Diaz-Garza *et al. Microbial Cell Factories* (2024) 23:286
https://doi.org/10.1186/s12934-024-02559-y

Microbial Cell Factories

RESEARCH

Open Access

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*]

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

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

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 [1, 2]. *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 [3]. In addition, the availability of the complete and annotated genome assembly from telomere to telomere [4, 5] has positioned *P. tricornutum* as one of the leading photosynthetic eukaryotic chassis for metabolic engineering [6].

Conjugation-based episome delivery is the most efficient method to transform diatoms [7]. It has facilitated the characterization of new promoter elements [8, 9], the development of tunable and reversible dose- and time-dependent transcriptional systems [10], and the delivery of gene editing elements [11–13]. For instance, *P. tricornutum* episomal expression system has been used to produce the receptor-binding domain of the SARS-CoV-2 spike protein [14] and specialized metabolites of pharmaceutical interest, i.e., monoterpenoids [15], cannabinoids [16], and cannabinoid precursors [17]. 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 [18], 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 [19]. 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. [20] 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 [21]. 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 [21] 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.

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}) 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 [21] 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* (40 S 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. (2024) with a plasmid harboring no expression cassette besides *sh ble*. Episomes were transformed into wild type *P. tricornutum* by bacterial conjugation as previously described [21]. 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. S1). 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. [21]. 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 ($OD_{730} \sim 0.6$) 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. S2), 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) [22]. 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" [23]. Violin and scatter plots were designed with the package matplotlib (v. 3.8.0) [24].

Episome DNA isolation and sequencing

Episome isolation from *P. tricornutum* was conducted as described in Diamond et al. [21], using Large Plasmid Mini Kit (Geneaid Biotech Ltd., Taiwan), with approximately 9×10^7 cells of *P. tricornutum*. Expression cassettes were amplified by PRIMESTAR GXL (Takara Bio) using the primer pair pPtGE30 Bb F and pPtGE30 Bb R (Supplementary Table 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)25 . 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 50 µl 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* [26]. 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). 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 100 bp 25 M 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 27. Good quality reads were mapped using HISAT2 (v 2.2.1) 28 to the latest version of the genome of *Phaeodactylum tricornutum* (Phatr3) [5] downloaded from Ensembl Protists [29] 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) 30. Posterior analyses were carried in R (v 4.3.1) [31].

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) [32], 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 [29] and used to identified enriched protein families in the deregulated genes using clusterProfiler (v. 4.8.3) [33]. Results of different subpopulations were compared visually with Venn diagrams using ggvenn (v. 0.1.10) [34].

Gene co-expression analysis, without bait genes, was done to highlight specific expression patterns following the workflow from Li et al. [35] powered by tidyverse (v2.0.0) [36] and igraph (v2.0.2) [37] 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 selfcleavable 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. 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. 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. 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. S4).

We proceeded to enrich the subpopulations in individual cultures by FACS, as shown in Fig. 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 and 2045 (Fig. S5). In the case of GFP^{++} sorted cells, the corresponding subpopulation represented 89–92% of the total population

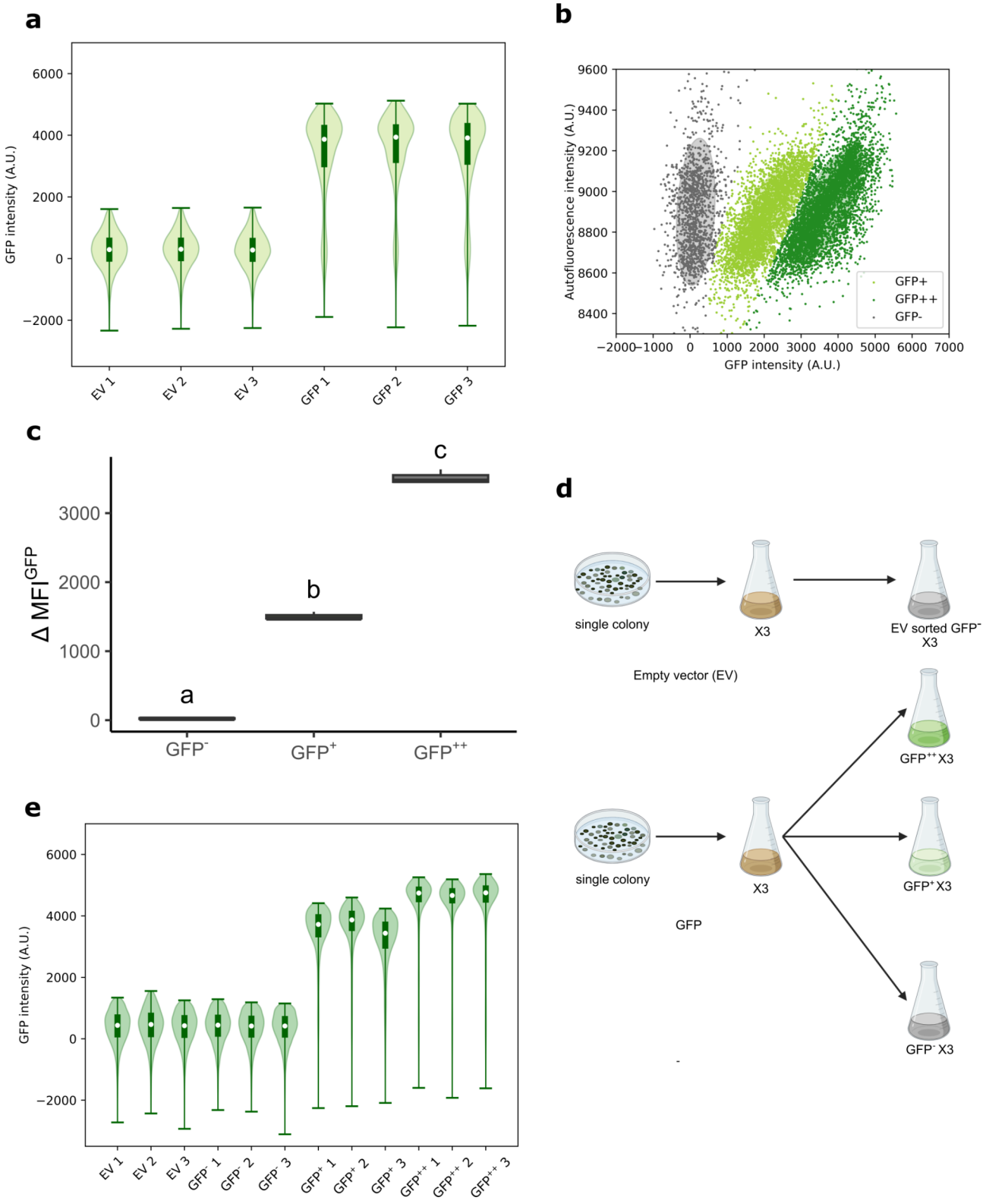

Fig. 1 Subpopulations are efficiently enriched through cell sorting. **(a)** Violin plots of per cell GFP fluorescence intensity of exconjugants harboring *eGFP* (GFP 1–3) and empty vector (EV 1–3) analyzed by flow cytometry. **(b)** Scatter plot of GFP cell line (GFP 1) before sorting with three populations 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)

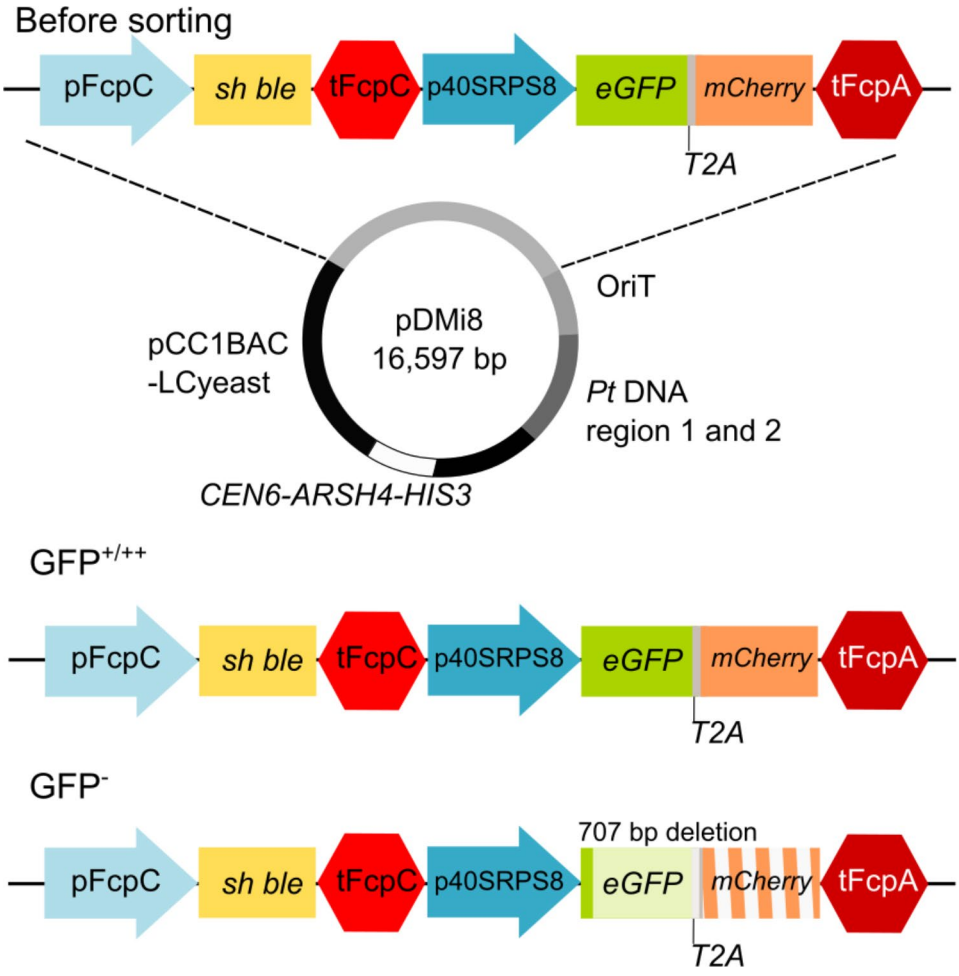

Fig. 2 Enriched GFP⁻ subpopulation exhibits episome rearrangement. Schematic representation of expression cassettes sequenced from the pDMi8 before and after sorting. The 707 bp deletion causes a frame shift of the expression cassette

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 and 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. 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. 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 4 kb and a lower band of 3.3 kb (Fig. S6).

Sequencing of the episomal expression cassettes showed that both GFP^{+} and GFP^{++} enriched cultures had identical sequences to the non-sorted culture (Fig. 2), while GFP⁻ 3.3 kb sequence was the product of a deletion

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

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

* 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. ($n = 3$)

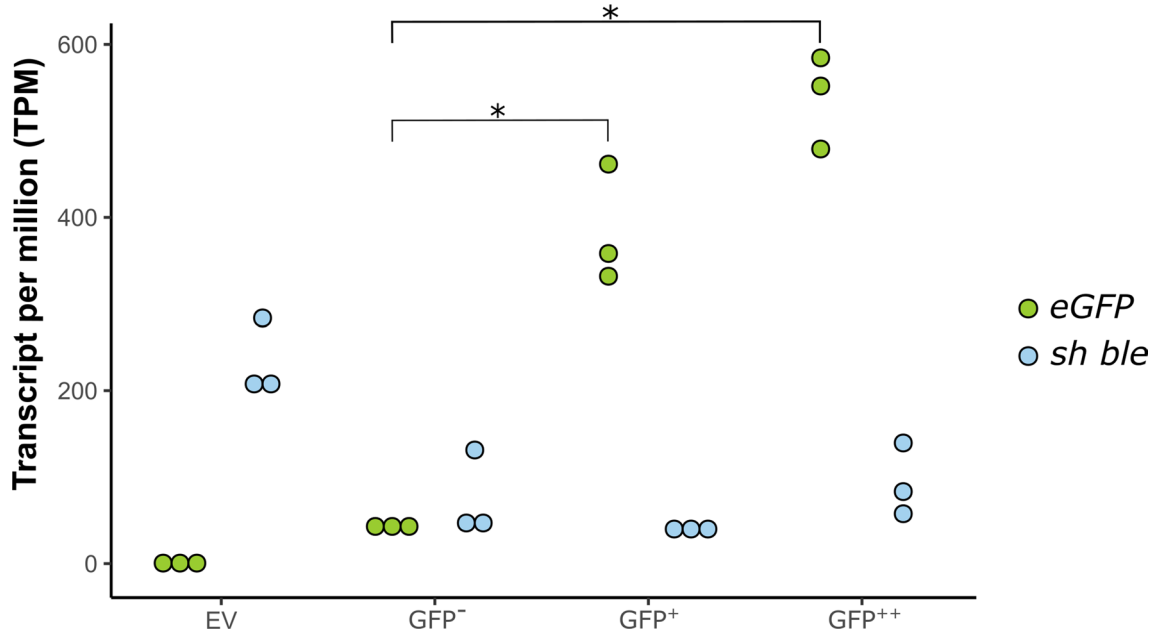

Fig. 3 *eGFP*, but not *sh ble*, is expressed at higher level in GFP positive subpopulations. Dot plot showing differences in *eGFP* expression between sorted subpopulations measured in transcript per million (TPM). Statistically significant differences were calculated using Kruskal-Wallis and Dunn's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

of 707 bp starting in *eGFP* which caused a frame shift of the *T2A* and *mCherry* sequences (Fig. S7). This is consistent with the GFP^{-} phenotype and its persistence following culture after sorting. Interestingly the GFP^{-} enriched culture also contained the 4 kb 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 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

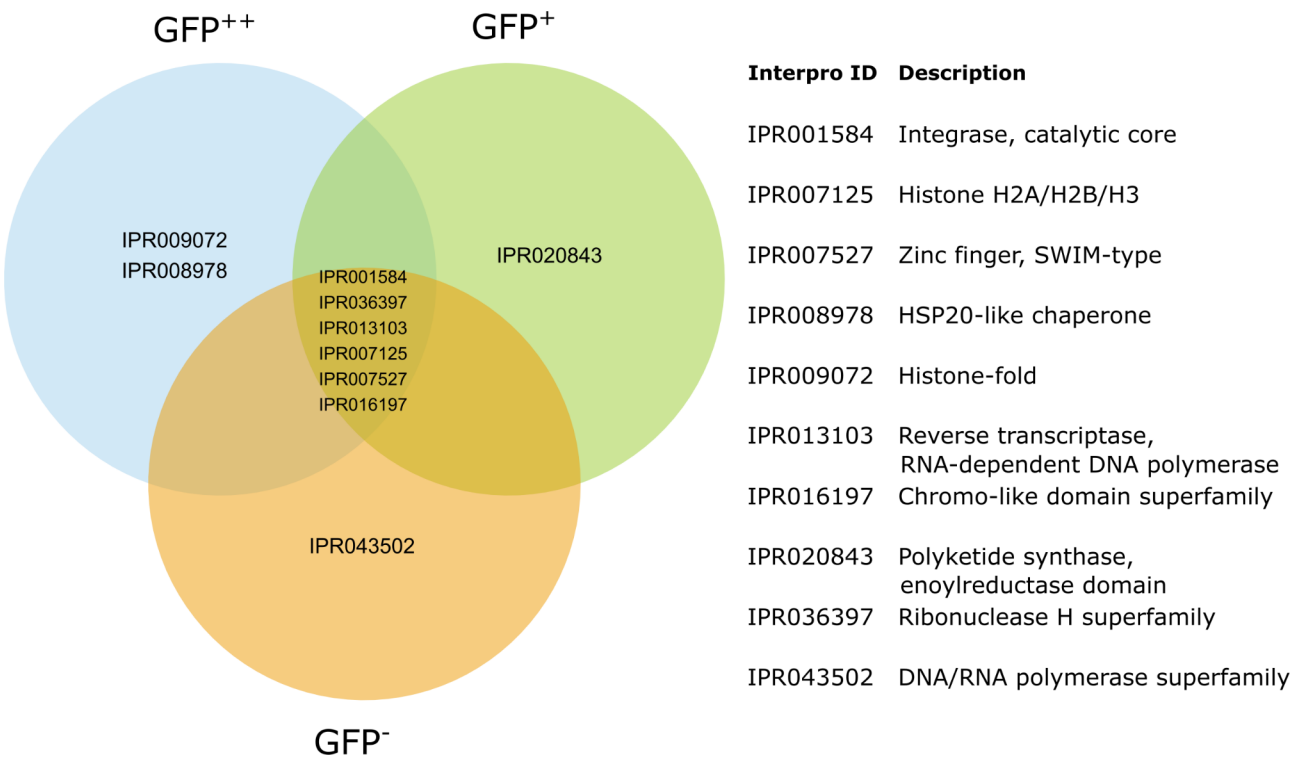

Fig. 4 Most of the protein families enriched in the down-regulated genes are shared between the three subpopulations. Venn diagram representing the overlap between subpopulations of InterPro protein families enriched in down-regulated genes compared to empty vector strain (EV). Genes down-regulated correspond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

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 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. S8), 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 ($\rho = 0.85$ and p-value < 0.001) with the fluorescence intensity observed in flow cytometry (Fig. 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

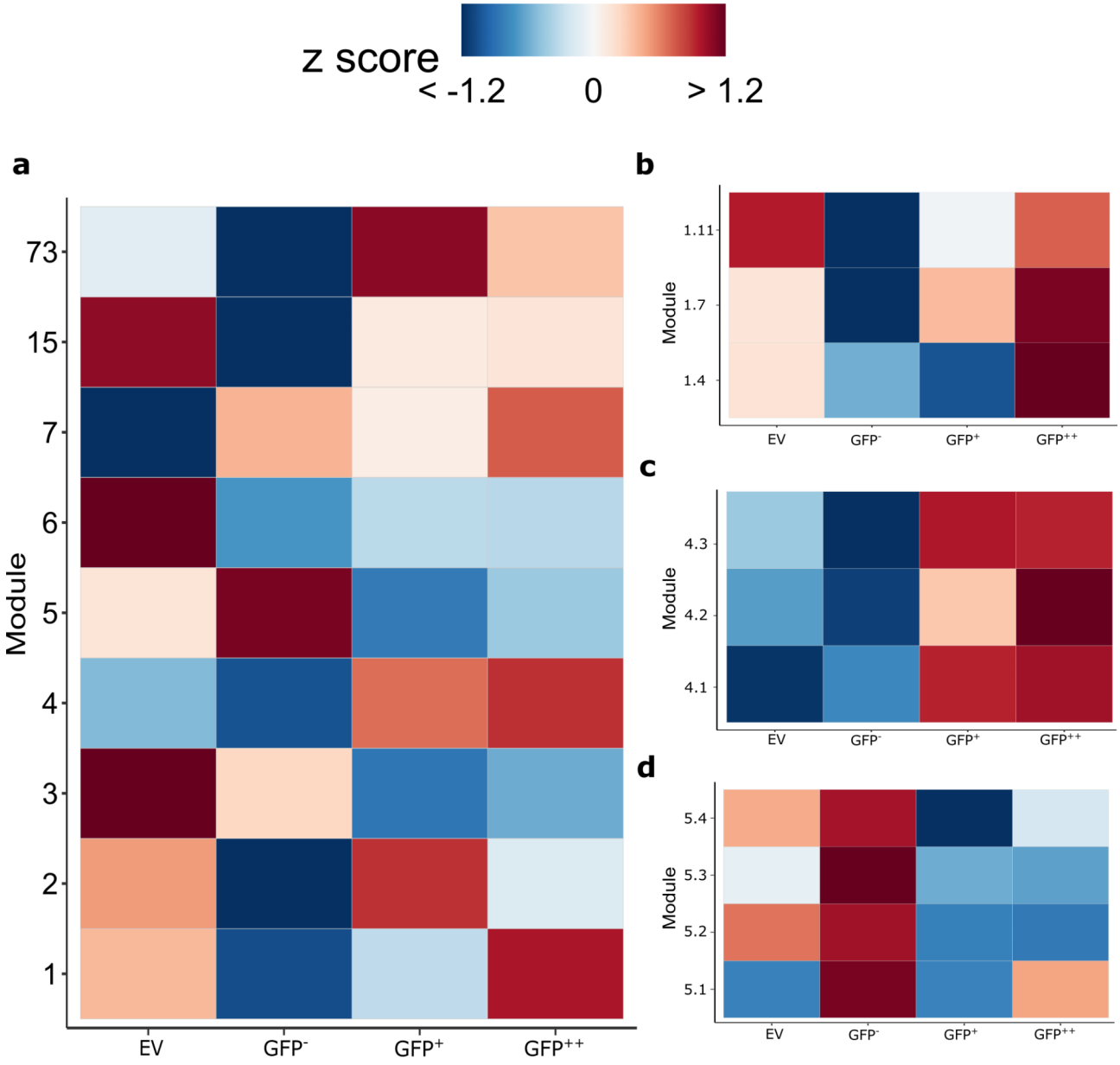

Fig. 5 Nested co-expression analysis highlights specific expression patterns in the subpopulations. Co-expression analysis of **(a)** all the highly correlated genes expressed in *P. tricornutum* enriched cultures, **(b)** genes from module 1, **(c)** genes from module 4, and **(d)** genes from module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

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. 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

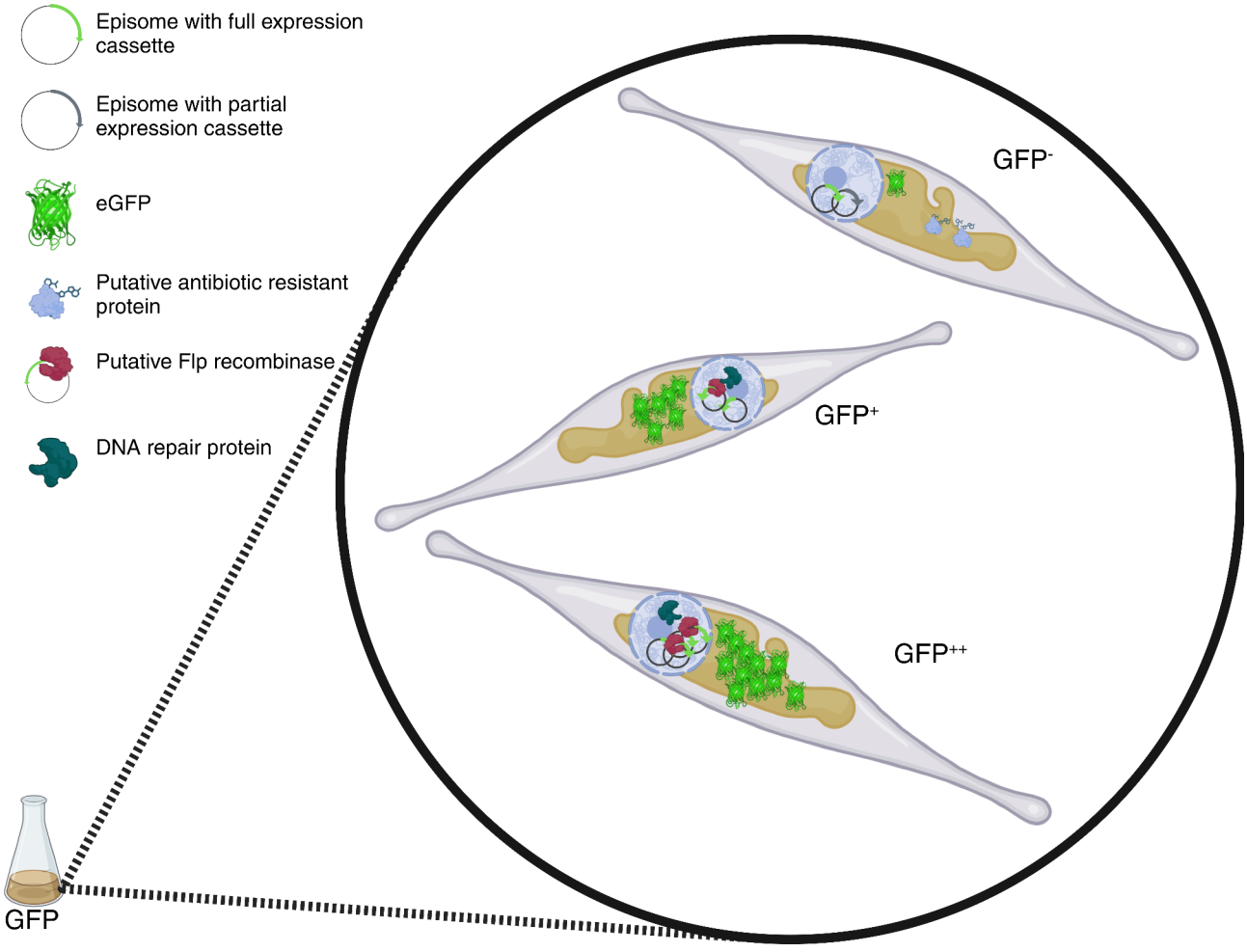

Fig. 6 Speculative mechanisms of *P. tricornutum* cells in a clonally propagated culture creating differences in genetic content and gene expression. Schematic representation of the proposed mechanisms causing phenotypic subpopulation in a single culture: GFP$^-$ down-regulation of DNA repair proteins 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

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. 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. 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 (Supplementary Table 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

Table 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 binding 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 (SRA1);
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	Gyrl-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

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), 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 [19]. 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 intensity 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 [20]. 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 [16, 21, 38]. 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 [39]. 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 [40]. 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 [41]. Sequence analysis revealed three identical nucleotides (GGA) at each extremity of the deletion, however only the two guanines were conserved after the repair (Fig. S7). 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 [42]. 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 4 kb sequence after amplification.

While based on the literature, big deletions are not common in episomes, rearrangements have been reported by Diamond et al. [21]. 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 [43]. 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 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 [21].

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 [11, 21, 44]. 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 [45]. 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 [46]. Yeast episomal plasmids, such as the 2micron, segregate by chromosome "hitchhiking": localizing at the telomeres of sister chromatids during mitosis to utilize spindle forces and localize to opposite cell poles [47]. In *S. cerevisiae*, centromeric plasmids containing centromeric (CEN) sequence and autonomously replicating sequence (ARS) elements segregate using chromosome machinery 48. Centromeric plasmid copy number can vary in a population, averaging one copy per haploid genome [49]. 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* [50]. 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 [7, 49], 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* [51].

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. 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. 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 [52, 53]. 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) [54, 55]. 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 [56], 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 [57]. 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 [58]. Also in this cluster was the SAM-dependent RNA methyltransferase, involved in rRNA methylation during ribosome biogenesis in humans [59]. A predicted recombinase Flp protein,

responsible for plasmid amplification and preventing copy number drop of the 2-micron plasmid in yeast [60], 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 [61]. 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 [61]. 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 [62], 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. 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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-024-02559-y.

Supplementary Material 1

Supplementary Material 2

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). 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.

Author 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.

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.

Data availability

Data availabilityRaw reads from RNA-seq are available at NCBI Sequence Read Archive BioProject ID PRJNA1108718.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 2 August 2024 / Accepted: 7 October 2024
Published online: 21 October 2024

References

1. Huang W, Daboussi F. Genetic and metabolic engineering in diatoms. Philos Trans R Soc B Biol Sci. 2017;372:20160411.
2. Sharma N, Simon DP, Diaz-Garza AM, Fantino E, Messaabi A, Meddeb-Mouelhi F, et al. Diatoms Biotechnology: various industrial applications for a greener tomorrow. Front Mar Sci. 2021;8:636613.
3. Russo MT, Rogato A, Jaubert M, Karas BJ, Falciatore A. Phaeodactylum tricornutum: an established model species for diatom molecular research and an emerging chassis for algal synthetic biology. J Phycol. 2023;59:1114–22.
4. Giguere DJ, Bahcheli AT, Slattery SS, Patel RR, Browne TS, Flatley M, et al. Telomere-to-telomere genome assembly of Phaeodactylum tricornutum. PeerJ. 2022;10:e13607.
5. Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, 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;8:4834.
6. Pampuch M, Walker EJL, Karas BJ. Towards synthetic diatoms: the *Phaeodactylum tricornutum* Pt-syn 1.0 project. Curr Opin Green Sustain Chem. 2022;35:100611.
7. Karas BJ, Diner RE, Lefebvre SC, McQuaid J, Phillips APR, Noddings CM, et al. Designer diatom episomes delivered by bacterial conjugation. Nat Commun. 2015;6:6925.
8. Windhagauer M, Abbriano RM, Ashworth J, Barolo L, Jaramillo-Madrid AC, Pernice M, et al. Characterisation of novel regulatory sequences compatible with modular assembly in the diatom *Phaeodactylum tricornutum*. Algal Res. 2021;53:102159.
9. Garza EA, Bielinski VA, Espinoza JL, Orlandi K, Alfaro JR, Bolt TM, et al. Validating a promoter Library for application in plasmid-based Diatom Genetic Engineering. ACS Synth Biol. 2023;12:3215–28.
10. Kassaw TK, Paton AJ, Peers G. Episome-based gene expression modulation platform in the Model Diatom Phaeodactylum tricornutum. ACS Synth Biol. 2022;11:191–204.

11. Slattery SS, Diamond A, Wang H, Therrien JA, Lant JT, Jazey T, 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;7:328–38.
12. Taparia Y, Dolui AK, Boussiba S, 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 [Internet]. 2022 [cited 2024 Mar 18];12. https://www.frontiersin.org/journals/plant-science/articles/https://doi.org/10.3389/fpls.2021.784780
13. Gao S, Zhou L, Yang W, Wang L, Liu X, Gong Y, et al. Overexpression of a novel gene (Pt2015) endows the commercial diatom Phaeodactylum tricornutum high lipid content and grazing resistance. Biotechnol Biofuels Bioprod. 2022;15:131.
14. Slattery SS, Giguere DJ, Stuckless EE, Shrestha A, Briere L-AK, Galbraith A, et al. Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom Phaeodactylum tricornutum for pandemic diagnostics. Sci Rep. 2022;12:7010.
15. Fabris M, George J, Kuzhiumparambil U, Lawson CA, Jaramillo-Madrid AC, Abbriano RM, et al. Extrachromosomal Genetic Engineering of the Marine Diatom Phaeodactylum tricornutum enables the Heterologous production of Monoterpenoids. ACS Synth Biol. 2020;9:598–612.
16. Fantino E, Awwad F, Merindol N, Diaz Garza AM, Gélinas S-E, Gajón Robles GC, et al. Bioengineering *Phaeodactylum tricornutum*, a marine diatom, for cannabinoid biosynthesis. Algal Res. 2024;77:103379.
17. Awwad F, Fantino EI, Héneault M, Diaz-Garza AM, Merindol N, Custeau A, 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.
18. Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD. Refinement of the Diatom Episome Maintenance Sequence and improvement of conjugation-based DNA delivery methods. Front Bioeng Biotechnol. 2016;4:65.
19. Tomoiaga D, Bubnell J, Herndon L, Feinstein P. High rates of plasmid cotransformation in E. Coli overturn the clonality myth and reveal colony development. Sci Rep. 2022;12:11515.
20. George J, Kahlke T, Abbriano RM, Kuzhiumparambil U, Ralph PJ, Fabris M. Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology. Front Bioeng Biotechnol [Internet]. 2020 [cited 2020 Dec 2];8. https://www.frontiersin.org/articles/https://doi.org/10.3389/fbioe.2020.00513/full
21. Diamond A, Diaz-Garza AM, Li J, Slattery SS, Merindol N, Fantino E, et al. Instability of extrachromosomal DNA transformed into the diatom Phaeodactylum tricornutum. Algal Res. 2023;70:102998.
22. Yurtsev E, Friedman J. FlowCytometryTools [Internet]. [object Object]; 2015 [cited 2024 Mar 18]. https://zenodo.org/record/32992
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
24. Hunter JD, Matplotlib. A 2D Graphics Environment. Comput Sci Eng. 2007;9:90–5.
25. Filloramo GV, Curtis BA, Blanche E, Archibald JM. Re-examination of two diatom reference genomes using long-read sequencing. BMC Genomics. 2021;22:379.
26. Lee C, Kim J, Shin SG, Hwang S. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. J Biotechnol. 2006;123:273–80.
27. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
28. Zhang Y, Park C, Bennett C, Thornton M, Kim D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3 N. Genome Res. 2021;31:1290–5.
29. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011;2011:bar030.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
31. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing. 2023. https://www.R-project.org/
32. Dos Santos KCG, Desgagné-Penix I, Germain H. Correction to: Custom selected reference genes outperform pre-defined reference genes in transcriptomic analysis. BMC Genomics. 2021;22:607.
33. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innov. 2021;2:100141.
34. Linlin Y, ggvenn. Draw Venn Diagram by ggplot2 [Internet]. 2023. https://CRAN.R-project.org/package=ggvenn
35. Li C, Deans NC, Buell CR. Simple tidy GeneCoEx: a gene co-expression analysis workflow powered by tidyverse and graph-based clustering in R. Plant Genome. 2023;16:e20323.
36. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4:1686.
37. Csárdi G, Nepusz T, Müller K, Horvát S, Traag V, Zanini F et al. igraph for R: R interface of the igraph library for graph theory and network analysis [Internet]. [object Object]; 2024 [cited 2024 Mar 18]. https://doi.org/10.5281/zenodo.7682609
38. Faessler AC. Optimising tools for metabolic engineering in the marine diatom Phaeodactylum tricornutum. 2024.
39. Bulankova P, Sekulić M, Jallet D, Nef C, Van Oosterhout C, Delmont TO, et al. Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. Curr Biol. 2021;31:3221–e32329.
40. LaFave MC, Sekelsky J. Mitotic recombination: why? When? How? Where? PLoS Genet. 2009;5:e1000411.
41. Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation. Trends Genet. 2014;30:85–94.
42. Matsui H, Harada H, Maeda K, Sugiyama T, Fukuchi Y, Kimura N, et al. Coordinated phosphate uptake by extracellular alkaline phosphatase and solute carrier transporters in marine diatoms. New Phytol. 2024;241:1210–21.
43. Snoeck S, Guidi C, 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.
44. Baiden N, Gandini C, Goddard P, 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.
45. Liu Z, Chen O, Wall JBJ, Zheng M, Zhou Y, Wang L, et al. Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. Sci Rep. 2017;7:2193.
46. Million-Weaver S, Camps M. Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? Plasmid. 2014;0:27–36.
47. Sau S, Ghosh SK, Liu Y-T, Ma C-H, Jayaram M. Hitchhiking on chromosomes: a persistence strategy shared by diverse selfish DNA elements. Plasmid. 2019;102:19–28.
48. Gnügge R, Rudolf F. Saccharomyces cerevisiae shuttle vectors. Yeast. 2017;34:205–21.
49. Tschumper G, Carbon J. Copy number control by a yeast centromere. Gene. 1983;23:221–32.
50. Diner RE, Noddings CM, Lian NC, Kang AK, McQuaid JB, Jablanovic J, et al. Diatom centromeres suggest a mechanism for nuclear DNA acquisition. Proc Natl Acad Sci U S A. 2017;114:E6015–24.
51. Rouches MV, Xu Y, Cortes LBG, Lambert G. A plasmid system with tunable copy number. Nat Commun. 2022;13:3908.
52. Liu K, Wessler SR. Transposition of mutator–like transposable elements (MULEs) resembles hAT and transib elements and V(D)J recombination. Nucleic Acids Res. 2017;45:6644–55.
53. Novikova O. Chromodomains and LTR retrotransposons in plants. Commun Integr Biol. 2009;2:158–62.
54. Longworth J, Wu D, Huete-Ortega M, Wright PC, Vaidyanathan S. Proteome response of Phaeodactylum tricornutum, during lipid accumulation induced by nitrogen depletion. Algal Res. 2016;18:213–24.
55. Avilan L, Lebrun R, Puppo C, Citerne S, Cuiné S, Li-Beisson Y, et al. ppGpp influences protein protection, growth and photosynthesis in Phaeodactylum tricornutum. New Phytol. 2021;230:1517–32.
56. Addinall SG, Downey M, Yu M, Zubko MK, Dewar J, Leake A, et al. A genome-wide suppressor and enhancer analysis of cdc13-1 reveals varied Cellular processes influencing Telomere Capping in Saccharomyces cerevisiae. Genetics. 2008;180:2251–66.
57. Qiu X-B, Shao Y-M, Miao S, Wang L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell Mol Life Sci CMLS. 2006;63:2560–70.
58. Zharikova NV, Iasakov TR, Bumazhkin BK, Patutina EO, Zhurenko EI, Korobov VV, et al. Isolation and sequence analysis of pCS36-4CPA, a small plasmid from *Citrobacter* sp. 36-4CPA. Saudi J Biol Sci. 2018;25:660–71.
59. Shen H, Stoute J, Liu KF. Structural and catalytic roles of the human 18S rRNA methyltransferases DIMT1 in ribosome assembly and translation. J Biol Chem. 2020;295:12058–70.

60. Sau S, Liu Y-T, Ma C-H, 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;5:21–8.
61. Ma C-H, Su B-Y, Maciaszek A, Fan H-F, Guga P, Jayaram M. A Flp-SUMO hybrid recombinase reveals multi-layered copy number control of a self-ish DNA element through post-translational modification. PLOS Genet. 2019;15:e1008193.
62. Buck JM, Río Bártulos C, Gruber A, Kroth PG. Blasticidin-S deaminase, a new selection marker for genetic transformation of the diatom Phaeodactylum tricornutum. PeerJ. 2018;6:e5884.

Publisher's note