Title: Endogenous GFP tagging in the diatom *Thalassiosira pseudonana*

Onyou Nam[1]*, Irina Grouneva[1]*, Luke C. M. Mackinder[1]

[1]Centre for Novel Agricultural Products, Department of Biology, University of York, York, YO10 5DD, UK
*Equal contribution
Corresponding author: L.C.M.M

Abstract:

The regulated abundance and spatial distribution of proteins determines cellular structure and function. The discovery of green fluorescent protein (GFP) and fusing it to a target protein to determine subcellular localization revolutionized cell biology. Most localization studies involve introducing additional copies of a target gene genetically fused to GFP and under the control of a constitutive promoter, resulting in the expression of the GFP-fusion protein at non-native levels. Here we have developed a single vector CRISPR/Cas9 guided GFP knock-in strategy in the diatom *Thalassiosira pseudonana*. This enables precise and scarless knock-in of GFP at the endogenous genomic location to create GFP fusion proteins under their native *cis* and *trans* regulatory elements with knock-in efficiencies of over 50%. We show that a previously uncharacterized bestrophin-like protein localizes to the CO_2-fixing pyrenoid and demonstrate that by measuring GFP fluorescence we can track relative protein abundance in response to environmental change. To enable endogenous tagging, we developed a Golden Gate Molecular Cloning system for the rapid assembly of episomes for transformation into *Thalassiosira pseudonana* via bacterial conjugation. In addition, this versatile toolbox enables CRISPR/Cas9 gene editing, provides a broad range of validated fluorophores and enables future large-scale functional studies in diatoms.

Significance statement

Fluorescent protein (FP) tagging is a widely utilized technique for understanding the spatial distribution of proteins. However, introducing extra gene copies under constitutive promoters that randomly integrate into the genome can result in non-biologically relevant expression levels, unwanted genomic mutations and localization artefacts. To overcome this, we developed a novel single vector system capable of CRISPR/Cas9-guided endogenous GFP tagging in a globally important model diatom. This allows scarless GFP knock-in at precise genomic locations resulting in GFP fusions regulated by native promoters/terminators, which facilitates accurate localization and determination of relative protein abundance. Moreover, the developed modular cloning framework is user-friendly and opens the door for high throughput large-scale studies, including FP tagging, knock-out, and knock-in.

Introduction

Diatoms are a diverse unicellular phytoplankton group that evolved as a consequence of a secondary endosymbiosis event of a red alga (1). They are key marine primary producers responsible for approximately 20% of global net primary production (2) and play central roles in the global carbon and silicate cycles (3–5). In a drive to understand the molecular processes that underpin the ecological success of diatoms, recent years have seen the development of a broad range of genetic tools for two model species, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. These include a range of transformation systems such as biolistics, electroporation and bacterial conjugation (6, 7) as well as an array of expression vectors for CRISPR/Cas9 gene editing, heterologous gene expression and fluorescent protein tagging (7–10).

Central to understanding gene function and cellular processes is to know the spatial organization of proteins in the cell. Localizing proteins by tagging the corresponding gene of interest with a fluorescent protein (FP) followed by microscopy to determine the subcellular localization is routinely completed in diatoms (11, 12). However, localization studies often rely on introducing additional copies of the gene under a strong constitutive promoter with these being randomly integrated into the genome. This potentially leads to overexpression of the target gene that might result in incorrect folding and targeting and nonspecific protein-protein interactions (13). In addition, random integration into the genome can result in unpredictable background genotypes and phenotypes (7). An alternative, yet typically more challenging, strategy is to insert a FP seamlessly into the genome in frame with the gene of interest. Known as “endogenous tagging” this approach maintains native promoter and terminator sequences typically resulting in expression levels comparable to native levels (reviewed (13)). In model organisms where homology-directed recombination (HDR) is dominant (e.g. yeast) endogenous tagging has enabled large-scale proteome tagging studies (14). However, in many eukaryotes a double stranded break is required for efficient HDR. This has been shown to be the case for diatoms where targeted genomic integration via HDR only occurs if (i) a double-stranded break has been introduced by a targeted endonuclease (15, 16) or (ii) components of the non-homologous end joining pathway have been knocked down (17). Previous reports have indicated a higher occurrence of HDR in *T. pseudonana* compared to *P. tricornutum* (15–18). The now routine application of CRISPR/Cas9 gene editing in diatoms (7, 19) opens up the potential for endogenous tagging. CRISPR/Cas engineering was first demonstrated in 2016 (8, 20) and can be achieved via stable integration into the genome of Cas9 and sgRNA (20), editing using RNP complexes (21) or expression of Cas9 and sgRNA(s) from an episome (9).

In this study we develop a seamless CRISPR/Cas9 guided knock-in strategy that enables endogenous GFP-tagging of proteins expressed from their native genomic locus and without the co-integration of an antibiotic marker. To achieve this, we first

built a Golden Gate based modular cloning (MoClo) framework for episomal delivery via conjugation in *T. pseudonana*. This framework enables rapid and scalable cloning for a broad range of applications and requires no specialist equipment for transformation. We demonstrate versatility by validating a range of promoter-terminator pairs, testing a broad spectrum of fluorophores, and performing CRISPR/Cas9 gene editing. We then expand the methodology to enable precise genomic knock-in of GFP for endogenous protein tagging using a single vector system.

Results

Development of a Modular Cloning framework in *T. pseudonana*. To enable the versatile application of fluorescent protein tagging and CRISPR/Cas9 gene editing combined with the ease of bacterial conjugation we developed an episomal based Golden Gate MoClo framework (Fig. 1A; see SI appendix Table S1 for vectors). Golden Gate cloning uses Type IIS restriction enzymes that cut outside of their recognition sites allowing user defined overhangs (syntax). Assemblies are hierarchical, typically individual level 0 parts (i.e. promoter, coding sequence, FP, terminator) are assembled into level 1 transcriptional units, that are further assembled into a level 2 transformation vector. To be compatible with previously developed MoClo parts (8) and maintain some compatibility with the universal Loop (uLoop) assembly framework (10) we used the Plant/*Chlamydomonas* MoClo syntax (22, 23). Our standard level 2 episome used for diatom transformations contains: an episome maintenance part, *CEN6-ARSH4-HIS3,* that prevents random integration into the genome and enables it to replicate and be maintained as an episome inside *T. pseudonana* (6); the origin of Transfer (*oriT*) required for vector transfer during conjugation (6) and; the nourseothricin N-acetyl transferase (NAT) gene conferring resistance to the antibiotic nourseothricin. As the *CEN6-ARSH4-HIS3* and *oriT* are always required in combination these were combined to form a single level 1 part and occupy position 1 in our L2 assemblies with the *NAT* marker typically position 2. Depending on application, positions 3-5 of our L2 assemblies can be occupied with different L1 plasmids assembled from L0 parts or custom made with the correct flanking syntax e.g. sgRNAs (Fig. 1A).

We initially tested our assembly framework by assembling an episome for CRISPR/Cas9 editing of the urease gene using previously validated sgRNAs (8) (Fig. 1B). After transformation and selection of *T. pseudonana* colonies the urease locus was PCR amplified and cloned. Sequencing indicated that we could efficiently edit the urease gene using our episome assembly and delivery framework (Fig. 1C). We next tested our Golden Gate episome system for fluorescent protein tagging by expressing an additional copy of our gene of interest (GOI) C-terminally fused to monomeric enhanced GFP (mEGFP; GFP from here on; Fig. 1D). We first validated this on the previously localized (24) mitochondrial gamma carbonic anhydrase 3 (γCA3) using the fucoxanthin chlorophyll a/c-binding protein (FCP) promoter and FCP terminator (Fig. 1E). We next applied it to a previously unlocalized target gene, bestrophin-like protein

(THAPSDRAFT_4820), which we termed BST2. BST2 has previously been shown to be transcriptionally and translationally induced at low CO_2 levels (25, 26). Bestrophin-like proteins have been shown to play an important role in the CO_2 concentrating mechanism of *Chlamydomonas reinhardtii*. They are proposed to channel HCO_3^- into the thylakoids that traverse the Rubisco containing pyrenoid, enabling its sequential dehydration to CO_2 and fixation by Rubisco (27). We showed that BST2-GFP localized to a defined elongated spot at the center of the chloroplast, and we tentatively assign this to the membranes that traverse the pyrenoid (Fig. 1F). To expand the available fluorophores and regulatory regions of our system we tested and validated the fluorophores mNeonGreen, mScarlet-i and mTurquoise2 (Fig. 1E and F) and the previously developed but untested promoter/terminator pairs of elongation factor 2 (EF2) and light harvesting complex f 6 (LHCF6) from the uLoop system (10) (SI appendix Fig. S1).

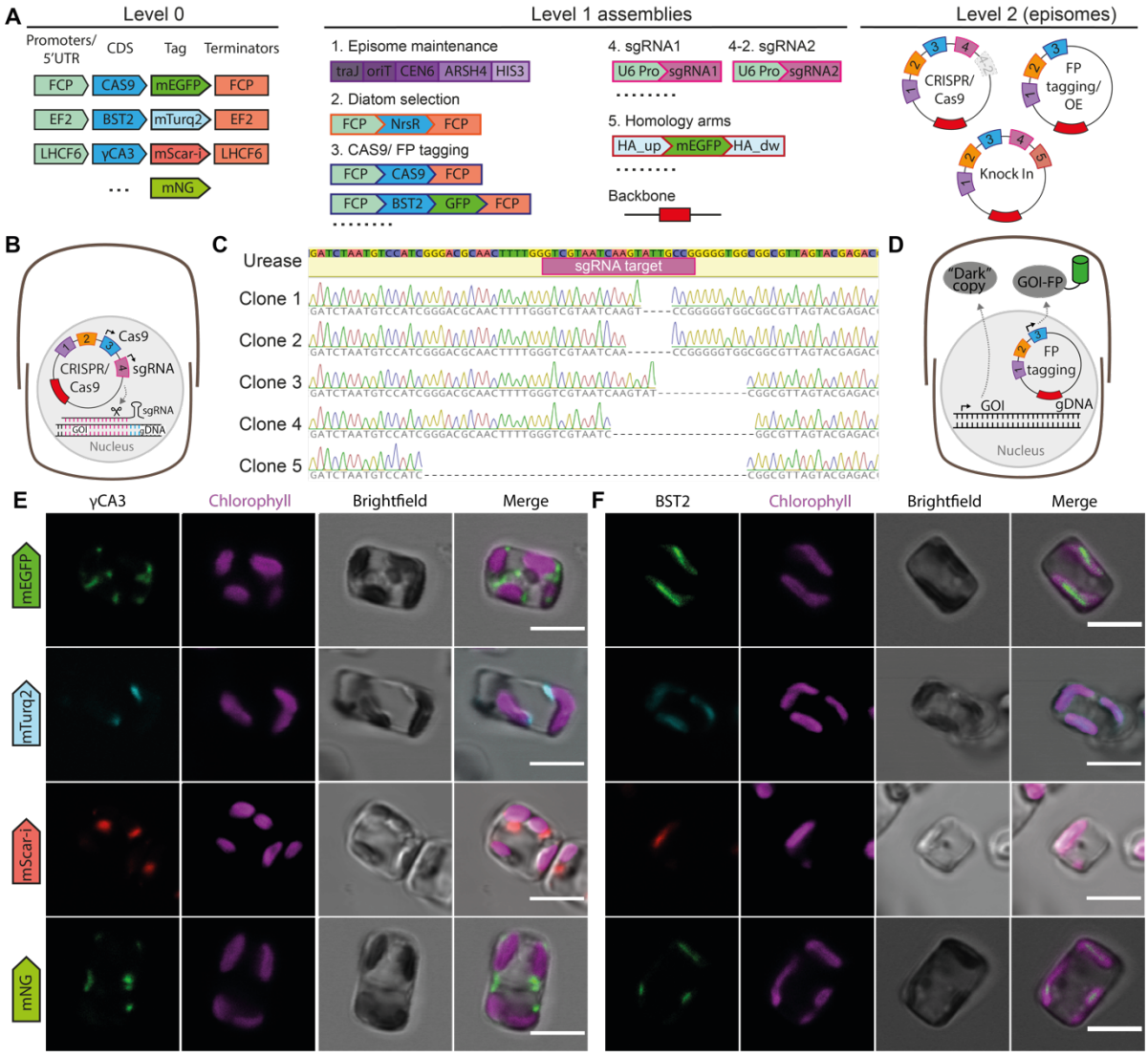


Fig. 1. Establishment of a Golden Gate based cloning framework for episome delivery of constructs in *T. pseudonana*. *(A)* Overview of Golden Gate assembly parts developed. These parts

can be combined to assembly episomes for CRISPR/Cas9 gene editing, fluorescent protein (FP) tagging and FP knock-in. *(B)* Cas9 and sgRNA targeting the Urease gene were expressed from an assembled episome. *(C)* After CRISPR/Cas9 gene editing the Urease locus was PCR amplified, cloned and sequenced from episome containing colonies. *(D)* Episomes expressing a gene of interest (GOI) C-terminally fused to a fluorescent protein were assembled and transformed. This results in a tagged version and the native untagged ("dark") version of the protein. *(E)* and *(F)* Confocal imaging of developed fluorophores validated on γCA3 and BST2. Scale bars: 5 µm.

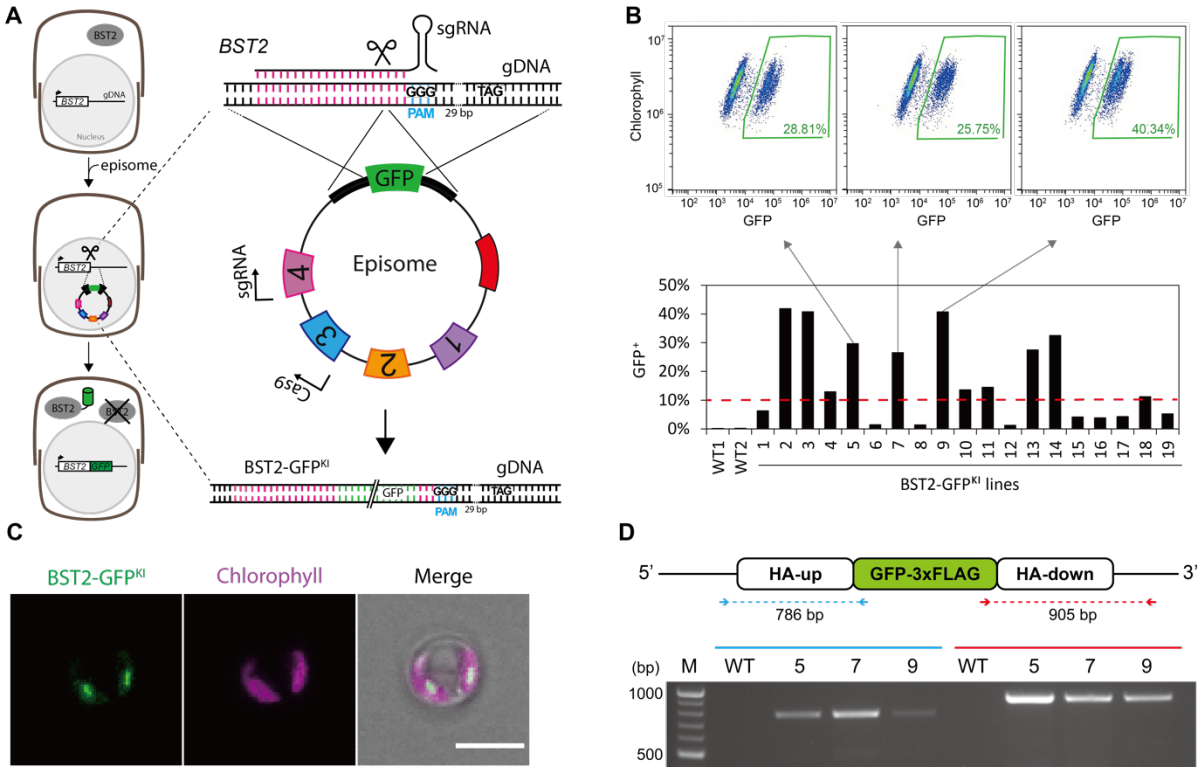
Development of a scarless endogenous GFP tagging strategy. The "gold standard" of fluorescent protein tagging is to scarlessly insert a fluorescent protein into the genome resulting in the target protein being labelled whilst still under the control of its native *cis* and *trans* regulatory elements (13). To achieve this, we set out to establish a CRISPR/Cas9 guided HDR GFP knock-in approach that could be delivered on a single episome (Fig. 2A). To test our approach, we attempted to endogenously GFP tag BST2, which we previously localized via expression as a fusion protein directly from episomes (Fig. 1F).

Using CRISPick (28) (see methods) we selected a sgRNA with a predicted on-target score of 0.7 that cut 40 bp upstream of the BST2 stop codon. We designed homology arms that covered both sides of the predicted Cas9 cut site (Fig. 2A). In this case successful BST2-GFP knock-in (BST2-GFP^{KI}) would lead to the disruption of the sgRNA target region preventing re-cutting by Cas9. GFP knock-in would result in the GFP being inserted 13 amino acids upstream from the stop codon (Fig. 2A). AlphaFold modeling (29) of BST2 and BST2-GFP^{KI} indicated that this would be in a region of low predicted structure and unlikely to impact BST2 folding (SI appendix Fig. S2). Homology arms were either PCR amplified from genomic DNA or custom synthesized to enable domestication (removal of BsaI and BpiI restriction sites) when required. The domestication of any homology arms was done via synonymous mutations to avoid changes at the amino acid level. Homology arms were assembled into a L1 plasmid separated by the in-frame GFP sequence. The L0 GFP vector used contained no stop codon, with the stop codon coming from the native stop codon within the downstream homology arm. The resulting homology arm GFP flanked L1 construct (GFP repair template) was assembled into an L2 plasmid. This L2 plasmid contained five L1 parts including [1] *CEN6-ARSH4-HIS3* and *oriT*, [2] *NAT* selectable marker, [3] Cas9, [4] sgRNA and [5] GFP repair template (Fig. 2A). Assembled episomes were transformed into *E. coli* EPI300 cells carrying the pTA_Mob mobility plasmid that enables transfer of the episome to the target organism. Episome and pTA_Mob containing *E. coli* were incubated with *T. pseudonana* and selected on nourseothricin containing plates.

Identifying and genotyping endogenously tagged GFP knock-in lines. Nourseothricin resistant colonies were transferred to 24-well plates for growth in liquid media. Liquid grown transformants were screened via flow cytometry using a CytoFLEX that enables rapid screening in a 96-well plate format. Initial data for our BST2-GFP^{KI} construct indicated that 58% (11/19) of transformants were GFP positive when a stringent 10% cut-off for GFP positive cells was used (Fig. 2B and SI appendix

188 Fig. S3). Lines that showed a high population of GFP cells were selected for
189 microscopy (Fig. 2C) with the localization of BST2-GFP^{KI} showing a linear traversion
190 at the center of the chloroplast most likely within the thylakoid membrane traversing
191 the pyrenoid. This data corresponded to when BST2-GFP was expressed from an
192 episome (Fig. 1F), indicating that in this case neither the native expression levels nor
193 insertion of GFP 13 amino acids upstream of the C-terminal affected localization. We
194 confirmed scarless insertion of *GFP* via amplification and sequencing of the insertion
195 flanking regions using oligos located outside the homology arms to exclude the
196 possibility of amplicons originating from the episome (Fig. 2D). To ensure that Cas9
197 mediated DNA cutting was required for GFP insertion, and to check that GFP could
198 not be expressed from the episome prior to HDR, we assembled an episome that
199 lacked the sgRNA targeting *BST2*. Screening of 15 colonies via flow cytometry showed
200 no GFP positive colonies (SI appendix Fig. S4A). The presence of the episome was
201 confirmed by PCR against the *NAT* gene (SI appendix Fig. S4B). This indicates that
202 Cas9 DNA cleavage is required for efficient GFP knock-in and GFP is not expressed
203 without successful integration into the genome.

204



205

Fig. 2. Endogenous GFP tagging strategy overview and transformant screening. *(A)* Endogenous GFP tagging strategy. *T. pseudonana* is transformed with an episome containing (1) *CEN6-ARSH4-HIS3* and *oriT*, (2) *NAT* selectable marker, (3) Cas9, (4) sgRNA and the repair template containing GFP. Cas9 cutting enables homology-directed repair by the repair template generating a GFP gene fusion. To prevent Cas9 re-cutting after homology directed repair the repair template disrupts the sgRNA recognition region. The layout of the schematics and indicated base pairs represent what was used for the BST2-GFP^{KI}. *(B)* Screening colonies for fluorescence: Nourseothricin resistant colonies were picked, grown in liquid, and then screened for fluorescence using a CytoFLEX flow cytometer. The green box represents the percentage of GFP positive cells. Lines with >10% (dashed red line) of GFP positive cells were classified as successful BST2-GFP^{KI} lines. *(C)* Confocal images for a BST2-GFP^{KI} line identified from flow cytometry. Scale bar: 5 µm**.** (D) Genotyping of BST2-GFP^{KI} lines showing the presence of GFP at the targeted genomic locus of BST2.

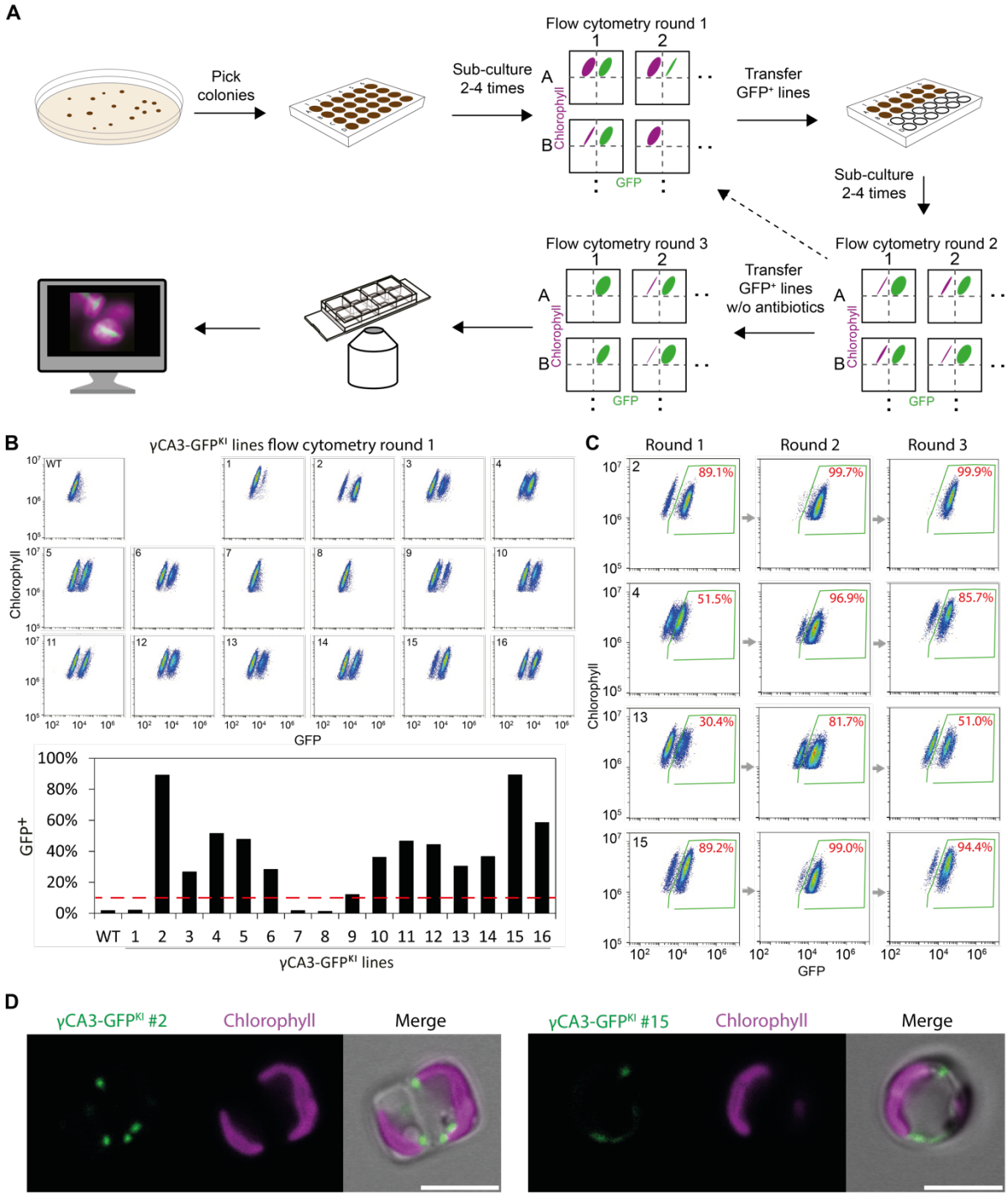


Fig. 3. Development of a flow cytometry screening pipeline. *(A)* Developed pipeline to identify and select colonies with high GFP knock-in fluorescence. *(B)* Round 1 flow cytometry counting of γCA3 transformants. Bar chart corresponds to the above plots. Lines with >10% (dashed red line) of GFP positive cells were classified as successful γCA3-GFP^{KI} lines. *(C)* Four lines from round 1 flow cytometry counting were subcultured on antibiotics for ~3-4 generations and the percentage of GFP positive cells counted (Round 2). Lines were transferred to antibiotic free media for episome curing, subcultured for ~3-4 generations and then counted (Round 3). *(D)* Confocal microscopy of two γCA3-GFP^{KI} lines. Scale bar 5 µm.

Determining knock-in efficiency and isolating lines with close to 100% GFP positive cells. From our flow cytometry data (Fig. 2B) and GFP imaging we typically

232 saw a mixed population of GFP positive and GFP negative cells suggesting the
233 presence of mosaic colonies, where different genotypes are present within the same
234 colony (30). To isolate pure clonal lines, we developed a flow cytometry fluorescence
235 cell counting pipeline and optimized this for the GFP knock-in of γCA3 (Fig. 3A and
236 B). Our pipeline typically involves three rounds of flow cytometry. Transformants are
237 picked and grown in the presence of antibiotics in 24-well plates with weekly
238 subculturing. After 2-3 weeks (~6-9 generations) these are then screened via flow
239 cytometry. Lines with the highest percentage of GFP positive cells are further
240 propagated for ~3-4 generations and re-screened under antibiotic pressure. Once
241 lines are close to 100% GFP positive they are transferred to antibiotic free media to
242 drive episome loss. First round flow cytometry data gave us the percentage
243 transformants with positive GFP signal. From this data we conclude that knock-in
244 efficiency of at least one allele being successfully tagged in *T. pseudonana* for BST2
245 is 58% (Fig. 2B) and for γCA3 is 81% (Fig. 3B). Further rounds of flow cytometry
246 screening indicated that multiple lines with close to 100% GFP positive cells could be
247 isolated using this approach (Fig. 3C). Confocal microscopy of these lines showed a
248 typical mitochondrial localization pattern (Fig. 3D).

249 **Screening enables biallelic editing.** Vegetative *T. pseudonana* is diploid. GFP
250 knock-in at only a single allele potentially is not stable due to chromosomal cross-over
251 (31) that could result in GFP removal and revertant back to a wild-type genotype. In
252 addition, in a hetero-allelic knock-in GFP fluorescence would not represent native
253 protein abundance due to presence of an untagged copy. To examine if GFP has been
254 knocked into both alleles in our clonal lines we PCR amplified the knock-in genomic
255 locus with primers that flank the homology arms. In a biallelic knock-in we would expect
256 a single band corresponding to the locus containing GFP, whereas in a non-clonal or
257 mono-allelic line we would expect an additional lower molecular weight wild-type band.
258 In both BST2-GFP^{KI} and γCA3-GFP^{KI} lines with close to 100% GFP positive cells we
259 could detect only a single band indicating successful GFP integration into both alleles
260 (Fig. 4A and B). Sequencing across the homology arm and genomic DNA junctions
261 and across the GFP homology arm junctions indicates scarless GFP insertion (Fig. 4C
262 and D). Immunoblotting against the GFP tag of the confirmed biallelic lines indicated
263 that both BST2-GFP^{KI} and γCA3-GFP^{KI} were at the expected size (Fig. 4E and F).

264

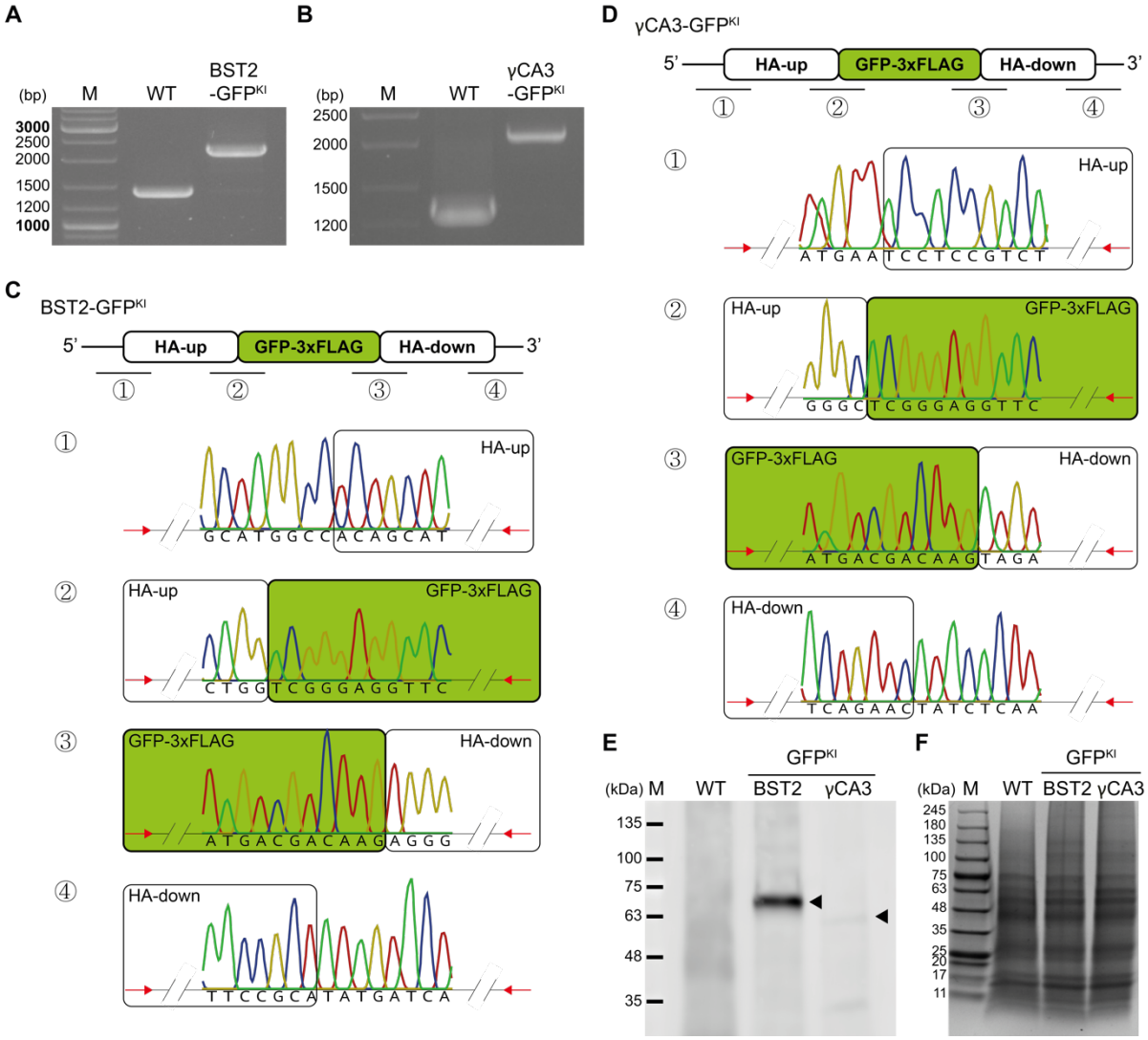


Fig. 4. Confirmation of biallelic GFP knock-ins. *(A)* and *(B)* PCR to confirm biallelic editing of clonal lines. *(C)* and *(D)* Sequencing of junctions confirming scarless integration of GFP into the genome at the target location. *(E)* Immunoblot against GFP in BST2-GFP^{KI} and γCA3-GFP^{KI}. Arrow heads indicate the expected sizes of BST2-GFP (73 kDa) and γCA3-GFP (57 kDa). *(F)* Coomassie stained SDS-PAGE loading control.

Relative protein abundance can be tracked by fluorescence intensity. As GFP knock-in lines are under their native regulatory elements we tested if changes in relative protein abundance could be tracked in response to an environmental change. As BST2 was previously shown to increase in abundance ~4-fold at low CO_2 (26) we grew a BST2-GFP^{KI} line at both high and low CO_2 and then measured GFP intensity per cell using flow cytometry (Fig. 5A). We saw an ~2-fold increase in GFP fluorescence in low CO_2 grown cells demonstrating that relative protein abundance can be tracked using our GFP knock-in lines (Fig 5B and C).

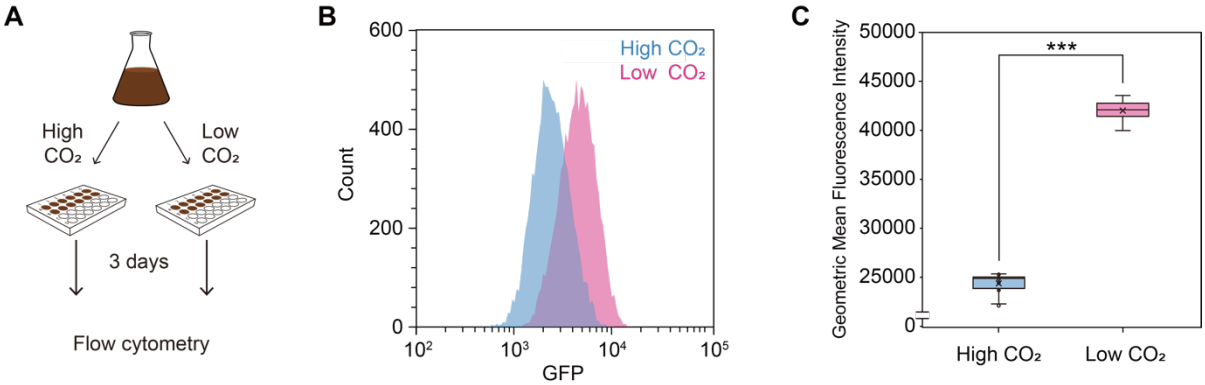


Fig. 5. Relative protein abundance can be tracked in GFP knock-in lines. *(A)* A biallelic BST2-GFP^{KI} line was grown in ambient CO_2, transferred to both high CO_2 and low CO_2 for three days then GFP intensity measured by flow cytometry. *(B)* and *(C)* Low CO_2 grown cells show a clear shift in fluorescence intensity (***p<0.001, Students *t* test).

Discussion

We developed a single plasmid-based approach for high efficiency endogenous GFP tagging in the diatom *T. pseudonana*. Our method enables the scarless insertion of GFP at a precise genomic locus to generate a fluorescent protein fusion without the co-integration of a selection marker. This ensures that native *cis* and *trans* regulatory regions are maintained enabling relative protein abundance to be tracked in response to perturbations. To demonstrate the power of our approach we applied it to BST2, a bestrophin-like protein. The localization of endogenously GFP tagged BST2 to an elongated central chloroplast region corresponds to the thylakoids traversing the pyrenoid. Tracking BST2-GFP^{KI} relative abundance using flow-cytometry showed an increase at low CO_2 supporting previous transcriptional (25) and proteomic (26) data. As all of the native protein copies are GFP fusions in our biallelic BST2-GFP^{KI} lines, we envision that single molecule counting approaches can be used in the future to determine absolute protein numbers (32). Collectively, the data supports an analogous role of BST2 to *Chlamydomonas* bestrophin-like proteins implicated in the CO_2 concentrating mechanism (27). In *Chlamydomonas*, BST1-3 are low CO_2 upregulated and localized to the pyrenoid periphery with a proposed role to deliver HCO_3^- from the stroma to the thylakoid lumen where it is subsequently dehydrated by a carbonic anhydrase to release CO_2 within the pyrenoid for fixation by Rubisco (27). Further evidence supporting an analogous mechanism in diatoms is the presence of a θ-carbonic anhydrase in the lumen of pyrenoid-traversing thylakoids in *P. tricornutum* that is required for efficient CO_2 fixation (33). As algal bestrophin-like proteins are further characterized it will be interesting to see if they play a central role in evolutionary diverse pyrenoid based CO_2 concentrating mechanisms.

Whilst HDR endogenous FP knock-in approaches have been developed for several model systems, the seamless knock-in of FPs has yet to be achieved for eukaryotic algae. In photosynthetic eukaryotes a GFP knock-in strategy has been established in the model plant *Arabidopsis thaliana* by sequential expression of a sgRNA and GFP

318 donor template in a parental line expressing Cas9 (34). In-line with our single plasmid
319 approach, the authors also tested the presence of all components on a single T-DNA
320 construct, however this failed to result in heritable GFP knock-in (34). As *T.*
321 *pseudonana* is diploid, an additional challenge associated with endogenous tagging is
322 the knock-in of the GFP at both alleles. To overcome this, we maintained
323 transformants under antibiotic selection pressure and developed a rapid and scalable
324 flow cytometry screening pipeline. This resulted in the generation of homoplasmic
325 knock-in lines.
326
327 During the establishment of our GFP knock-in approach we only tested homology
328 arms of ~600 bp and the knock-in of GFP (~800 bp). Further tests to identify the
329 minimum homology arm size requirement could lower homology arm synthesis costs
330 in the future. In addition, exploring the limitations of knock-in fragment size would be
331 useful to determine if the approach could be used for multi-gene/pathway knock-in at
332 neutral genomic sites for biotechnology applications. We also only tested guide block
333 knock-ins where the GFP is inserted into the Cas9 cut-site. It has been demonstrated
334 in other organisms that HDR can be offset from the cut-site which would then enable
335 GFP to be inserted directly prior to the stop codon and increase sgRNA flexibility.
336 However, knock-in efficiency can rapidly drop off as the cut-site to HDR-site distance
337 increases (35).
338
339 Our knock-in efficiency of 58-81% appears to be high compared to other systems (34,
340 36, 37). This reduces the effort required for screening and selection of successful GFP
341 knock-in lines. One potential explanation for this high efficiency is the presence of all
342 the necessary engineering elements on a single episome that can be continually
343 maintained through antibiotic selection. The high efficiency of our system indicates
344 that the same framework could be used for insertions, substitutions and functional
345 knock-outs at specific target sites. However, for biallelic editing additional screening
346 strategies, other than fluorescence-based flow cytometry, would need to be
347 established.
348
349 To enable the successful and easy application of our system we first developed and
350 validated a Golden Gate based episome assembly framework for multi-purpose
351 genetic engineering in *T. pseudonana* (Fig. 1). This included the episomal testing of
352 two additional promoter/terminator pairs, the validation of three additional fluorophores
353 (mTurquoise2, mScarlet-i and mNeonGreen) and the establishment of episomal based
354 CRISPR/Cas9 gene editing. The Golden Gate based episomal assembly framework
355 allows large, multi-part constructs to be rapidly and easily assembled. Episomal
356 delivery by bacterial conjugation enables the delivery of large constructs, avoids
357 random integration into the genome and requires no specialized equipment such as a
358 Gene gun. This engineering framework now opens the door for large-scale fluorescent
359 protein tagging and mutant library generation in diatoms. Similar studies in other
360 systems have proven to be extremely powerful for understanding gene function and
361 biological processes (14, 38–41).

362 **Methods**

363 *Strains and growth conditions*

364 *Thalassiosira pseudonana* CCAP1085/12 (equivalent strain CCMP1335) was
365 obtained from the Scottish Culture Collection of Algae and Protozoa (CCAP) and
366 grown axenically at 20°C with a continuous illumination of ~50 µmol photons m^{-2} s^{-1}.
367 in F/2 medium (42, 43) prepared with artificial sea water (33 g L^{-1}, Instant Ocean SS15-
368 10).

369 *Episome assemblies using Golden Gate cloning*

370 Level 0 (L0), Level 1 (L1) and Level 2 (L2) plasmids were assembled by Golden Gate
371 (GG) cloning (44) using a range of custom parts, uLoop (10) parts and Golden Gate
372 MoClo Plant Kit parts (45) (SI appendix Table S1). The assembly framework and
373 syntax was according to Patron et al. 2015 (22). Available *T. pseudonana* expression
374 cassettes for Cas9, the *NAT* selection marker, *U6* promoter and sgRNA backbone
375 were used to construct L1 vectors (8). With both Cas9 and the *NAT* selection marker
376 under the expression of the *FCP* promoter and terminator. The yeast element *CEN6-*
377 *ARSH4-HIS3* (6) was domesticated for GG cloning by removing BsaI and BpiI
378 restriction sites and cloning it into a L1 plasmid (pICH47732) together with oriT and
379 traJ. The resulting L1 part permits low-copy episomal replication in diatoms without
380 random integration into the genome (6). Depending on application (i.e. gene knock-
381 out, FP-tagging or GFP knock-in) 3-5 L1 parts can be assembled into an L2 plasmid
382 (termed episome) for transformation into *T. pseudonana*.

383 Each one-tube, one-step, restriction-ligation GG assembly contained 40 fmols of each
384 component. Final reaction volumes of 20 µl were incubated in 10x ligase buffer
385 (Promega) with 10 units restriction enzyme (BsaI-HFv2, NEB or BpiI, Thermo Fisher
386 Scientific) and 10 units T4 DNA ligase (HC, 20 units $µl^{-1}$, Promega). In a thermocycler,
387 the reaction was switched between 16°C and 37°C at 5 min intervals for 30 cycles,
388 followed by 5 min at 37°C, then increased to 65°C for 20 min. 2.5 µl were transformed
389 into 50 µl DH5α *E.coli* cells and selected on antibiotic containing plates.

390 *Transformation of* T. pseudonana

391 Golden Gate assembled episomes were transformed into electrocompetent *E.coli*
392 (TransforMax EPI300) already carrying the mobility plasmid pTA_Mob (46) (gift from
393 R. Lale). *E.coli* colonies carrying both mobility and cargo plasmid were used for
394 subsequent conjugation. Bacterial conjugation was performed according to (6) with
395 the minor adjustment that post conjugation *T. pseudonana* cells were left to recover
396 under standard growth conditions (20°C, 50 µmole photons m^{-2} s^{-1}) for 24 hours
397 instead of four hours before spreading onto selective media containing 100 µg ml^{-1}
398 Nourseothricin.

399 *Gene targets and homology arm design*

400 The following genes were targeted for GFP knock-in: BST2 (THAPSDRAFT_4820)
401 and γCA3 (THAPSDRAFT_265081).

402 A template for HDR was integrated into an episome as a further L1 component. It
403 consisted of i) a domesticated 600 bp sequence upstream of the stop codon of the
404 targeted gene (HA_up), ii) in-frame GFP without a stop codon and iii) a domesticated
405 600 bp sequence downstream of but including the stop codon (HA_down). sgRNAs
406 used in this experiment were selected so that the cut site they induced would be as
407 close to the stop codon as possible. For both genes a guide blocking approach was
408 used whereas after insertion of the GFP via HDR the sgRNA recognition site would be
409 split preventing further cutting. Domesticated 600 bp homology arms (upstream and
410 downstream of the selected Cas9 cut site) containing restriction sites and junctions for
411 GG cloning were either gene synthesized (Twist Bioscience) or amplified from
412 genomic DNA. GFP was fitted with a short flexible linker and a 3xFLAG tag.

413 *Genotyping/screening for HDR*

414 Fifty microliters were harvested from each transformant line grown in liquid selection
415 medium. Pellets were resuspended in 20 µl of lysis buffer (20 mM Tris-HCl pH 8.0, 10
416 mM EDTA, 10% Triton X-100) and kept on ice for 10-20 min, followed by incubating
417 at 95°C for 10 min. Cell debris was removed by brief centrifugation and 1 µl of
418 supernatant used as PCR template. The OneTaq Hot Start Quick-load 2x Master Mix
419 with GC buffer (M0485L, New England BioLabs) was used for PCR. All primers used
420 in this study are listed in SI appendix Table S2.

421 *Confocal microscopy*

422 Images were taken with a laser scanning microscope (LSM880, Zeiss) using 63x
423 objective with a 1.4 numerical aperture. Excitation lasers and emission filters were as
424 follows: mTurquoise2 excitation 458 nm, emission 463 - 535 nm; GFP and
425 mNeonGreen excitation 488 nm, emission 481 - 541 nm; mScarlet-i excitation 561 nm,
426 emission 561 - 633 nm. For chlorophyll excitation we used the same excitation as the
427 corresponding fluorophore with emission 642 - 712 nm.

428 *Flow cytometry*

429 GFP expression was analyzed by flow cytometry using the CytoFLEX LX355
430 (Beckman Coulter) analyzer. Forward scattered (FSC) and side scattered photons by
431 the 488 nm laser were used to determine diatoms from culture debris. FSC-height
432 versus FSC-area signal was used to isolate single events from sample aggregates.
433 Chlorophyll autofluorescence excited by 561 nm laser and emitted photons detected
434 with 675/25 filter was used to ensure all samples were fully intact. GFP fluorescence
435 excited by the 488 nm laser was detected by avalanche photo diode detector with
436 band pass filter 525/40. CytExpert software (Beckman Coulter) was used for analysis.

437 *SDS-PAGE and immunoblotting*

438 Cells were harvested by centrifugation (3,000 *g*, 10 min). Pellets were resuspended in
439 cell lysis buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 0.1 mM EDTA, 12.5% glycerol,
440 1x protease inhibitor cocktail [Pierce]) and sonicated for 1 min for cell disruption. The
441 protein extracts were separated by SDS-PAGE using 10% Protean TGX, 50 ul wells
442 gel (BioRad). Subsequently, gels were stained with coomassie blue or transferred to
443 PVDF membrane for immunoblotting. The GFP protein was probed with an anti-GFP
444 antibody at 1:1,000 dilution (G10362, Thermofisher) and visualized using a goat anti-
445 rabbit IgG (H+L) cross-adsorbed secondary antibody, Alexa Flour 488 (Invitrogen).
446 The fluorescence was imaged using Typhoon-5 (Amersham).

447 *Vector and strain distribution*

448 Upon publication all developed vectors will be made freely available via Addgene, and
449 all developed strains available from the authors.

450

451 **Acknowledgements:** The funding was supported through a United Kingdom
452 Research and Innovation Future Leaders Fellowship to L.C.M.M (MR/T020679/1),
453 Biotechnology and Biological Sciences Research Council Grants (BB/T017589/1,
454 BB/S015337/1, BB/R001014/1) and Engineering and Physical Research Council
455 Grant (EP/W024063/1). The authors would like to thank the University of York
456 Biosciences Technology Facility for confocal microscopy and flow cytometry
457 equipment access and support.
458

459 **Author Contributions:** L.C.M.M. and I.G. conceived the knock-in strategy. I.G., O.N.
460 and L.C.M.M. designed the experiments and wrote the manuscript. O.N. and I.G.
461 carried out the experiments.
462

463
464 **References**
465

466 1. P. J. Keeling, The endosymbiotic origin, diversification and fate of plastids. *Philos.*
467 *Trans. R. Soc. Lond. B Biol. Sci.* **365**, 729–748 (2010).

468 2. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production
469 of the biosphere: integrating terrestrial and oceanic components. *Science* **281**,
470 237–240 (1998).

471 3. P. G. Falkowski, J. A. Raven, *Aquatic Photosynthesis: Second Edition* (Princeton
472 University Press, 2013).

473 4. P. G. Falkowski, R. T. Barber, V. Smetacek V., Biogeochemical Controls and
474 Feedbacks on Ocean Primary Production. *Science* **281**, 200–207 (1998).

475 5. P. J. Tréguer, C. L. De La Rocha, The world ocean silica cycle. *Ann. Rev. Mar.*
476 *Sci.* **5**, 477–501 (2013).

477 6. B. J. Karas, *et al.*, Designer diatom episomes delivered by bacterial conjugation.
478 *Nat. Commun.* **6**, 6925 (2015).

479 7. A. Falciatore, M. Jaubert, J.-P. Bouly, B. Bailleul, T. Mock, Diatom Molecular
480 Research Comes of Age: Model Species for Studying Phytoplankton Biology and
481 Diversity. *Plant Cell* **32**, 547–572 (2020).

482 8. A. Hopes, V. Nekrasov, S. Kamoun, T. Mock, Editing of the urease gene by
483 CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods* **12**, 49
484 (2016).

485 9. S. S. Slattery, *et al.*, An Expanded Plasmid-Based Genetic Toolbox Enables Cas9
486 Genome Editing and Stable Maintenance of Synthetic Pathways in
487 *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **7**, 328–338 (2018).

488 10. B. Pollak, *et al.*, Universal loop assembly: open, efficient and cross-kingdom DNA
489 fabrication. *Synth. Biol.* **5**, ysaa001 (2020).

490 11. D. Ewe, *et al.*, The intracellular distribution of inorganic carbon fixing enzymes
491 does not support the presence of a C4 pathway in the diatom *Phaeodactylum*
492 *tricornutum*. *Photosynth. Res.* **137**, 263–280 (2018).

493 12. M. Samukawa, C. Shen, B. M. Hopkinson, Y. Matsuda, Localization of putative
494 carbonic anhydrases in the marine diatom, *Thalassiosira pseudonana*.
495 *Photosynth. Res.* **121**, 235–249 (2014).

496 13. H. Bukhari, T. Müller, Endogenous Fluorescence Tagging by CRISPR. *Trends*
497 *Cell Biol.* **29**, 912–928 (2019).

498 14. W.-K. Huh, *et al.*, Global analysis of protein localization in budding yeast. *Nature*
499 **425**, 686–691 (2003).

500 15. F. Daboussi, *et al.*, Genome engineering empowers the diatom *Phaeodactylum*
501 *tricornutum* for biotechnology. *Nat. Commun.* **5**, 3831 (2014).

502 16. N. Belshaw, *et al.*, Efficient CRISPR/Cas-mediated homologous recombination in
503 the model diatom *Thalassiosira pseudonana*. *bioRxiv*, 215582 (2017).

504 17. M. Angstenberger, J. Krischer, O. Aktaş, C. Büchel, Knock-Down of a ligIV
505 Homologue Enables DNA Integration via Homologous Recombination in the
506 Marine Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **8**, 57–69 (2019).

507 18. P. D. Weyman, *et al.*, Inactivation of *Phaeodactylum tricornutum* urease gene
508 using transcription activator-like effector nuclease-based targeted mutagenesis.
509 *Plant Biotechnol. J.* **13**, 460–470 (2015).

510 19. P. G. Kroth, *et al.*, Genome editing in diatoms: achievements and goals. *Plant Cell*
511 *Rep.* **37**, 1401–1408 (2018).

512 20. M. Nymark, A. K. Sharma, T. Sparstad, A. M. Bones, P. Winge, A CRISPR/Cas9
513 system adapted for gene editing in marine algae. *Sci. Rep.* **6**, 24951 (2016).

514 21. M. Serif, *et al.*, One-step generation of multiple gene knock-outs in the diatom
515 *Phaeodactylum tricornutum* by DNA-free genome editing. *Nat. Commun.* **9**, 3924
516 (2018).

517 22. N. J. Patron, *et al.*, Standards for plant synthetic biology: a common syntax for
518 exchange of DNA parts. *New Phytol.* **208**, 13–19 (2015).

519 23. P. Crozet, *et al.*, Birth of a Photosynthetic Chassis: A MoClo Toolkit Enabling
520 Synthetic Biology in the Microalga *Chlamydomonas reinhardtii*. *ACS Synth. Biol.*
521 **7**, 2074–2086 (2018).

522 24. M. Tachibana, *et al.*, Localization of putative carbonic anhydrases in two marine
523 diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Photosynth.*
524 *Res.* **109**, 205–221 (2011).

525 25. G. M. M. Hennon, *et al.*, Diatom acclimation to elevated CO2 via cAMP signalling
526 and coordinated gene expression. *Nat. Clim. Chang.* **5**, 761 (2015).

527 26. A. B. Kustka, *et al.*, Low CO2 results in a rearrangement of carbon metabolism to
528 support C4 photosynthetic carbon assimilation in *Thalassiosira pseudonana*. *New*
529 *Phytol.* **204**, 507–520 (2014).

530 27. A. Mukherjee, *et al.*, Thylakoid localized bestrophin-like proteins are essential for
531 the CO2 concentrating mechanism of *Chlamydomonas reinhardtii*. *Proc. Natl.*
532 *Acad. Sci. U. S. A.* **116**, 16915–16920 (2019).

533 28. J. G. Doench, *et al.*, Optimized sgRNA design to maximize activity and minimize
534 off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

535 29. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold.
536 *Nature* **596**, 583–589 (2021).

30. Huang Weichao, Daboussi Fayza, Genetic and metabolic engineering in diatoms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160411 (2017).

31. P. Bulankova, *et al.*, Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. *Curr. Biol.* (2021) https:/doi.org/10.1016/j.cub.2021.05.013.

32. S. Shashkova, M. C. Leake, Single-molecule fluorescence microscopy review: shedding new light on old problems. *Biosci. Rep.* **37** (2017).

33. S. Kikutani, *et al.*, Thylakoid luminal θ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom *Phaeodactylum tricornutum*. *Proceedings of the National Academy of Sciences* **113**, 9828–9833 (2016).

34. D. Miki, W. Zhang, W. Zeng, Z. Feng, J.-K. Zhu, CRISPR/Cas9-mediated gene targeting in Arabidopsis using sequential transformation. *Nat. Commun.* **9**, 1967 (2018).

35. D. Kwart, D. Paquet, S. Teo, M. Tessier-Lavigne, Precise and efficient scarless genome editing in stem cells using CORRECT. *Nat. Protoc.* **12**, 329–354 (2017).

36. M. Banan, Recent advances in CRISPR/Cas9-mediated knock-ins in mammalian cells. *J. Biotechnol.* **308**, 1–9 (2020).

37. J. A. Bosch, R. Colbeth, J. Zirin, N. Perrimon, Gene Knock-Ins in Drosophila Using Homology-Independent Insertion of Universal Donor Plasmids. *Genetics* **214**, 75–89 (2020).

38. X. Li, *et al.*, A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. *Nat. Genet.* (2019) https:/doi.org/10.1038/s41588-019-0370-6.

39. N. H. Cho, *et al.*, OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).

40. L. C. M. Mackinder, *et al.*, A Spatial Interactome Reveals the Protein Organization of the Algal CO_2-Concentrating Mechanism. *Cell* **171**, 133-147.e14 (2017).

41. E. A. Winzeler, *et al.*, Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).

42. R. R. Guillard, J. H. Ryther, Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (cleve) Gran. *Can. J. Microbiol.* **8**, 229–239 (1962).

43. R. R. L. Guillard, “Culture of Phytoplankton for Feeding Marine Invertebrates” in *Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of Marine Invertebrate Animals Greenport*, W. L. Smith, M. H. Chanley, Eds. (Springer US, 1975), pp. 29–60.

44. J. Peccoud, *et al.*, A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, e16765 (2011).

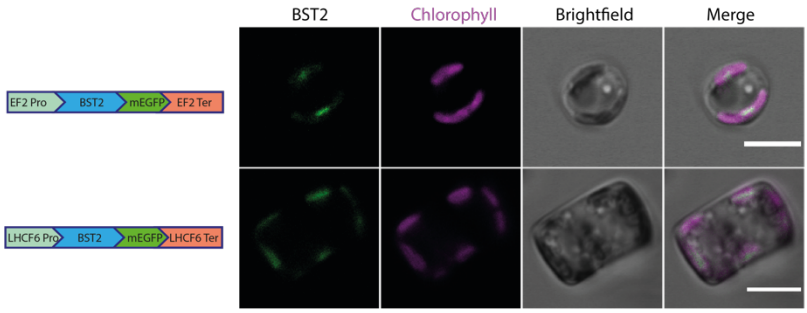
575 45. C. Engler, *et al.*, A golden gate modular cloning toolbox for plants. *ACS Synth.*
576 *Biol.* **3**, 839–843 (2014).

577 46. T. A. Strand, R. Lale, K. F. Degnes, M. Lando, S. Valla, A new and improved host-
578 independent plasmid system for RK2-based conjugal transfer. *PLoS One* **9**,
579 e90372 (2014).

580

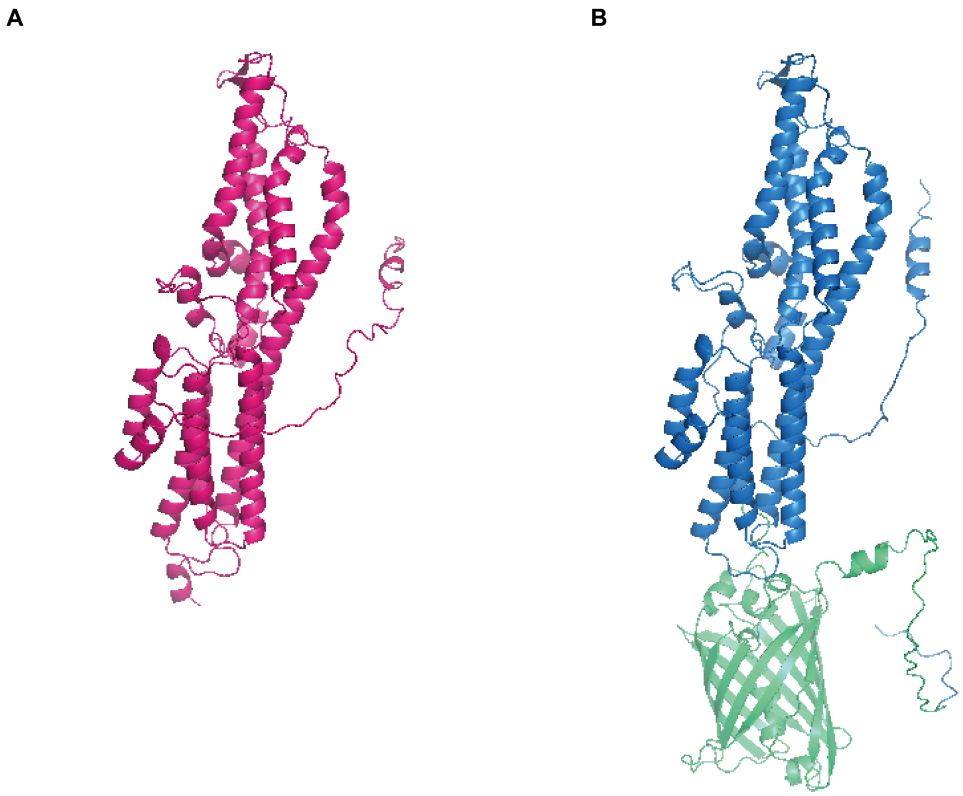
581

582 **Supplementary Figures**



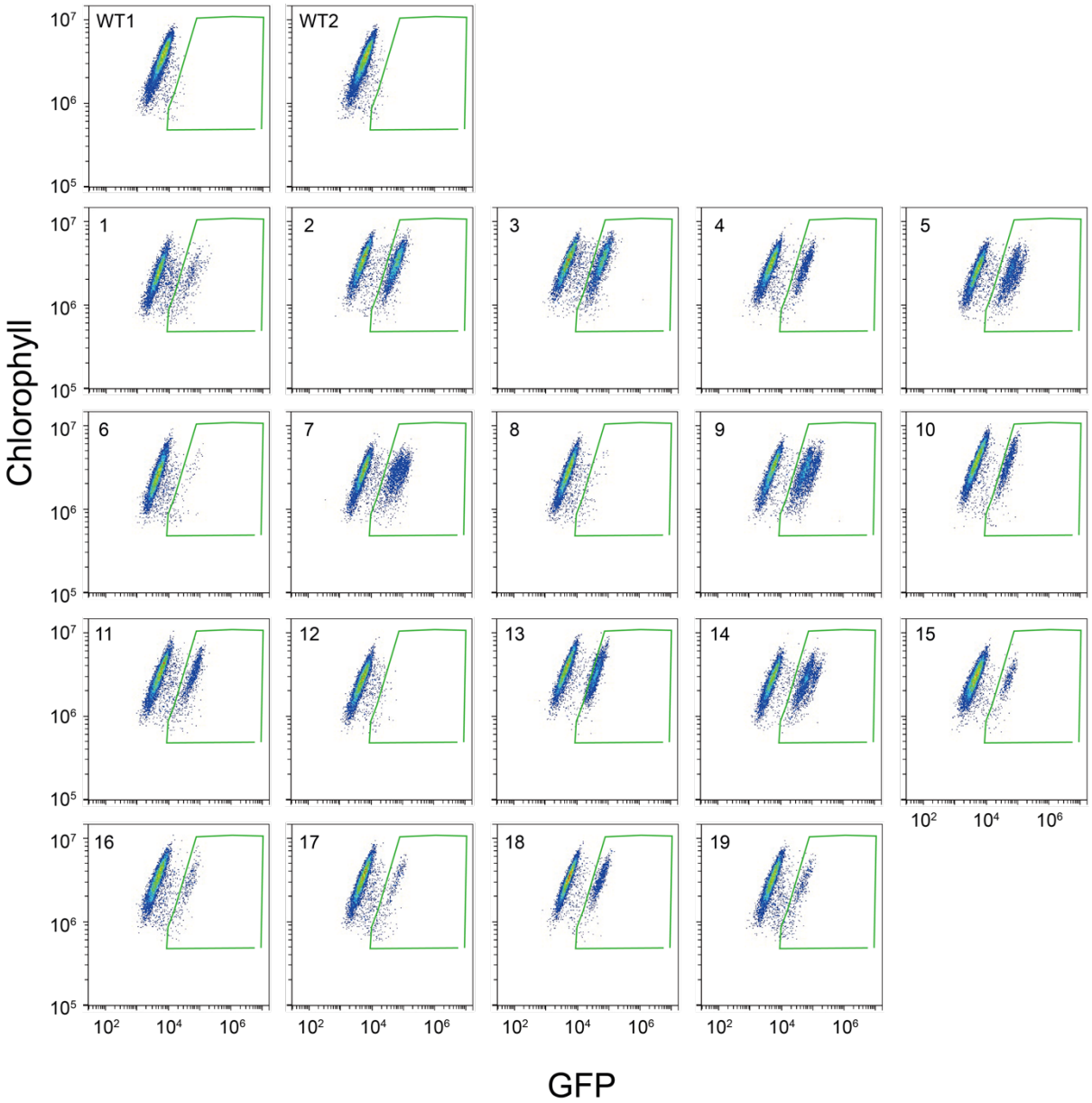
583

584 **Fig. S1. Validation of two additional promoter and terminator pairs.** The
585 elongation factor 2 (EF2) promoter and terminator and the light harvesting complex f
586 6 (LHCF6) promoter and terminator were validated for fusion protein expression in *T.*
587 *pseudonana.* Scale bar: 5 µm.



588
589
590 **Fig. S2. AlphaFold model of BST2 and BST2-GFP^{KI}.** *(A)* AlphaFold model of BST2.
591 *(B)* AlphaFold model of BST2-GFP^{KI} showing that GFP knock-in has minimal predicted
592 structural changes. Blue: BST2, Green: mEGFP-3xFLAG.

593



594
595
596 **Fig. S3. Flow Cytometry screening of BST2-GFP^{KI} lines.** Round 1 flow cytometry
597 counting of BST GFP knock-in transformants. Green boxes are gates used for
598 counting GFP positive cells.

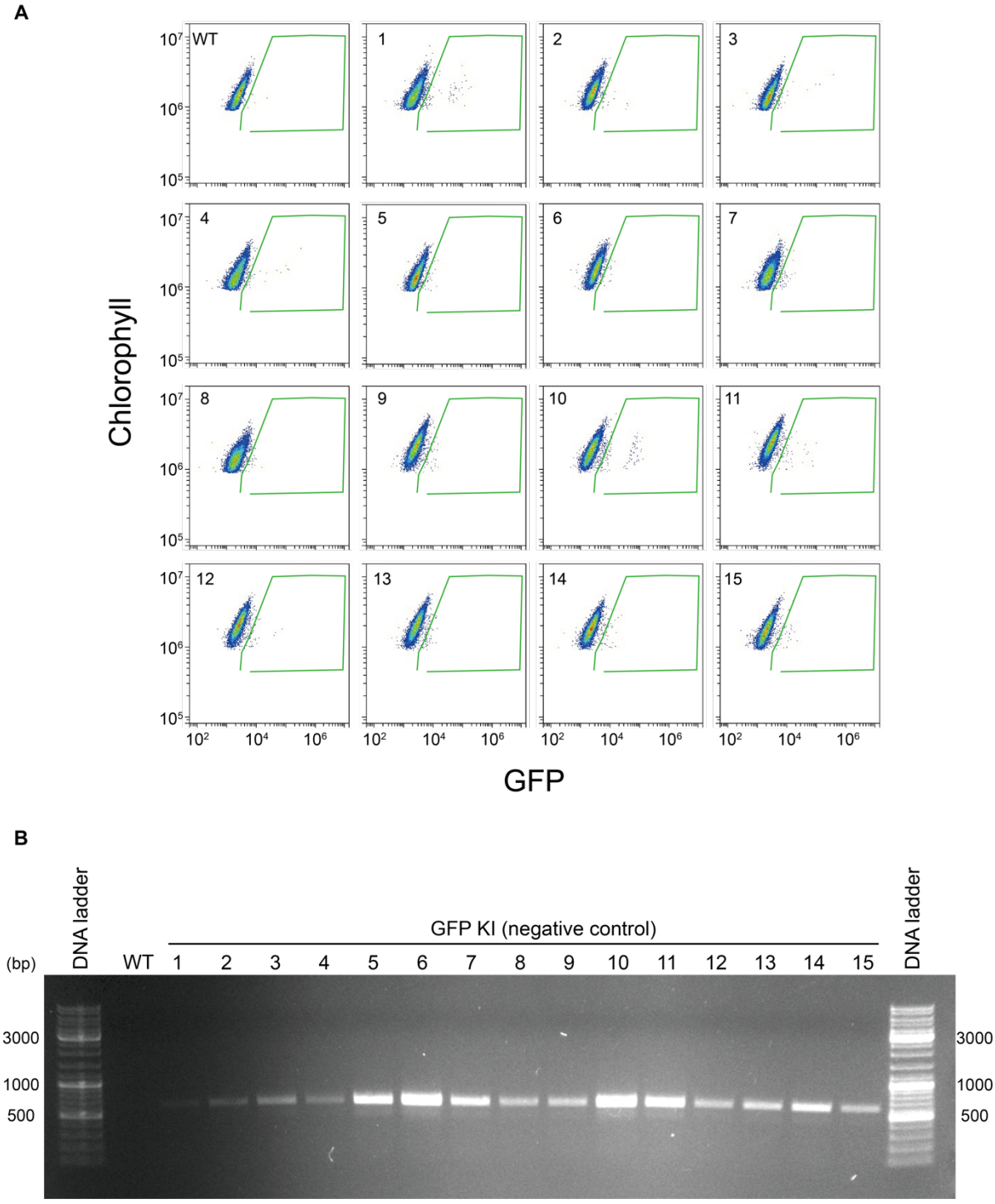


Fig. S4. The sgRNA is required for GFP knock-in and fluorescence. *(A)* Flow cytometry screening of 15 colonies picked after transformation with a GFP knock-in episome containing homology arms for BST2 GFP knock-in but that lacked the sgRNA. No GFP positive colonies were identified. *(B)* Colony PCR against the *NAT* gene in the episome backbone to confirm the presence of the episome in the flow cytometry screened colonies.