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. (2023). Instability of extrachromosomal DNA transformed into the diatom Phaeodactylum tricornutum. Algal Research, 70, Article 102998. https://doi.org/10.1016/j.algal.2023.102998 CC BY-NC-ND 4.0

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
2211-9264/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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,

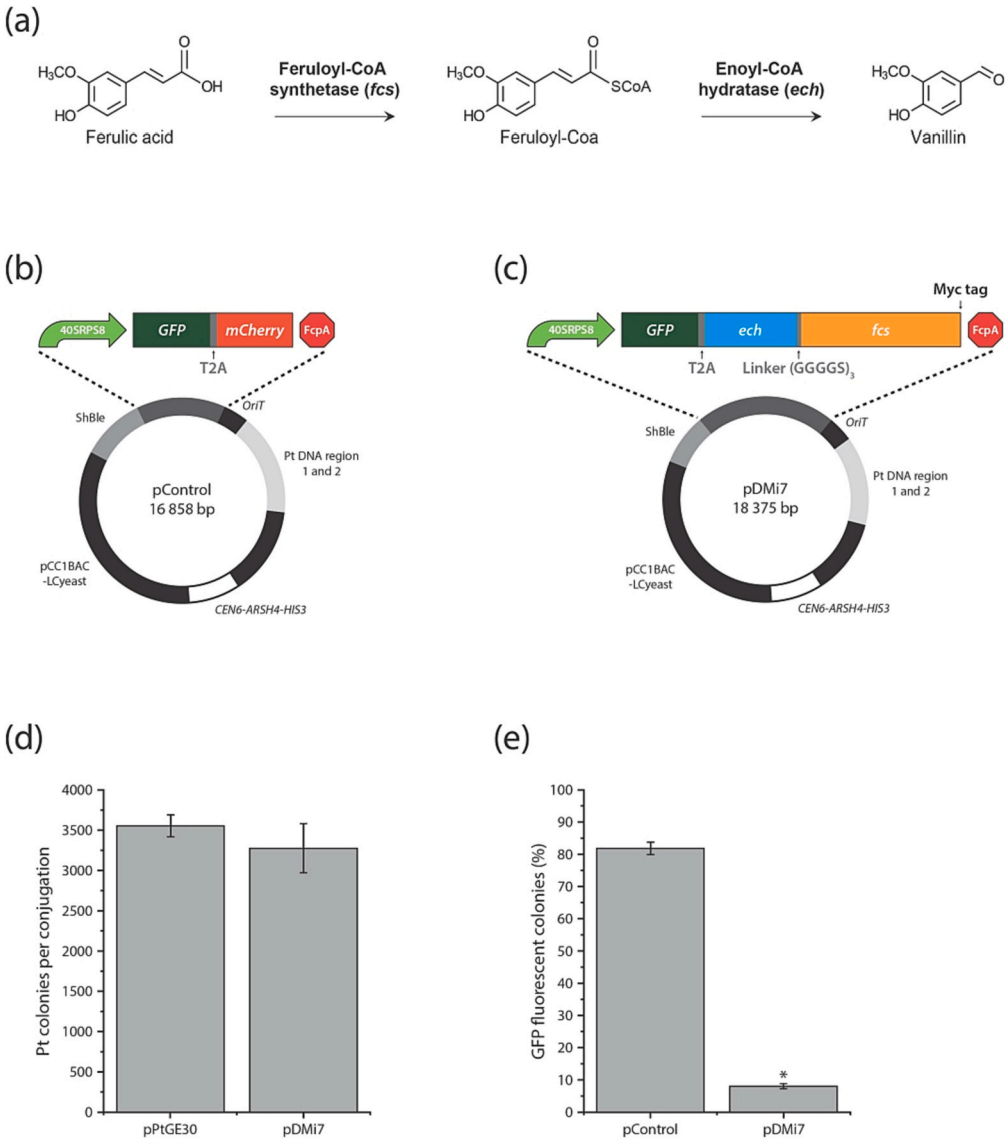


Fig. 1. Transformation of *P. tricornutum* cells with a vanillin biosynthetic pathway. (a) Biosynthetic pathway of two bacterial enzymes from *Streptomyces* sp. that convert ferulic acid into vanillin. (b) Plasmid scheme of pControl. The scheme is on scale, with the exceptions of the size of the promoter 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

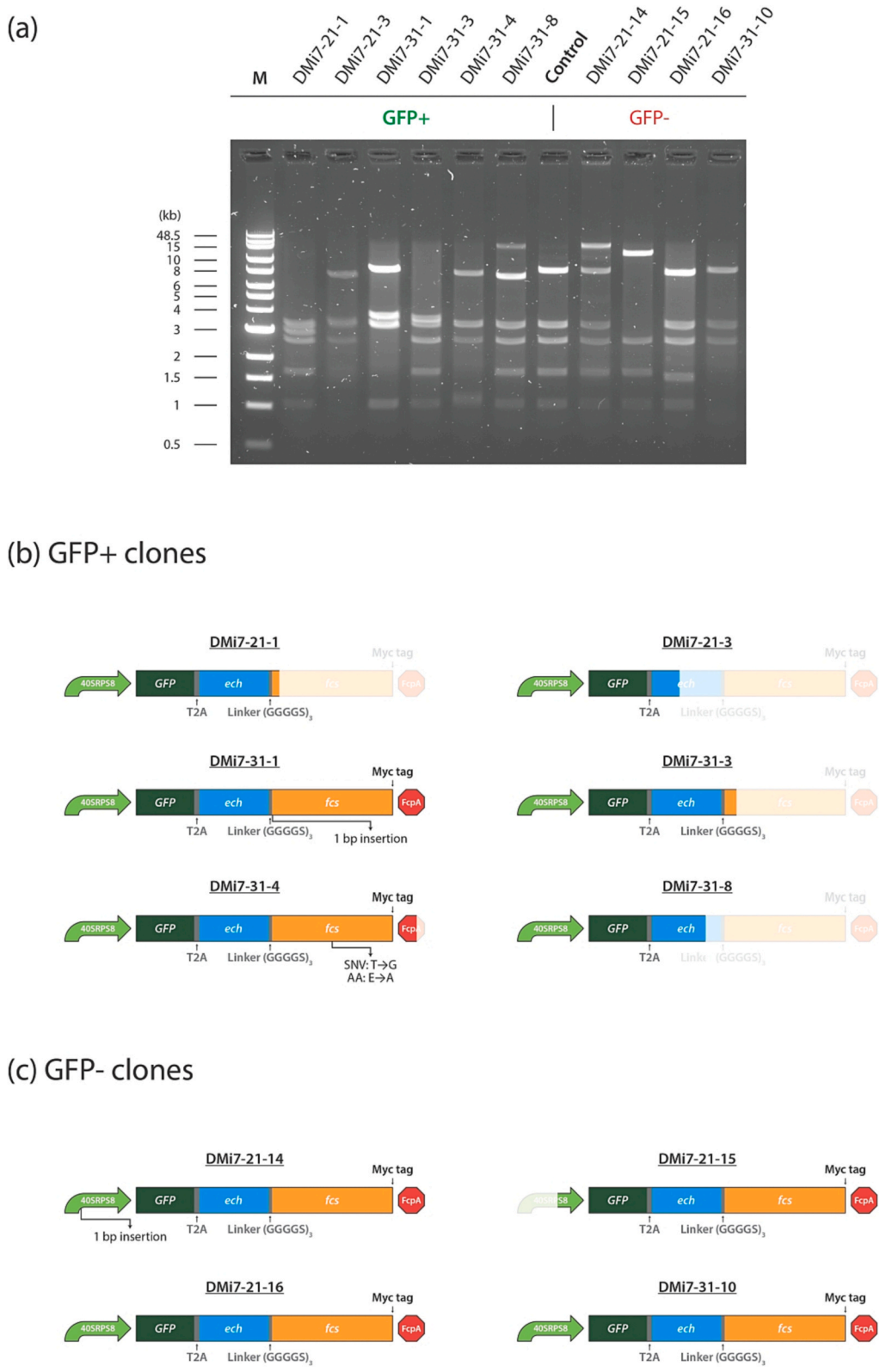


Fig. 2. DNA digestion profiles and representation of the rearrangements of the recovered episomes from pDMi7 *P. tricornutum* transconjugants
(a) Double digestion by *Pst*I and *Spe*I of the episomes recovered from *P. tricornutum* transconjugants showing rearrangement in their sequence. The control 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

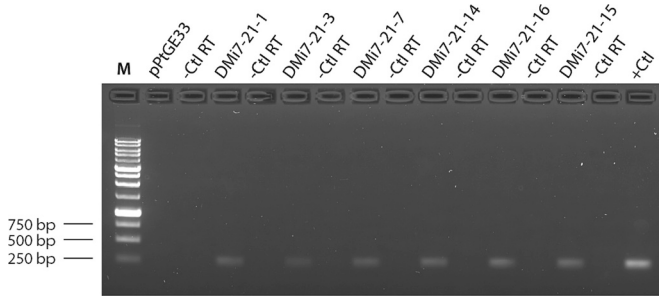


Fig. 3. Expression of the cassette of interest from pDMi7 *P. tricornutum* transconjugants
Endpoint RT-PCR of *P. tricornutum* transconjugants that have been observed by fluorescence microscopy as fluorescent GFP+ (DMi7-21-1, -3, and -7) or non-fluorescent GFP- (DMi7-21-14, -15, -16). The plasmid pDMi7 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

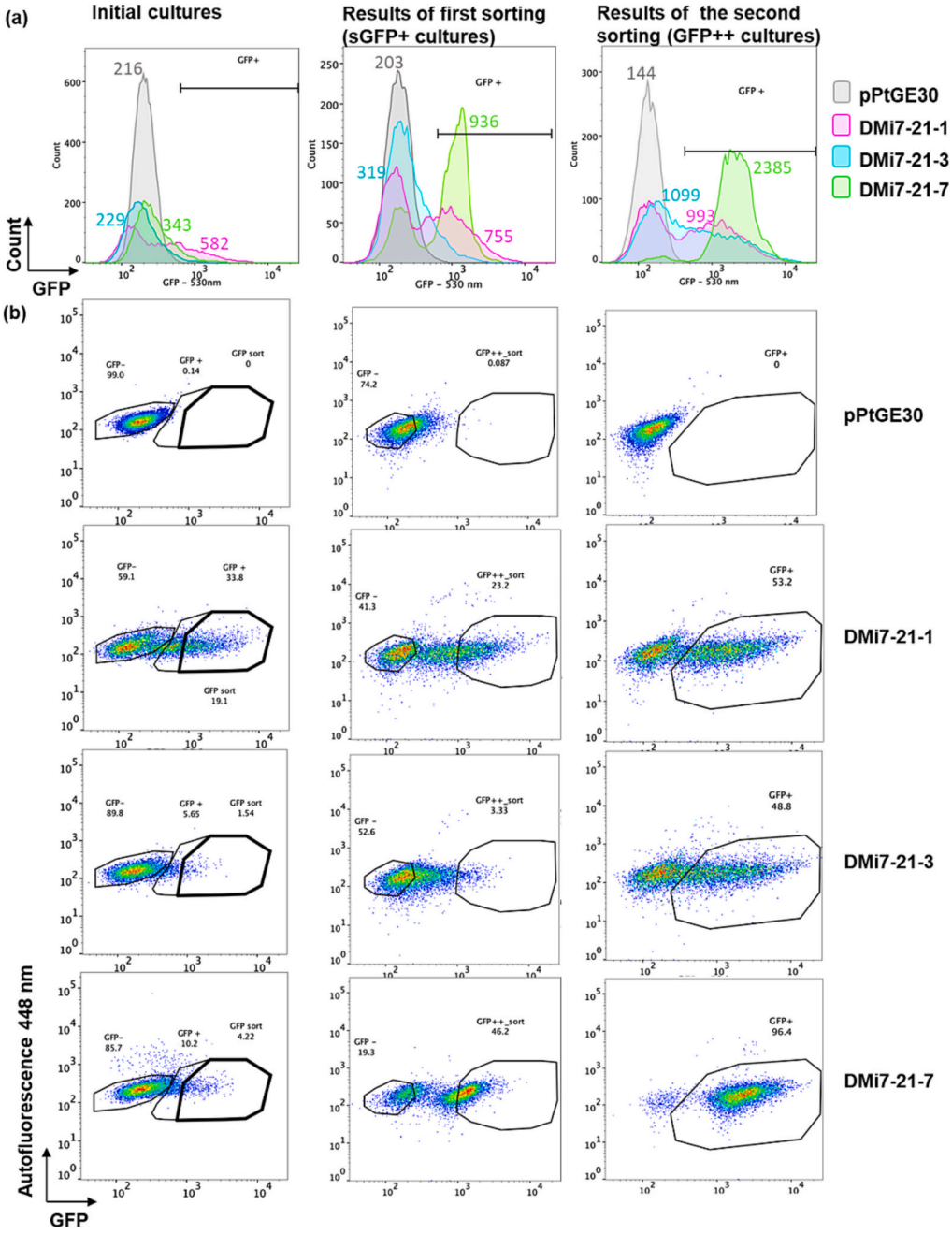


Fig. 4. Enrichment of *P. tricornutum* GFP+ through fluorescence activated cell sorting.
(a). Histogram plots of GFP intensity profile in cultures of DMi7-21-1 (pink), DMi7-21-3 (turquoise) and DMi7-21-7 (green) compared to control pPtGE30 (grey). Background autofluorescence across days following two 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

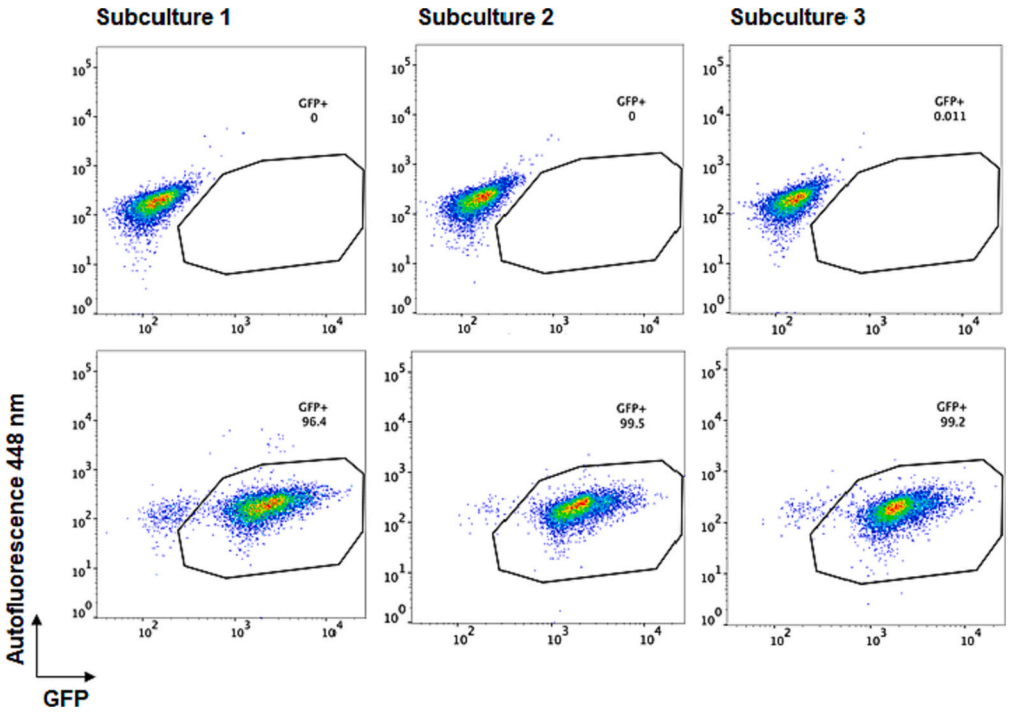
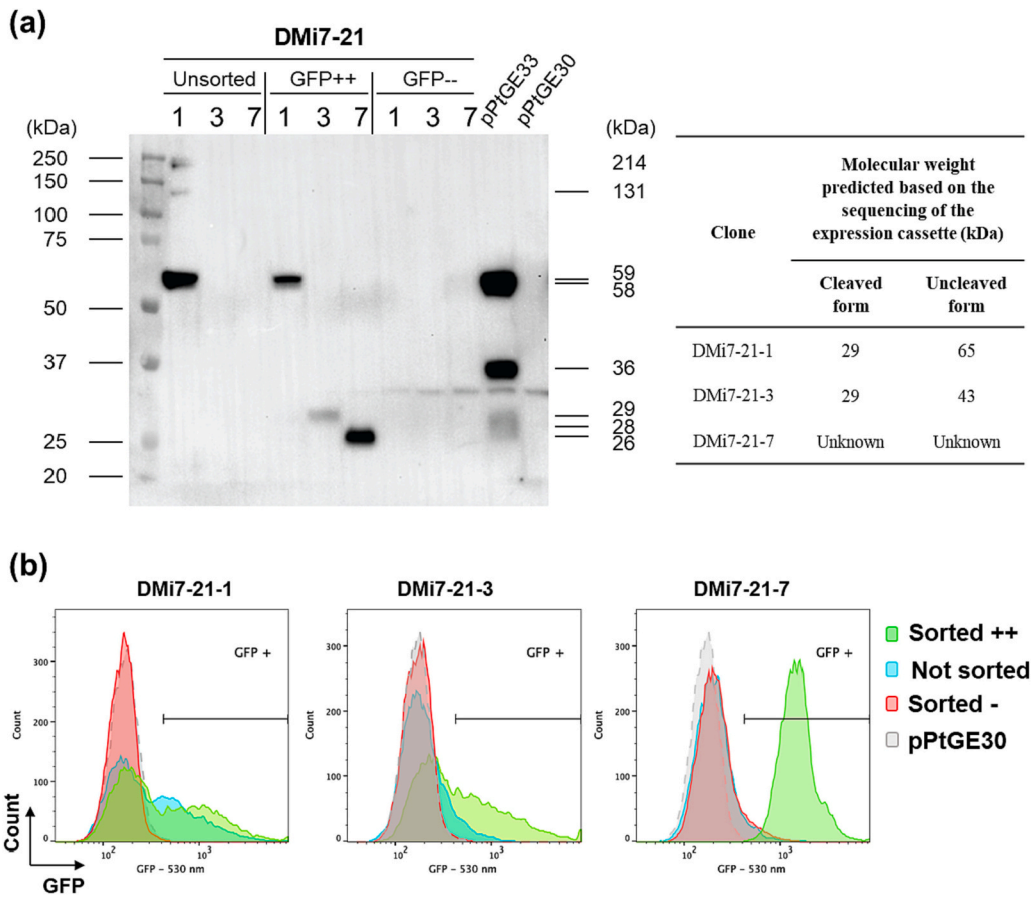


Fig. 5. Stable and high level of GFP expression in DMi7-21-7 across subculturing.
Pseudocolor dot plots of pPtGE30 (upper lane) and DMi7-21-7GFP++ (lower lane) cultures with 530 nm (GFP) in the *x* axis and autofluorescence at 448 nm in the *y* axis of the first three subculturing after the second sorting 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.