGAGAAAATACC GCATCAGGCGCCATTCGCCATTCAGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAA AGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTGT AATACGACTCACTATAGGGCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGAGTATTCT ATAGTCTCACCTAAATAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACA TACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCT TTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGAACCCCTTGCGGCCGCCCGGGCCGTCGACCAATT CTCATGTTTGACAGCTTATCATCGAATTTCTGCCATTCATCCGCTTATTATCACTTATTCAGGCGTAGCAACCAGGCGTTTAAGGG CACCAATAACTGCCTTAAAAAAATTACGCCCCGCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGG AAGCCATCACAAACGGCATGATGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAA AACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACA TATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAACTGCCGGA AATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCAT ATCACCAGCTCACCGTCTTTCATTGCCATACGAAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGAATAAAGGCCGGATA AAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACT GACTGAAATGCCTCAAAATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGC TTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTT ACGTGCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGC	Slattery et al., 2018 [61]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
p*Pt*GE30-part 2	GAAGTGATCTTCCGTCACAGGTATTTATTCGCGATAAGCTCATGGAGCGGCGTAACCGTCGCACAGGAAGGACAGAGAAAGCGCGGA TCTGGGAAGTGACGGACAGAACGGTCAGGACCTGGATTGGGGAGGCGGTTGCCGCCGCTGCTGCTGACGGTGTGACGTTCTCTGTT CCGGTCACACCACATACGTTCCGCCATTCCTATGCGATGCACATGCTGTATGCCGGTATACCGCTGAAAGTTCTGCAAAGCCTGAT GGGACATAAGTCCATCAGTTCAACGGAAGTCTACACGAAGGTTTTTGCGCTGGATGTGGCTGCCCGGCACCGGGTGCAGTTTGCGA TGCCGGAGTCTGATGCGGTTGCGATGCTGAAACAATTATCCTGAGAATAAATGCCTTGGCCTTTATATGGAAATGTGGAACTGAGTG GATATGCTGTTTTTGTCTGTTAAACAGAGAAGCTGGCTGTTATCCACTGAGAAGCGAACGAAACAGTCGGGAAAATCTCCCATTAT CGTAGAGATCCGCATTATTAATCTCAGGAGCCTGTGTAGCGTTTATAGGAAGTAGTGTTCTGTCATGATGCCTGCAAGCGGTAACGA AAACGATTTGAATATGCCTTCAGGAACAATAGAAATCTTCGTGCGGTGTTACGTTGAAGTGGAGCGGATTATGTCAGCAATGGACA GAACAACCTAATGAACACAGAACCATGATGTGGTCTGTCCTTTTACAGCCAGTAGTGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCC	Slattery et al., 2018 [61]
ShBle cassette	TCGAGCTGGTTGCCCTCGCCGCTGGGCTGGCGGCCGTCTATGGCCCTGCAAACGCGCCAGAAACGCCGTCGAAGCCGTGTGCGAGA CACCGCGGCCGGCCGCCGGCGTTGTGGATACCTCGCGGAAAACTTGGCCCTCACTGACAGATGAGGGGCGGACGTTGACACTTGAG GGGCCGACTCACCCGGCGCGGCGTTGACAGATGAGGGGCAGGCTCGATTTCGGCCGGCGACGTGGAGCTGGCCAGCCTCGCAAATC GGCGAAAACGCCTGATTTTACGCGAGTTTCCCACAGATGATGTGGACAAGCCTGGGGATAAGTGCCCTGCGGTATTGACACTTGAG GGGCGCGACTACTGACAGATGAGGGGCGCGATCCTTGACACTTGAGGGGCAGAGTGCTGACAGATGAGGGGCGCACCTATTGACAT TTGAGGGGCTGTCCACAGGCAGAAAATCCAGCATTTGCAAGGGTTTCCGCCCGTTTTTCGGCCACCGCTAACCTGTCTTTTAACCT GCTTTTAAACCAATATTTATAAACCTTGTTTTTAACCAGGGCTGCGCCCTGTGCGCGTGACCGCGCACGCCGAAGGGGGGTGCCCC CCCTTCTCGAACCCTCCCGGTCGAGTGAGCGAGGAAGCACCAGGGAACAGCACTTATATATTCTGCTTACACACGATGCCTGAAAA AACTTCCCTTGGGGTTATCCACTTATCCACGGGGATATTTTTATAATTATTTTTTTTATAGTTTTTAGATCTTCTTTTTTAGAGCG CCTTGTAGGCCTTTATCCATGCTGGTTCTAGAGAAGGTGTTGTGACAAATTGCCCTTTCAGTGTGACAAATCACCCTCAAATGACA GTCCTGTCTGTGACAAATTGCCCTTAACCCTGTGACAAATTGCCCTCAGAAGAAGCTGTTTTTTCACAAAGTTATCCCTGCTTATT GACTCTTTTTTATTTAGTGTGACAATCTAAAAACTTGTCACACTTCACATGGATCTGTCATGGCGGAAACAGCGGTTATCAATCAC AAGAAACGTAAAAATAGCCCGCGAATCGTCCAGTCAAACGACCTCACTGAGGCGGCATATAGTCTCTCCCGGGATCAAAAACGTAT GCTGTATCTGTTCGTTGACCAGATCAGAAAATCTGATGGCACCCTACAGGAACATGACGGTATCTGCGAGATCCATGTTGCTAAAT ATGCTGAAATATTCGGATTGACCTCTGCGGAAGCCAGTAAGGATATACGGCAGGCATTGAAGAGTTTCGCGGGGAAGGAAGTGGTT TTTTATCGCCCTGAAGAGGATGCCGGCGATGAAAAAGGCTATGAATCTTTTCCTTGGTTTATCAAACGTGCGCACAGTCCATCCAG AGGGCTTTACAGTGTACATATCAACCCATATCTCATTCCCTTCTTTATCGGGTTACAGAACCGGTTTACGCAGTTTCGGCTTAGTG AAACAAAAGAAATCACCAATCCGTATGCCATGCGTTTATACGAATCCCTGTGTCAGTATCGTAAGCCGGATGGCTCAGGCATCGTC TCTCTGAAAATCGACTGGATCATAGAGCGTTACCAGCTGCCTCAAAGTTACCAGCGTATGCCTGACTTCCGCCGCCGCTTCCTGCA GGTCTGTGTTAATGAGATCAACAGCAGAACTCCAATGCGCCTCTCATACATTGAGAAAAAGAAAGGCCGCCAGACGACTCATATCG TATTTTCCTTCCGCGATATCACTTCCATGACGACAGGATAGTCTGAGGGTTATCTGTCACAGATTTGAGGGTGGTTCGTCACATTT GTTCTGACCTACTGAGGGTAATTTGTCACAGTTTTGCTGTTTCCTTCAGCCTGCATGGATTTTCTCATACTTTTTGAACTGTAATT TTTAAGGAAGCCAAATTTGAGGGCAGTTTGTCACAGTTGATTTCCTTCTCTTTCCCTTCGTCATGTGACCTGATATCGGGGGTTAG TTCGTCATCATTGATGAGGGTTGATTATCACAGTTTATTACTCTGAATTGGCTATCCGCGTGTGTACCTCTACCTGGAGTTTTTCC CACGGTGGATATTTCTTCTTGCGCTGAGCGTAAGAGCTATCTGACAGAACAGTTCTTCTTTGCTTCCTCGCCAGTTCGCTCGCTAT GCTCGGTTACACGGCTGCGGCGAGCATCACGTGCTATAAAAATAATTATAATTTAAATTTTTTAATATAAATATATAAATTAAAAA TAGAAAGTAAAAAAAGAAATTAAAGAAAAAATAGTTTTTGTTTTCCGAAGATGTAAAAGACTCTAGGGGGATCGCCAACAAATACT ACCTTTTACCTTGCTCTTCCTGCTCTCAGGTATTAATGCCGAATTGTTTCATCTTGTCTGTGTAGAAGACCACACACGAAAATCCT GTGATTTTACATTTTACTTATCGTTAATCGAATGTATATCTATTTAATCTGCTTTTCTTGTCTAATAAATATATATGTAAAGTACGC TTTTTGTTGAAATTTTTTAAACCTTTGTTTATTTTTTTTTCTTCATTCCGTAACTCTTCTACCTTCTTTATTTACTTTCTAAAATC	Karas et al., 2015 [38]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
ShBle cassette	CAAATACAAAACATAAAAATAAATAAACACAGAGTAAATTCCCAAATTATTCCATCATTAAAAGATACGAGGCGCGTGTAAGTTAC AGGCAAGCGATCCTAGTACACTCTATATTTTTTTATGCCTCGGTAATGATTTTCATTTTTTTTTTCCACCTAGCGGATGACTCTTT TTTTTTCTTAGCGATTGGCATTATCACATAATGAATTATACATTATATAAAGTAATGTGATTTCTTCGAAGAATATACTAAAAAAT GAGCAGGCAAGATAAACGAAGGCAAAGATGACAGAGCAGAAAGCCCTAGTAAAGCGTATTACAAATGAAACCAAGATTCAGATTGC GATCTCTTTAAAGGGTGGTCCCCTAGCGATAGAGCACTCGATCTTCCCAGAAAAAGAGGCAGAAGCAGTAGCAGAACAGGCCACAC AATCGCAAGTGATTAACGTCCACACAGGTATAGGGTTTCTGGACCATATGATACATGCTCTGGCCAAGCATTCCGGCTGGTCGCTA ATCGTTGAGTGCATTGGTGACTTACACATAGACGACCATCACACCACTGAAGACTGCGGGATTGCTCTCGGTCAAGCTTTTAAAGA GGCCCTAGGGGCCGTGCGTGGAGTAAAAAGGTTTGGATCAGGATTTGCGCCTTTGGATGAGGCACTTTCCAGAGCGGTGGTAGATC TTTCGAACAGGCCGTACGCAGTTGTCGAACTTGGTTTGCAAAGGGAGAAAGTAGGAGATCTCTCTTGCGAGATGATCCCGCATTTT CTTGAAAGCTTTGCAGAGGCTAGCAGAATTACCCTCCACGTTGATTGTCTGCGAGGCAAGAATGATCATCACCGTAGTGAGAGTGC GTTCAAGGCTCTTGCGGTTGCCATAAGAGAAGCCACCTCGCCCAATGGTACCAACGATGTTCCCTCCACCAAAGGTGTTCTTATGT AGTTTTACACAGGAGTCTGGACTTGACGCTAGTGATAATAAGTGACTGAGGTATGTGCTCTTCTTATCTCCTTTTGTAGTGTTGCT CTTATTTTAAACAACTTTGCGGTTTTTTGATGACTTTGCGATTTTGTTGTTGCTTTGCAGTAAATTGCAAGATTTAATAAAAAAAC GCAAAGCAATGATTAAAGGATGTTCAGAATGAAACTCATGGAAACACTTAACCAGTGCATAAACGCTGGTCATGAAATGACGAAGG CTATCGCCATTGCACAGTTTAATGATGACAGCCCGGAAGCGAGGAAAATAACCCGGCGCTGGAGAATAGGTGAAGCAGCGGATTTA GTTGGGGTTTCTTCTCAGGCTATCAGAGATGCCGAGAAAGCAGGGCGACTACCGCACCCGGATATGGAAATTCGAGGACGGGTTGA GCAACGTGTTGGTTATACAATTGAACAAATTAATCATATGCGTGATGTGTTTGGTACGCGATTGCGACGTGCTGAAGACGTATTTC CACCGGTGATCGGGGTTGCTGCCCATAAAGGTGGCGTTTACAAAACCTCAGTTTCTGTTCATCTTGCTCAGGATCTGGCTCTGAAG GGGCTACGTGTTTTGCTCGTGGAAGGTAACGACCCCCAGGGAACAGCCTCAATGTATCACGGATGGGTACCAGATCTTCATATTCAT GCAGAAGACACTCTCCTGCCTTTCTATCTTGGGGAAAAGGACGATGTCACTTATGCAATAAAGCCCACTTGCTGGCCGGGGCTTGAC ATTATTCCTTCCTGTCTGGCTCTGCACCGTATTGAAACTGAGTTAATGGGCAAATTTGATGAAGGTAAACTGCCCACCGATCCACA CCTGATGCTCCGACTGGCCATTGAAACTGTTGCTCATGACTATGATGTCATAGTTATTGACAGCGCGCCTAACCTGGGTATCGGCAC GATTAATGTCGTATGTGCTGCTGATGTGCTGATTGTTCCCACGCCTGCTGAGTTGTTTGACTACACCTCCGCACTGCAGTTTTTCG ATATGCTTCGTGATCTGCTCAAGAACGTTGATCTTAAAGGGTTCGAGCCTGATGTACGTATTTTGCTTACCAAATACAGCAATAGC AATGGCTCTCAGTCCCCGTGGATGGAGGAGCAAATTCGGGATGCCTGGGGAAGCATGGTTCTAAAAAATGTTGTACGTGAAACGGA TGAAGTTGGTAAAGGTCAGATCCGGATGAGAACTGTTTTTGAACAGGCCATTGATCAACGCTCTTCAACTGGTGCCTGGAGAAATGC TCTTTCTATTTGGGAACCTGTCTGCAATGAAATTTTCGATCGTCTGATTAAACCACGCTGGGAGATTAGATAATGAAGCGTGCGCCT GTTATTCCAAAACATACGCTCAATACTCAACCGGTTGAAGATACTTCGTTATCGACACCAGCTGCCCCGATGGTGGATTCGTTAATT GCGCGCGTAGGAGTAATGGCTCGCGGTAATGCCATTACTTTGCCTGTATGTGGTCGGGATGTGAAGTTTACTCTTGAAGTGCTCCGG GGTGATAGTGTTGAGAAGACCTCTCGGGTATGGTCAGGTAATGAACGTGACCAGGAGCTGCTTACTGAGGACGCACTGGATGATCTC ATCCCTTCTTTTCTACTGACTGGTCAACAGACACCGGCGTTCGGTCGAAGAGTATCTGGTGTCATAGAAATTGCCGATGGGAGTCGC CGTCGTAAAGCTGCTGCACTTACCGAAAGTGATTATCGTGTTCTGGTTGGCGAGCTGGATGATGAGCAGATGGCTGCATTATCCAGA TTGGGTAACGATTATCGCCCAACAAGTGCTTATGAACGTGGTCAGCGTTATGCAAGCCGATTGCAGAATGAATTTGCTGGAAATATT TCTGCGCTGGCTGATGCGGAAAATATTTCACGTAAGATTATTACCCGCTGTATCAACACCGCCAAATTGCCTAAATCAGTTGTTGCTC TTTTTTCTCACCCCGGTGAACTATCTGCCCGGTCAGGTGATGCACTTCAAAAAGCCTTTACAGATAAAGAGGAATTACTTAAGCAGCA GGCATCTAACCTTCATGAGCAGAAAAAAGCTGGGGTGATATTTGAAGCTGAAGAAGTTATCACTCTTTTAACTTCTGTGCTTAAAACG TCATCTGCATCAAGAACTAGTTTAAGCTCACGACATCAGTTTGCTCCTGGAGCGACAGTATTGTATAAGGGCGATAAAATGGTGCTTA ACCTGGACAGGTCTCGTGTTCCAACTGAGTGTATAGAGAAAATTGAGGCCATTCTTAAGGAACTTGAAAAGCCAGCACCCTGATGCGAC CTCGTTTTAGTCTACGTTTATCTGTCTTTACTTAATGTCCTTTGTTACAGGCCAGAAAGCATAACTGGCCTGAATATTCTCTCTGGGC CCACTGTTCCACTTGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGAC	Karas et al., 2015 [38]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
ShBle cassette	CACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATAATCAGACTGGGACC ACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCATGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCA CGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGAACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCAC GGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGATCCCACTCGTGTTGTCGGTCTGATTATCGGTCTGGGACCACG GTCCCACTTGTATTGTCGATCAGACTATCAGCGTGAGACTACGATTCCATCAATGCCTGTCAAGGGCAAGTATTGACATGTCGTCGT AACCTGTAGAACGGAGTAACCTCGGTGTGCGGTTGTATGCCTGCTGTGGATTGCTGCTGTGTCCTGCTTATCCACAACATTTTGCGC ACGGTTATGTGGACAAAATACCTGGTTACCCAGGCCGTGCCGGCACGTTAACCGGGCTGCATCCGATGCAAGTGTGTCGCTGTCGAC GAGCTCGCGAGCTCGGACATGAGGTTGCCCCGTATTCAGTGTCGCTGATTTGTATTGTCTGAAGTTGTTTTTACGTTAAGTTGATGC AGATCAATTAATACGATACCTGCGTCATAATTGATTATTTGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGATGATAA TCATTATCACTTTACGGGTCCTTTCCGGTGATCCGACAGGTTACGGGGCGGCGACCTCGCGGGTTTTCGCTATTTATGAAAATTTTCC GGTTTAAGGCGTTTCCGTTCTTCTTCGTCATAACTTAATGTTTTTATTTAAAATACCCTCTGAAAAGAAAGGAAACGACAGGTGCTGA AAGCGAGCTTTTTGGCCTCTGTCGTTTCCTTTCTCTGTTTTTGTCCGTGGAATGAACAATGGAAGTCCGAGCTCATCGCTAATAAC TTCGTATAGCATACATTATACGAAGTTATATTCGATGCGGCCGCAAGGGGTTCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGC TTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCACACAGATGCGTAAGGAGAAAATAC CGCATCAGGCGCCATTCGCCATTCAGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCG AAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATT GTAATACGACTCACTATAGGGCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGAGTATT CTATAGTCTCACCTAAATAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAA CATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCG CTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGAACCCCTTGCGGCCGCCCGGGCCGTCGACCAA TTCTCATGTTTGACAGCTTATCATCGAATTTCTGCCATTCATCCGCTTATTATCACTTATTCAGGCGTAGCAACCAGGCGTTTAAG GGCACCAATAACTGCCTTAAAAAAATTACGCCCCGCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACA TGGAAGCCATCACAAACGGCATGATGAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGT GAAAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAA ACATATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAACTGCC GGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCC ATATCACCAGCTCACCGTCTTTCATTGCCATACGAAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGAATAAAGGCCGGA TAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACT GACTGAAATGCCTCAAAATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGCT TCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTTACG TGCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAG TGATCTTCCGTCACAGGTATTTATTCGCGATAAGCTCATGGAGCGGCGTAACCGTCGCACAGGAAGGACAGAGAAAGCGCGGATCTG GGAAGTGACGGACAGAACGGTCAGGACCTGGATTGGGGAGGCGGTTGCCGCCGCTGCTGCTGACGGTGTGACGTTCTCTGTTCCGGT CACACCACATACGTTCCGCCATTCCTATGCGATGCACATGCTGTATGCCGGTATACCGCTGAAAGTTCTGCAAAGCCTGATGGGACA TAAGTCCATCAGTTCAACGGAAGTCTACACGAAGGTTTTTGCGCTGGATGTGGCTGCCCGGCACCGGGTGCAGTTTGCGATGCCGGA GTCTGATGCGGTTGCGATGCTGAAACAATTATCCTGAGAATAAATGCCTTGGCCTTTATATGGAAATGTGGAACTGAGTGGATATGC TGTTTTTGTCTGTTAAACAGAGAAGCTGGCTGTTATCCACTGAGAAGCGAACGAAACAGTCGGGAAAATCTCCCATTATCGTAGAGA TCCGCATTATTAATCTCAGGAGCCTGTGTAGCGTTTATAGGAAGTAGTGTTCTGTCATGATGCCTGCAAGCGGTAACGAAAACGATT TGAATATGCCTTCAGGAACAATAGAAATCTTCGTGCGGTGTTACGTTGAAGTGGAGCGGATTATGTCAGCAATGGACAGAACAACCT AATGAACACAGAACCATGATGTGGTCTGTCCTTTTACAGCCAGTAGTGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCC	Karas et al., 2015 [38]

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
40SRPS8 promoter	CCCTGCGATAGACCTTTTCCAAACTCACGCAGTCCAAGAAAACAAAGGGGTGAGAAGTATACGCACCTTTCGGTTTCGGCATAATTCT TAAACTCTTGTGGTCACTTTCTTGTGAAGAAGCTAGGGGCACTCGTTTTCCCTCAGAGCCTGCAAACACAAAATTCCTGCAGTCAAT TGTCCCAACACTCGGCAAACCGTATGCGCAAGCAACGATGCGCAGAAGGCCGTGGATGGATGGCGACTCGCGATATGGCTTCTTGGG TCGCCAGTGTGGTACGTCCGGCGTATGTCAATACGCGAATTCGGACGACTGGCATCTCTAGGAGGAGGATTCCTTCTTTTATGACAT GTTTATTTTATATACATTGATGCTTTCCGACAGTCGGAAGTAATAAATGAATTTATTTCAAGACTACCTATACTCCTTTGACTTGTT CGACTAATCTTACCGCTTACTAAAATCTCGAAATCACGCTTGACCTCTCGCACGCAAATTTTTGCTGCTGGACGCTACGCACTCGGC CCAATTCTTCTCGGTCCTCGTCGTCGCAATTGTCGTTGCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACC	Slattery et al., 2018 [61]
FcpA terminator	CCGCAACAACTACCTCGACTTTGGCTGGGACACTTTCAGTGAGGACAAGAAGCTTCAGAAGCGTGCTATCGAACTCAACCAGGGACGTG CGGCACAAATGGGCATCCTTGCTCTCATGGTGCACGAACAGTTGGGAGTCTCTATCCTTCCTTAAAAATTTAATTTTCATTAGTTGCAGTCA- CTCCGCTTTGGTTT	Niu et al., 2013 [63]
OA1	ATGAACCACCTCCGTGCCGAAGGACCCGCCTCCGTCCTCGCCATTGGAACCGCCAACCCCGAAAACATTCTCCTCCAGGACGAATTCCC GGACTACTACTTCCGCGTCACCAAGTCGGAACACATGACCCAGCTCAAGGAAAAGTTCCGTAAGATCTGTGACAAGTCGATGATTCG TAAGCGCAACTGCTTCCTCAACGAAGAACACTTGAAGCAGAACCCCCGTCTCGTCGAACACGAAATGCAGACCCTCGACGCCCGTCA GGACATGCTCGTCGTCGAAGTCCCCAAGCTCGGAAAGGATGCCTGCGCCAAGGCCATCAAGGAATGGGGTCAGCCCAAGTCCAAGAT CACCCACCTCATTTTCACCTCCGCCTCCACCACCGACATGCCCGGAGCTGACTACCACTGCGCCAAGTTGCTCGGCCTCTCCCCCTC CGTCAAGCGCGTCATGATGTACCAGCTTGGCTGCTACGGCGGAGGTACCGTCTTGCGTATTGCCAAGGACATCGCCGAAAACAACAA GGGAGCCCGTGTCTTGGCCGTCTGCTGTGACATTATGGCCTGCCTCTTCCGTGGCCCCTCCGAATCCGACCTCGAACTCTTGGTCGGT CAGGCCATCTTTGGCGATGGAGCCGCCGCCGTCATTGTCGGTGCCGAACCCGACGAATCCGTCGGAGAACGTCCCATTTTTGAACTT GTCTCCACCGGACAGACCATTCTCCCCAACTCCGAAGGAACCATTGGCGGCCACATTCGTGAAGCCGGACTCATCTTTGACCTCCAC AAGGACGTCCCGATGCTCATCTCCAACAACATCGAAAAGTGCCTCATTGAAGCCTTCACTCCGATCGGAATTTCCGACTGGAACTCC ATCTTCTGGATTACCCACCCCGGCGGAAAGGCCATTTTGGACAAGGTCGAAGAAAAGCTCCACCTCAAGTCCGATAAGTTTGTCGAC TCCCGTCACGTCCTCTCCGAACACGGTAACATGTCCTCCTCCACCGTCCTCTTTGTCATGGACGAACTCCGTAAGCGTTCCTTGGAA GAAGGCAAGTCGACCACCGGCGACGGCTTCGAATGGGGAGTCTTGTTTGGCTTCGGTCCCGGATTGACCGTCGAACGTGTCGTCGTC CGTTCCGTTCCCATCAAGTACGGATCCGGAGAAGGACGCGGTTCGTTGCTCACCTGCGGAGACGTCGAAGAAAACCCCGGACCCATGG CCGTCAAGCACCTCATTGTCTTGAAGTTCAAGGATGAAATCACCGAAGCTCAGAAGGAAGAATTCTTCAAGACCTACGTCAACCTCG TCAACATCATTCCCGCCATGAAGGACGTCTACTGGGGAAAGGACGTCACCCAGAAGAACAAGGAAGAAGGATACACCCACATTGTCG AAGTCACCTTTGAATCCGTCGAAACCATCCAGGACTACATCATCCACCCCGCCCACGTCGGATTCGGTGACGTCTACCGCTCCTTCT GGGAAAAGCTCTTGATTTTCGACTACACCCCCCGTAAATAA	This study
OA2	ATGAACCACCTCCGTGCCGAAGGTCCCGCCTCCGTCCTCGCCATTGGAACCGCCAACCCCGAAAACATTCTCCTCCAAGACGAATTCCC TGACTACTACTTCCGCGTGACTAAGAGTGAGCACATGACGCAATTGAAAGAAAAGTTCCGTAAGATCTGTGATAAGTCCATGATCCG TAAGCGTAACTGTTTCCTGAATGAGGAACACCTTAAGCAGAACCCCCGCTTGGTCGAACATGAAATGCAGACCTTGGATGCCCGCCA GGACATGCTCGTCGTTGAAGTCCCTAAGTTGGGTAAAGACGCTTGCGCCAAGGCCATTAAGGAATGGGGACAGCCCAAGAGTAAGAT CACCCATTTGATTTTCACATCCGCCTCTACGACCGATATGCCTGGAGCCGATTACCACTGCGCCAAGCTTCTGGGACTGTCCCCCTC CGTTAAGCGCGTCATGATGTACCAGCTTGGATGTTACGGAGGTGGCACCGTCCTCCGTATTGCCAAGGACATTGCCGAAAACAACAA GGGTGCCCGTGTCCTCGCCGTCTGCTGCGACATCATGGCCTGCCTCTTCCGCGGACCCTCCGAATCCGACTTGGAACTCCTCGTGGG ACAGGCCATTTTTGGTGACGGCGCCGCCGCCGTCATTGTCGGAGCTGAACCCGACGAATCCGTCGGAGAAAGGCCCATTTTTGAATT GGTCTCCACCGGACAGACCATTCTCCCGAACTCCGAAGGAACCATCGGCGGACACATTCGTGAAGCCGGTCTCATCTTTGACTTGCA CAAGGACGTTCCCATGCTCATTTCCAACAACATTGAAAAGTGCCTCATTGAAGCCTTCACCCCCATTGGAATCTCCGACTGGAATAG TATCTTTTGGATTACCCACCCCGGAGGAAAGGCCATCTTGGATAAGGTGGAAGAAAAGCTCCACCTCAAGTCCGACAAGTTCGTCGA	This study

Table A2. *Cont.*

Name	Nucleotide Sequence (Codon Optimized)	Description and/or References
OA2	CTCCCGTCACGTCTTGTCCGAACACGGAAACATGTCGTCCTCCACCGTCCTCTTTGTCATGGATGAATTGCGTAAGCGCTCGCTCGA AGAAGGAAAGTCCACCACCGGCGATGGTTTCGAATGGGGCGTTCTCTTCGGATTTGGACCCGGTCTCACCGTCGAACGTGTCGTCGTC CGTTCTGTCCCGATTAAGTACCACCATCACCACCACCACGGATCGGGCGAAGGACGTGGTTCCCTCCTTACGTGCGGAGACGTCGAA GAAAACCCCGGTCCCATGGCCGTCAAGCACTTGATCGTTCTCAAGTTCAAGGACGAAATCACCGAAGCCCAGAAGGAAGAATTCTTC AAGACCTACGTCAACCTCGTCAACATCATTCCCGCCATGAAGGATGTTTACTGGGGAAAGGACGTCACCCAGAAGAACAAGGAAGAA GGTTACACCCACATCGTCGAAGTCACCTTTGAATCCGTTGAAACCATCCAGGACTACATTATCCACCCGGCCCACGTCGGCTTCGGA GACGTCTACCGTTCCTTTTGGGAAAAGTTGTTGATTTTCGACTACACCCCCCGTAAGGAACAGAAGTTGATTTCCGAAGAAGACCTCTAA	This study
OA3	ATGCGTAAGGGTGAAGAACTCTTCACCGGAGTCGTCCCCATCTTGGTGGAATTGGACGGAGACGTCAACGGCCATAAGTTCTCGGTT TCCGGTGAAGGTGAAGGTGACGCCACTAACGGAAAGCTCACCCTCAAGTTCATTTGCACCACCGGCAAGCTCCCGGTCCCCTGGCCC ACCCTCGTCACCACCTTGACTTACGGTGTCCAGTGCTTCGCCCGTTACCCCGACCACATGAAGCAGCACGACTTCTTTAAGTCCGCC ATGCCCGAAGGATACGTCCAGGAACGCACCATTTCCTTCAAGGACGACGGAACCTACAAGACCCGTGCCGAAGTCAAGTTTGAAGGAG ATACCCTCGTCAACCGCATTGAACTCAAGGGCATTGATTTCAAAGAAGATGGCAACATTCTCGGACACAAGCTCGAATACAACTTCAA CTCCCACAACGTCTACATCACCGCCGATAAGCAAAAGAACGGTATTAAGGCCAACTTCAAGATTCGCCACAACGTCGAAGACGGTTC CGTCCAGTTGGCCGACCACTACCAGCAGAACACCCCGATTGGTGACGGTCCGGTCCTCCTCCCCGACAACCACTACCTCTCGTACCAG TCCGCCCTCTCCAAGGATCCGAACGAGAAGCGTGACCACATGGTCCTCTTGGAATTCGTCACCGCCGCCGGAATCACCCACGGCATGG ACGAACTCTACAAGCGTCCCGCCGCCAACGACGAGAACTACGCCGCCTCCGTCGGTATGAACCACTTGCGCGCCGAAGGACCGGCTTC CGTCTTGGCCATTGGTACCGCCAACCCGGAAAACATCCTCCTCCAGGACGAATTCCCCGACTACTACTTCCGTGTCACCAAGTCCGAA CACATGACCCAGTTGAAGGAAAAATTCCGTAAGATTTGTGATAAGTCCATGATTCGTAAGCGTAACTGCTTCTTGAACGAAGAACACC TCAAGCAAAACCCCCGTCTCGTCGAACACGAAATGCAAACCCTCGACGCCCGTCAGGACATGCTCGTTGTCGAAGTCCCGAAGCTTGG CAAGGACGCCTGCGCCAAGGCCATTAAGGAATGGGGCCAGCCCAAGAGCAAGATTACCCACTTGATTTTCACCTCCGCCTCCACCACC GATATGCCCGGCGCTGACTACCACTGCGCTAAGCTCCTCGGTCTCTCCCCCTCCGTCAAGCGTGTCATGATGTACCAGCTCGGATGCT ACGGAGGTGGAACCGTCCTCCGCATTGCCAAGGACATTGCTGAAAACAACAAGGGCGCCCGTGTCCTCGCCGTCTGCTGTGACATTAT GGCCTGCTTGTTTCGTGGTCCCTCCGAATCCGATTTGGAACTCCTCGTCGGCCAGGCCATTTTTGGTGACGGTGCCGCCGCCGTCATC GTCGGCGCCGAACCGGACGAATCCGTCGGTGAACGTCCGATTTTTGAACTCGTCTCCACCGGTCAAACCATTCTCCCCAACTCCGAAG GAACCATCGGAGGACACATCCGTGAAGCCGGACTCATTTTTGACCTCCACAAGGATGTCCCCATGCTCATCTCCAACAACATCGAAAA GTGCCTCATCGAAGCCTTCACCCCCATTGGCATCTCCGACTGGAACTCCATTTTCTGGATTACCCACCCCGGTGGAAAGGCCATTCTC GACAAGGTCGAAGAAAAGCTCCACTTGAAGTCCGATAAGTTCGTCGACTCCCGCCACGTCCTCTCCGAACACGGAAACATGTCCTCCT CCACCGTCCTCTTCGTCATGGACGAACTCCGTAAGCGTTCCCTCGAAGAAGGAAAGTCCACGACGGGTGACGGATTCGAGTGGGGAGT CCTCTTCGGCTTCGGTCCCGGACTCACCGTCGAACGTGTCGTCGTCCGTTCCGTGCCCATCAAGTACGGAGGAGGAGGATCCGGTGGT GGCGGATCCGGAGGTGGTGGATCGATGGCCGTCAAGCACCTCATTGTCTTGAAGTTTAAGGACGAAATCACCGAAGCCCAGAAGGAAG AATTCTTCAAGACTTACGTCAACTTGGTCAACATCATCCCCGCCATGAAGGACGTCTACTGGGGAAAGGACGTCACCCAAAAGAACAA GGAAGAAGGATACACCCACATTGTCGAAGTCACCTTTGAATCCGTCGAAACCATTCAGGACTACATTATTCACCCCGCCCACGTCGGTT TCGGTGACGTCTACCGTTCCTTCTGGGAAAAGCTCTTGATTTTTGACTACACCCCCCGTAAGGAACAGAAGTTGATTTCCGAAGAAGACTTGTAA	This study

Table A3. List of the primers used in this study for the *P. tricornutum* episome, with the design of and description of the obtained product.

Primer Name	Sequence (5′ to 3′)	Amplicon Size (bp)	Description of Product
bk1-FP	CAGGGTAATATAGATCTTCCGCTGCATAACCCTGCTTCGGGGTCATTATAGCGATTTTTT	6158	pPTGE30 backbone part 1
bk1-RP	TTTGCAAACCAAGTTCGACAACTGCGTACGGCCTGTTCGAAAGATCTACCACCGCTCTGG		
bk1-FP	GGGCCGTGCGTGGAGTAAAAAGGTTTGGATCAGGATTTGCGCCTTTGGATGAGGCACTTT	6318	pPTGE30 backbone part 2
bk2-RP	ATGGGCTTCGCCCTGTCGCTCGACTGCGGCGAGCACTACTGGCTGTAAAAGGACAGACCA		
ShBle-FP	TGGTCTGTCCTTTTACAGCCAGTAGTGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCCATGAGCACAAGAGGTGACAAAA	1466	*P. t* selection marker region
ShBle-RP	ATACTTCTCACCCCTTTGTTTTCTTGGACTGCGTGAGTTTGGAAAAGGTCTATCGCAGGGATTGCAGCTTGTTGCAGAAG		
40SRPS8-FP	TTGCATGGTTAGACCTCCTTGACGACTGTGAGCCTACATCCTTCTGCAACAAGCTGCAATCCCTGCGATAGACCTTTTCC	700	40SRPS8 Promoter region
40SRPS8-RP	CGAGGACGGAGGCGGGTCCTTCGGCACGGAGGTGGTTCATGGTATTCTATTCTCTGATTC		
FcpA terminator-FP	GGAAAAGCTCTTGATTTTCGACTACACCCCCCGTAAATAACCGCAACAACTACCTCGACT	275	FcpA Terminator region
FcpA terminator-RP	CCGAAGCAGGGTTATGCAGCGGAAGATCTATATTACCCTGAAACCAAAGCGGAGTGACTG		
OA1-FP	GCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACCATGAACCACCTCCGTGCCGA	1524	OA1 pathway genes
OA1-RP	ACTGAAAGTGTCCCAGCCAAAGTCGAGGTAGTTGTTGCGGTTATTTACGGGGGGTGTAGT		
OA2-FP	GCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACCATGAACCACCTCCGTGCCGA	1572	OA2 pathway genes
OA2-RP	AGTCGAGGTAGTTGTTGCGGTTAGAGGTCTTCTTCGGAAATCAACTTCTGTTCCTTACGG		
OA3-FP	GCGTTGATCTTGCACCGAAGGAATCAGAGAATAGAATACCATGCGTAAGGGTGAAGAACT	2352	OA3 pathway genes
OA3-RP	AGTCGAGGTAGTTGTTGCGGTTACAAGTCTTCTTCGGAAATCAACTTCTGTTCCTTACGG		
Q-TKS1-Gn2-F	CGA CTG GAA CTC CAT CTT CTG	527	Pt OA1 colony PCR
Q-OAC1-Gn2-R	GGGTGTATCCTTCTTCCTTGTT		
Q-TKS3-Gn2-F	ATTGGAATCTCCGACTGGAATAG	594	Pt OA2 colony PCR
Q-OAC3-Gn2-R	GATGGTTTCAACGGATTCAAAGG		
Q-TKS7-Gn2-F	GTCAGGACATGCTCGTTGT	1211	Pt OA3 colony PCR
Q-OAC7-Gn2-R	TTCCCAGAAGGAACGGTAGA		

P. t: Phaeodactylum tricornutum; FcpA: fucoxanthin-chlorophyll binding protein.

Table A4. Olivetolic acid biosynthesis construct analysis. The plasmid constructs were rescued from positive *P. tricornutum*-transformed clones and amplified in *E. coli* before being extracted and fully sequenced at the Massachusetts General Hospital (MGH), Center for Computational and Integrative Biology (CCIB) DNA Core, using Illumina MiSeq technology. ON: original nucleotide; SN substitute nucleotide.

Sequence ID	Identity	Mismatch Position (ON-SN)	Effect on Protein	Number of Insertion
*Pt*OA1	>99%	9305 (T-C)	No	60 bp insertion (after terminator, no effect)
*Pt*OA2	>99%	8526 (C-T) 9304 (T-C)	No	0
*Pt*OA3	>99%	8192 (G-A) 9304 (T-C)	No	0

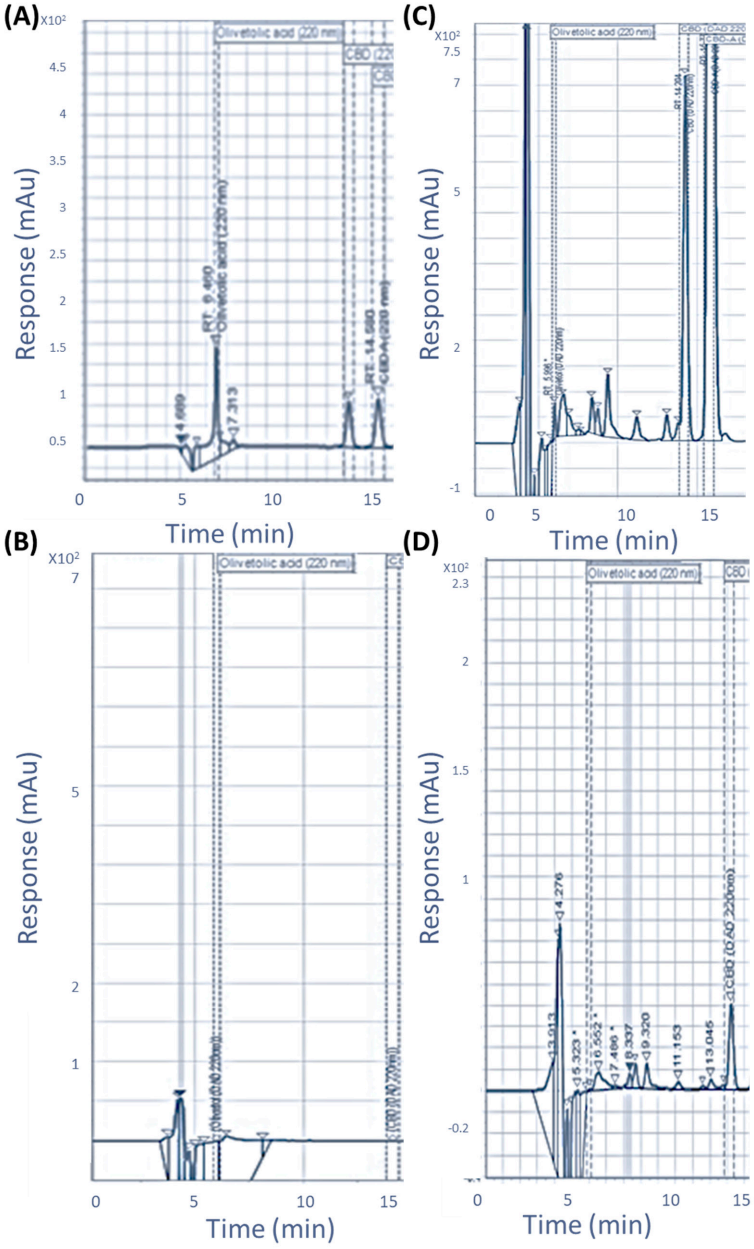
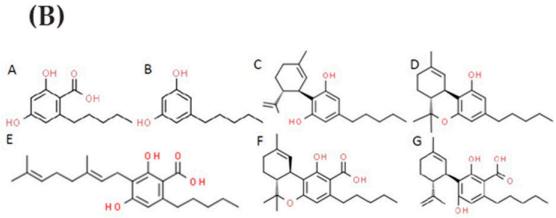
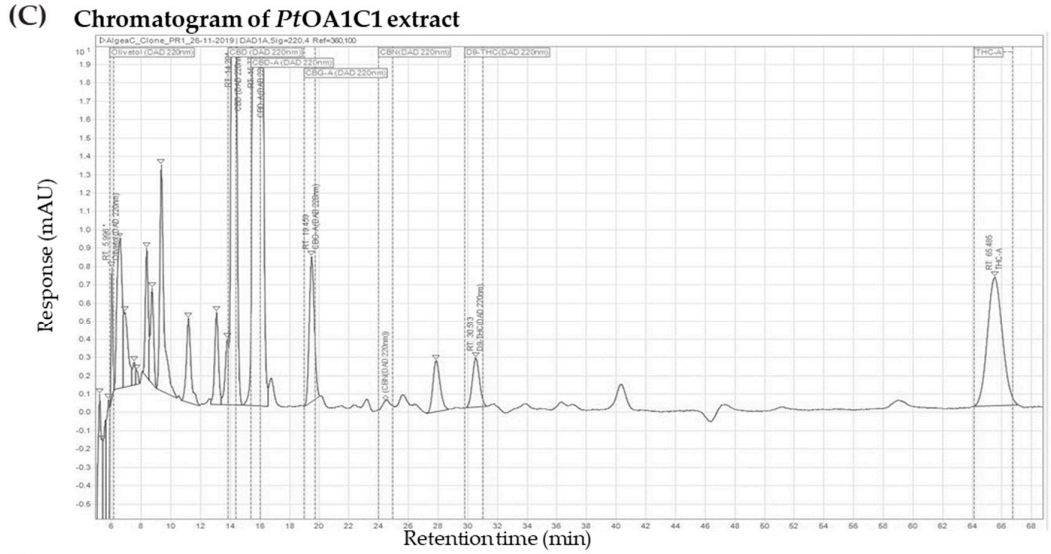


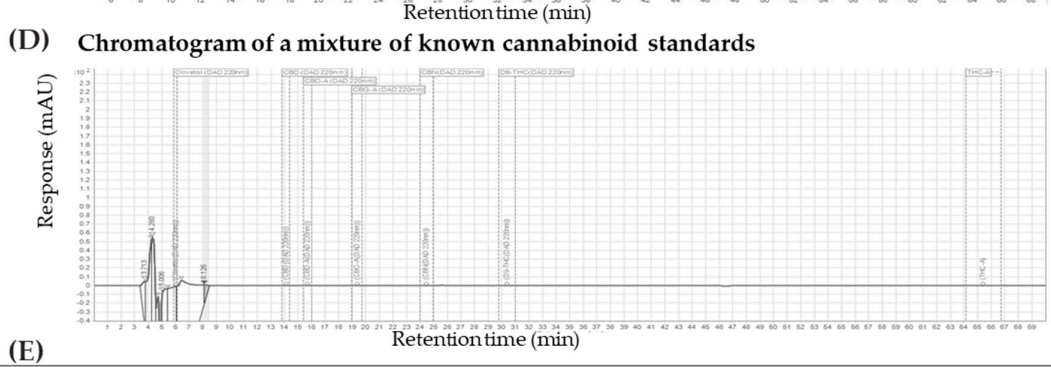
Figure A3. Olivetolic acid detection by HPLC-UV. Curves obtained by the injection of 10 µL of each sample. A concentration of 10 mg/kg of commercial standards (**A**); metabolite extraction from the *P. tricornutum* wild type (**B**); *Pt*OA1 (**C**) and *Pt*OA2 (**D**).

(A)

Compound	Molecular formula	Structure assignment	Parent m/z
Olivetolic acid	$C_{12}H_{16}O_4$	A	225.104
Olivetol	$C_{11}H_{16}O_2$	B	181.115
Cannabidiol	$C_{21}H_{30}O_2$	C	315.224
Δ-9-Tetrahydrocannabinol	$C_{21}H_{30}O_2$	D	315.224
Cannabigerolic acid	$C_{22}H_{32}O_4$	E	361.230
Δ-9-Tetrahydrocannabinolic acid	$C_{22}H_{30}O_4$	F	319.214
Cannabidiolic acid	$C_{22}H_{30}O_4$	G	319.214







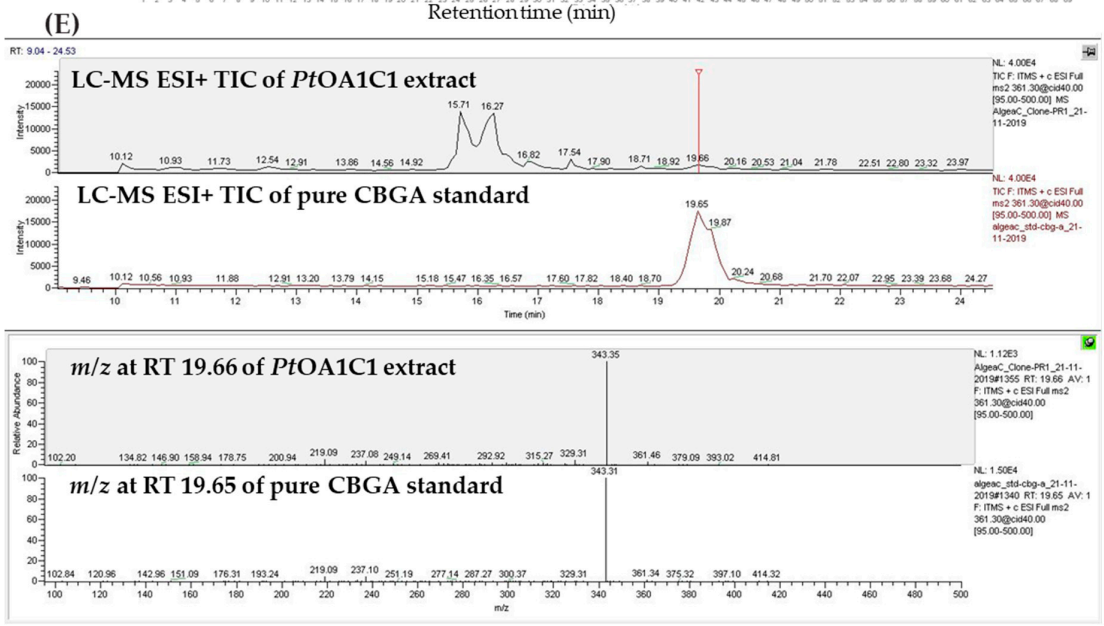


Figure A4. *Cont.*

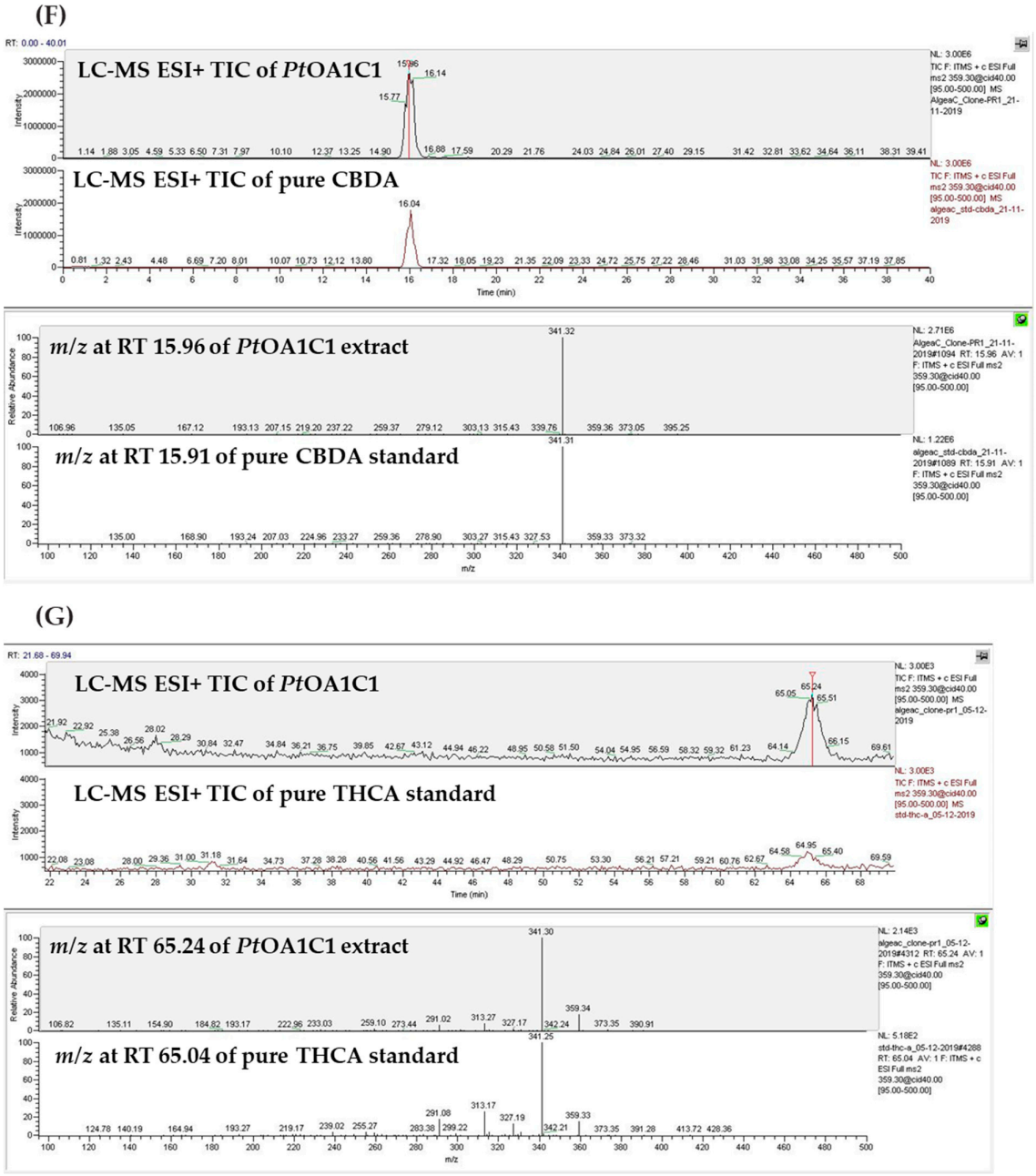


Figure A4. The cannabinoid-like metabolites detected via HPLC-UV and LC-MS analyses. (**A**) The list and (**B**) chemical structure of the targeted metabolites. A representative LC chromatogram of (**C**) 10 µL of metabolite extraction from *P. tricornutum Pt*OA1C1 and of (**D**) 10 mg/kg of a mixture of known cannabinoid commercial standards, both showing peaks corresponding to the retention times (RT) of cannabinoid standards. LC-MS in ESI+ analyses. Total ion chromatograph (TIC) of selected peaks from the corresponding cannabinoids, such as *P. tricornutum Pt*OA1C1 (**E**) peak at RT 19.66 min, with an m/z similar to the CBGA standard (RT 19.65 min). (**F**) The peak at RT 15.96 min, with an m/z similar to the CBDA standard (RT 15.91 min). (**G**) The peak at RT 19.66 min, with an m/z similar to the THCA standard (RT 65.04 min).

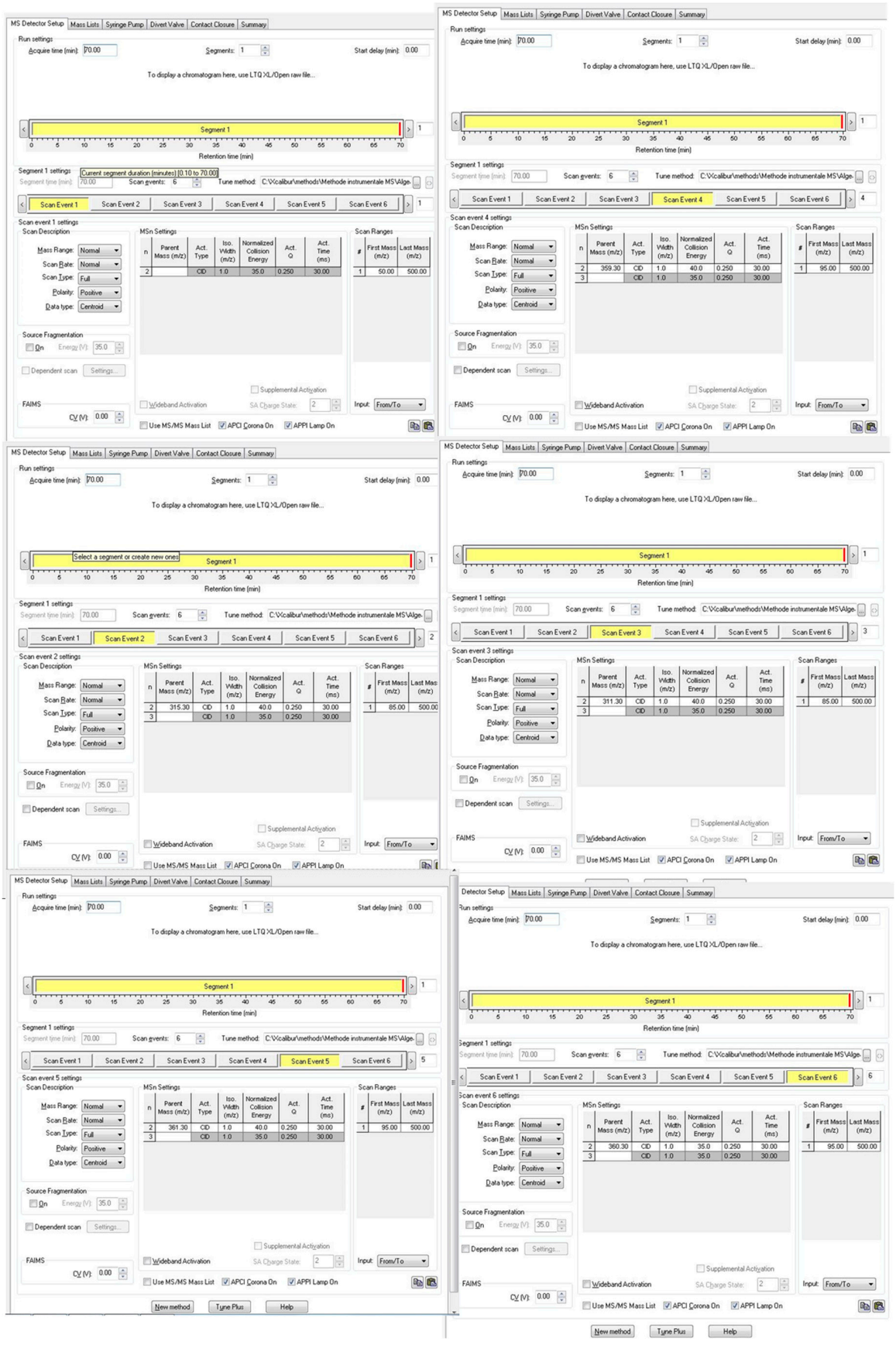


Figure A5. Details of the events used in LC-MS to confirm the identity of the cannabinoids.

Appendix B. The In Silico Analyses Identified Putative Endogenous *P. tricornutum* Candidates for CB-like Biosynthesis

To investigate the hypothesis that the endogenous enzyme could yield cannabinoids, characterization of the enzyme candidates was performed using bioinformatic tools. Putative candidate sequences for endogenous *P. tricornutum* enzymes that were able to convert OA and GPP into CBGA were obtained from the HMMER web server [64], using the CsAPT4 amino acid sequence as a query against the Protist database. We obtained five hits where the protein B7G488 (gene ID Phatr3_J37858) (57% similarity) and the protein B7FT51 (gene ID: Phatr3_J2738.t1) (51% similarity) had been annotated as the predicted protein members of the UbiA prenyltransferase family, to which CsAPT4 belongs [49]. The phylogenetic analysis indicated that these two proteins are clustered closer, according to sequence similarity, to the *C. sativa* APTs group (Appendix A, Figure A5). The soluble aromatic prenyltransferase from *Streptomyces* sp., NphB, had already been characterized by Kuzuyama et al. in 2005 [65], in a cluster with that from *Aspergillus terreus*, a fungal APT. No hits were found by using the same approach to identify endogenous diatom-soluble APT candidates, using *Streptomyces* sp. NphB as a query. Endogenous *P. tricornutum* candidate enzymes that were similar to THCA/CBDA synthase were not found in the database. One putative hypothetical candidate (with a 25.50% sequence identity compared to THCAS) from the diatom database was the violaxanthin deepoxidase-like protein (VDL), identified for its putative function in pigment deepoxidation and its requirement of a specific pH, which is similar to THCAS.

Figure A6. The phylogenetic tree of a putative candidate of endogenous *P. tricornutum* enzymes (in orange), which is similar to that of APT from cannabis and other species.

References

1. Aizpurua-Olaizola, O.; Soydaner, U.; Öztürk, E.; Schibano, D.; Simsir, Y.; Navarro, P.; Etxebarria, N.; Usobiaga, A. Evolution of the Cannabinoid and Terpene Content during the Growth of *Cannabis sativa* Plants from Different Chemotypes. *J. Nat. Prod.* **2016**, *79*, 324–331. [CrossRef] [PubMed]
2. Borrelli, F.; Pagano, E.; Romano, B.; Panzera, S.; Maiello, F.; Coppola, D.; De Petrocellis, L.; Buono, L.; Orlando, P.; Izzo, A.A. Colon Carcinogenesis Is Inhibited by the TRPM8 Antagonist Cannabigerol, a Cannabis-Derived Non-Psychotropic Cannabinoid. *Carcinogenesis* **2014**, *35*, 2787–2797. [CrossRef] [PubMed]
3. Chakravarti, B.; Ravi, J.; Ganju, R.K. Cannabinoids as Therapeutic Agents in Cancer: Current Status and Future Implications. *Oncotarget* **2014**, *5*, 5852–5872. [CrossRef] [PubMed]
4. Eibach, L.; Scheffel, S.; Cardebring, M.; Lettau, M.; Özgür Celik, M.; Morguet, A.; Roehle, R.; Stein, C. Cannabidivarin for HIV-Associated Neuropathic Pain: A Randomized, Blinded, Controlled Clinical Trial. *Clin. Pharmacol. Ther.* **2021**, *109*, 1055–1062. [CrossRef] [PubMed]
5. Brunt, T.M.; Bossong, M.G. The Neuropharmacology of Cannabinoid Receptor Ligands in Central Signaling Pathways. *Eur. J. Neurosci.* **2022**, *55*, 909–921. [CrossRef] [PubMed]

6. Schachtsiek, J.; Warzecha, H.; Kayser, O.; Stehle, F. Current Perspectives on Biotechnological Cannabinoid Production in Plants. *Planta Medica* **2018**, *84*, 214–220. [CrossRef] [PubMed]
7. Cohen, K.; Weinstein, A. The Effects of Cannabinoids on Executive Functions: Evidence from Cannabis and Synthetic Cannabinoids—A Systematic Review. *Brain Sci.* **2018**, *8*, 40. [CrossRef] [PubMed]
8. Vadivelu, N.; Kai, A.M.; Kodumudi, G.; Sramcik, J.; Kaye, A.D. Medical Marijuana: Current Concepts, Pharmacological Actions of Cannabinoid Receptor Mediated Activation, and Societal Implications. *Curr. Pain Headache Rep.* **2018**, *22*, 3. [CrossRef]
9. Gagne, S.J.; Stout, J.M.; Liu, E.; Boubakir, Z.; Clark, S.M.; Page, J.E. Identification of Olivetolic Acid Cyclase from *Cannabis sativa* Reveals a Unique Catalytic Route to Plant Polyketides. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 12811–12816. [CrossRef]
10. Valliere, M.A.; Korman, T.P.; Woodall, N.B.; Khitrov, G.A.; Taylor, R.E.; Baker, D.; Bowie, J.U. A Cell-Free Platform for the Prenylation of Natural Products and Application to Cannabinoid Production. *Nat. Commun.* **2019**, *10*, 565. [CrossRef]
11. Garfinkel, A.R.; Otten, M.; Crawford, S. SNP in Potentially Defunct Tetrahydrocannabinolic Acid Synthase Is a Marker for Cannabigerolic Acid Dominance in *Cannabis sativa* L. *Genes* **2021**, *12*, 228. [CrossRef] [PubMed]
12. Sirikantaramas, S.; Taura, F.; Tanaka, Y.; Ishikawa, Y.; Morimoto, S.; Shoyama, Y. Tetrahydrocannabinolic Acid Synthase, the Enzyme Controlling Marijuana Psychoactivity, Is Secreted into the Storage Cavity of the Glandular Trichomes. *Plant Cell Physiol.* **2005**, *46*, 1578–1582. [CrossRef]
13. Morita: How Structural Subtleties Lead to Molecular...—Google Scholar. Available online: https://scholar.google.com/scholar_lookup?title=How%20structural%20subtleties%20lead%20to%20molecular%20diversity%20for%20the%20type%20III%20polyketide%20synthases&journal=J%20Biol%20Chem&doi=10.1074%2Fjbc.REV119.006129&volume=294&pages=15121-15136&publication_year=2019&author=Morita%2CH&author=Wong%2CCP&author=Abe%2CI (accessed on 8 August 2023).
14. Abe, I. Biosynthesis of Medicinally Important Plant Metabolites by Unusual Type III Polyketide Synthases. *J. Nat. Med.* **2020**, *74*, 639–646. [CrossRef] [PubMed]
15. Pandith, S.A.; Ramazan, S.; Khan, M.I.; Reshi, Z.A.; Shah, M.A. Chalcone Synthases (CHSs): The Symbolic Type III Polyketide Synthases. *Planta* **2019**, *251*, 15. [CrossRef] [PubMed]
16. Nivina, A.; Yuet, K.P.; Hsu, J.; Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem. Rev.* **2019**, *119*, 12524–12547. [CrossRef] [PubMed]
17. Tian, B.; Liu, J. Resveratrol: A Review of Plant Sources, Synthesis, Stability, Modification and Food Application. *J. Sci. Food Agric.* **2020**, *100*, 1392–1404. [CrossRef] [PubMed]
18. Degenhardt, F.; Stehle, F.; Kayser, O. The Biosynthesis of Cannabinoids. In *Handbook of Cannabis and Related Pathologies*; Elsevier: London, UK, 2017; pp. 13–23.
19. Gaoni, Y.; Mechoulam, R. Isolation, Structure, and Partial Synthesis of an Active Constituent of Hashish. *J. Am. Chem. Soc.* **1964**, *86*, 1646–1647. [CrossRef]
20. Welling, M.T.; Deseo, M.A.; Bacic, A.; Doblin, M.S. Biosynthetic Origins of Unusual Cannabimimetic Phytocannabinoids in *Cannabis sativa* L: A Review. *Phytochemistry* **2022**, *201*, 113282. [CrossRef]
21. Taura, F.; Tanaya, R.; Sirikantaramas, S. Recent Advances in Cannabinoid Biochemistry and Biotechnology. *ScienceAsia* **2019**, *45*, 399. [CrossRef]
22. Livingston, S.J.; Quilichini, T.D.; Booth, J.K.; Wong, D.C.; Rensing, K.H.; Laflamme-Yonkman, J.; Castellarin, S.D.; Bohlmann, J.; Page, J.E.; Samuels, A.L. Cannabis Glandular Trichomes Alter Morphology and Metabolite Content during Flower Maturation. *Plant J.* **2020**, *101*, 37–56. [CrossRef]
23. Günnewich, N.; Page, J.E.; Köllner, T.G.; Degenhardt, J.; Kutchan, T.M. Functional Expression and Characterization of Trichome-Specific (-)-Limonene Synthase and (+)-α-Pinene Synthase from *Cannabis Sativa*. *Nat. Prod. Commun.* **2007**, *2*, 223–232.
24. Grof, C.P. Cannabis, from Plant to Pill. *Br. J. Clin. Pharmacol.* **2018**, *84*, 2463–2467. [CrossRef]
25. Fellermeier, M.; Zenk, M.H. Prenylation of Olivetolate by a Hemp Transferase Yields Cannabigerolic Acid, the Precursor of Tetrahydrocannabinol. *FEBS Lett.* **1998**, *427*, 283–285. [CrossRef]
26. Frontiers | Cannabis sativa: The Plant of the Thousand and One Molecules | Plant Science. Available online: https://www.frontiersin.org/articles/10.3389/fpls.2016.00019/full (accessed on 28 January 2022).
27. Balthazar, C.; Novinscak, A.; Cantin, G.; Joly, D.L.; Filion, M. Biocontrol Activity of *Bacillus* spp. and *Pseudomonas* spp. against Botrytis Cinerea and Other Cannabis Fungal Pathogens. *Phytopathology* **2022**, *112*, 549–560. [CrossRef]
28. Poulos, J.L.; Farnia, A.N. Production of Cannabidiolic Acid in Yeast 2018. U.S. Patent US20180073043A1, 15 March 2018.
29. Poulos, J.L.; Farnia, A.N. Production of Tetrahydrocannabinolic Acid in Yeast 2019. U.S. Patent US10392635B2, 27 August 2019.
30. Thomas, F.; Schmidt, C.; Kayser, O. Bioengineering Studies and Pathway Modeling of the Heterologous Biosynthesis of Tetrahydrocannabinolic Acid in Yeast. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 9551–9563. [CrossRef]
31. Luo, X.; Reiter, M.A.; d'Espaux, L.; Wong, J.; Denby, C.M.; Lechner, A.; Zhang, Y.; Grzybowski, A.T.; Harth, S.; Lin, W. Complete Biosynthesis of Cannabinoids and Their Unnatural Analogues in Yeast. *Nature* **2019**, *567*, 123–126. [CrossRef]
32. Kufs, J.E.; Reimer, C.; Steyer, E.; Valiante, V.; Hillmann, F.; Regestein, L. Scale-up of an Amoeba-Based Process for the Production of the Cannabinoid Precursor Olivetolic Acid. *Microb. Cell Fact.* **2022**, *21*, 217. [CrossRef]
33. Tan, Z.; Clomburg, J.M.; Gonzalez, R. Synthetic Pathway for the Production of Olivetolic Acid in *Escherichia coli*. *ACS Synth. Biol.* **2018**, *7*, 1886–1896. [CrossRef]
34. Zirpel, B.; Degenhardt, F.; Martin, C.; Kayser, O.; Stehle, F. Engineering Yeasts as Platform Organisms for Cannabinoid Biosynthesis. *J. Biotechnol.* **2017**, *259*, 204–212. [CrossRef]

35. Dhaouadi, F.; Awwad, F.; Diamond, A.; Desgagné-Penix, I. Diatoms' Breakthroughs in Biotechnology: *Phaeodactylum tricornutum* as a Model for Producing High-Added Value Molecules. *Am. J. Plant Sci.* **2020**, *11*, 1632–1670. [CrossRef]
36. Fabris, M.; George, J.; Kuzhiumparambil, U.; Lawson, C.A.; Jaramillo-Madrid, A.C.; Abbriano, R.M.; Vickers, C.E.; Ralph, P. Extrachromosomal Genetic Engineering of the Marine Diatom *Phaeodactylum tricornutum* Enables the Heterologous Production of Monoterpenoids. *ACS Synth. Biol.* **2020**, *9*, 598–612. [CrossRef]
37. Kassaw, T.K.; Paton, A.J.; Peers, G. Episome-Based Gene Expression Modulation Platform in the Model Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **2022**, *11*, 191–204. [CrossRef]
38. Karas, B.J.; Diner, R.E.; Lefebvre, S.C.; McQuaid, J.; Phillips, A.P.; Noddings, C.M.; Brunson, J.K.; Valas, R.E.; Deerinck, T.J.; Jablanovic, J. Designer Diatom Episomes Delivered by Bacterial Conjugation. *Nat. Commun.* **2015**, *6*, 6925. [CrossRef]
39. Karas, B.J.; Molparia, B.; Jablanovic, J.; Hermann, W.J.; Lin, Y.-C.; Dupont, C.L.; Tagwerker, C.; Yonemoto, I.T.; Noskov, V.N.; Chuang, R.-Y. Assembly of Eukaryotic Algal Chromosomes in Yeast. *J. Biol. Eng.* **2013**, *7*, 30. [CrossRef]
40. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; Kothari, A.; Keseler, I.M.; Krummenacker, M.; Midford, P.E.; Ong, Q. The BioCyc Collection of Microbial Genomes and Metabolic Pathways. *Brief. Bioinform.* **2019**, *20*, 1085–1093. [CrossRef]
41. Athey, J.; Alexaki, A.; Osipova, E.; Rostovtsev, A.; Santana-Quintero, L.V.; Katneni, U.; Simonyan, V.; Kimchi-Sarfaty, C. A New and Updated Resource for Codon Usage Tables. *BMC Bioinform.* **2017**, *18*, 391. [CrossRef]
42. Diamond, A.; Diaz-Garza, A.M.; Li, J.; Slattery, S.S.; Merindol, N.; Fantino, E.; Meddeb-Mouelhi, F.; Karas, B.J.; Barnabé, S.; Desgagné-Penix, I. Instability of Extrachromosomal DNA Transformed into the Diatom *Phaeodactylum tricornutum*. *Algal Res.* **2023**, *70*, 102998. [CrossRef]
43. Li-Beisson, Y.; Thelen, J.J.; Fedosejevs, E.; Harwood, J.L. The Lipid Biochemistry of Eukaryotic Algae. *Prog. Lipid Res.* **2019**, *74*, 31–68. [CrossRef]
44. Liu, B.; Benning, C. Lipid Metabolism in Microalgae Distinguishes Itself. *Curr. Opin. Biotechnol.* **2013**, *24*, 300–309. [CrossRef]
45. Chng, J.; Wang, T.; Nian, R.; Lau, A.; Hoi, K.M.; Ho, S.C.; Gagnon, P.; Bi, X.; Yang, Y. Cleavage Efficient 2A Peptides for High Level Monoclonal Antibody Expression in CHO Cells. *MAbs* **2015**, *7*, 403–412. [CrossRef]
46. Liu, Z.; Chen, O.; Wall, J.B.J.; Zheng, M.; Zhou, Y.; Wang, L.; Ruth Vaseghi, H.; Qian, L.; Liu, J. Systematic Comparison of 2A Peptides for Cloning Multi-Genes in a Polycistronic Vector. *Sci. Rep.* **2017**, *7*, 2193. [CrossRef]
47. Ma, J.; Gu, Y.; Xu, P. Biosynthesis of Cannabinoid Precursor Olivetolic Acid in Genetically Engineered Yarrowia Lipolytica. *Commun. Biol.* **2022**, *5*, 1239. [CrossRef]
48. George, J.; Kahlke, T.; Abbriano, R.M.; Kuzhiumparambil, U.; Ralph, P.J.; Fabris, M. Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations with Implications for Algal Genetics and Synthetic Biology. *Front. Bioeng. Biotechnol.* **2020**, *8*, 513. [CrossRef]
49. Munakata, R.; Kitajima, S.; Nuttens, A.; Tatsumi, K.; Takemura, T.; Ichino, T.; Galati, G.; Vautrin, S.; Bergès, H.; Grosjean, J.; et al. Convergent Evolution of the UbiA Prenyltransferase Family Underlies the Independent Acquisition of Furanocoumarins in Plants. *New Phytol.* **2020**, *225*, 2166–2182. [CrossRef]
50. Wiles, D.; Shanbhag, B.K.; O'Brien, M.; Doblin, M.S.; Bacic, A.; Beddoe, T. Heterologous Production of *Cannabis sativa*-Derived Specialised Metabolites of Medicinal Significance—Insights into Engineering Strategies. *Phytochemistry* **2022**, *203*, 113380. [CrossRef]
51. Chuberre, C.; Chan, P.; Walet-Balieu, M.-L.; Thiébert, F.; Burel, C.; Hardouin, J.; Gügi, B.; Bardor, M. Comparative Proteomic Analysis of the Diatom *Phaeodactylum tricornutum* Reveals New Insights Into Intra- and Extra-Cellular Protein Contents of Its Oval, Fusiform, and Triradiate Morphotypes. *Front. Plant Sci.* **2022**, *13*, 673113. [CrossRef]
52. Hussain, M.H.; Hong, Q.; Zaman, W.Q.; Mohsin, A.; Wei, Y.; Zhang, N.; Fang, H.; Wang, Z.; Hang, H.; Zhuang, Y.; et al. Rationally Optimized Generation of Integrated *Escherichia coli* with Stable and High Yield Lycopene Biosynthesis from Heterologous Mevalonate (MVA) and Lycopene Expression Pathways. *Synth. Syst. Biotechnol.* **2021**, *6*, 85–94. [CrossRef]
53. Taylor, G.M.; Hitchcock, A.; Heap, J.T. Combinatorial Assembly Platform Enabling Engineering of Genetically Stable Metabolic Pathways in Cyanobacteria. *Nucleic Acids Res.* **2021**, *49*, e123. [CrossRef]
54. Hensing, M.C.M.; Rouwenhorst, R.J.; Heijnen, J.J.; van Dijken, J.P.; Pronk, J.T. Physiological and Technological Aspects of Large-Scale Heterologous-Protein Production with Yeasts. *Antonie Leeuwenhoek* **1995**, *67*, 261–279. [CrossRef]
55. Okorafor, I.C.; Chen, M.; Tang, Y. High-Titer Production of Olivetolic Acid and Analogs in Engineered Fungal Host Using a Nonplant Biosynthetic Pathway. *ACS Synth. Biol.* **2021**, *10*, 2159–2166. [CrossRef]
56. Trantas, E.; Panopoulos, N.; Ververidis, F. Metabolic Engineering of the Complete Pathway Leading to Heterologous Biosynthesis of Various Flavonoids and Stilbenoids in Saccharomyces Cerevisiae. *Metab. Eng.* **2009**, *11*, 355–366. [CrossRef]
57. Favero, G.R.; de Melo Pereira, G.V.; de Carvalho, J.C.; de Carvalho Neto, D.P.; Soccol, C.R. Converting Sugars into Cannabinoids—The State-of-the-Art of Heterologous Production in Microorganisms. *Fermentation* **2022**, *8*, 84. [CrossRef]
58. Barbosa, M.J.; Janssen, M.; Südfeld, C.; D'Adamo, S.; Wijffels, R.H. Hypes, Hopes, and the Way Forward for Microalgal Biotechnology. *Trends Biotechnol.* **2023**, *41*, 452–471. [CrossRef]
59. Sharma, N.; Fantino, E.I.; Awwad, F.; Mérindol, N.; Augustine, A.; Meddeb, F.; Desgagné-Penix, I. Impact of Different Light Characteristics on the Growth and Lipid Content of Diatom *Phaeodactylum tricornutum* Transconjugant Strains. *Am. J. Plant Sci.* **2023**, *14*, 41–63. [CrossRef]
60. Kouprina, N.; Larionov, V. Recent Advances in Chromosome Engineering. *Chromosome Res.* **2015**, *23*, 1–5. [CrossRef]

61. Slattery, S.S.; Diamond, A.; Wang, H.; Therrien, J.A.; Lant, J.T.; Jazey, T.; Lee, K.; Klassen, Z.; Desgagné-Penix, I.; Karas, B.J. An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **2018**, *7*, 328–338. [CrossRef]
62. Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B.; et al. Fiji: An Open-Source Platform for Biological-Image Analysis. *Nat. Methods* **2012**, *9*, 676–682. [CrossRef]
63. Niu, Y.F.; Zhang, M.H.; Li, D.W.; Yang, W.D.; Liu, J.S.; Bai, W.B.; Li, H.Y. Improvement of neutral lipid and polyunsaturated fatty acid biosynthesis by overexpressing a type 2 diacylglycerol acyltransferase in marine diatom Phaeodactylum tricornutum. *Mar Drugs* **2013**, *11*, 4558–4569. [CrossRef]
64. Shoichet, B.; Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
65. Kuzuyama, T.; Noel, J.P.; Richard, S.B. Structural Basis for the Promiscuous Biosynthetic Prenylation of Aromatic Natural Products. *Nature* **2005**, *435*, 983–987. [CrossRef]

Algal Research 77 (2024) 103379

Contents lists available at ScienceDirect

Algal Research

journal homepage: www.elsevier.com/locate/algal

Bioengineering *Phaeodactylum tricornutum*, a marine diatom, for cannabinoid biosynthesis

Elisa Fantino [a,b,1], Fatima Awwad [a,1], Natacha Merindol [a], Aracely Maribel Diaz Garza [a], Sarah-Eve Gélinas [a], Gabriela Carolina Gajón Robles [a], Alexandre Custeau [a], Fatma Meddeb-Mouelhi [a,b], Isabel Desgagné-Penix [a,b,*]

[a] Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada
[b] Plant Biology Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

ARTICLE INFO

Keywords:
Metabolic engineering
Aromatic prenyltransferase
Biomanufacturing
Synthetic biology
Light-driving chassis
Phytocannabinoids

ABSTRACT

Over the last few years, phytocannabinoids have been studied for their bioactive properties as potential drug candidates to treat or alleviate symptoms of several diseases. The isolation of single and pure cannabinoid (CB) from their natural source, *i.e., Cannabis* plants results in a low yield, diluted among hundreds of other plant metabolites. The use of biotechnological platforms to produce single CBs is appealing to the pharmaceutical industry. Therefore, our aim was to develop a sustainable system using the model diatom *Phaeodactylum tricornutum* to produce cannabigerolic acid (CBGA), the precursor to several CBs such as well-known cannabidiol (CBD) and delta-9-tetrahydrocannabinol (THC). We engineered *P. tricornutum* to express a mutant version of the *Streptomyces sp.* strain CL190's naptherpin biosynthetic cluster gene B *NphB* (Y288A/G286S, nphB), a non-*Cannabis* aromatic prenyltransferase enzyme, either by random integrated chromosomal expression (RICE) or extrachromosomal expression (EE), to maximize the success of protein production. The gene of *nphB* was linked to the reporter cyan fluorescent protein (*CFP*) and introduced in *P. tricornutum*. Clones were characterized by CFP fluorescence intensity, protein synthesis, enzymatic activity, and production of CBGA. We present, for the first time in diatoms, the successful production of a CB, CBGA up to 4.1 (± 0.2) mg/kg of microalgal fresh biomass weight. This work shows the potential of *P. tricornutum* as a sustainable controlled heterologous platform for CBs production, plant bioactive compounds, and relevant pharmaceuticals.

1. Introduction

Cannabinoids (CBs) are bioactive metabolites from the terpeno-phenolic group produced in the glandular trichomes of *Cannabis,* as well as in some liverworts and fungi [1]. Although they have been used for centuries in traditional medicine such as Ayurveda and traditional Chinese medicine for pain relief, inflammation, and gastrointestinal disorders [2], their mode of action was only deeply understood after the discovery and characterization of endocannabinoid receptors CB1 in humans [3]. They have been described in preclinical studies to be potential treatments for chronic pain [4], multiple sclerosis [5], and neurodegenerative disorders, such as Parkinson's disease [6], to name a few. Although CBs were used for different applications, their biosynthetic pathway *in planta* was not known until two decades ago [7–11]. CBs biosynthesis starts with the carboxylation of acetyl-CoA into malonyl-CoA by the action of acetyl-CoA carboxylase. Malonyl-CoA, which is a building block of fatty acids, is further converted into olivetolic acid (OA) through a series of polyketide synthesis reactions followed by a C12-polyketide cyclization that allows the formation of OA, the precursor of various CBs. OA undergoes fusion with the prenyl moiety of geranyl pyrophosphate (GPP) to form cannabigerolic acid (CBGA) (Fig. 1A), catalyzed by plastid membrane-localized aromatic

Abbreviations: Au, Absorbance units; CB, Cannabinoid; CBD, Cannabidiol; CBDA, Cannabidiolic acid; CBGA, Cannabigerolic acid; CFP, Cyan fluorescent protein; EE, Episomal expression; FACS, Fluorescence-activated cell sorting; Gly, Glycine; GPP, Geranyl pyrophosphate; NphB, *Streptomyces sp.* strain CL190's naphterpin biosynthetic cluster gene B; OL, Olivetol; OA, Olivetolic acid; OAC, Olivetolic acid cyclase; OD, Optical density; RICE, Random integrated chromosomal expression; THCA, Delta-9-tetrahydrocannabinolic acid.

* Corresponding author at: Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada.
E-mail address: Isabel.Desgagne-Penix@uqtr.ca (I. Desgagné-Penix).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.algal.2023.103379
Received 13 October 2023; Received in revised form 23 November 2023; Accepted 26 December 2023
Available online 30 December 2023

prenyltransferases (*Cs*PT) in *C. sativa* [12–14]. CBGA is the branching point for the synthesis of delta-9-tetrahydrocannabinolic acid (THCA), cannabidiolic acid (CBDA), and cannabichromic acid (CBCA) *via* THCA synthase (THCAS), CBDA synthase (CBDAS), and CBCA synthase (CBCAS), respectively [8,9,15,16]. CBs accumulate at low levels *in planta,* and their single isolation from the complex extract of *C. sativa* is challenging [17], prompting the interest in alternative approaches for their synthesis.

The formation of CBGA is considered as a central step in the biosynthesis of diverse CBs. Among the eleven candidates of *Cs*PT, only *Cs*PT1 and *Cs*PT4 were shown to be active for this reaction [11,18–20]. Multiple attempts to replicate *C. sativa* CBGA production in heterologous systems have been done using biotechnological approaches. The expression of *CsPT1* in *Escherichia coli* yielded detectable levels of CBGA when feeding OA and GPP [21] as substrates, but supplementation with such precursors is costly. In *Saccharomyces cerevisiae, Cs*PT1 showed low or non-specific activity when compared to *Cs*PT4, which yielded 136 mg/L. [19]. Also, *S. cerevisiae* strains were engineered with multiple enzymes to produce the major CBs found in *C. sativa* including CBGA, using a more affordable substrate, galactose [19]. This was done by introducing over 30 genes such as a multi-organism-derived hexanoyl-CoA pathway and engineering the mevalonate pathway to increase the flux of GPP. Besides performing metabolic engineering in yeast to ensure the presence of precursors, higher quantities have been achieved by overexpressing endogenous farnesyl diphosphate synthase, increasing the native mevalonate pathway, establishing orthogonal monoterpenoid biosynthesis [22–24]. In plants, transient transformation of *Nicotiana benthamiana* with *Cs*PT4 showed accumulation of CBGA and CBGA-diglucoside [25].

One of the major limitations of the heterologous production of CBGA is due to the membrane-bound properties of *Cs*PTs, as enzymatic activity could be affected by protein folding and incorrect subcellular localization [19,26]. To overcome this issue, *Cs*PT4 has been replaced by a soluble prenyltransferase (NphB) from *Streptomyces sp.* strain CL190, involved in the biosynthesis of naphtherpin [27–29]. This hydroxynaphthalene prenyltransferase, is able to produce trace amounts of CBGA, as well as a side product 2-O-geranyl-olivetolic acid (2-O-GOA), in *S. cerevisiae* and *Komagataella phaffii* [30]. NphB is substrate-promiscuous for its prenyl acceptors, hence it is able to catalyze similar prenyltransferase reactions with a number of different aromatic substrates, including OA [31,32]. To increase its specificity to produce CBGA, a few groups have engineered NphB, producing diverse mutant variants [32,33]. Valliere et al. [32] constructed two libraries of mutants, from which they identified the variant NphB (Y288A/G286S) that produced almost exclusively CBGA and had a turnover number (k_{cat}) 1000-fold higher than the wild-type in a cell-free platform [32].

Alternatively, to the well-characterized chassis for metabolic engineering that are *E. coli* and *S. cerevisiae*, the intrinsic metabolic traits of the photosynthetic diatoms *Phaeodactylum tricornutum* are promising for the production of high-value compounds [34–36]. Diatoms have the ability to fix carbon, they are robust and can be grown at industrial-scale [37]. *P. tricornutum* has been successfully engineered to enhance triacylglycerol synthesis [38], to produce monoterpenoids, such as geraniol [39] and OA precursor for CBs biosynthesis [40]. Unlike yeast and *E. coli* [41–44], *P. tricornutum* naturally produces prenylphosphate precursors such as GPP [39]. In this study, we use diatoms as an expression system to accumulate soluble prenyltransferases, NphB. We engineered *P. tricornutum* to produce CBGA, the first CB and a major biosynthetic precursor molecule of several other CBs by introducing the double mutant NphB (Y288A/G286S), *nphB* (Fig. 1). We expressed *nphB* either by extrachromosomal expression (EE) or randomly integrated chromosomal expression (RICE). EE is convenient for delivery of large DNA constructs, allowing consistent transgene expression [39,45,46]. Nevertheless, little is known about extrachromosomal transcription mechanism, stability and copy number [37,47,48]. RICE promises higher expression level depending on the genomic integration region [37]. In this study, both strategies were used to increase the chances of obtaining strains that successfully express the transgene. Moreover, *in vitro* assays at different temperatures revealed that nphB was active at *P. tricornutum*'s growth temperature. We performed *in vivo* supplementation of OA and GPP using DMSO and hypo-osmotic shock to ensure the substrates permeate the cells. Here, we show for the first time, the successful production of CBGA and EE/RICE of CB metabolic enzymes in *P. tricornutum.* Our results suggest that both methods of transformation can be used efficiently. This study is a proof of concept that *P. tricornutum* can serve as a photosynthetic platform for the production of single phytocannabinoids.

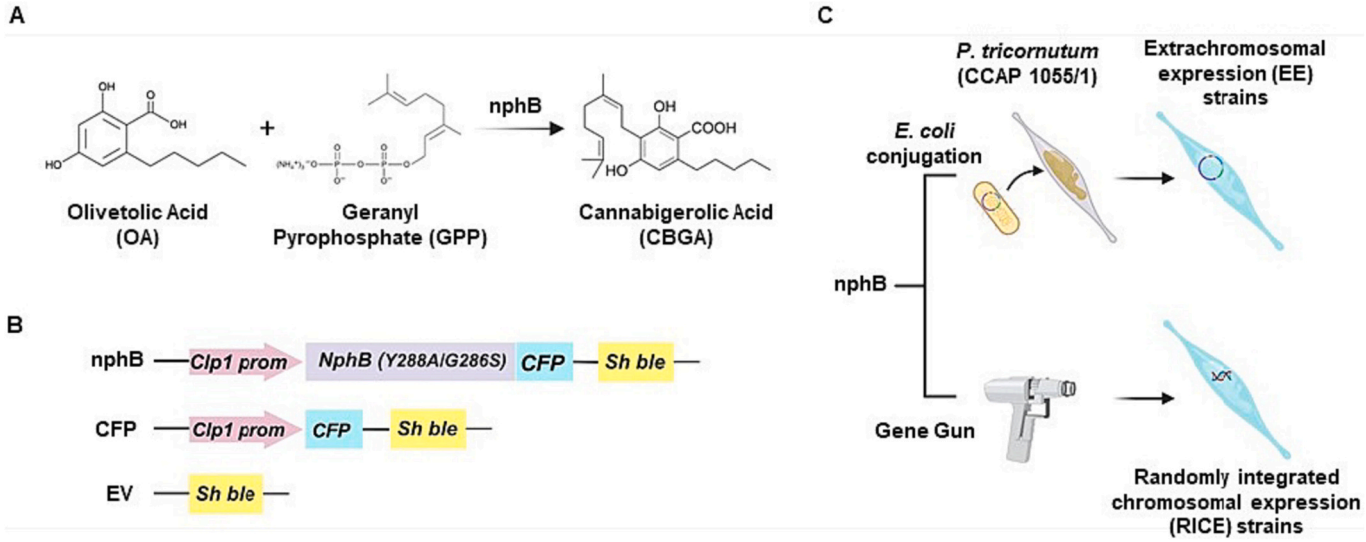


Fig. 1. *P. tricornutum* nphB transformant strains. A. CBGA Enzymatic reaction; CBGA forms from the C—C prenylation of olivetolic acid (OA) by geranyl diphosphate (GPP). B. Scheme of the recombinant cassettes expressing *Streptomyces* strain CL109 NphB (G286S/Y288A) driven by the diatom-infecting virus promoter from *Chaetoceros protobacilla* DNA virus (*Clp1*) and tagged with *cyan fluorescence protein* (*CFP*) reporter gene. Moreover, two negative controls were used such as *CFP* gene driven by *Clp1 promoter* and Empty vector (EV) expression only *Sh ble* gene conferring zeocin resistance. C. *P tricornutum* transformants strains were generated by the *E. coli* conjugation of nphB Extrachromosomal expression (EE) and by biolistic for Randomly integrated chromosomal expression (RICE) of the wild type *P. tricornutum* strain CCAP 1055/1. The figure was created with BioRender.com [49].

2. Materials and methods

2.1. Microbial strains and growth conditions

E. coli strains: NEB® 10-beta (New England Biolabs, Canada) and Epi300 (Epicenter) were grown in Luria Broth (LB) supplemented with appropriate antibiotics: chloramphenicol (30 mg/L), and gentamicin (40 mg/L) for Epi300 strain. *P. tricornutum* (CCAP 1055/1, Culture Collection of Algae and Protozoa; kindly provided by Prof. Bogumil Karas) was grown in a modified L1 medium without silica [48]. The culture conditions were at 18 °C under cool white fluorescent lights (75 $\mu E\ m^{-2}\ s^{-1}$) and a photoperiod of 16 h light 8 h dark with an agitation of 130 rpm for liquid cultures.

2.2. Plasmid constructions

All plasmid constructs were done by Gibson assembly using the NEBuilder® HiFi DNA Assembly Bundle for Large Fragments (New England Biolabs, Canada). Fragments used for the assemblies were amplified by PCR with PrimeSTAR GXL DNA Polymerase (Takara Bio, Japan) following the manufacturer's protocol. The three episomes: NphB (Y288A/G286S):CFP named hereafter nphB, CFP, and empty vector only harboring the *Shble* gene that confers antibiotic resistance to zeocin, were assembled by replacing the *40SRPS8* promoter and *YFP* from *Pt*YFP episome [50] with *Chaetoceros lorenzianus*-infecting DNA virus (ClorDNAV) promoter (*Clp1*) [51] and *nphB* fused to *CFP* with a $(GGGGS)_3$ peptide linker, respectively. Both coding sequences (*nphB* and *CFP)* were codon optimized and synthesized by Genewiz from Azenta Life Sciences (Plainfield, NJ, USA). The *Shble* selection cassette was designed with fucoxanthin chlorophyll *a*/c binding protein C (*FcpC*) promoter and terminator sequences [52]. All DNA and primers (forward and reverse) sequences used in this study are listed in Supplementary Table A.1 and A.2, respectively. All plasmids were sequenced by next-generation sequencing Illumina MiSeq technology at the Massachusetts General Hospital Centre of Computational and Integrative MGH CCIB DNA Core (Massachusetts, USA), to verify the integrity of the DNA.

2.3. Generation of episomal expression (EE) strains

Episomes were transformed into the *E. coli* Epi 300 strain containing pTA-MOB plasmid to allow conjugation with wild-type diatoms, as described in the literature [45]. Briefly, 1 mL of wild type *P. tricornutum* was seeded on 0.5× L1, 1 % agar plates and grown at 18 °C on a light/dark cycle of 16/8 h for 4 days. Prior to transformation, 1 mL of L1 media was added to each agar plate, cells were scraped and recovered by pipetting in a sterile tube. Cell concentration was then adjusted to 5.0×10^8 cells/mL. A volume of 25 mL *E. coli* culture containing the assembled plasmid and pTA-MOB was grown at 37 °C under agitation to OD_{600} of 0.9, then centrifuged at 3000 x*g* for 10 min and resuspended in 250 μL of SOC media. Conjugation was initiated by adding 200 μL of *P. tricornutum* to 200 μL of *E. coli* cells. The cell mixture was plated on 0.5× L1, 5 % LB, ~1 % agar plates, incubated at 30 °C for 90 min in the dark, and transferred to 18 °C in the light and grown for 2 days. After the recovery period, 1 mL of L1 media was added to the plates to collect cells by scraping. Then, cells were plated on 0.5× L1, 1 % agar plates supplemented with zeocin 50 μg/mL for selection and incubated at 18 °C. Transformed colonies appeared after 2 weeks.

2.4. Genome integration transformation

Biolistic transformation protocol to integrate nphB:CFP sequence was adapted from Nymark, [53]. *P. tricornutum* was pelleted from exponentially grown liquid culture, resuspended, and plated onto agar growth plates at a cell concentration of 1×10^8 cells/mL, for two days. The genomic material used consisted of the selection and expression cassettes plus 40 bp long arms amplified by PCR using PrimeSTAR GXL DNA Polymerase (Takara Bio, Japan). Next, 2.5 μg of DNA was hybridized to 0.6 μm gold beads and introduced to the plated *P. tricornutum* at high velocity using the Bio-Rad Helios Gene Gun instrument Handheld (Bio-Rad, Hercules, CA), with a distance of 10 cm and a pressure of 500 kPa, as described by the manufacturer. The transformed *P. tricornutum* plates were then allowed to recover for 48 h at 18 °C. Cells were re-plated on agar selection plates, zeocin at 50 μg/mL. Transformed colonies appeared after 2–3 weeks of incubation.

2.5. Flow cytometry and fluorescence-activated cell sorting (FACS)

EE and RICE positive colonies were screened by measuring fluorescence emission using the Synergy H1 BioTek microplate reader (Agilent, Santa Clara, CA, USA). In brief, colonies were transferred from the agar plates to clear 96-well plates with to 200 μL of L1 in each well, after 8 days old, 50 μL of liquid culture, were pipetted into black 96-well plates containing 200 μL of L1, to measure CFP fluorescence at Ex/Em wavelengths of 430/491 nm ($n = 3$ for each screened transformant) and OD730nm. EV and CFP strains were used as negative and positive controls, respectively. Strains with higher fluorescence ratio (Ex/Em430/491/OD730) than EV were selected and analyzed with a CytoFLEX S flow cytometer (Beckman) equipped with violet (405 nm), blue (488 nm), yellow-green (561 nm) and red (638 nm) lasers. In this case, 75 μL of 8 days old cultures were filtered and transferred to clear 96-well plates with 200 μL of L1 in each well. Chlorophyll autofluorescence was detected in the PerCP channel (690/50 nm), while CFP fluorescence was detected in the KO525 channel (525/40 nm).

The BD FACS Melody (BD Biosciences, La Jolla, CA, USA) equipped with violet (405 nm), blue (488 nm), yellow-green (561 nm) lasers were used to sort nphB EE, nphB RICE, and CFP cells according to CFP emission. Selected cells were grown in L1 liquid medium supplemented with zeocin (50 μg/mL) and grown for 8 days. *P. tricornutum* cultures were filtered with Falcon™ Cell Strainers (Fisher Scientific, USA) and diluted to an $OD_{730\ nm} = 0.1$ in fresh L1 media prior to sorting. Events were acquired at a fixed flow rate and at least 10,000 events were analyzed. Cells were gated according to FSC-A (forward scatter area) and SSC-A (side scatter area) parameters and doublons were excluded according to further gating on homogeneous FSC-H (height) *vs.* FSC-W (width) and SSC-H *vs.* SSC-W populations. Chloroplast autofluorescence was gated in the PerCP channel (700/54 nm, 665 LP). Cells with non-specific autofluorescence detected in the PE channel (582/15 filter, 582 long pass filter mirror - LP) were excluded from sorting. CFP fluorescence intensity was further analyzed on the CFP channel (528/45 nm,500 LP). Sorted cells (1.0×10^5 cells/mL) were collected in a 1.5 mL tube containing 500 μL of L1 media without antibiotics, centrifuged for 10 min at 3500 ×*g*. The supernatant was removed and replaced by 1 mL of L1 media supplemented with zeocin 50 μg/mL and ampicillin 100 μg/mL and the culture was grown for one week. This was used as an inoculum for enriched 20 mL cultures in L1 media supplemented with zeocin (50 μg/mL), cultures were grown as mentioned in section 2.1 and sorted a second round, after 14 days of growth, as described above.

2.6. Protein extraction

Twenty mL of eight-days-old cultures were centrifuged at 3500 x*g* for 10 min at 4 °C. Pellets were weighed and resuspended in 500 μL of extraction buffer (EB), containing 51.4 mM Tris pH 8, 0.75 mM SDS, 10 % Glycerol, 0.02 mM EDTA, 10 mM PMFS and 2 μL of protease inhibitor cocktail. Sonication was performed 6 times at 35 % amplitude, with a 30 s pulse on, and 30 s off for 6 min total using FisherbrandTM Model 505 Sonic Dismembrator (Thermo Fisher Scientific). Protein extracts were centrifuged at 20000 x*g* for 30 min at 4 °C. The supernatants containing the total soluble protein fractions were kept at 4 °C to be used for western blot and CBGAS *in vitro* enzymatic assay. The proteins were quantified with the RC DC™ Protein Assay Kit I (Bio-Rad).

2.7. Western blot

For protein detection, 25 μg of total extracted protein were loaded and run into a 10 % SDS-PAGE. The proteins were then transferred to a 0.2 μm PVDF membrane, using 100 V constant and 400 mA for 2 h. The transfer was done on ice using a transfer buffer composed of 25 mM Tris, 192 mM glycine and 20 % methanol. Anti-CFP antibody, purchased from Cedarlane (Ontario, Canada), was used at a 1:1000 dilution in 3 % bovine serum albumin and incubated overnight (ON) at 4 °C. After three washes with Tris-buffered saline, 0.1 % Tween 20 (TBST) solution, the blot was incubated for 1 h in a 1:20000 dilution, in 5 % milk, of Immun-Star Goat Anti-Mouse (GAM)-HRP conjugate from Bio-Rad (Ontario Canada). Membranes were again washed three times using the TBST solution. Protein detection was done using Clarity Max Western ECL Substrate-Luminol solution from Bio-Rad. Chemiluminescence detection and Ponceau S stained (Glacial Acetic Acid 5 % *v*/v, Ponceau Red dye 0.1 % m/v) of the blots were visualized using ChemiDoc Imaging System with Image Lab™ Software (Bio-Rad). The molecular weights of the proteins corresponding to the detected bands were confirmed with protein markers (Precision Plus Protein Dual Color Standards). NphB: CFP native sequence was used as positive control.

2.8. CBGAS in vitro enzymatic assay

Enzymatic assays were carried out using 500 μg of total protein extraction, from 3 biologically independent samples, in a 150 μL volume reaction. The reaction was adapted from Lim et al. [33], composed of 100 mM HEPES (pH 7.5), 25 mM $MgCl_2$, 2 mM OA, and 2 mM GPP, and incubated ON at 30 °C. The reactions were then extracted with 3 volumes of methanol. Metabolite extracts were filtered (0.2 μm PTFE, Agilent Technologies, cat. no. 5190–5265), then dried in a SpeedVac concentrator and resuspended in 100 μL methanol, HPLC grade, and stored at −20 °C for high-performance liquid chromatography with diode-array detection (HPLC-DAD) analyses. Metabolite detection was confirmed by HPLC coupled with tandem mass spectrometry (MS/MS). For the CBGAS *in vitro* assay under different temperatures, 3 biologically independent samples of 30 mL liquid cultures of *P. tricornutum* were grown for 6 days, under the conditions defined above. Cells were harvested by centrifugation at 3500 x*g* at 4 °C temperature for 10 min. The supernatants were discarded, and the biomass was resuspended in 500 μL of elution buffer, sonication was performed as previously described. To test for CBGAS activity at different temperatures, 2 mM of OA and GPP were added, and tubes were incubated ON at 18, 24, 30, and 40 °C. The reactions were then extracted, as previously mentioned and analyzed by HPLC-DAD.

2.9. Cannabinoids and precursors standards

Olivetolic acid (OA, CAS 491–72-5) and olivetol (OL, CAS 500–66-3) were purchased from Santa Cruz biotechnologies (Dallas United states). Delta-9-tetrahydrocannabinol (THC, CAS 1972-08-3), cannabidiol (CBD, CAS 13956–29-1), cannabinol (CBN, CAS 521–35-7), delta-9-tetrahydrocannabinolic acid (THCA, CAS 23978–85-0), cannabidiolic acid (CBDA, CAS 1244-58-2), cannabigerolic acid (CBGA, CAS 25555–57-1), cannabichromene (CBC, CAS 20675–51-8), cannabigerol (CBG, CAS 25654–31-3), tetrahydrocannabivarin (THCV, CAS 31262–37-0) and cannabidivarin (CBDV, CAS 24274–48-4) were purchased from Agilent Technologies (QC, Canada). Cannabinolic acid (CBNA, CAS 2808-39-1) was purchased from Sigma-Aldrich (ON, Canada).

2.10. HPLC-DAD and HPLC-MS/MS analysis

Analyses were conducted using high-performance liquid chromatography (HPLC) with diode-array detection (DAD). Chromatographic separation of analytes was performed using an InfinityLab Poroshell 120 EC-C18 column (4.6 × 100 mm, 2.7 mm; Agilent Technologies, QC, Canada) maintained at 30 °C. Ten microliters of sample were injected into the analytical device. Mobile phases used during analysis were made of (A) formic acid 0.1 % *v*/v in milli-Q water and (B) formic acid 0.1 % v/v in methanol with a flow rate of 1 mL/min. The HPLC gradient program was set as follows: 0 min, 70 % B; 1.0 min, 70 % B; 6.0 min, 77 % B; 15.0 min, 90 % B; 15.1 min, 70 % B and 18.0 min, 70 % B. The total run time per sample was 18.5 min to allow the reconditioning of the column prior to the next injection. The diode array detector was set to acquire the wavelength range of 190 to 400 nm with a deuterium (D2) lamp. All the analyses were done using a UV wavelength of 220 nm. Compounds were identified comparing retention time and maximum absorption wavelengths obtained with the ones of reference standards (Table A.3). Standard calibration curves were prepared as follows; two working solutions were prepared containing OA, OL and CBGA at 10 mg/L and 100 mg/L each in HPLC grade methanol. These solutions were further diluted to prepare calibration solutions with the following concentrations in triplicate: 0.5, 1, 2, 4, 5, 10, 25, 50 and 100 mg/L. These standard solutions were injected into the HPLC-DAD system and used to generate calibration curve regressions. Fig. A.1 shows the calibration curves obtained by plotting the area under the curve obtained as a function of the analyte's concentration, which allowed OA, OL and CBGA quantification for the enzymatic and supplementation assays.

Confirmatory analyses were performed using high-performance liquid chromatography (HPLC) coupled with tandem mass spectrometry (MS/MS) (Agilent, QC, Canada). This system is equipped with an Agilent Jet Stream ionization source, a binary pump, an autosampler, and a column compartment. Compounds separation was achieved using an InfinityLab Poroshell 120 EC-C18 column (4.6 × 100 mm, 2.7 mm; Agilent Technologies, QC, Canada). *In vitro* and supplementation assays samples were centrifuged for 10 min at 12000 rpm and diluted 10-fold in the mobile phase (*i.e.,* formic acid 0.1 % *v*/v in milli-Q water and formic acid 0.1 % v/v in methanol 30:70). Five μL of each sample were injected onto the column that was set at 50 °C. A gradient method made of (A) formic acid 0.1 % v/v in milli-Q water and (B) formic acid 0.1 % v/v in methanol with a flow rate of 0.5 mL/min was used to achieve chromatographic separation. The HPLC elution program was as follows: 0 min, 70 % B; 7.0 min, 100 % B; 10 min, 100 % B; 12.0 min, 70 % B. The total run time was 14 min per sample. The parameters used in the MS/MS source were set as follows: gas flow rate 8 L/min, gas temperature 220 °C, nebulizer 55 psi, sheath gas flow 12 L/min, sheat gas temperature 380 °C, capillary voltage 4500 V and nozzle voltage 0 V. Agilent MassHunter Data Acquisition (version 1.2) and MassHunter Qualitative Analysis (version 10.0) softwares were used for data acquisition and processing respectively. Samples analyses were carried out in triggered multiple reaction monitoring (tMRM) acquisition mode allowing compounds identification using authentic standards. Table A.4 shows MRM transitions and MS/MS parameters used for targeted compounds identification.

2.11. Cell membrane permeabilization and olivetolic acid supplementation assay

To ensure that the substrate OA penetrated inside the cells, six different permeabilization treatments were performed. A positive *P. tricornutum* nphB RICE strain showing active CBDAS enzyme was used. Liquid cultures of 10 mL (in triplicates) were grown for 7 days under the conditions defined above. At day 7, 2 mM of OA was supplied and the following treatments were conducted: i) sonication using an ultrasonic bath treatment with the following settings; 50/60 HZ, 117 Volts, 1.0 Amps (Branson, ultrasonic cleaner) for 8 min [54], and kept ON; ii) dark treatment, where the culture was incubated at 18 °C in the dark ON; iii) cold treatment, incubating the culture for 1 h at 4 °C, before adding OA and transferring it to the growth chamber; iv) coadjuvant treatment, supplementing DMSO (2 %) with OA, and then incubating the culture ON under standard conditions; v) hypo-osmotic shock,

centrifuging the cultures at 3500 x g for 10 min, then washing with L1 10%), and resuspending in 10 mL of L1 50% [55] followed by the addition of OA and incubation in the growth chamber ON. To perform the vi) nitrate-phosphate starvation assay, 10 mL of culture from day 5 were centrifuged at 3500 x*g* for 10 min. Then washed with 10 % L1, and resuspended in 10 mL of L1 (without NaNO3 and NaH_2PO_4-H_2O, [56]). The cultures were kept in the growth chamber for 2 days. On day 7th, OA was supplemented and incubated ON. Cultures supplemented with OA 2 mM without being subjected to any treatment and DMSO (2 %) were used as authentic standard and negative control respectively. In all the treatments, cells were harvested on day 8, by centrifugation at 3500 x*g* at 4 °C temperature for 10 min, the supernatants were discarded. Pellets were washed with 10 mL of L1 10%. And they were resuspended in 100 % methanol (1 mL per 100 mg of the fresh-weight pellet). The mix was vortexed for 10 min and prepared for HPLC-DAD analysis.

2.12. CBGA supplementation assay

In order to check the production of the metabolites inside the *P. tricornutum* cells, we performed CBGA supplementation assay. *P. tricornutum* nphB EE, nphB RICE, and CFP strains were used. Liquid cultures of 30 mL (triplicates) were grown for 6 days, under the conditions defined above. On day 6, the cells were harvested, by centrifugation at 3500 x*g* at 4 °C temperature for 10 min. The supernatants were discarded, then the biomass was resuspended in 1.5 mL of L1 or L1 50%. Coadjuvant and hypo-osmotic shock treatments were performed as mentioned above. OA (0.25 and 2 mM) and GPP (0.25 and 2 mM) were added, then the cultures were kept ON under standard conditions. In addition, supplementation was conducted by adding OA 2 mM and GPP 2 mM and adjusting the pH to 8. The negative controls were performed by adding 1.2 μL of DMSO and 3 μL of H2O (as solvent controls). The biomass was harvested on day 7, 2 mL of methanol were used as extraction solvant, and extracts were vortexed for 30 s and prepared for the HPLC-DAD and HPLC-MS/MS analysis, as referred to in section 2.8.

2.13. Substrates and product cytotoxicity assay

To evaluate the cytotoxicity of the substrates, CBGA, solvents, and permeabilization treatments in nphB RICE *P. tricornutum* cells; 200 μL of 6 days-old cultures were incubated ON with serial concentrations of OA (0.1, 0.25, 0.5, 0.75, 1, 1.25, 1,5 and 2) mM, GPP (0.5, 0.75, 1, 1.25, 1.5, 2) mM, CBGA (3, 6, 13, 25, 50, 100, 200) μM, methanol (0.1, 0.2, 0.5, 0.9, 1.8, 3.6, 7.2) %, DMSO (0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2) % and Hypo-osmotic shock (L1 50%) treatment. CellTiter-Glo® Luminescent Cell Viability Assay (Promega) and Propidium Iodide (PI, Sigma Aldrich) were used to analyze the cytotoxic effect of CBs precursors and CBGA. CellTiter-Glo® reagent was added to each well of the treatments, the luminescence was then recorded according to manufacturers' instructions, using the Synergy H1 BioTek microplate reader. PI was added at a final concentration (3 μg/mL), and incubated at room temperature for 10 min. PI emission was measured by CytoFlex in the ECD channel (610/20 nm). The percentages of live and necrotic cells were calculated relative to the untreated control wells.

2.14. Statistics and reproducibility

General data analysis (means and standard deviation) was performed primarily by GraphPad Prism 10.0.2. All experiments were performed with three biological replicates and values were expressed as means ± standard errors.

3. Results and discussion

3.1. Selection of nphB P. tricornutum transformants

In order to metabolically engineer the pennate diatom *P. tricornutum* to produce CBGA, we designed an episome harboring the gene encoding for the double mutant variant *NphB* (Y288A/G286S; nphB; Valliere et al. [32]) (Fig. 1B). Gene expression was driven by the *Chaetoceros protobacilla*-infecting DNA virus (ClorDNAV) promoter (*Clp1*) [51], and nphB was tagged with CFP at its C-terminal. Additionally, two episomes were constructed as negative controls: the empty vector (EV) containing only a zeocin resistance cassette with *Sh ble* gene; and a plasmid with the reporter gene CFP and *Sh ble* only (Fig. 1B). These three constructs were successfully transformed into *P. tricornutum* by *E. coli* conjugation (EE) or particle bombardment (RICE) (Fig. 1C). Both methods were included, as EE transgenes are known to be less prone to genomic control mechanisms such as silencing or regiospecific expression [37,39,45]; while RICE transgenes might show overall higher gene expression levels but with stochastic outcomes depending on integration site [37].

Initial screening of the transformant lines was performed by measuring CFP fluorescence intensity by plate reader, normalized to the OD730 nm and compared with EV strain autofluorescence. For the EE lines, 36 colonies were analyzed, while for the RICE lines, a higher number of colonies (52) was screened to take into consideration random integration effects. Consistently with previous reports [37], the percentage of positive clones in this study was higher in the EE lines (61 % of clones were CFP+), compared to the RICE lines (15 %) (Table 1). Clones with higher (Ex/Em 430/491 nm/ OD730 nm) than EV strain were selected and confirmed by flow cytometry. Three independent clones of each EE ($n = 22$) and RICE ($n = 8$) strains presented a percentage of CFP+ cells higher than 2 %. These six strains were selected to be enriched through two rounds of fluorescence-activating cell sorting, as in [48,50,57]. At last, we continued working with the clones with the highest % of CFP+ cells. For nphB EE selected strain, the percentage of CFP+ cells before enrichment was 7 %, it scaled to 13 %, and then 59 % after the first and second sorting, respectively (Table 1, Fig. 2A). The percentage of CFP+ cells in nphB RICE selected strain was 2.1 %, increased to 5 % after the first enrichment, and to 90 % after the second sorting (Table 1, Fig. 2A). Post-sorting, both EE and RICE strains were enriched in CFP expression, which theoretically reflect nphB levels.

The presence of the full size tagged nphB in the EE, RICE and control strains was confirmed by western blot analysis (Fig. 2B) from cells grown to late exponential phase (8 days-old cultures). As expected, the negative control (EV) showed no CFP specific band (Fig. 3B). Positive control (CFP) showed the predicted band at 27 kDa. Native NphB was used as positive control to verify full size protein (NphB:CFP). The recombinant strains (EE and RICE) produced the predicted size protein of nphB:CFP (67 kDa). Uncropped blot and red ponceau stain are shown in Fig. A.2.

3.2. Characterization of nphB P. tricornutum transformant cell lines

The growth and percentage of CFP+ cells (EE and RICE) were monitored for 14 days (Fig. 3A). CFP and nphB EE and RICE strains

Table 1
Transformant strains screening by cyan-fluorescent protein (CFP) fluorescence. Percentage of CFP+ clones obtained from Extrachromosomal expression (EE), $n = 36$, or randomly integrated chromosomal expression (RICE), $n = 52$, detected with Synergy H1 BioTek microplate reader (excitation at 430 nm and emission at 491 nm). Empty vector (EV) strain was used as negative control to settle the basal autofluorescence. All the positive clones were analyzed by Flow Cytometry (FC), only three of them presented a percentage of CFP+ cells higher than 2 %. The third column shows the percentage of CFP+, after two rounds of sorting of the selected EE and RICE strains.

Strain	CFP positive clones (ex 430 nm/em 491 nm)	No. of clones CFP positive cells (FC)	% of CFP+ cells of selected strains
nphB EE	61 % (n = 36)	3 (n = 22)	59 ± 4
nphB RICE	15 % (n = 52)	3 (n = 8)	90 ± 1

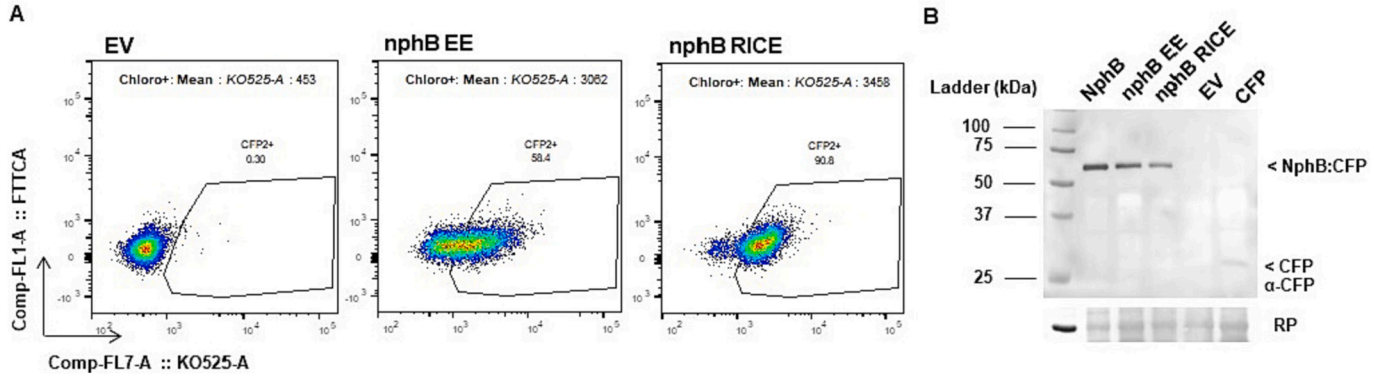


Fig. 2. CFP emission by nphB EE and RICE cell populations and nphB:CFP protein detection. A. Representative dot plots of two-round sorted selected strains. Pseudocolor dot plots of empty vector (EV) and transformants strains; nphB EE and nphB RICE, CFP fluorescence was detected in the KO525 channel (525 /40 nm), shown in the x axis, and chlorophyll autofluorescence was detected in the PerCP (690/50 nm) in the y axis, from CFP positive cells enriched cultures after two rounds of sorting by Fluorescence-Activated Cell Sorting (FACS). Gates and fre-quencies of total CFP populations were designed according to the autofluorescence of the negative control EV shown as reference. B. Western blot with an anti-CFP antibody of total cell extracts from NphB EE, nphB EE, nphB RICE (67 kDa) EV and CFP (27 kDa). Lower panel, stained blot with red ponceau solution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

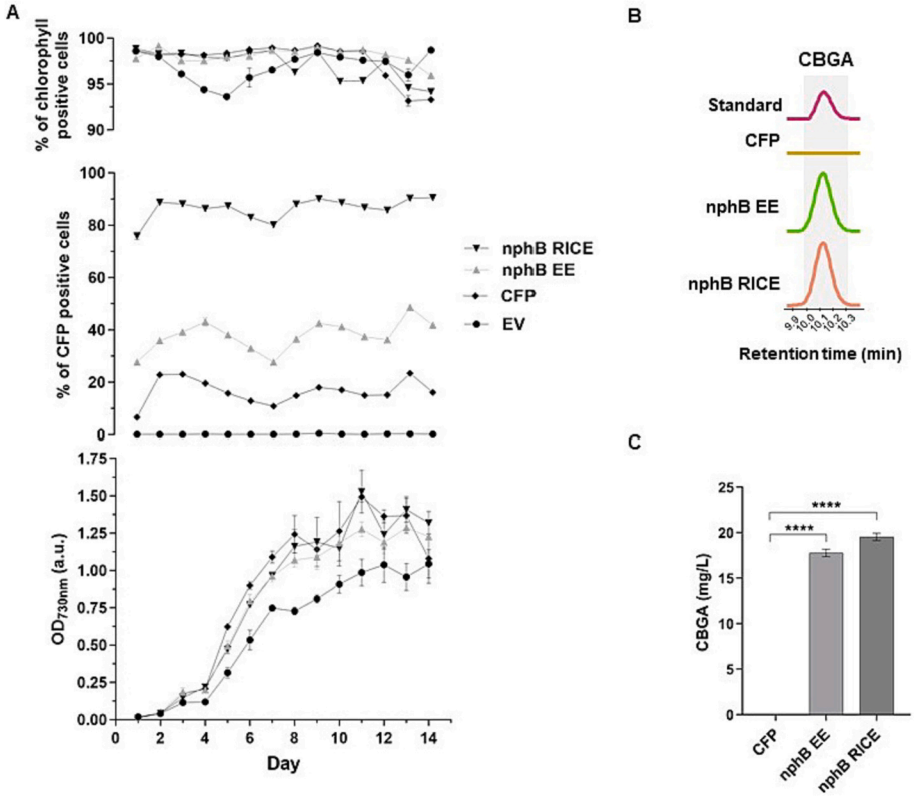


Fig. 3. nphB strains growth curves and *in vitro* production of CBGA. A. Upper panel, percentage of *P. tricornutum* chlorophyll positive cells followed for 14 days. Middle panel, the percentage of CFP positive cells were monitored for each strain, and all CFP positive cultures were normalized to EV autofluorescence. Lower panel, growth curves for each strain, the optical density (OD) at 730 nm was followed for14 days. B. HPLC spectra of *in vitro* CBGA production. Extracts were analyzed by HPLC-DAD and signals were compared to authentic CBGA standard (Fig. A.3). C. CBGA quantification determined using CBGA standard curve (Fig. A.4). Enzymatic assays were performed by adding olivetolic acid and GPP 2 mM to total protein extracts, at 30 °C during 16 h. Mean ± SD; $n = 3$ biologically independent samples for each strain were plotted, *t*-test (*$p < 0.05$) was performed.

showed a similar cell growth pattern with no significant differences (*$p > 0.05$, one way ANOVA) among them, but higher growth compared to the EV strain at days 7 and 11 (Fig. 3 A lower panel), which indicates that the expression of *nphB* did not compromise cell growth. From the population of the chlorophyll positive cells (Fig. 3A, upper panel), which was stable for 2 weeks, we tracked the percentage of CFP+ cells (Fig. 3A, middle panel). The percentages significantly varied between strains (*$p > 0.05$, one way ANOVA), with the RICE line presenting higher values, starting with 76 (± 1) % at day 1 and showing the highest at day 13, 90.3 (± 0.2) %; while nphB EE started with 27.7 (± 0.4) %, reaching to 48 (± 1) % at day 13 and the CFP line began at 6.7 (± 0.3) % finishing with 23.4 (± 0.6) % (Fig. 3A, middle panel). According to Kadono et al. [51], high levels of expression were reported in cells in the stationary phase compared to the exponential phase of growth using this promoter. When

P. tricornutum was engineered, with the fungal *Aspergillus niger* PhyA or the bacterial *E. coli* AppA phytases, to accumulate eicosapentaenoic and docosahexaenoic acids together, highest levels of phytase activity were achieved at late exponential phase [58]. Our results show that Clp1 promoter activity levels increase and stabilize after the 2nd day of culture, with low fluctuations during the two weeks of 20 mL culture. We did not observe expression peaks like it has been seen in inducible or strong promoters during the exponential phase [37,39,46,52]. Even though the percentage of CFP+ cells is lower in the CFP line, this value, 23.4 (± 0.6) %, allowed us to continue using this strain as negative control for further assays. These results show that nphB and CFP lines were stably expressed under Clp1 promoter in *P. tricornutum* using both transformation methods in the framework of this study, in accordance with what was shown before [51,58,59].

3.3. In vitro synthesis of CBGA

To assess if the marine diatom could produce active nphB, total soluble protein was extracted from nphB EE, nphB RICE, EV and CFP strains, grown to late exponential phase. Enzymatic assays were performed using OA and GPP as substrates at 30 °C for 16 h. The products of the reaction were analyzed by HPLC-DAD (Fig. 3B) and confirmed by HPLC-MS/MS (Fig. A.3 & A4). A peak at 10.146 min retention time corresponding to the CBGA standard was detected when nphB EE and RICE transformants protein extracts were incubated with substrates. This peak was not present when substrates were incubated with CFP strain or native NphB enzyme protein extracts or in the absence of substrate for all extracts. The reaction of nphB EE strain produced 17.8 (± 0.3) mg/L, while crude extract from nphB RICE transformant yielded 19.6 (± 0.3) mg/L under our assay conditions (Fig. 3C). Thus, despite a higher frequency of CFP+ cells in RICE transformant, both yielded comparable levels of CBGA. The same nphB mutant showed higher CBGA production when it was expressed and purified from *E. coli* [32,33,60]. In this study, the native enzyme did not display detectable activity; similar to what was previously shown in yeast [19,30].

3.4. nphB is active at P. tricornutum growth temperature

To confirm that nphB is active at *P. tricornutum* growth temperature range (from 18° to 24 °C) or at higher temperature, enzyme activity was performed at 18 °C, 24 °C, 30 °C and 40 °C for 16 h (Fig. 4A). All HPLC chromatograms (Fig. A.5 to A.8) were compared to CBGA standard and CFP strain chromatograms. Both nphB EE and RICE crude extracts produced similar amount of CBGA (*$p > 0.05$, one way ANOVA) at 18 °C, 24 °C or 30 °C; yielding 7 (± 3), 11 (± 2), and 11 (± 5) mg/mL for nphB EE; and 11 (± 1), 13 (± 3), and 12 (± 3) for nphB RICE, respectively. At 40 °C, nphB showed significantly less activity in comparison with the assays realized at lower temperatures, the production of CBGA decreased to 1.8 (± 0.4) and 3.6 (± 0.3) mg/mL for nphB EE and RICE respectively, being higher in the RICE line (Fig. 4A) (*$p > 0.05$, one way

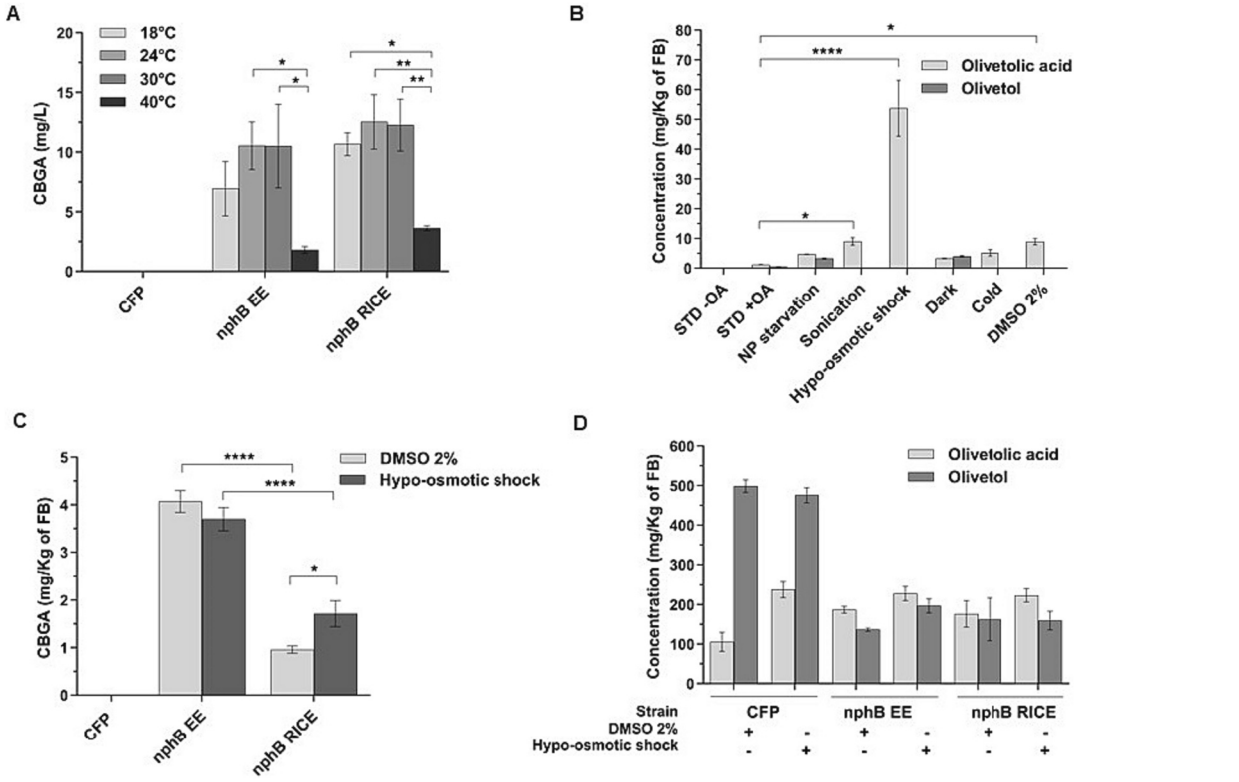


Fig. 4. Optimization of CBGA production A. Crude extracts of CFP, nphB EE and RICE strains were tested at different temperatures; 18 °C, 24 °C, 30 °C and 40 °C. Enzymatic assays were performed by incubating at different temperatures the total crude extracts with olivetolic acid (OA) and GPP 2 mM during 16 h. Mean ± SD; $n = 3$ biologically independent samples for each strain were plotted, one way ANOVA and Tukey's multiple comparisons test (*$p < 0.05$) were performed. B. OA and olivetol (OL) detection in *P. tricornutum* nphB RICE biomass extracts. Cell cultures underwent permeabilization treatments before supplementation with 2 mM OA and incubated during 16 h. Standard conditions (STD) −/+ Olivetolic Acid (OA), nitrate and phosphate starvation (NP). Mean ± SD; $n = 3$ biologically independent samples for each strain were plotted, *t*-test (*$p < 0.05$) was performed. C. CBGA production was determined in nphB EE and nphB RICE strains and in negative control CFP strain treated with DMSO 2 % as coadjuvant or hypo-osmotic shock, and by adding 2 mM of OA and GPP substrates. The cultures were kept 16 h under standard conditions, then metabolites were extracted and detected by HPLC-DAD and confirmed by HPLC-MS/MS. Mean ± SD; $n = 3$ biologically independent samples for each strain were plotted, two ways ANOVA and Tukey's multiple comparisons test (*$p < 0.05$) were performed. D. OA and OL present in *P. tricornutum* CFP, nphB EE and nphB RICE biomass after OA and GPP 2 mM supplementation and CBGA production (C) under DMSO 2 % and hypo-osmotic shock treatments. Mean ± SD; n = 3 biologically independent samples for each strain were plotted. The different letter (a, b) or (a*, b*) indicate significant differences between strains among the same treatment, with *$p < 0.05$ as determined by two ways ANOVA and Tukey's multiple comparisons test.

ANOVA). These results indicate that *P. tricornutum* was successfully transformed to produce nphB enzyme, which is active *in vitro* at two growth temperatures (18 and 24 °C). Since *P. tricornutum* grows between 18 and 24 °C, these results open the possibility of *in vivo* CBGA production. They differ from enzymatic assays performed with native NphB produced in yeast, which were successful at 37 °C, pH = 7.5 [30], but are in agreement with Luo et al. and Lim et al. that performed the reaction at room temperature [19,33]. Other enzymes of the pathway, like THCAS produced in *K. phaffii*, present higher activity at 20 °C compared to 30 °C [61]. Zirpel et al. [16] reported a 5-fold increased activity of THCAS by reducing the cultivation temperature from 25 to 15 °C, possibly due to a more efficient folding of the THCAS enzyme at lower temperatures [30]. Taken together, this shows that *P. tricornutum* growth temperature range is compatible with CBs-producing enzyme activity.

3.5. OA and GPP supplementation in microalgal culture

Enzymatic assays suggested that CBGA could be produced *in vivo*, without activity restrictions related to thermal growth conditions. Prior to testing *in vivo* production of CBGA in *P. tricornutum*, we established the proper conditions for OA supplementation assay. Cell cultures of nphB RICE strain underwent permeabilization treatments during OA feeding, in order to increase the uptake of OA by the cells. To limit substrate and consumable resource uses, tests were performed on nphB RICE considering that all the strains were behaving similarly (Fig. 3A). As *P. tricornutum* produces and accumulates endogenous GPP [39], we began with OA supplementation only. Several permeabilization treatments were performed to analyze OA uptake, such as nitrate and phosphate starvation (NP) [56]. Under N starvation, dual affinity transporters with a wide substrate range and probable roles in sensing external NO_3 modulate membrane permeability [62]. While the full mechanism explaining membrane behavior under P starvation is yet to be elucidated, in general, this condition triggers diatoms to synthesize less phosphoglycerolipids, members of the phospholipid bilayer, which provokes remodeling of the membrane [56]. Moreover, mechanical disruption of the membrane by sonication allows its permeabilization [54]. On the other hand, dark treatment was shown to increase carbon uptake in microalgae, including fatty acids such as OA, as the cell starts heterotrophic growth [63]. Hypo-osmotic shock [55], 1 h of cold treatment at 4 °C [64,65], and DMSO (2 %) treatment, as a coadjuvant were assayed to enhance OA internalization [66].

Before metabolite extraction, cells were washed to remove OA that was not internalized or that could be associated with the cell wall or plasma membrane. OA was detected at low levels in metabolite extracts of cells that received OA in standard conditions (STD), 1.19 (± 0.03) mg/kg fresh biomass (FB) (Fig. 4B). OA significantly increased (*$p < 0.05$, *t*-test) when cell cultures were subjected to sonication 9 (± 1) mg/kg FB, hypo-osmotic shock 54 (± 8) mg/kg FB and DMSO 9.0 (± 0.9) mg/kg FB treatments (Fig. 4B). Moreover, Olivetol (OL), a by-product from OA, was only detected in FB from cell cultures which underwent NP starvation and dark conditions. From these results, we concluded that DMSO and hypo-osmotic shock treatments present the optimal OA-internalizing conditions, and they were retained for further assays.

When the cultures were supplemented with 2 mM of OA and 2 mM of GPP under the up listed permeabilization treatments, CBGA was detected in both strains (EE and RICE) (Fig. 4C). The nphB EE line produced 4.1 (± 0.2) mg/kg FB and 3.7 (± 0.2) mg/kg FB upon DMSO and hypo-osmotic shock treatments, respectively. Meanwhile, nphB RICE strain produced 1.0 (± 0.1) mg/kg FB in presence of DMSO and 1.7 (± 0.2) mg/kg FB under hypo-osmotic shock treatment. CBGA was detected by HPLC-DAD (Fig. A.9 and A.10) and confirmed by HPLC-MS/MS, as was done for the enzymatic assay (Fig. A.4). Overall, *P. tricornutum* transformed lines produced similar yields than the yeast species *Candida viswanathii*, bioengineered with the β-oxidation and CBs pathways, which produced 1.51 mg/L of CBGA [67]. These results suggest that *P. tricornutum* EE and RICE transformants are able to yield CBGA following supplementation with precursors.

The amount of OA and OL present in all the samples was analyzed (Fig. 4D). CFP strain, which does not produce CBGA, shows a significantly higher amount of OL compared to nphB EE and RICE strains (*$p < 0.05$ two ways ANOVA and Tukey's multiple comparisons test). CFP presented 499 (± 13) and 476 (± 16) mg/kg FB of OL when DMSO and hypo-osmotic shock treatments are performed, comparing with nphB strains that show 197 (± 15) mg/kg FB or 137 (± 4) mg/kg FB for nphB EE and 159 (± 19) or 163 (± 45) mg/kg FB for nphB RICE (Fig. 4D. CFP line under both treatments present between 3.7- and 2.4-times higher concentration of OL, while OA concentrations do not present significant differences among treatments. These results showed that the presence of nphB in *P. tricornutum* is leading to a consumption of OA/OL compared to CFP strain.

3.6. Cytotoxicity of treatments

As the permeabilization methods could be cytotoxic, viability was measured in cells treated with DMSO and hypo-osmotic. The percentage of necrotic cells was measured by propidium iodide (PI) stain (Fig. A.11, lower panel, in gray), and the percentage of cell viability was obtained by monitoring the amount of ATP present in the sample (Fig. A.11, lower panel, in black). PI is impermeable to live cells with intact cell membranes, but it can enter dead or dying cells with compromised cell membranes [68]. Once inside the cell, PI binds to DNA, and its fluorescence increases upon binding which could be registered by a flow cytometer, allowing the quantification of dead or damaged cells in a population. On the other hand, ATP is an indicator of metabolically active cells. The assay provides a luminescent readout that is proportional to the amount of ATP present in the sample [69]. Both methodologies are complementary and give an estimation of the percentage of viable and necrotic cells in the population. >65 % of cells (66.5–81.1 %) were alive following the 2 % DMSO treatment, and >95 % of cells were alive following hypo-osmotic treatment. In addition, the percentage of positive CFP fluorescence cells in each treatment remained steady (Fig. A.12A, upper panel). This meant that permeabilization treatments were not cytotoxic (Fig. A.12A).

Then, we tested the toxicity of increasing OA, GPP and CBGA concentrations, as well as DMSO and methanol concentrations since they are the solvents used to dissolve OA and CBGA respectively. The aims were to establish the highest OA and GPP concentrations that could be tolerated by cells without generating cell death, as well as uncover the toxicity threshold of CBGA on *P. tricornutum* transformed cells. Results showed that after 16 h of incubation, increasing the concentrations of GPP, DMSO up to 2 % (*v*/v), or methanol up to 7 % (v/v) does not compromise cell viability (Fig. A.11 B, D and E), nor the percentage of CFP+ cells (Fig. A.11B, D and E, upper panel). Similar results were obtained when cultures underwent OA-internalization conditions, DMSO 2 % and hypo-osmotic shock treatments (Fig. A.12A). However, the substrate OA and the product CBGA generated cell death at concentrations higher than 250 μM and 13 μM, respectively (Fig. A.11A, C and Fig. A.12A). Simultaneously, the percentage of CFP positive cells decreased abruptly under higher OA and CBGA concentration, which indicates a possible protein loss or degradation (Fig. A.11A, C and Fig. A.12A, upper panels). Bright-field light microscopy images of nphB RICE cells treated with OA 0.25 and 2 mM, and CBGA 200 μM (Fig. A.12B), show that the shape of the cells was not altered by the treatments. However, cells treated with OA 2 mM and CBGA 200 μM present vesicles and chloroplasts granularity, which represent a classic behavior of cell death [70] compared to the incubation cells with OA at 0.25 mM or control. As the addition of 2 mM OA to nphB EE and RICE cultures, from 6 days old, caused a decrease in the pH from 8.7 to 6, we hypothesized that the acidification of the media could induce the cell death. However, supplying OA and GPP at 2 mM concentration to the media while adjusting the pH to 8 did not lead to a decrease in the percentage of necrotic cells in both strains (Fig. A.13). In addition, no

CBGA or other CBs were detected after pH adjustment. The decrease of pH in presence of OA has been observed as well in the oleaginous yeast *Yarrowia lipolytica* genetically engineered with the OA biosynthesis pathway [71]. Authors have seen that during fermentation, the culture pH dramatically dropped from 6.8 to below 3.5. As this strong acidity may negatively affect membrane permeability and strain performance, they controlled the medium pH by using either Phosphate Buffer Saline (PBS) buffer or $CaCO_3$. The supplementation of $CaCO_3$ maintained stable pH and increased OA titer by 3 times fold, whereas PBS failed to improve OA production [71].

Independently of pH, our results suggest that CBs cause cytotoxicity in *P. tricornutum* when applied in large amounts. This observation is consistent with several previous studies using other biosynthesis platforms. When CBGA cytotoxicity was tested in 10-day-old suspension-cultured cells of *C. sativa* with 50 μM, it caused 100 % cell death after 24 h treatment, whereas OA did not affect the cells [9]. The same result was observed in 7-day-old tobacco BY-2 cells treated with CBGA and OA. Meanwhile, CBGA induced apoptosis not only in plant cells but also in insect (*Spodoptera frugiperda*, Sf9) cells [9], suggesting that CBs could act as plant defense compounds, like many specialized metabolites biosynthesized in the glandular trichome [72,73]. To avoid CBs cytotoxicity in diatoms, CBs glycosylation [25] could offer a solution by making CBs water-soluble and allow higher accumulation [74] as demonstrated for other plant compounds [75,76].

When CBs pathway is engineered in new heterologous biosynthetic platforms, like the filamentous fungi *Penicillium chrysogenum*, only small detectable levels of THCA were found *in vitro*, using 1 mM OA and 0.1 mM GPP as substrates [77]. This shows that CBs yields remain considerably low, and inconsistent [78]. Yet, despite the cytotoxicity concern, our production yields exceeded those reported by Dusséaux et al. [44]. Cytotoxicity could be addressed using a cell-free system. In a prior study by Valliere et al. [60], researchers expressed and purified 12 enzymes, a new stabilized version of the nphB among them, subsequently they were supplemented with hexanoate, malonate, acetyl phosphate and isoprenol yielding to 570 (± 60) mg/L of CBGA [60]. Alternatively, to produce higher CBs titers in a cell chassis, subcellular compartmentalization approach could be performed to allow the accumulation of the cytotoxic compounds in specific cell organelles and avoid cell death. For instance, Dusséaux et al. [44] targeted *Cs*PT4 and the *S. cerevisiae* GPPS, Erg20p (N127W), to the yeast peroxisome, resulting in an active *Cs*PT4, leading to OA geranylation conversion into CBGA (0.82 mg/L, Dusséaux et al. [44]). In future studies, such compartmentalization approach could lead to higher titers of CBGA in *P. tricornutum* upon reduction of cytotoxicity. In addition, improvement of the metabolic flux into the CBs biosynthesis through remodeling of the terpenoids and other competitive pathways could be performed. Simultaneously designing enhanced synthetic enzymes and optimizing fermentation processes to boost lipid metabolism towards OA and other precursors are other strategies that could help increase the CBs yield in diatoms.

4. Conclusion

As the first publication on cannabinoid engineering in diatoms, this paper offers an assessment of *P. tricornutum*'s proof-of-concept for the heterologous production of CBs. We provide evidence that *P. tricornutum*'s endogenous metabolism is permissive to cannabinoids synthesis. Our results showed that CBGA, the first and essential precursor CB, is produced in *P. tricornutum* following supplementation. Overall, the diatom cell physiology was similar in presence or absence of nphB transgene, both in EE and RICE strains, and the current results show CBGA production in competitive levels compared to classic heterologous systems. However, the system would need further improvement to overcome the cytotoxicity of CBs. This shows that diatoms and *P. tricornutum* present potential as a platform for the production of phytocannabinoids.

CRediT authorship contribution statement

Elisa Fantino: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Fatima Awwad:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Natacha Merindol:** Conceptualization, Data curation, Project administration, Writing – review & editing. **Aracely Maribel Diaz Garza:** Software, Writing – original draft. **Sarah-Eve Gélinas:** Formal analysis, Writing – review & editing. **Gabriela Carolina Gajón Robles:** Formal analysis, Writing – review & editing. **Alexandre Custeau:** Formal analysis, Validation. **Fatma Meddeb-Mouelhi:** Project administration, Writing – review & editing. **Isabel Desgagné-Penix:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

All authors declare that there are no conflicts of interest, informed consent was not required, and there were no human or animal rights considerations applicable. All authors agreed on the authorship and submission of the manuscript for peer review.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge that financial support for this research was funded by the Natural Sciences and Engineering Research Council of Canada through the Alliance program Award No ALLRP 570476-2021 to IDP. Additional support in the form of scholarships to EF, FA, AMDG and GCGR from Mitacs-Acceleration/Globalink program grants no IT12310, IT16463 and IT19432 to IDP is also acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.algal.2023.103379.

References

[1] T. Gülck, et al., Phytocannabinoids: Origins and Biosynthesis, 2020, pp. 985–1004.
[2] R. Russo, et al., Gut-brain Axis: role of lipids in the regulation of inflammation, pain and CNS diseases, Curr. Med. Chem. 25 (32) (2018) 3930–3952.
[3] W.A. Devane, et al., Determination and characterization of a cannabinoid receptor in rat brain, Mol. Pharmacol. 34 (5) (1988) 605–613.
[4] N.J. van den Hoogen, et al., Cannabinoids in chronic pain: therapeutic potential through microglia modulation, Frontiers in Neural Circuits (2022) 15.
[5] Z.Z. Al-Ghezi, et al., Combination of cannabinoids, Δ9- tetrahydrocannabinol and Cannabidiol, ameliorates experimental multiple sclerosis by suppressing Neuroinflammation through regulation of miRNA-mediated signaling pathways, Front. Immunol. 10 (2019).
[6] B. Abd-Nikfarjam, et al., Cannabinoids in neuroinflammatory disorders: focusing on multiple sclerosis, Parkinsons, and Alzheimers diseases, Biofactors 49 (3) (2023) 560–583.
[7] F. Taura, et al., Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa L.* Biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid. 1996, Â©, ASBMB, 1996, pp. 17411–17416. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology.
[8] S. Sirikantaramas, et al., The gene controlling marijuana psychoactivity. Molecular cloning and heterologous expression of Δ1-tetrahydrocannabinolic acid synthase from *Cannabis sativa L.* 2004, Â©, ASBMB, 2004, pp. 39767–39774. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology.
[9] S. Sirikantaramas, et al., Tetrahydrocannabinolic Acid Synthase, the Enzyme Controlling Marijuana Psychoactivity, is Secreted into the Storage Cavity of the Glandular Trichomes, 2005, pp. 1578–1582.
[10] S.J. Gagne, et al., Identification of Olivetolic Acid Cyclase from *Cannabis sativa* Reveals a Unique Catalytic Route to Plant Polyketides, 2012, pp. 12811–12816.

[11] F. Taura, et al., Phytocannabiniods in *Cannabis sativa*: Recent Studies on Biosynthetic Enzymes, 2007, pp. 1649–1663.
[12] W. Eisenreich, F. Rohdich, A. Bacher, Deoxyxylulose phosphate pathway to terpenoids, Trends Plant Sci. 6 (2) (2001) 78–84.
[13] M. Fellermeier, M.H. Zenk, Prenylation of olivetolate by a hemp transferase yields cannabigerolic acid, the precursor of tetrahydrocannabinol, FEBS Lett. 427 (2) (1998) 283–285.
[14] M. Nagia, et al., Sequential regiospecific gem-diprenylation of tetrahydroxyxanthone by prenyltransferases from Hypericum sp, New Phytol. 222 (1) (2019) 318–334.
[15] F. Taura, et al., Cannabidiolic-acid Synthase, the Chemotype-Determining Enzyme in the Fiber-type *Cannabis sativa*, 2007, pp. 2929–2934.
[16] B. Zirpel, F. Stehle, O. Kayser, Production of Delta9-tetrahydrocannabinolic acid from cannabigerolic acid by whole cells of Pichia (Komagataella) pastoris expressing Delta9-tetrahydrocannabinolic acid synthase from *Cannabis sativa L.* Biotechnol. Lett. (2015) 1869–1875.
[17] M.M. Lewis, et al., Chemical profiling of medical Cannabis extracts, ACS Omega 2 (9) (2017) 6091–6103.
[18] S. Sirikantaramas, et al., Recent advances in Cannabis sativa research: biosynthetic studies and its potential in biotechnology, Curr. Pharm. Biotechnol. 8 (4) (2007) 237–243.
[19] X. Luo, et al., Complete Biosynthesis of Cannabinoids and their Unnatural Analogues in Yeast, Springer US, 2019, pp. 123–126.
[20] P.V. Apicella, et al., Delineating genetic regulation of cannabinoid biosynthesis during female flower development in Cannabis sativa, Plant Direct 6 (6) (2022) e412.
[21] Sonal R. Ayakar, et al., Metabolic Engineering of *E. coli* for the Biosynthesis of Cannabinoid Products, United States, 2020.
[22] J. Zhao, et al., Improving monoterpene geraniol production through geranyl diphosphate synthesis regulation in Saccharomyces cerevisiae, Appl. Microbiol. Biotechnol. (2016) 4561–4571.
[23] Y. Ma, et al., Engineering a universal and efficient platform for terpenoid synthesis in yeast, Proc. Natl. Acad. Sci. U. S. A. 120 (1) (2023) e2207680120.
[24] C. Ignea, et al., Orthogonal monoterpenoid biosynthesis in yeast constructed on an isomeric substrate, Nat. Commun. 10 (1) (2019) 3799.
[25] T. Gülck, et al., Synthetic Biology of Cannabinoids and Cannabinoid Glucosides in Nicotiana benthamiana and *Saccharomyces cerevisiae*, 2020, pp. 2877–2893.
[26] W.J. de Bruijn, et al., Plant aromatic prenyltransferases: tools for microbial cell factories, Trends Biotechnol. 38 (8) (2020) 917–934.
[27] T. Kuzuyama, et al., Structural Basis for the Promiscuous Biosynthetic Prenylation of Aromatic Natural Products, 2005, pp. 983–987.
[28] T. Bonitz, et al., Evolutionary relationships of microbial aromatic prenyltransferases, PloS One 6 (11) (2011) e27336.
[29] Y. Yang, et al., Catalytic mechanism of aromatic prenylation by NphB, Biochemistry 51 (12) (2012) 2606–2618.
[30] B. Zirpel, et al., Engineering Yeasts as Platform Organisms for Cannabinoid Biosynthesis, Elsevier, 2017, pp. 204–212.
[31] B.P. Johnson, et al., Acceptor substrate determines donor specificity of an aromatic prenyltransferase: expanding the biocatalytic potential of NphB, Appl. Microbiol. Biotechnol. 104 (2020) 4383–4395.
[32] M.A. Valliere, et al., A cell-free platform for the prenylation of natural products and application to cannabinoid production, in: Nature Communications, Springer US, 2019, pp. 1–9.
[33] K.J.H. Lim, et al., Structure-guided engineering of prenyltransferase NphB for high-yield and regioselective cannabinoid production, ACS Catalysis (2022) 4628–4639.
[34] F. Dhaouadi, et al., Diatoms' breakthroughs in biotechnology: *Phaeodactylum tricornutum* as a model for producing high-added value molecules, Am. J. Plant Sci. 11 (10) (2020) 1632–1670.
[35] N. Sharma, et al., Diatoms biotechnology: various industrial applications for a greener tomorrow, Front. Mar. Sci. 8 (2021) 106.
[36] T. Butler, et al., *Phaeodactylum tricornutum*: A Diatom Cell Factory, Elsevier Ltd., 2020, pp. 1–17.
[37] J. George, et al., Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology, Frontiers in Bioengineering and Biotechnology (2020) 1–19.
[38] X. Hao, et al., Enhanced triacylglycerol production in the diatom Phaeodactylum tricornutum by inactivation of a hotdog-fold thioesterase gene using TALEN-based targeted mutagenesis, Biotechnol. Biofuels 11 (1) (2018) 312.
[39] M. Fabris, et al., Extrachromosomal genetic engineering of the marine diatom *Phaeodactylum tricornutum* enables the heterologous production of monoterpenoids, ACS Synthetic Biology (2020) 598–612.
[40] F. Awwad, et al., Bioengineering of the marine diatom Phaeodactylum tricornutum with Cannabis genes enables the production of the cannabinoid precursor, Olivetolic acid, Int. J. Mol. Sci. 24 (23) (2023) 16624.
[41] S. Brown, et al., De novo production of the plant-derived alkaloid strictosidine in yeast, Proc. Natl. Acad. Sci. 112 (11) (2015) 3205–3210.
[42] A. Campbell, et al., Engineering of a Nepetalactol-producing platform strain of Saccharomyces cerevisiae for the production of plant Seco-Iridoids, ACS Synth. Biol. 5 (5) (2016) 405–414.
[43] C.E. Vickers, et al., Recent advances in synthetic biology for engineering isoprenoid production in yeast, Curr. Opin. Chem. Biol. 40 (2017) 47–56.
[44] S. Dusséaux, et al., Transforming Yeast Peroxisomes Into Microfactories for the Efficient Production of High-Value Isoprenoids, 2020, pp. 31789–31799.
[45] B.J. Karas, et al., Designer diatom episomes delivered by bacterial conjugation, in: Nature Communications, Nature Publishing Group, 2015, pp. 1–10.
[46] S.S. Slattery, et al., Phosphate - Regulated Expression of the SARS - CoV - 2 Receptor - Binding Domain in the Diatom *Phaeodactylum tricornutum* for Pandemic Diagnostics, Nature Publishing Group UK, 2022, pp. 1–15.
[47] M.A. Scaife, A.G. Smith, Towards developing algal synthetic biology, Biochem. Soc. Trans. 44 (3) (2016) 716–722.
[48] A. Diamond, et al., Instability of Extrachromosomal DNA Transformed into the Diatom *Phaeodactylum tricornutum*, Algal Res. 70 (2023) 102998.
[49] BioRender, BioRender, 2023.
[50] N. Sharma, et al., Impact of different light characteristics on the growth and lipid content of diatom *Phaeodactylum tricornutum* transconjugant strain, Am. J. Plant Sci. 14 (1) (2023) 41–63.
[51] T. Kadono, et al., Characterization of marine diatom-infecting virus promoters in the model diatom Phaeodactylum tricornutum, in: Scientific Reports, Nature Publishing Group, 2015, pp. 1–13.
[52] S.S. Slattery, et al., An expanded plasmid-based genetic toolbox enables Cas9 genome editing and stable maintenance of synthetic pathways in Phaeodactylum tricornutum, ACS Synth. Biol. 7 (2) (2018) 328–338.
[53] M. Nymark, et al., A CRISPR/Cas9 System Adapted for Gene Editing in Marine Algae, Nature Publishing Group, 2016, pp. 6–11.
[54] M. Zolghadrnasab, et al., Ultrasound-mediated gene delivery into suspended plant cells using polyethyleneimine-coated mesoporous silica nanoparticles, Ultrason. Sonochem. 73 (2021) 105507.
[55] K.E. Helliwell, et al., Spatiotemporal patterns of intracellular Ca2+ signalling govern hypo-osmotic stress resilience in marine diatoms, New Phytol. (2021) 155–170.
[56] H. Abida, et al., Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in Phaeodactylum tricornutum, Plant Physiology (2015) 118–136.
[57] A.E. Sproles, et al., Improved high-throughput screening technique to rapidly isolate Chlamydomonas transformants expressing recombinant proteins, Appl. Microbiol. Biotechnol. 106 (4) (2022) 1677–1689.
[58] A. Pudney, et al., Multifunctionalizing the marine diatom Phaeodactylum tricornutum for sustainable co-production of omega-3 long chain polyunsaturated fatty acids and recombinant phytase, Sci. Rep. 9 (1) (2019) 11444.
[59] T. Kadono, et al., The Possibility of Using Marine Diatom-Infecting Viral Promoters for the Engineering of Marine Diatoms, Elsevier, 2020, p. 110475.
[60] M.A. Valliere, et al., A bio-inspired cell-free system for cannabinoid production from inexpensive inputs, Nat. Chem. Biol. 16 (12) (2020) 1427–1433.
[61] F. Taura, et al., Production of Δ1-Tetrahydrocannabinolic Acid by the Biosynthetic Enzyme Secreted from Transgenic Pichia pastoris, 2007, pp. 675–680.
[62] C. Brownlee, et al., Regulation and integration of membrane transport in marine diatoms, Semin. Cell Dev. Biol. 134 (2023) 79–89.
[63] V. Turon, et al., Potentialities of dark fermentation effluents as substrates for microalgae growth: a review, Process Biochem. 51 (11) (2016) 1843–1854.
[64] J.M. Hyman, et al., A molecular method for the delivery of small molecules and proteins across the cell wall of algae using molecular transporters, Proc. Natl. Acad. Sci. U. S. A. 109 (33) (2012) 13225–13230.
[65] S.A. Kranz, et al., Low temperature reduces the energetic requirement for the CO2 concentrating mechanism in diatoms, New Phytol. 205 (1) (2015) 192–201.
[66] R. Wray, I. Iscla, P. Blount, Curcumin activation of a bacterial mechanosensitive channel underlies its membrane permeability and adjuvant properties, PLoS Pathog. 17 (12) (2021) e1010198.
[67] T.A. Beardslee, Biosynthetic Cannabinoid Production in Engineered Microorganisms, United States, 2020.
[68] Y. Wei, et al., Copper toxicity to Phaeodactylum tricornutum: a survey of the sensitivity of various toxicity endpoints at the physiological, biochemical, molecular and structural levels, BioMetals 27 (3) (2014) 527–537.
[69] T.L. Riss, et al., Cell viability assays, in: S. Markossian, et al. (Eds.), Assay Guidance Manual, 2004. Bethesda (MD).
[70] J. Wertman, et al., The pathway of cell dismantling during programmed cell death in lace plant (Aponogeton madagascariensis) leaves, BMC Plant Biol. 12 (1) (2012) 115.
[71] J. Ma, et al., Biosynthesis of Cannabinoid Precursor Olivetolic Acid in Genetically Engineered Yarrowia Lipolytica, Springer US, 2022, pp. 1–7.
[72] Y. Zhang, et al., Formation mechanism of glandular trichomes involved in the synthesis and storage of terpenoids in lavender, BMC Plant Biol. 23 (1) (2023) 307.
[73] B. Hua, et al., Mediation of JA signalling in glandular trichomes by the woolly/SlMYC1 regulatory module improves pest resistance in tomato, Plant Biotechnol. J. 19 (2) (2021) 375–393.
[74] K. Blatt-janmaat, et al., The Biochemistry of Phytocannabinoids and Metabolic Engineering of their Production in Heterologous Systems, Int. J. Mol. Sci. 22 (5) (2021) 2454.
[75] F. Thomas, et al., Bioengineering studies and pathway modeling of the heterologous biosynthesis of tetrahydrocannabinolic acid in yeast, Appl. Microbiol. Biotechnol. (2020) 9551–9563.
[76] E.H. Hansen, et al., De novo biosynthesis of vanillin in fission yeast (Schizosaccharomyces pombe) and baker's yeast (Saccharomyces cerevisiae), Appl. Environ. Microbiol. 75 (9) (2009) 2765–2774.
[77] K. Kosalková, et al., Biotechnological Fungal Platforms for the Production of Biosynthetic Cannabinoids, MDPI, 2023.
[78] Z. Xie, et al., Cannabis sativa: origin and history, glandular trichome development, and cannabinoid biosynthesis, Hortic Res 10 (9) (2023) p. uhad150.