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 and provides a range of validated fluorophores for robust fluorescent protein tagging.

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[11]. 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[20,21], 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 Supplementary Table 1 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[23] we used the Plant/*Chlamydomonas* MoClo syntax[22]. 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*[24]; the origin of Transfer (*oriT*) required for vector transfer during conjugation[24] 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[25] 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[28] (Supplementary Fig. 1).

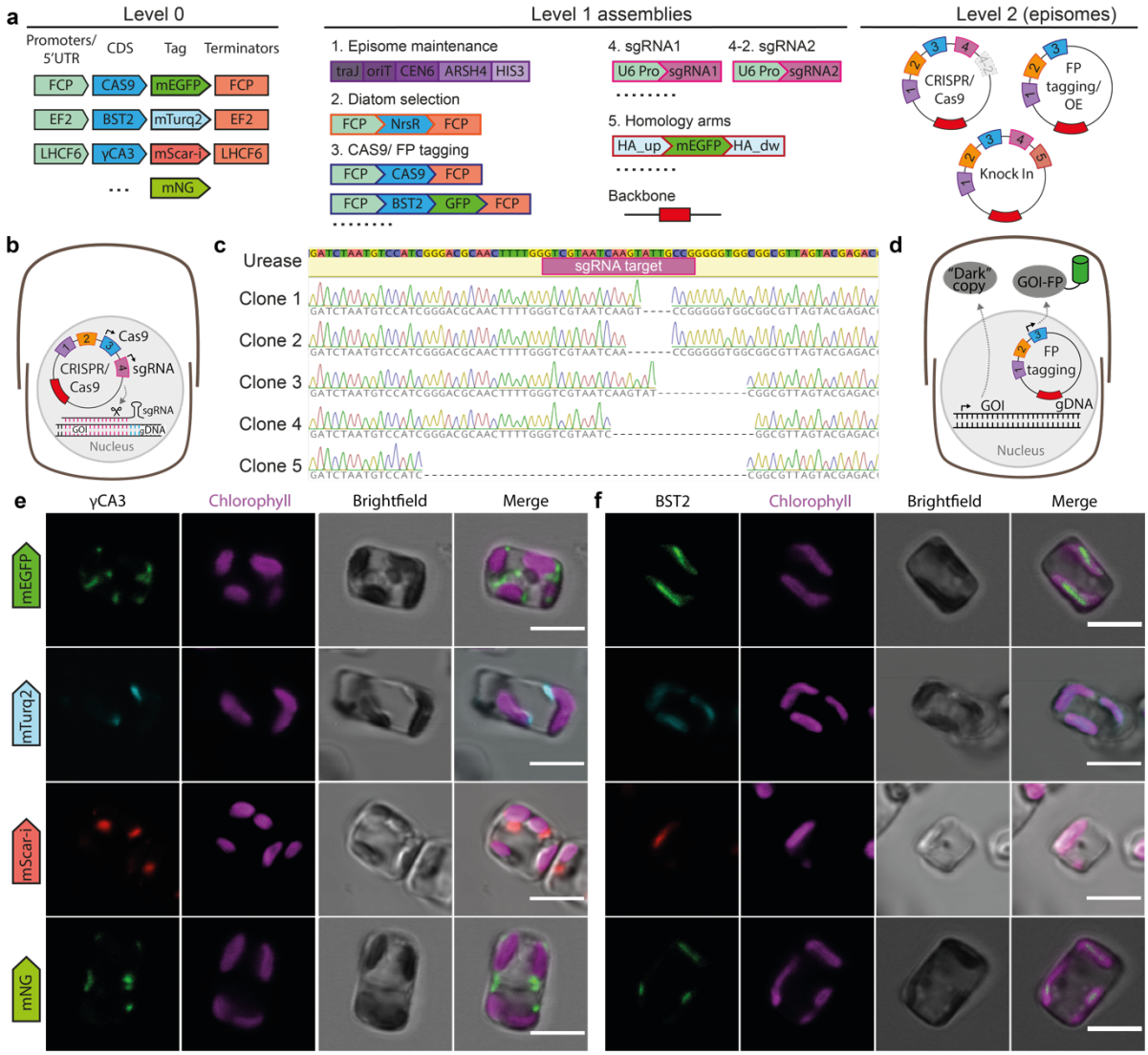


Fig. 1 Establishment of a Golden Gate based cloning framework for episome delivery of constructs in *T. pseudonana.* **a** Overview of Golden Gate assembly parts developed. These parts can be combined to assembly episomes for CRISPR/Cas9 gene editing, fluorescent protein (FP) tagging and FP knock-in. **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[27]. 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

143 tag BST2, which we previously localized via expression as a fusion protein directly
144 from episomes (Fig. 1f).

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

166 **Identifying and genotyping endogenously tagged GFP knock-in lines.**
167 Nourseothricin resistant colonies were transferred to 24-well plates for growth in liquid
168 media. Liquid grown transformants were screened via flow cytometry using a
169 CytoFLEX that enables rapid screening in a 96-well plate format. Initial data for our
170 BST2-GFP^{KI} construct indicated that 58% (11/19) of transformants were GFP positive
171 when a stringent 10% cut-off for GFP positive cells was used (Fig. 2b and
172 Supplementary Fig. 3). Lines that showed a high population of GFP cells were selected
173 for microscopy (Fig. 2c) with the localization of BST2-GFP^{KI} showing a linear
174 traversion at the center of the chloroplast most likely within the thylakoid membrane
175 traversing the pyrenoid. This data corresponded to when BST2-GFP was expressed
176 from an episome (Fig. 1f), indicating that in this case neither the native expression
177 levels nor insertion of GFP 13 amino acids upstream of the C-terminal affected
178 localization. We confirmed scarless insertion of *GFP* via amplification and sequencing
179 of the insertion flanking regions using oligos located outside the homology arms to
180 exclude the possibility of amplicons originating from the episome (Fig. 2d). To ensure
181 that Cas9 mediated DNA cutting was required for GFP insertion, and to check that
182 GFP could not be expressed from the episome prior to HDR, we assembled an
183 episome that lacked the sgRNA targeting *BST2*. Screening of 15 colonies via flow
184 cytometry showed no GFP positive colonies (Supplementary Fig. 4a). The presence
185 of the episome was confirmed by PCR against the *NAT* gene (Supplementary Fig. 4b).

This indicates that Cas9 DNA cleavage is required for efficient GFP knock-in and GFP is not expressed without successful integration into the genome.

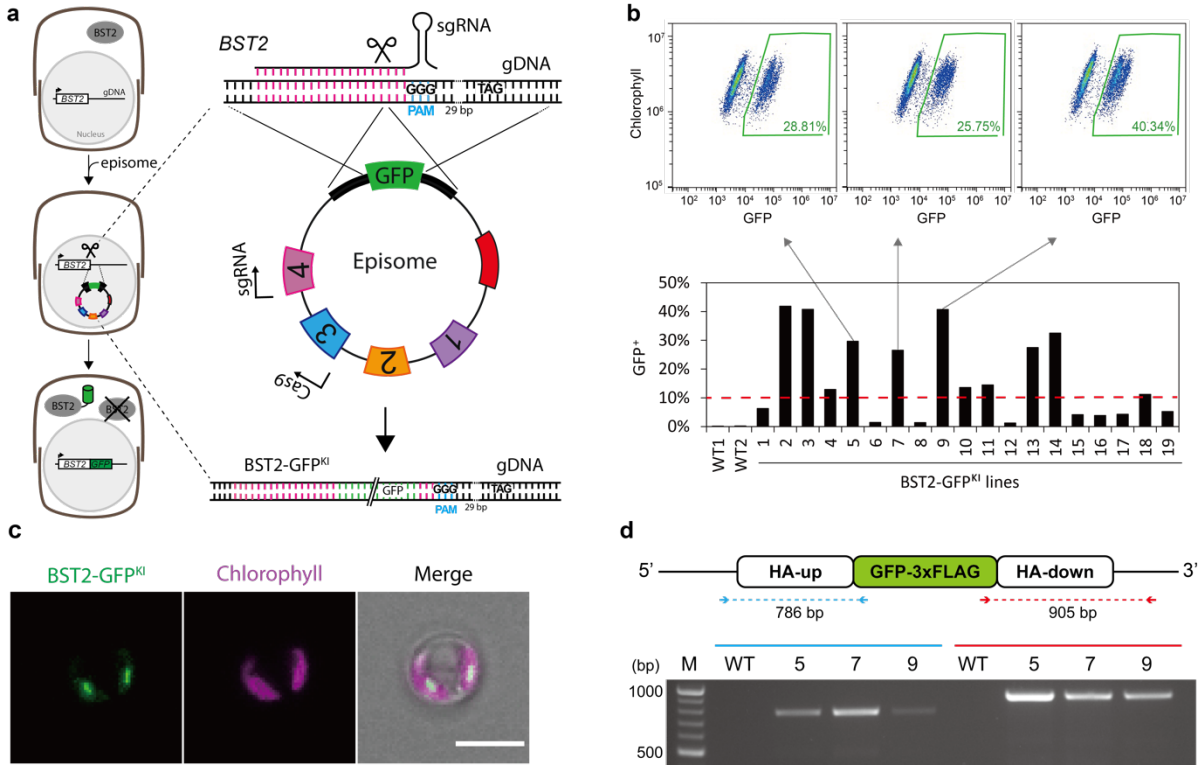


Fig. 2 Endogenous GFP tagging strategy overview and transformant screening. a Endogenous GFP tagging strategy. *T. pseudonana* is transformed with an episome containing (1) *CEN6-ARSH4-HIS3* and *oriT*, (2) *NAT* selectable marker, (3) Cas9, (4) sgRNA and the repair template containing GFP. Cas9 cutting 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.

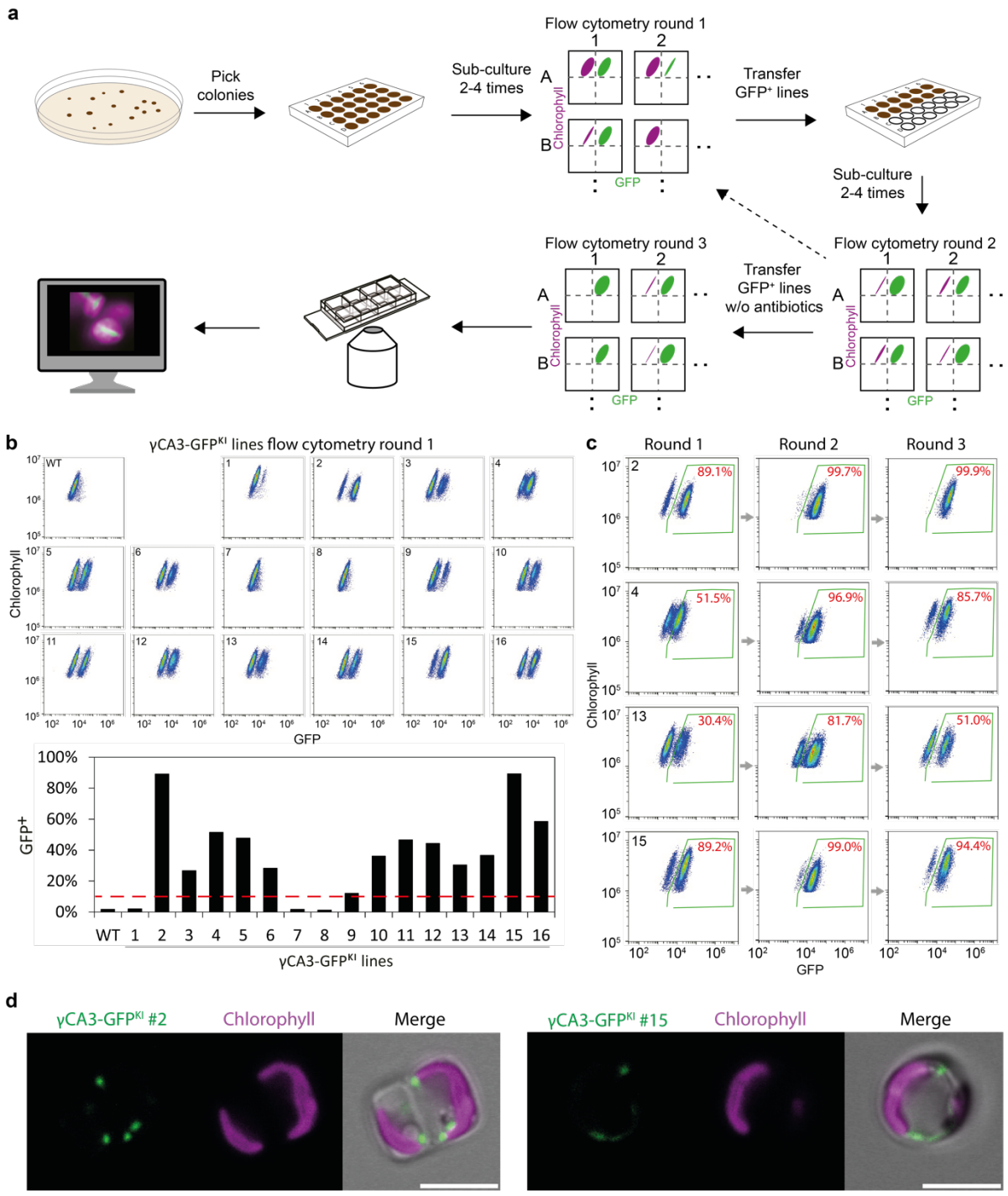


Figure 3. Development of a flow cytometry screening pipeline. a Developed pipeline to identify and select colonies with high GFP knock-in fluorescence. **b** Round 1 flow cytometry counting of γCA3 transformants. Bar chart corresponds to the above plots. Lines with >10% (dashed red line) of GFP positive 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 saw a mixed population of GFP positive and GFP negative cells suggesting the presence of mosaic colonies, where different genotypes are present within the same colony[30]. To isolate pure clonal lines, we developed a flow cytometry fluorescence cell counting pipeline and optimized this for the GFP knock-in of γCA3 (Fig. 3a and b). Our pipeline typically involves three rounds of flow cytometry. Transformants are picked and grown in the presence of antibiotics in 24-well plates with weekly subculturing. After 2-3 weeks (~6-9 generations) these are then screened via flow cytometry. Lines with the highest percentage of GFP positive cells are further propagated for ~3-4 generations and re-screened under antibiotic pressure. Once lines are close to 100% GFP positive they are transferred to antibiotic free media to drive episome loss. First round flow cytometry data gave us the percentage transformants with positive GFP signal. From this data we conclude that knock-in efficiency of at least one allele being successfully tagged in *T. pseudonana* for BST2 is 58% (Fig. 2b) and for γCA3 is 81% (Fig. 3b). Further rounds of flow cytometry screening indicated that multiple lines with close to 100% GFP positive cells could be isolated using this approach (Fig. 3c). Confocal microscopy of these lines showed a typical mitochondrial localization pattern (Fig. 3d).

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

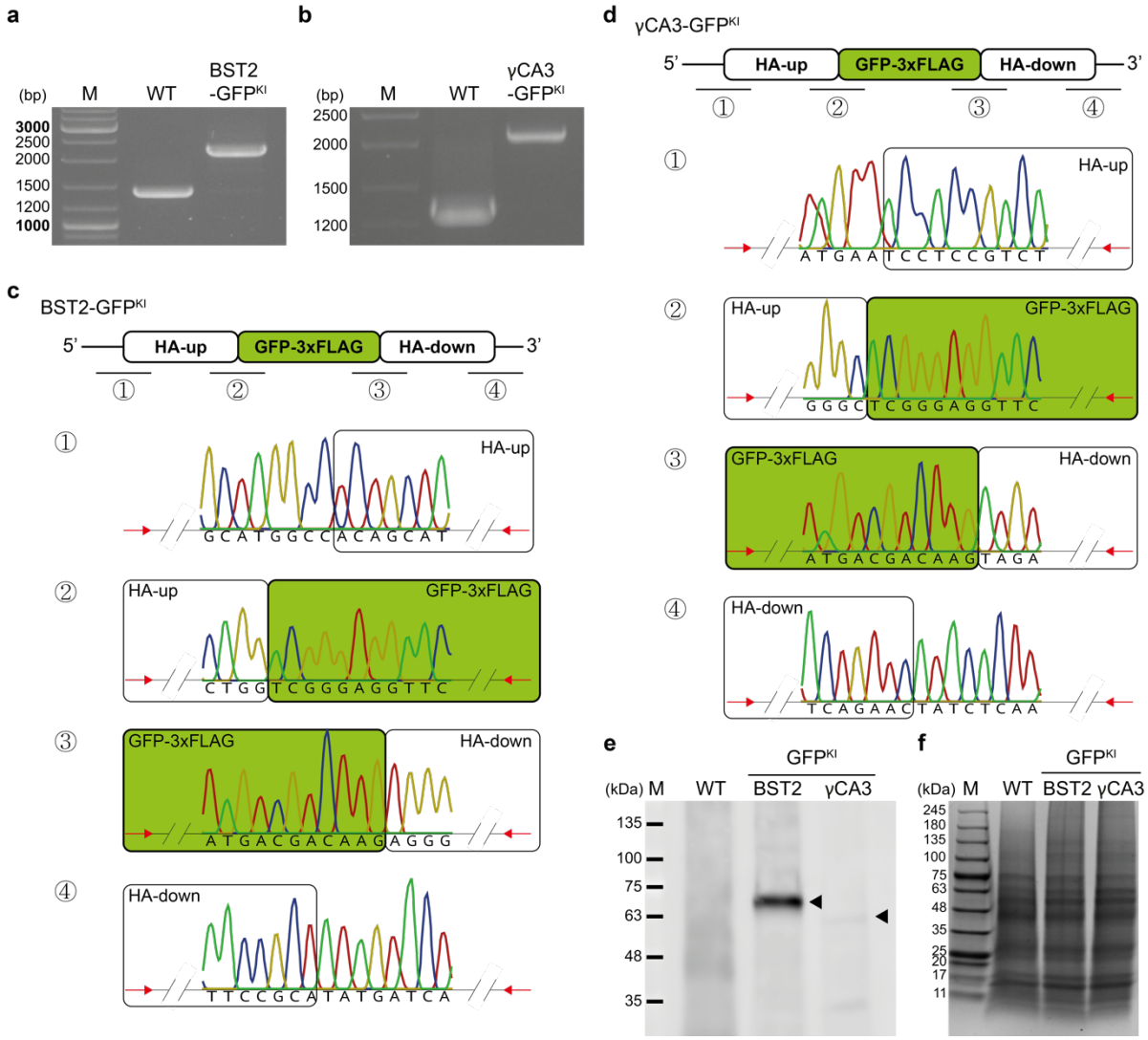


Fig. 4 Confirmation of biallelic GFP knock-ins. a and **b** PCR to confirm biallelic editing of clonal lines. **c** and **d** Sequencing of junctions confirming scarless integration of GFP into the genome at the target location. **e** Immunoblot against GFP in BST2-GFP^{KI} and γCA3-GFP^{KI}. Arrow heads 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).

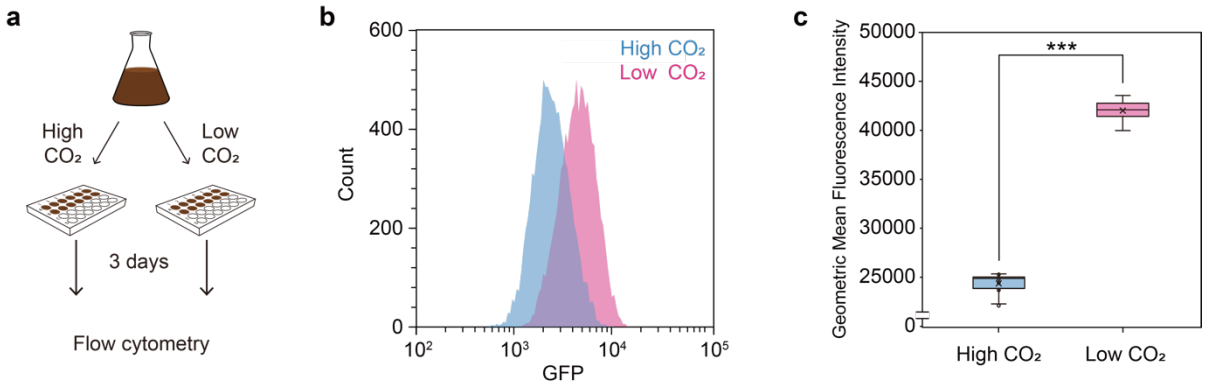


Fig. 5 Relative protein abundance can be tracked in GFP knock-in lines. a A biallelic BST2-GFP^{KI} line was grown in ambient CO_2, transferred to both high CO_2 and low CO_2 for three days then GFP intensity measured by flow cytometry. **b** and **c** Low CO_2 grown cells show a clear shift in fluorescence 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[32] 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[34]. 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[34,36,37]. 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

304 donor template in a parental line expressing Cas9[38]. In-line with our single plasmid
305 approach, the authors also tested the presence of all components on a single T-DNA
306 construct, however this failed to result in heritable GFP knock-in[38]. As *T. pseudonana*
307 is diploid, an additional challenge associated with endogenous tagging is the knock-in
308 of the GFP at both alleles. To overcome this, we maintained transformants under
309 antibiotic selection pressure and developed a rapid and scalable flow cytometry
310 screening pipeline. This resulted in the generation of homoplasmic knock-in lines.
311
312 During the establishment of our GFP knock-in approach we only tested homology
313 arms of ~600 bp and the knock-in of GFP (~800 bp). Further tests to identify the
314 minimum homology arm size requirement could lower homology arm synthesis costs
315 in the future. In addition, exploring the limitations of knock-in fragment size would be
316 useful to determine if the approach could be used for multi-gene/pathway knock-in at
317 neutral genomic sites for biotechnology applications. We also only tested guide block
318 knock-ins where the GFP is inserted into the Cas9 cut-site. It has been demonstrated
319 in other organisms that HDR can be offset from the cut-site which would then enable
320 GFP to be inserted directly prior to the stop codon and increase sgRNA flexibility.
321 However, knock-in efficiency can rapidly drop off as the cut-site to HDR-site distance
322 increases[39].
323
324 Our knock-in efficiency of 58-81% appears to be high compared to other
325 systems[34,36,37]. This reduces the effort required for screening and selection of
326 successful GFP knock-in lines. One potential explanation for this high efficiency is the
327 presence of all the necessary engineering elements on a single episome that can be
328 continually maintained through antibiotic selection. The high efficiency of our system
329 indicates that the same framework could be used for insertions, substitutions and
330 functional knock-outs at specific target sites. However, for biallelic editing additional
331 screening strategies, other than fluorescence-based flow cytometry, would need to be
332 established.
333
334 To enable the successful and easy application of our system we first developed and
335 validated a Golden Gate based episome assembly framework for multi-purpose
336 genetic engineering in *T. pseudonana* (Fig. 1). This included the episomal testing of
337 two additional promoter/terminator pairs, the validation of three additional fluorophores
338 (mTurquoise2, mScarlet-i and mNeonGreen) and the establishment of episomal based
339 CRISPR/Cas9 gene editing. The Golden Gate based episomal assembly framework
340 allows large, multi-part constructs to be rapidly and easily assembled. Episomal
341 delivery by bacterial conjugation enables the delivery of large constructs, avoids
342 random integration into the genome and requires no specialized equipment such as a
343 Gene gun. This engineering framework now opens the door for large-scale fluorescent
344 protein tagging and mutant library generation in diatoms. Similar studies in other
345 systems have proven to be extremely powerful for understanding gene function and
346 biological processes[14,38–41].

347 **Methods**

348 *Strains and growth conditions*

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

353 *Episome assemblies using Golden Gate cloning*

354 Level 0 (L0), Level 1 (L1) and Level 2 (L2) plasmids were assembled by Golden Gate
355 (GG) cloning[44] using a range of custom parts, uLoop[43] parts and Golden Gate MoClo
356 Plant Kit parts[45] (Supplementary Table 1). The assembly framework and syntax was
357 according to Patron et al. 2015[8]. Available *T. pseudonana* expression cassettes for
358 Cas9, the *NAT* selection marker, *U6* promoter and sgRNA backbone were used to
359 construct L1 vectors[8]. With both Cas9 and the *NAT* selection marker under the
360 expression of the *FCP* promoter and terminator. The yeast element *CEN6-ARSH4-*
361 *HIS3*[44] was domesticated for GG cloning by removing BsaI and BpiI restriction sites
362 and cloning it into a L1 plasmid (pICH47732) together with oriT and traJ. The resulting
363 L1 part permits low-copy episomal replication in diatoms without random integration
364 into the genome[46]. Depending on application (i.e. gene knock-out, FP-tagging or GFP
365 knock-in) 3-5 L1 parts can be assembled into an L2 plasmid (termed episome) for
366 transformation into *T. pseudonana*.

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

374 *Transformation of* T. pseudonana

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

383 *Gene targets and homology arm design*

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

386 A template for HDR was integrated into an episome as a further L1 component. It
387 consisted of i) a domesticated 600 bp sequence upstream of the stop codon of the
388 targeted gene (HA_up), ii) in-frame GFP without a stop codon and iii) a domesticated
389 600 bp sequence downstream of but including the stop codon (HA_down). sgRNAs
390 used in this experiment were selected so that the cut site they induced would be as
391 close to the stop codon as possible. For both genes a guide blocking approach was
392 used whereas after insertion of the GFP via HDR the sgRNA recognition site would be
393 split preventing further cutting. Domesticated 600 bp homology arms (upstream and
394 downstream of the selected Cas9 cut site) containing restriction sites and junctions for
395 GG cloning were either gene synthesized (Twist Bioscience) or amplified from
396 genomic DNA. GFP was fitted with a short flexible linker and a 3xFLAG tag.

397 *Genotyping/screening for HDR*

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

405 *Confocal microscopy*

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

412 *Flow cytometry*

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

421 *SDS-PAGE and immunoblotting*

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

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

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

References

1. Keeling, P. J. The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 729–748 (2010).
2. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
3. Falkowski, P. G. & Raven, J. A. *Aquatic Photosynthesis: Second Edition*. (Princeton University Press, 2013).
4. Falkowski, P. G., Barber, R. T. & Smetacek, V., V. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200–207 (1998).
5. Tréguer, P. J. & De La Rocha, C. L. The world ocean silica cycle. *Ann. Rev. Mar. Sci.* **5**, 477–501 (2013).
6. Karas, B. J. *et al.* Designer diatom episomes delivered by bacterial conjugation. *Nat. Commun.* **6**, 6925 (2015).
7. Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. & Mock, T. Diatom Molecular Research Comes of Age: Model Species for Studying Phytoplankton Biology and Diversity. *Plant Cell* **32**, 547–572 (2020).
8. Hopes, A., Nekrasov, V., Kamoun, S. & Mock, T. Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods* **12**, 49 (2016).
9. Slattery, S. S. *et al.* An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **7**, 328–338 (2018).

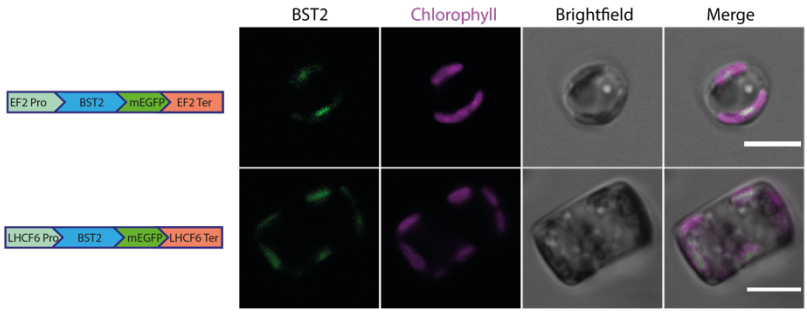
469 10. Pollak, B. *et al.* Universal loop assembly: open, efficient and cross-kingdom DNA
470 fabrication. *Synth. Biol.* **5**, ysaa001 (2020).
471 11. Ewe, D. *et al.* The intracellular distribution of inorganic carbon fixing enzymes
472 does not support the presence of a C4 pathway in the diatom *Phaeodactylum*
473 *tricornutum*. *Photosynth. Res.* **137**, 263–280 (2018).
474 12. Samukawa, M., Shen, C., Hopkinson, B. M. & Matsuda, Y. Localization of putative
475 carbonic anhydrases in the marine diatom, *Thalassiosira pseudonana*.
476 *Photosynth. Res.* **121**, 235–249 (2014).
477 13. Bukhari, H. & Müller, T. Endogenous Fluorescence Tagging by CRISPR. *Trends*
478 *Cell Biol.* **29**, 912–928 (2019).
479 14. Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature*
480 **425**, 686–691 (2003).
481 15. Daboussi, F. *et al.* Genome engineering empowers the diatom *Phaeodactylum*
482 *tricornutum* for biotechnology. *Nat. Commun.* **5**, 3831 (2014).
483 16. Belshaw, N. *et al.* Efficient CRISPR/Cas-mediated homologous recombination in
484 the model diatom *Thalassiosira pseudonana*. *bioRxiv* 215582 (2017)
485 doi:10.1101/215582.
486 17. Angstenberger, M., Krischer, J., Aktaş, O. & Büchel, C. Knock-Down of a ligIV
487 Homologue Enables DNA Integration via Homologous Recombination in the
488 Marine Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **8**, 57–69 (2019).
489 18. Weyman, P. D. *et al.* Inactivation of *Phaeodactylum tricornutum* urease gene
490 using transcription activator-like effector nuclease-based targeted mutagenesis.
491 *Plant Biotechnol. J.* **13**, 460–470 (2015).
492 19. Kroth, P. G. *et al.* Genome editing in diatoms: achievements and goals. *Plant Cell*
493 *Rep.* **37**, 1401–1408 (2018).

494 20. Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M. & Winge, P. A
495 CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* **6**,
496 24951 (2016).
497 21. Serif, M. *et al.* One-step generation of multiple gene knock-outs in the diatom
498 *Phaeodactylum tricornutum* by DNA-free genome editing. *Nat. Commun.* **9**, 3924
499 (2018).
500 22. Patron, N. J. *et al.* Standards for plant synthetic biology: a common syntax for
501 exchange of DNA parts. *New Phytol.* **208**, 13–19 (2015).
502 23. Crozet, P. *et al.* Birth of a Photosynthetic Chassis: A MoClo Toolkit Enabling
503 Synthetic Biology in the Microalga Chlamydomonas reinhardtii. *ACS Synth. Biol.*
504 **7**, 2074–2086 (2018).
505 24. Tachibana, M. *et al.* Localization of putative carbonic anhydrases in two marine
506 diatoms, Phaeodactylum tricornutum and Thalassiosira pseudonana. *Photosynth.*
507 *Res.* **109**, 205–221 (2011).
508 25. Hennon, G. M. M. *et al.* Diatom acclimation to elevated CO2 via cAMP signalling
509 and coordinated gene expression. *Nat. Clim. Chang.* **5**, 761 (2015).
510 26. Kustka, A. B. *et al.* Low CO2 results in a rearrangement of carbon metabolism to
511 support C4 photosynthetic carbon assimilation in *Thalassiosira pseudonana*. *New*
512 *Phytol.* **204**, 507–520 (2014).
513 27. Mukherjee, A. *et al.* Thylakoid localized bestrophin-like proteins are essential for
514 the CO2 concentrating mechanism of *Chlamydomonas reinhardtii*. *Proc. Natl.*
515 *Acad. Sci. U. S. A.* **116**, 16915–16920 (2019).
516 28. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize
517 off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

518 29. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold.
519 *Nature* **596**, 583–589 (2021).
520 30. Huang Weichao & Daboussi Fayza. Genetic and metabolic engineering in
521 diatoms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160411 (2017).
522 31. Bulankova, P. *et al.* Mitotic recombination between homologous chromosomes
523 drives genomic diversity in diatoms. *Curr. Biol.* (2021)
524 doi:10.1016/j.cub.2021.05.013.
525 32. Shashkova, S. & Leake, M. C. Single-molecule fluorescence microscopy review:
526 shedding new light on old problems. *Biosci. Rep.* **37**, (2017).
527 33. Kikutani, S. *et al.* Thylakoid luminal θ-carbonic anhydrase critical for growth and
528 photosynthesis in the marine diatom *Phaeodactylum tricornutum*. *Proceedings of*
529 *the National Academy of Sciences* **113**, 9828–9833 (2016).
530 34. Miki, D., Zhang, W., Zeng, W., Feng, Z. & Zhu, J.-K. CRISPR/Cas9-mediated
531 gene targeting in Arabidopsis using sequential transformation. *Nat. Commun.* **9**,
532 1967 (2018).
533 35. Kwart, D., Paquet, D., Teo, S. & Tessier-Lavigne, M. Precise and efficient scarless
534 genome editing in stem cells using CORRECT. *Nat. Protoc.* **12**, 329–354 (2017).
535 36. Banan, M. Recent advances in CRISPR/Cas9-mediated knock-ins in mammalian
536 cells. *J. Biotechnol.* **308**, 1–9 (2020).
537 37. Bosch, J. A., Colbeth, R., Zirin, J. & Perrimon, N. Gene Knock-Ins in Drosophila
538 Using Homology-Independent Insertion of Universal Donor Plasmids. *Genetics*
539 **214**, 75–89 (2020).
540 38. Li, X. *et al.* A genome-wide algal mutant library and functional screen identifies
541 genes required for eukaryotic photosynthesis. *Nat. Genet.* (2019)
542 doi:10.1038/s41588-019-0370-6.

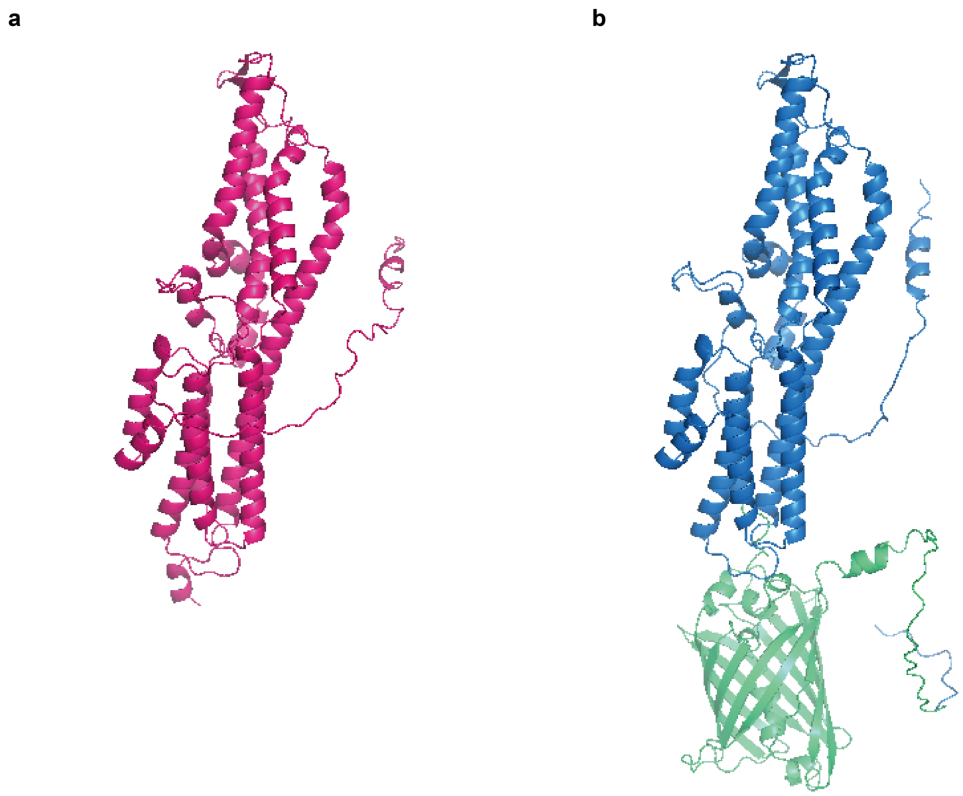
543 39. Cho, N. H. *et al.* OpenCell: Endogenous tagging for the cartography of human
544 cellular organization. *Science* **375**, eabi6983 (2022).
545 40. Mackinder, L. C. M. *et al.* A Spatial Interactome Reveals the Protein Organization
546 of the Algal CO_2-Concentrating Mechanism. *Cell* **171**, 133-147.e14 (2017).
547 41. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by
548 gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
549 42. Guillard, R. R. & Ryther, J. H. Studies of marine planktonic diatoms. I. *Cyclotella*
550 *nana* Hustedt, and *Detonula confervacea* (cleve) Gran. *Can. J. Microbiol.* **8**, 229–
551 239 (1962).
552 43. Guillard, R. R. L. Culture of Phytoplankton for Feeding Marine Invertebrates. in
553 *Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on*
554 *Culture of Marine Invertebrate Animals Greenport* (eds. Smith, W. L. & Chanley,
555 M. H.) 29–60 (Springer US, 1975).
556 44. Peccoud, J. *et al.* A modular cloning system for standardized assembly of
557 multigene constructs. *PLoS One* **6**, e16765 (2011).
558 45. Engler, C. *et al.* A golden gate modular cloning toolbox for plants. *ACS Synth.*
559 *Biol.* **3**, 839–843 (2014).
560 46. Strand, T. A., Lale, R., Degnes, K. F., Lando, M. & Valla, S. A new and improved
561 host-independent plasmid system for RK2-based conjugal transfer. *PLoS One* **9**,
562 e90372 (2014).
563
564

565 **Supplementary Figures**



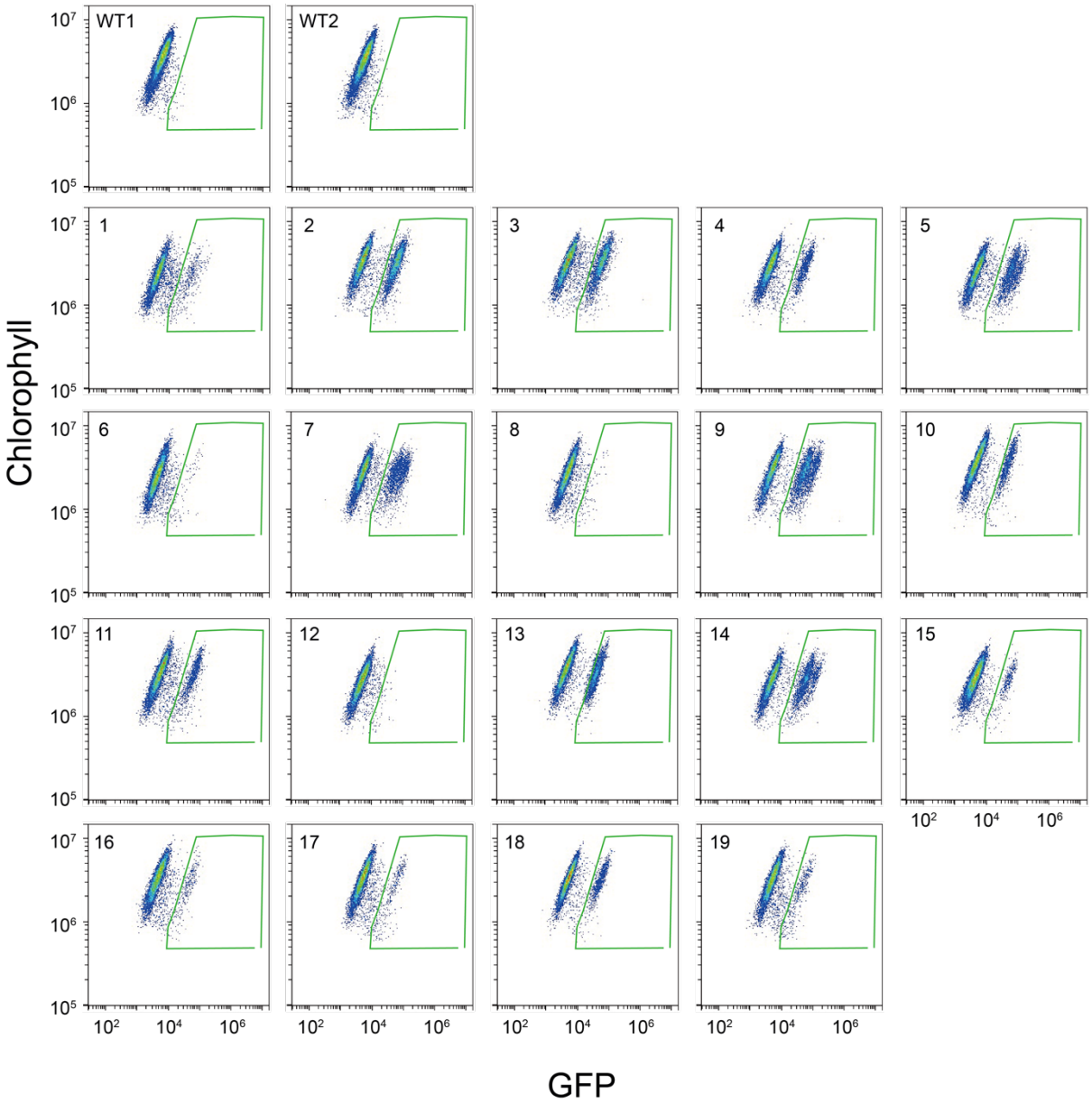
566

567 **Supplementary Fig. 1 Validation of two additional promoter and terminator pairs.**
568 The elongation factor 2 (EF2) promoter and terminator and the light harvesting
569 complex f 6 (LHCF6) promoter and terminator were validated for fusion protein
570 expression in *T. pseudonana.* Scale bar: 5 µm.

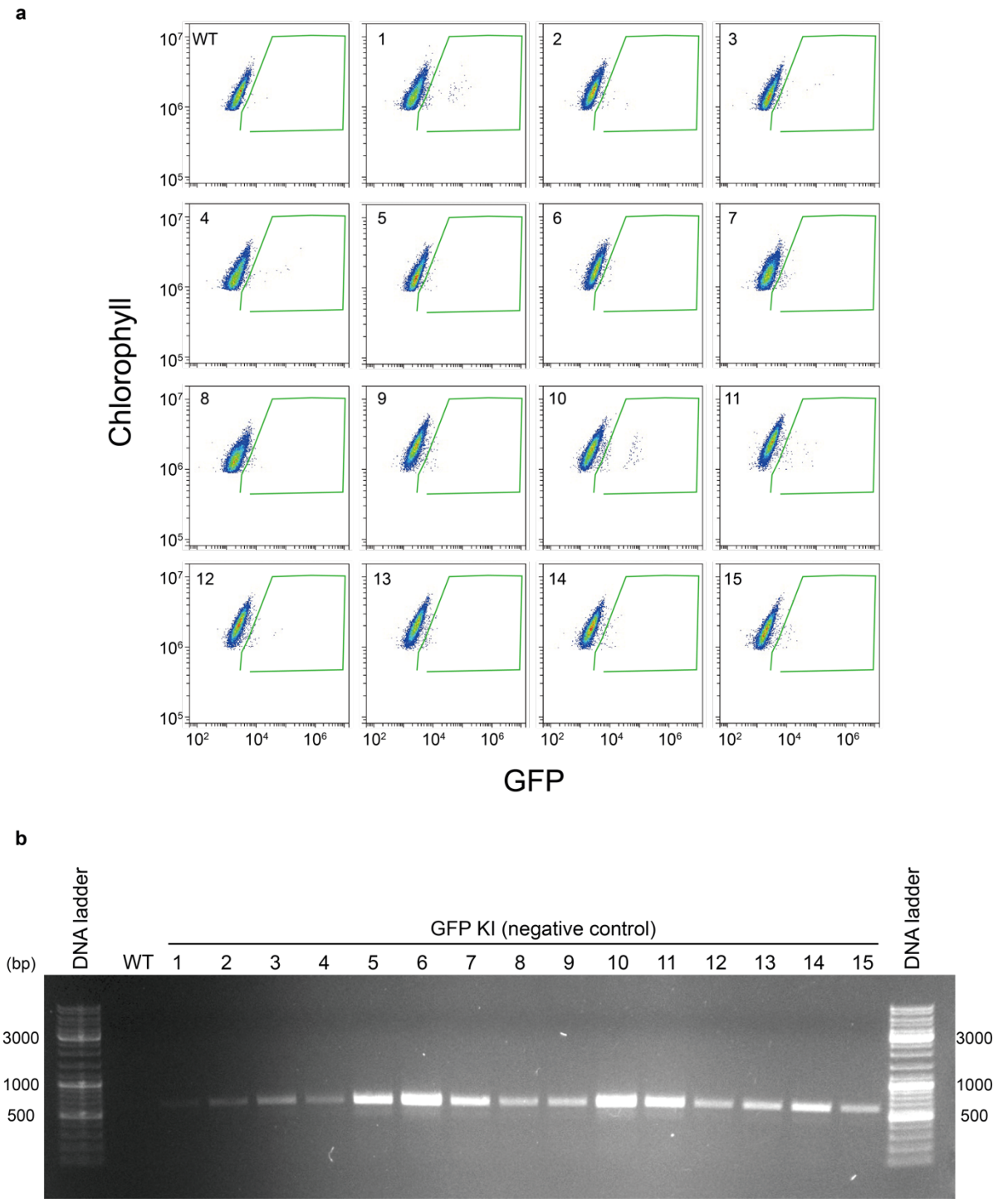


571
572
573 **Supplementary Fig. 2 AlphaFold model of BST2 and BST2-GFP^{KI}. a** AlphaFold
574 model of BST2. **b** AlphaFold model of BST2-GFP^{KI} showing that GFP knock-in has
575 minimal predicted structural changes. Blue: BST2, Green: mEGFP-3xFLAG.

576



577
578
579 **Supplementary Fig. 3 Flow Cytometry screening of BST2-GFP^{KI} lines.** Round 1
580 flow cytometry counting of BST GFP knock-in transformants. Green boxes are gates
581 used for counting GFP positive cells.



Supplementary Fig. 4 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.