Exploring genetic engineering strategies to enable heterologous monoterpenoid production in model microalgae, *Chlamydomonas reinhardtii* and *Phaeodactylum tricornutum*

Jestin George

PhD by Research

February 2021

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Climate Change Cluster

School of Life Sciences

University of Technology Sydney

Certificate of Original Authorship

I, Jestin George, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctorate of Philosophy, in the Faculty of Science at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Jestin George

19/02/2021

Date

Contents

Acknowledgements

I am a white, able bodied, queer woman who has benefited from the systemic racism that continues to run rampant across the globe, of which scientific academia is not immune and repeatedly still upholds. I would first and foremost like to acknowledge this privilege, having researched Molecular and Cellular Biology at universities in South Africa, United Kingdom and Australia: universities yet to be decolonised in countries where white privilege is deeply saturated in the culture. This thesis does not exist in a vacuum and its completion is intricately connected to that privilege and all it has afforded me in my life, including academically. This thesis is therefore dedicated to the activists whose work has directly impacted me; including but not limited to the South African Fees Must Fall Student Protestors (2015-2016), the George Floyd and Black Lives Matter Protestors (protesting at the time of this thesis write-up and submission), and activists Rachel Cargle, Erika Hart, Dr Senthorun Raj, Dr Priyamvada Gopal, Shay-Akil McLean and Ijeoma Oluo. This thesis is a reminder to me to do the challenging and uncomfortable work to drive change in the academic and social systems that I seamlessly move through every day.

*

To the Gadigal People and Aboriginal ancestors past, present and future. Thank you for hosting me and this research in the Eora Nation. Always way, always will be Aboriginal land.

To the Georges. Thank you for raising me, for allowing me to make mistakes, and for being infinitely proud of me.

To Tyler and Dylan. Thank you for supporting me financially and emotionally, and for being my escape; my brothers.

To Mom and Dad. Thank you for teaching me from a young age that my endeavours in education should not be for glory but for enriching my own life, for your selflessness and unconditional love, for who you are as people.

To Ebee. Thank you for being my second father, for being the foundation of the George Family.

To Mei. Thank you for bringing me to Sydney. Thank you for your patience, for loving me and repeatedly picking me up when I couldn’t, for coaching me to develop the Biodesign course, for believing in my dreams for me.

To the Maudlins. Thank you for housing me and welcoming me to Sydney, for loving me, laughing with me, for taking an interest in my PhD.

To Gabi, Tree, Rebecca, Jenna, Karena, Isabelle. Thank you for directly impacting the person I am today, for loving and supporting me across oceans, for believing in my abilities, for being excited about the research and work I am doing.

To The OG ABF. Thank you for improving many presentations and building my confidence, for sharing resources and ideas, for community, and for so much fun.

To Lorenzo. Thank you for being my work husband, day in and day out.

To Cristina. Thank you for advice and judgement-free support in and out of the lab. Thank you for being an inspirational woman in science and for sharing an abundance of resources.

To Audrey. Thank you for kindness, your empathy, your genuineness, and for much escapism. Thank you for being an inspirational woman in science.

To Raffie. Thank you for giving me an incredible amount of time, which always included your full effort and capability, you have always gone above and beyond. Thank you for your semi-private two-hour Python tutorship programme (with unlimited follow-up questions permitted). Thank you for being an inspirational woman in science.

To Michele. Thank you for your insight, guidance and leadership, for your support and encouragement. Thank you for the hours–if not days or weeks by the end–of editing, drafting manuscript titles, attending meetings and developing ideas with me. You introduced me to the world of synthetic biology, which helped trigger various other achievements throughout my PhD. Thank you for consistently contributing to help shape me to be a better scientist and certainly a better writer. Thank you for always being approachable and for always listening, for not centring yourself in a single moment of your mentorship.

To Peter. Thank you for the opportunity to do this research. Thank you for encouraging me to stay when I was ready to walk away, for being tolerant and understanding of times of difficulty and for helping me pursue my ambitions outside of the PhD.

To the assessors of this thesis. Thank you for your time and energy in evaluating my work and continued service to the field.

Thesis Abstract

This thesis focuses on next generation engineering strategies for *Chlamydomonas reinhardtii* and *Phaeodactylum tricornutum*, exploring aspects at the genomic and phenotypic level, to understand the biochemical implications and potential of heterologous monoterpenoid production in microalgae.

Chapter 1 outlines the ecological, and biotechnological relevance of microalgae, in the context of genetic engineering strategies for heterologous monoterpenoid production.

Chapter 2 investigates different strategies for delivering CRISPR-Cas9 ribonucleoprotein (RNP) into *C. reinhardtii* for targeted genome editing. This study highlighted major bottlenecks in CRISPR-Cas9 genome editing in this species, specifically low delivery efficiencies and unreliable endogenous markers.

Chapter 3 explores extrachromosomal expression (EE) and randomly integrated chromosomal expression (RICE) strategies to genetically engineer *P. tricornutum* to express *Catharanthus roseus* geraniol synthase (GES) for production of the monoterpenoid, geraniol. We identified superior RICE geraniol-yielding strains by developing a high-throughput phenotyping analysis and used long-read whole genome sequencing to interrogate the genomes of highly expressing cell lines. This revealed precise integration locations and unexpectedly large concatenated arrangements. We also demonstrated that exogenous DNA designed for EE does not inadvertently integrate into the nuclear genome.

Chapter 4 investigates CRISPR-Cas9 mediated targeted integration (TGI) for geraniol production in *P. tricornutum* in the genomic loci identified in Chapter 3. We showed that CRISPR-Cas9 RNP delivery is still inefficient in this species and that the recently

described endogenous marker gene uridine-5'-monophosphate synthase *(UMPS)* is unreliable in *P. tricornutum*, due to the highly mutagenic effect of 5-fluoroorotic acid, the selectable agent required to screen *UMPS* knock-out mutants.

Chapter 5 explores metabolic engineering approaches for increasing heterologous geraniol production in *P. tricornutum*. We fused two genes encoding adjacent enzymes in geraniol biosynthesis pathway, *GES* and *Abies grandii* geranyl pyrophosphate synthase, and showed that this approach decreased geraniol production, while constitutive expression of *GES* using a strong promoter resulted in a three times increased geraniol production. We used these strains to demonstrate that heterologous geraniol production in *P. tricornutum* did not perturb the native biosynthesis of major sterols and pigments.

Chapter 6 discusses why these findings are important for (1) providing insight as to why CRISPR-Cas9-based editing is still difficult to achieve in microalgae (2) improving *P. tricornutum's* status for heterologous terpenoid production with regard to its metabolic flexibility and capacity for high geraniol accumulation (3) characterising both well-established and new genetic engineering tools, including uncovering putative safe harbour loci for TGI required for developing more complex synthetic biology approaches in *P. tricornutum*.

General introduction

Model microalgal species are primed for exploring heterologous monoterpenoid production

GENERAL INTRODUCTION

Microalgae

Microalgae comprise of about 200,000 phylogenetically diverse species of photosynthetic unicellular eukaryotic microorganisms (Sharif et al., 2017) that span approximately eight phyla and occur in almost every aquatic environment on the planet (Figure 1) (Katz et al., 2004). They produce about 50% of the world's oxygen and play a crucial role in fixing inorganic carbon, consequently contributing to about 45% of Earth's net primary production (Field et al., 1998). They are a natural source of many useful chemicals for industry, including the pigments β-carotene, astaxanthin and phycocyanin used as colourants and nutraceuticals; and fatty acids, such as docosahexaenoic acid and eicosapentaenoic acid for food supplementation (Borowitzka, 2013).

Alongside their contemporary ecological and biotechnological importance (Armbrust, 2009), the earliest microalgal ancestors, originating from a primordial cyanobacterium, where able to perform oxygenic photosynthesis (Falkowski et al., 1998; Field et al., 1998). These cyanobacteria were responsible for oxygen accumulation so significant that a geochemical shift in Earth's atmosphere occurred (Bekker et al., 2004; Falkowski et al., 1998; Field et al., 1998). This Great Oxidation Event facilitated the evolution of new biochemical metabolic pathways and cellular compartmentalisation and later resulted in eukaryogenesis and the rise of complex aerobic organisms (Sharif et al., 2017). This is exemplified in the rise of complex eukaryotic sterol biosynthesis. Sterols are biochemicals that are particularly well known for their role in maintaining eukaryotic membrane fluidity and are produced in oxygen-dependent metabolic

pathways (Dufourc, 2008). Sterol biosynthesis genes found in both eukaryotes and bacteria are shown to have diverged around the time of

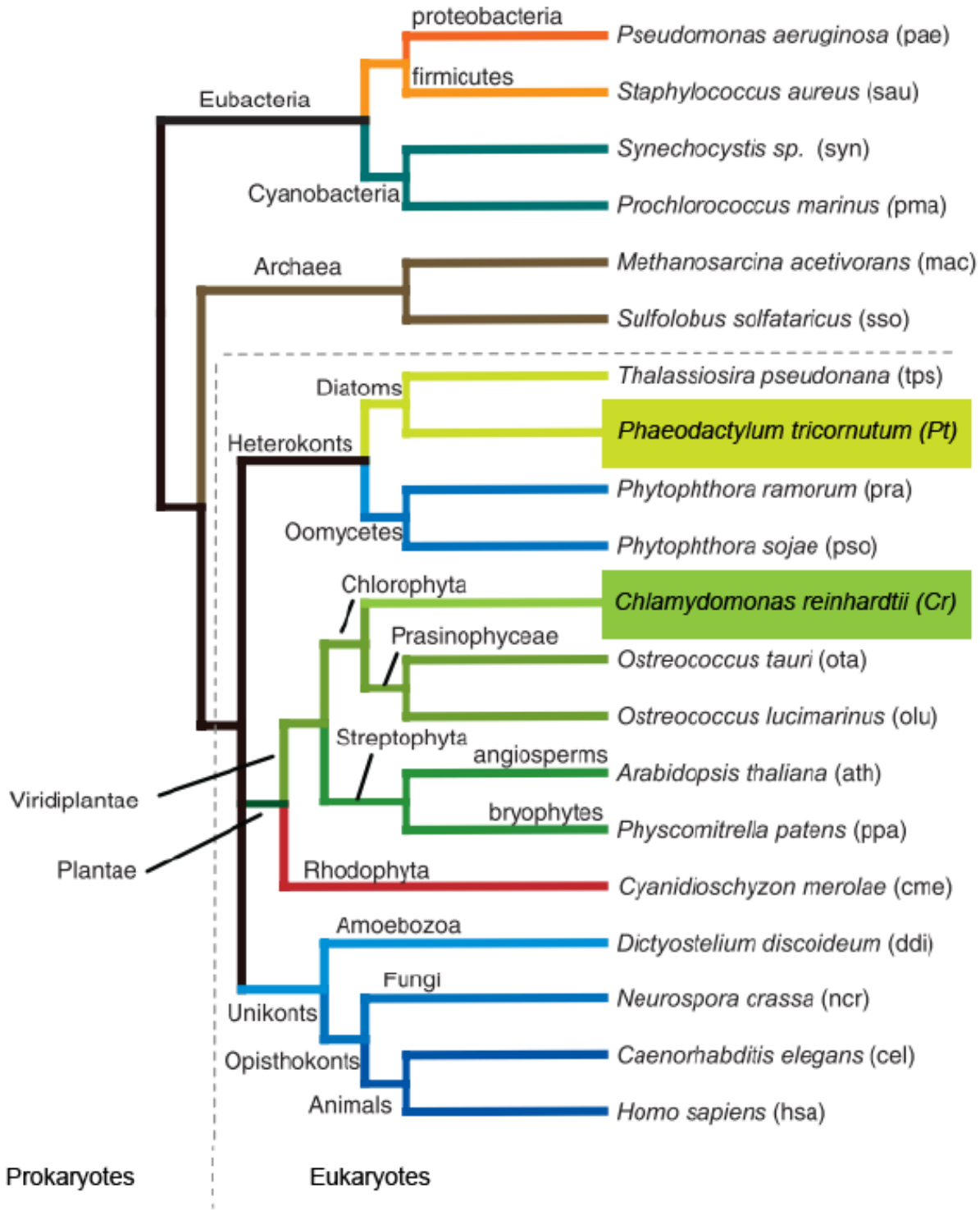


Figure 1. Evolutionary relationships of 20 species including cyanobacteria and non-photosynthetic eubacteria, archaea and eukaryotes from the oomycetes, diatoms, rhodophytes, plants, amoebae and opisthokonts. Endosymbiosis of a cyanobacterium by a eukaryotic protist gave rise to the green (green branches) and red (red branches)

plant lineages, respectively. The presence of motile or non-motile flagella is indicated at the right of the cladogram. Adapted from Merchant et al. (2007).

the Great Oxidation Event, suggesting that simpler sterol metabolic pathways would have originated well before this period, when oxygen was becoming more widespread across Earth (Gold et al., 2017). Over billions of years, dead carbon-sequestering microalgae also became trapped in the lithosphere, resulting in the origin of petroleum, the most widely used industrial feedstock today.

Given their ecological significance and polyphyletic origins, key microalgal species, namely *Chlamydomonas reinhardtii* and *Phaeodactylum tricornutum,* have been repeatedly studied to uncover important aspects of functional biology, as well as their potential role in biotechnology, ranging from food and feed to nutraceuticals, cosmetics and biofuels.

Chlorophyta and the model green alga *Chlamydomonas reinhardtii*

The first photosynthetic eukaryote arose from an endosymbiotic event in which an ocean-dwelling heterotrophic cell engulfed a prokaryotic, photoautotrophic cyanobacterium resulting in the evolution of chlorophyta, a diverse group of green algae, red algae (rhodophyta), land plants (streptophyta) (Figure 2). Most chlorophyta today are fresh water dwelling, occurring in small bodies of water and soils (Chapman & Chapman, 1973), and include species of both scientific and biotechnological value, namely *Chlamydomonas reinhardtii*.

C. reinhardtii is a particularly useful model Chlorophyta. This haploid, unicellular, mixotrophic microorganism still shares photosynthetic genes with the last common plant ancestor and can live in the dark autotrophically when supplied with an external carbon source. These features make it a very useful model for generating mutants to

study eukaryotic photosynthesis including biosynthesis and assembly of photosystems (Cahoon & Timko, 2000) and light perception (Jianming & Timko, 1996; White et al., 1996). *C. reinhardtii* also has a protruding motile appendage, a flagellum. The *C. reinhardtii* flagellum-associated genes can be traced back to the last common plant-animal ancestor, also known as the last eukaryotic common ancestor (LECA). While these genes were lost in land plants, they remained in animals, resulting in ciliated cells involved in reproduction (such as sperm), and respiration and digestion (such as epithelial cells). Numerous human diseases are associated with cilia dysfunction and consequently, studying cilia biology is imperative for developing medical interventions. Important studies in *C. reinhardtii* have led to better understandings of these disease models because of this genetic overlap and because this microalga can survive with or without its flagella (Pazour, Agrin, Walker, & Witman, 2006).

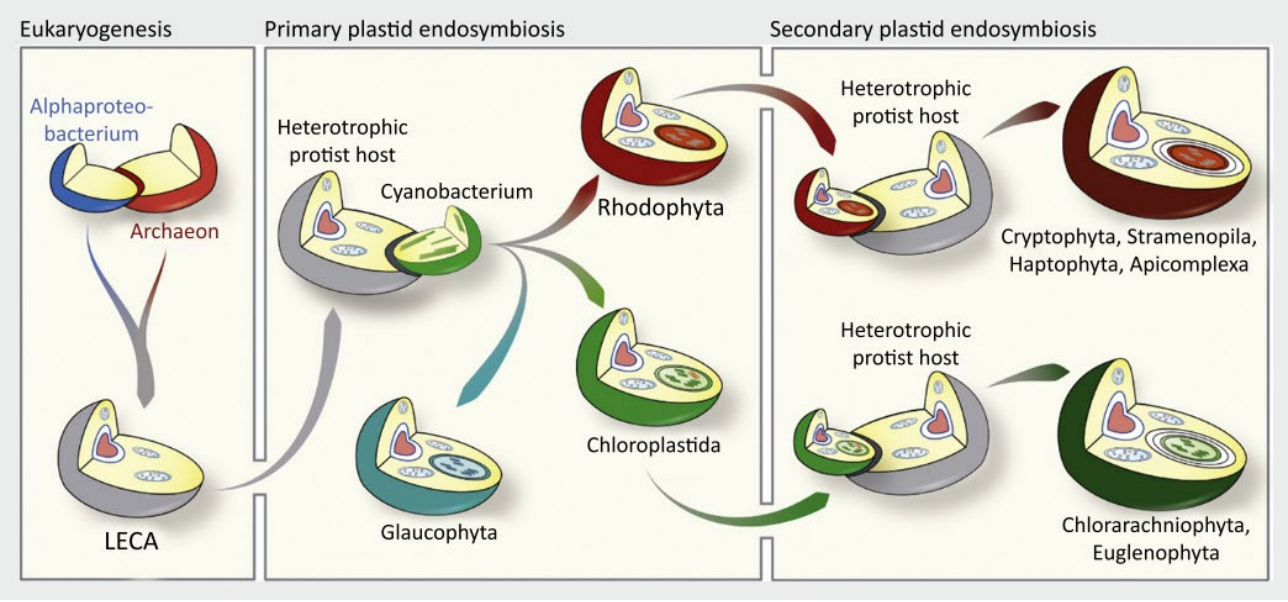


Figure 2. Eukaryogenesis and the primary and secondary endosymbiotic origin of plastids. Both the mitochondrion and plastids are of endosymbiotic origin, but the circumstances under which they evolved and their effects on evolution differ significantly. Eukaryogenesis describes the origin of eukaryotes and is the results of one prokaryote (an ancestor of extant alphaproteobacteria) coming to reside in another (an archaeal cell). The transition of the alphaproteobacterium to the

mitochondrion inside the host was key to the origin of eukaryotes themselves. With the endosymbiotic origin of the mitochondrion also the endomembrane system evolved in the last eukaryotic common ancestor (LECA), providing the blueprint for all eukaryotic cells. From Garg & Gould (2016).

C. reinhardtii also contains an array of genes involved in a diverse range of fermentation pathways, such as those involved with catabolism of pyruvate for generation of energy (Grossman et al., 2007; Merchant et al., 2007). This flexible metabolism is hypothesised to be a key reason for *C. reinhardtii's* ability to acclimate to niche high stress environments, such as anoxia (Dubini et al., 2009).

C. reinhardtii was the first alga to be reliably genetically manipulated (Boynton et al., 1998; Kindle et al., 1989) and consequently, most of the tools for microalgal genetic engineering have been developed for this particular species. Such tools include but are not limited to: various regulatory sequences for driving transgene expression as summarised by Jinkerson and Jonikas (2015), compatibility analyses for expression of various fluorescence proteins (Rasala et al., 2013), RNA interference strategies for gene silencing (Rohr et al., 2004), modular vectors (Lauersen et al., 2015) and protocols for engineering the nuclear, mitochondrial and plastidial genomes (Bateman & Purton, 2000; Kindle et al., 1989; Remacle et al, 2006). This has enabled a range of functional genetics studies over the years (Liu et al., 2013; Matsuura et al., 2004; Zorin et al., 2009) and caused *C. reinhardtii* to become a widely used model for genetic phycological research and for exploring its biotechnological potential (Radakovits et al., 2010; Rosales-Mendoza et al., 2012).

Diatoms and the model pennate species *Phaeodactylum tricornutum*

Diatoms are heterokonts that can be found in fresh-water sources and soils, and dominate oceanic phytoplankton populations and contribute to half of the total oxygen

produced by microalgae on Earth (Nelson et al., 1995). Because they are encased in silica frustules, they readily sink to the ocean floor after they die. Consequently, they are responsible for large silica deposits and carbon sequestration in the lithosphere (Amin et al, 2012; Sims et al., 2006). Diatoms evolved about 100-200 million years ago (Armbrust, 2009; Benoiston et al., 2017), following two endosymbiotic events in which a chloroplast containing eukaryote –generated during primary endosymbiosis– was engulfed by a eukaryotic heterotroph (Figure 2). This unique evolution resulted in extensive genetic and metabolic redundancy that, over time, saw gene transfers and gene losses that have left today's diatoms with a uniquely flexible metabolism. This has been repeatedly demonstrated in the model stramenopile pennate diatom, *Phaeodactylum tricornutum*, a diploid, unicellular phytoplankton (Allen et al., 2011; Fabris et al., 2014, 2012; Kroth et al., 2008; Smith et al., 2019). It is unsurprising that diatoms are able to adapt to various high stress conditions especially regarding nutrient availability, temperature and light availability and salinity which have likely contributed to its ecological advantage over other marine alga (Benoiston et al., 2017; Smith et al., 2019, 2016), allowing diatoms to often dominate phytoplankton consortia (Yool & Tyrrell, 2003).

Like *C. reinhardtii, P. tricornutum* boasts an advanced genetic engineering toolkit compared to other microalgae, despite being only recently developed. This includes nuclear and plastidial transformation protocols (Apt et al., 1996; Materna et al., 2009); various regulatory sequences, and reporter and selection genes as summarised by Huang and Daboussi (2017) and RNA interference strategies for gene silencing (De Riso et al., 2009). Consequently, *P. tricornutum* is being increasingly used as a model microalga to study and better elucidate its patchwork genetic evolution and ecological significance, as well as its potential for biotechnology.

Genetic engineering in microalgae

Genetic engineering is essential for both basic and applied research and describes the ability to add recombinant DNA, remove or silence endogenous genes and sequences, or randomly mutate or modify the genomic DNA of an organism. It can be deployed using many different molecular tools –including but not limited to recombinant DNA transformation (Apt et al., 1996; Kindle et al., 1989), gene silencing (Rohr et al., 2004), and genome editing (Daboussi et al., 2014; Greiner et al., 2017). It is an exceptionally powerful approach useful for both functional biology studies and biotechnological applications. The range of increasing technologies and approaches for genetic engineering can be viewed as two waves: first generation genetic engineering, which depends on random integration of transgenes into the nuclear genome, and next generation genetic engineering, which depends on non-random engineering.

First generation (random) genetic engineering in microalgae

First generation genetic engineering in microalgae involves a genetic modification that cannot be confined to a unique locus and consequently can occur anywhere in the genome following randomly integrated chromosomal expression (RICE). RICE has been exceptionally useful for obtaining a myriad of microalga knock-in and insertional knock out mutants over the past few decades. It depends on delivering recombinant DNA into the cell, which is integrated into the nuclear chromosomal genome at random location(s) for a heritable trait that can be expressed by the cell's native transcriptional machinery.

To date, it is generally accepted that random integration operates via aspects of the cell's natural DNA repair pathways, broadly categorised as non-homology end joining

(NHEJ). Although it occurs during all stages of cell cycle and is evolutionarily conserved among bacteria, archaea and eukaryotes (Demogines et al., 2010), the mechanistic details of how NHEJ occurs have been contradictory or elusive (Li et al., 2005; Park et al., 2015). These repair pathways are ubiquitous in biology and serve to preserve genome integrity and ensure the cell's survival from toxic double stranded breaks (DSB), which can occur throughout the genome daily (Deriano & Roth, 2013; Krejci et al., 2012). NHEJ-driven integration of transgenes has been described in algae (Cerutti et al., 1997; Cerutti et al., 2015) and non-algal species including plants (Li et al., 2005; Liu, 2018; Saika et al., 2014) and zebrafish (Dai et al., 2010), following various transformation techniques.

However, there is virtually no understanding of how NHEJ mechanisms drive RICE and consequently, it is not known how RICE impacts the chromosomal genome, nor exactly how exogenous DNA is configured once integrated into the genome (Figure 3). For example, recombinant DNA entering the cell is subject to a diverse range of

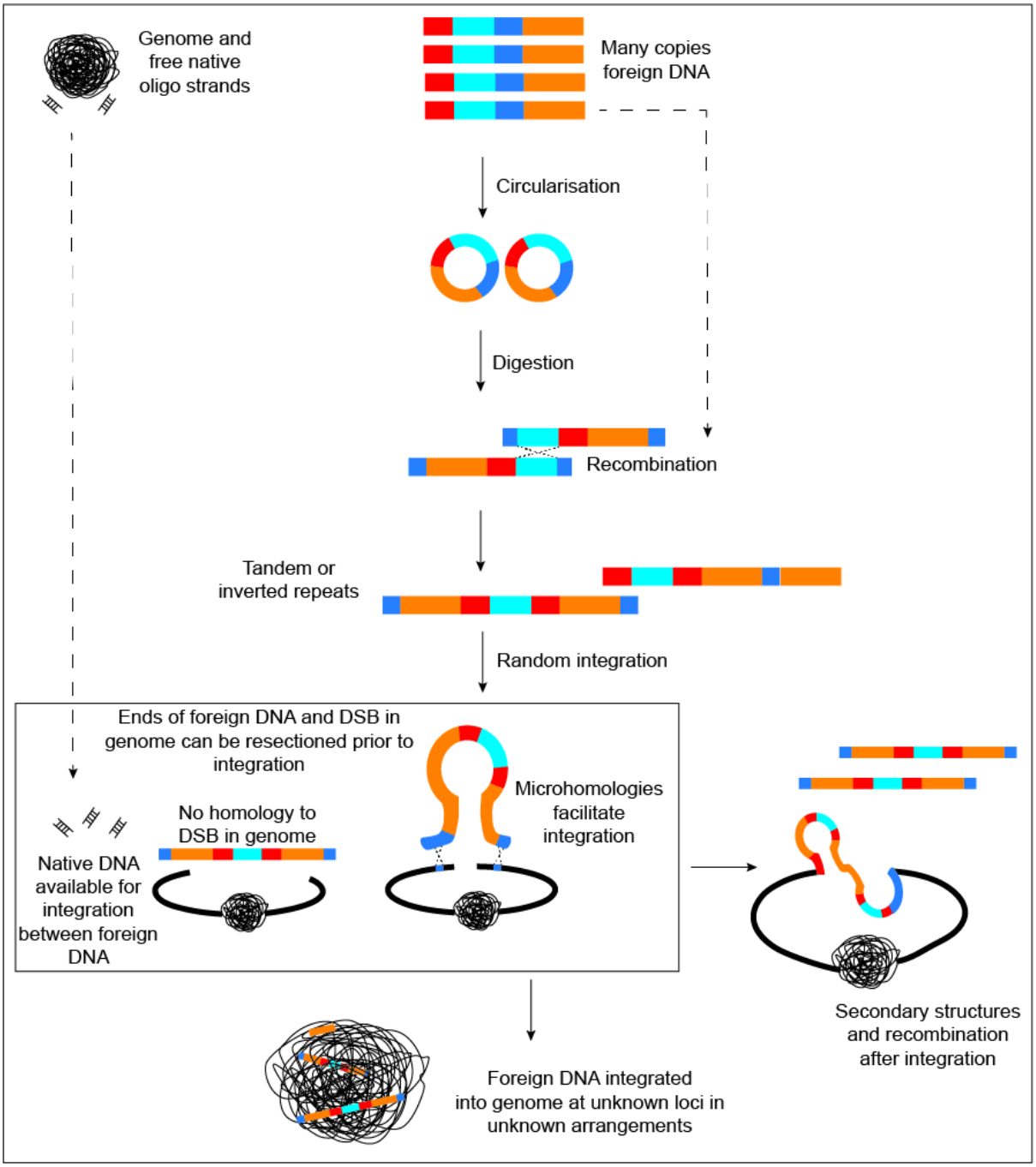


Figure 3. Schematic representation of integration mechanisms that may be at play following transformation with recombinant DNA. DNA entering the cell is subject to a diverse range of DNA-acting enzymes, such as those involved with DNA synthesis, replication, splicing, degradation, repairing, transcription and more (Kohli et al., 2006; Kohli et al., 2010). These enzymes are active during most stages of the cell cycle and may be acting on the transgene DNA before, during or after integration events (Kohli et al., 2006). Consequently, there are possibly different fates for various fragments and therefore various mechanisms by which fragments get integrated into the genome.

DNA-acting enzymes, such as those involved with DNA synthesis, replication, splicing, degradation, repairing, transcription and more (Kohli et al., 2006; Kohli et al., 2010) (Figure 3). These enzymes are active during most stages of the cell cycle and may be acting on the transgene DNA before, during or after integration events (Kohli et al., 2006). Consequently, there are possibly different fates for various fragments and therefore various mechanisms by which fragments get integrated into the genome. Furthermore, RICE is beset with issues regarding transgene expression and stability, even though it is still widely relied on today (Moosburner et al., 2020; Shahar et al., 2020; Taparia et al., 2019). Because the mechanisms driving RICE are not well elucidated, such complications are more generically attributed to 'positional effects', whereby transgenes are integrated into important endogenous regions, or transcriptionally repressed regions, or loci that induce epigenetic silencing of the transgene (Jupe et al., 2019). Other factors shown to affect transgene expression and stability include transgene copy number and transgene structure or rearrangement (Jeon et al., 2017; León-Bañares et al., 2004; Scaife & Smith, 2016). Overall, RICE results in genetically diverse transformants with unknown genomic modifications and arrangements. Transformants generated through this process therefore require extensive, laborious and costly screening to identify high expressing cell lines, which may not exhibit stable phenotypes or bio-product output over time (Hallmann, 2007) with deleterious phenotypes that may go unnoticed.

Next generation (non-random) genetic engineering in microalgae

Ideally RICE –and its associated problems– could be replaced by non-random technologies. Such alternatives would ideally (1) result in high, stable expression of transgenes over time without triggering transcriptional silencing (2) be efficient and easy to reproduce for reduced screening (3) cause no disruptions to important

genomic elements and (4) offer predictable, controllable genetic modifications to the species. With respect to both well established and recently developed engineering strategies, non-random microalgal mutagenesis can be categorised as targeted modification and extrachromosomal expression.

Targeted genomic integration

Targeted genomic integration (TGI) results in a gene knock-in or gene knock-out phenotype, and can be used for elucidating or validating a gene function in reverse genetics, or creating a transgenic cell line with novel or enhanced native traits (Huang & Daboussi, 2017). Targeted knock-in involves the integration of recombinant DNA of interest into a locus that is known to provide predictable and stable expression the transgene(s) to overcome position effects (Figure 4). On the other hand, targeted insertional gene knock-out involves integrating a reporter gene encoding antibiotic resistance or a fluorescence signal into an endogenous gene coding sequence, disrupting it for a knock-out genotype and generating a detectable phenotype in the process (Figure 4). Until the use of programmable endonucleases, both processes required the intracellular delivery of recombinant DNA designed to contain flanks that align to the desired location, as well as a high rate of a natural DNA repair pathway, homologous recombination or homology driven repair (HR or HDR) (Figure 4).

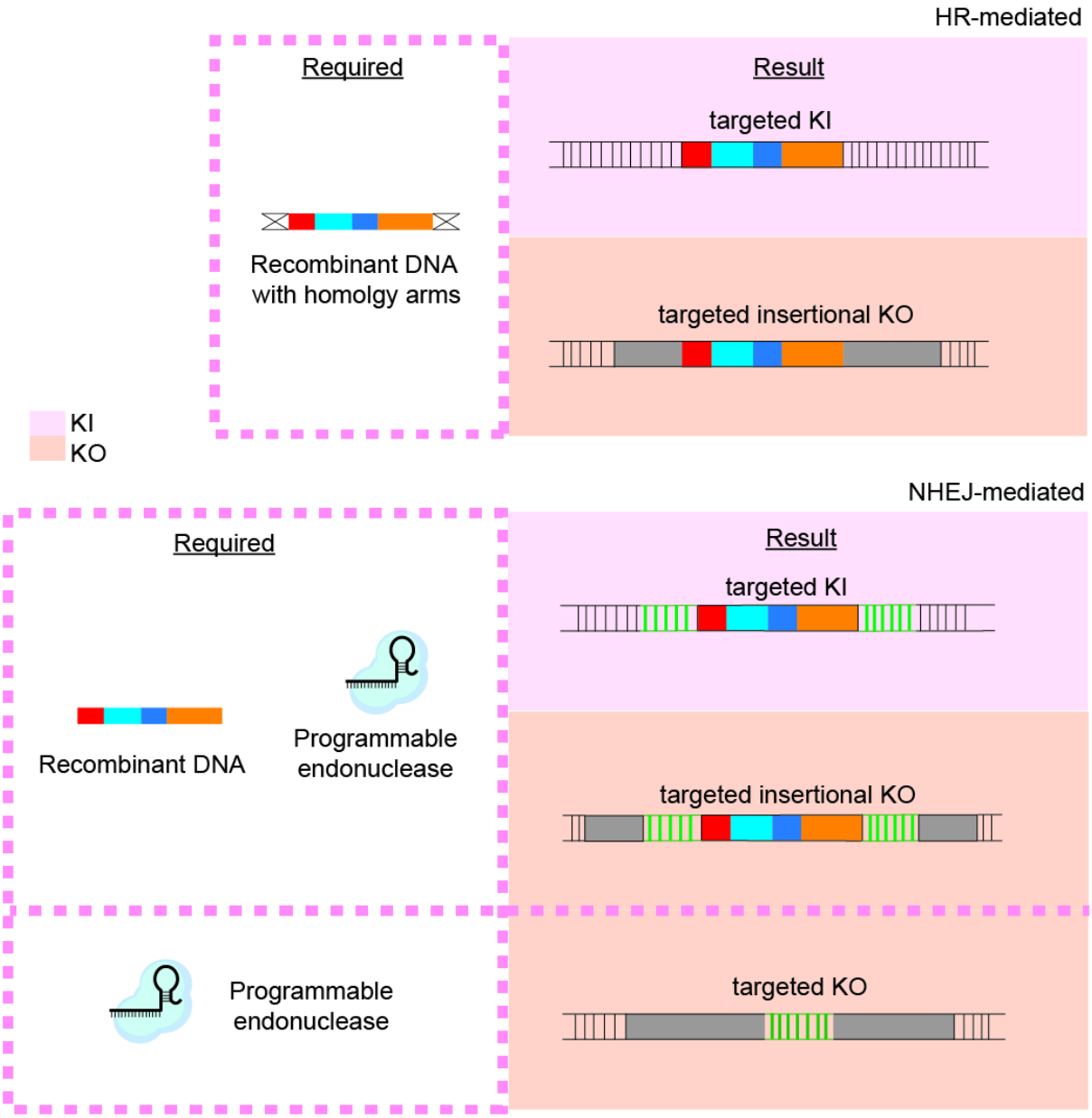


Figure 4. Graphic representation of targeted knock-in and knock-out strategies. When recombinant donor DNA (multi-coloured bar) designed to contain flanks that align to a desired location (marked with an 'X') is supplied intracellularly, cells with high rates of homologous recombination or homology driven repair (HR or HDR) are able to integrate the DNA. When the aim of the integration is gene knock-in or overexpression (targeted KI), donor DNA is usually targeted to a safe harbour locus. However, when the aim is to knock-out an endogenous gene, the donor DNA is targeted to the endogenous gene (grey bar). Herein, a programmable nuclease is not required, unless the organism performs HDR at a rate that is too low to be efficient without it. On the contrary, recombinant DNA without homology arms can also be integrated at a targeted location for knock-in (targeted KI) or for endogenous gene disruption for insertional knock-out (KO) by non-homologous mediated integration (NHEJ-mediated). Here, a programmable nuclease must be supplied to the cell. NHEJ-mediated targeted KO can also be achieved without recombinant DNA, where the

double stranded break generated by the endonuclease is repaired with INDELs, resulting in gene disruption.

HR occurs in cells naturally to repair breaks in the genome by using DNA with homologous flanks, such as a sister chromatid (Steinert et al., 2016). In some species, HR is efficient enough to warrant TGI, including *S. cerevisiae* (Giaever et al., 2002*;* Gietz & Woods, 2002), *E. coli* (Datsenko & Wanner, 2000), certain cyanobacterial strains (Eaton-Rye, 2011; Lan et al., 2015), the thraustochytrid *Schizochytrium* (Cheng et al., 2011)*,* the heterokont *Nannochloropsis oceanica* (Kilian et al., 2011) and the red alga *Cyanidioschyzon merolae 10D* (Minoda et al., 2004). Such engineering strategies have enabled targeted integration of multigene pathways (Bentley et al., 2014; Maury et al., 2008) and dual knock-in knock-out modifications (Li et al., 2016), as well as identified safe harbour regions or neutral sites in the genome which are appropriate for integration (Cantos et al, 2014; Hong et al., 2017; Papapetrou et al., 2011; Pinto et al., 2015; Sadelain et al., 2012). Such ‘safe harbours’ can support high, stable expression without disrupting any essential genetic elements, such as protein coding genes or regulatory regions (Sadelain et al., 2012; Tuggle & Waters, 2015), and have been demonstrated in human cell lines (e.g. *CCR5* and *ROSA26* loci) and mouse cell lines (e.g. *Rosa26 locus)*.

Unfortunately, TGI is not feasible to implement in many model species due to their inherently low rate of HR and consequently, these organisms have had to depend on RICE (Krejci et al., 2012; San Filippo et al., 2008). This includes the model microalgae *C. reinhardtii* and *P. tricornutum*. Numerous strategies have been explored to either increase the rate of HR or reduce the rate of NHEJ in various organisms (Choo et al., 2014; Evy, 1999; Zelensky et al., 2017), including *P. tricornutum* (Angstenberger et al., 2019). One approach which has shown success across most model species is to

cause a site-specific double stranded break at the integration locus of interest using a programmable endonuclease at the same time as recombinant DNA transformation (Lee et al., 2018; Steinert et al., 2016). This has shown to increase the likelihood of the break being repaired by HR in the presence of donor DNA.

In microalgae, endonuclease driven TGI has been validated using transcription activator-like effector nucleases (TALENs), meganucleases, and ribonucleoproteins such as CRISPR-Cas9 and CRISPR-Cpf1 (Daboussi et al., 2014; Ferenczi et al., 2017; Serif et al., 2017; Weyman et al., 2015). Of these endonucleases, ribonucleoproteins like CRISPR-Cas9 and Cpf1 have received the most attention because they are easiest and cheapest to design, as they rely on DNA:RNA base pairing instead of more complex protein-protein interactions, and consequently can target virtually any genomic location. CRISPR technology can also be adapted for many genome editing and DNA targeting goals beyond basic targeted gene knock-in and insertional gene knock-out; such as NHEJ mediated targeted gene knock-in in species with low HR and DNA-free targeted gene knock-out (Figure 4). Importantly, CRISPR technology can be adapted for epigenetic modification (Pflueger et al., 2018; Thakore et al., 2015), gene localisation (Chen et al., 2013; Roberts et al., 2017), genome-wide screening (Chen et al., 2015; Shalem, 2014), regulating gene circuits in synthetic biology (Kiani et al., 2014). Reviews have covered these types of applications and expand upon why the ‘CRISPR movement’ has been so significant in molecular research (Sternberg & Doudna, 2015) as well as in microalgae specifically (Jeon et al., 2017; Naduthodi et al., 2018; Spicer & Molnar, 2018).

Extrachromosomal expression

Another major breakthrough in next generation microalgal genetic engineering technology is extrachromosomal expression (EE). Here, potentially large episomes that contain various DNA parts can be maintained and expressed without requiring genomic integration, as recently demonstrated in the model diatoms, *P. tricornutum* and *Thalassiosira pseudonana* (Karas et al., 2015), the lipid-accumulating *Nannochloropsis oceanica* (Poliner et al., 2018) and the red-algae *Porphyridium purpureum* (Li & Bock, 2018). EE of transgenes present on non-integrative episomes have been validated in yeast (Heinemann, 1989; Sikorski et al., 1990) and mammalian cells (Waters, 2001). EE is well-suited to large, multigene DNA construct delivery. Unlike RICE, it is expected to offer predictable and consistent transgene expression (Karas et al., 2015; Scaife & Smith, 2016). Non-integrative episomes are extremely appealing for synthetic biology applications as backbones of self-maintaining mini-chromosomes.

However, there is little knowledge available regarding mechanisms of episomal maintenance (Diner et al., 2016) and expression, including the level of transgene expression that can be achieved by EE, and whether fragments of episomal DNA are inadvertently integrated into the nuclear genome. It is also not yet known how episomal copies are maintained within each cell, and how dynamic episomal copy number and segregation across individual cells at different stages of the life cycle, as has been uncovered in yeast. This is because EE technology has only recently been described in diatoms with limited examples of its use to express transgenes.

Terpenoids

Terpenoids are an exceptionally diverse (>50,000 compounds) group of natural compounds found ubiquitously throughout nature (Buckingham, 2004). In eukaryotic cells, terpenoids make up both primary metabolites, which are essential and required for growth and development, and secondary metabolites, which are not essential, but instead provide the organism with beneficial traits (Pichersky & Raguso, 2018; Pan el al., 2016). Two major groups of essential terpenoids include sterols and pigments. Sterols are components of eukaryotic membranes and regulate their function (Dufourc, 2008). In autotrophic organisms including plants and photosynthetic microbes, photosynthetic pigments (chlorophylls) are required for light capture, whereas accessory pigments (carotenoids) are necessary for photoprotection during light stress (Nagegowda & Gupta, 2020). In plants, carotenoids also provide protection from free radicals during oxidative stress and are important precursors for hormone biosynthesis (Wang et al., 2014).

Although essential terpenoids are highly diverse, the greatest array of terpenoids existing in nature are the secondary, non-essential metabolites found in plants (Dudareva et al., 2006). These trace compounds provide important roles for the plant's ecological interaction with its environment, for example, as cytotoxins for herbicides, pesticides and pathogen defence mechanisms (Pichersky & Raguso, 2018; Pan el al., 2016) and as chemical signals to encourage symbiotic pollinators and beneficial root-associating microbes (Pichersky & Raguso, 2018).

Humans have made use of plant terpenoids for thousands of years (Pichersky & Raguso, 2018); for example, in carving ornaments and religious artefacts from fossilised terpene exudates (Pichersky & Raguso, 2018) and preparing medicinal teas in indigenous cultures across the world (Chandler & Hooper, 1982). More recently,

many plant terpenoids are in demand by various industries for their functions as pharmaceuticals, food additives, fragrances, flavourants, cosmetics, pesticides/insect repellents and for chemical feedstock (Vickers et al., 2014; Zebec et al., 2016; Vavitsas et al., 2018). For example, cleaning agents often contain terpenoids such as α-pinene and limonene, which have fragrance and antimicrobial properties.

MIAs from the medicinal plant *Catharanthus roseus*

Monoterpenoid indole alkaloids (MIAs) are high-value plant secondary terpenoids with potent bioactive, therapeutic properties including anti-tumor, such as vinblastine, vincristine and ellipticine; anti-hypertensive such as ajmalicine and rescinnamine; and antimalarial such as quinine (Pan et al., 2016). A wide array of MIAs including ajmalicine, vinblastine and vincristine are naturally produced in the plant *Catharanthus roseus* and as such, this medicinal plant has been widely studied (Carqueijeiro et al., 2018; Barrales-Cureño et al., 2012; Miettinen et al., 2014; Oudin et al., 2007; Simkin et al., 2013; Thabet et al., 2012; Van Moerkercke et al., 2013; Yamamoto et al., 2016; Pan et al., 2016; Wang et al., 2010) (Figure 5).

All MIAs are produced from the common precursor strictosidine in the plant vacuole (Figure 5). Strictosidine is produced at the branch point of two metabolic pathways where the indole tryptamine–produced in the shikimate pathway –is condensed with the monoterpenoid secologanin–arising in the seco-iridoid pathway– by strictosidine synthetase (EC 4.3.3.2) (Figure 5) (Miettinen et al., 2014; Simkin et al., 2013). The first step of the seco-iridoid pathway involves the hydroxylation of the monoterpenoid geraniol (EC 1.14.13.152). Geraniol is therefore the first monoterpenoid produced in the biosynthesis of MIAs (Figure 5). While geraniol itself is exclusively produced in the plastidial compartment of internal phloem-associated parenchyma (IPAP) and

vascular cells, secologanin is produced in the vacuoles of epidermal cells (Pan et al., 2016) (Figure 1). Such cellular compartmentalisation and separation across tissues is a major aspect of terpenoid metabolism (De Luca et al, 2014; Pan et al., 2016). This highlights the importance of a detailed and widespread understanding of terpenoid metabolism; from central metabolic pathways involved in producing the universal terpenoid precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), to specialised terpenoid pathways required to modify these precursors and chemically ‘decorate’ their end products (Vavitsas et al., 2018).

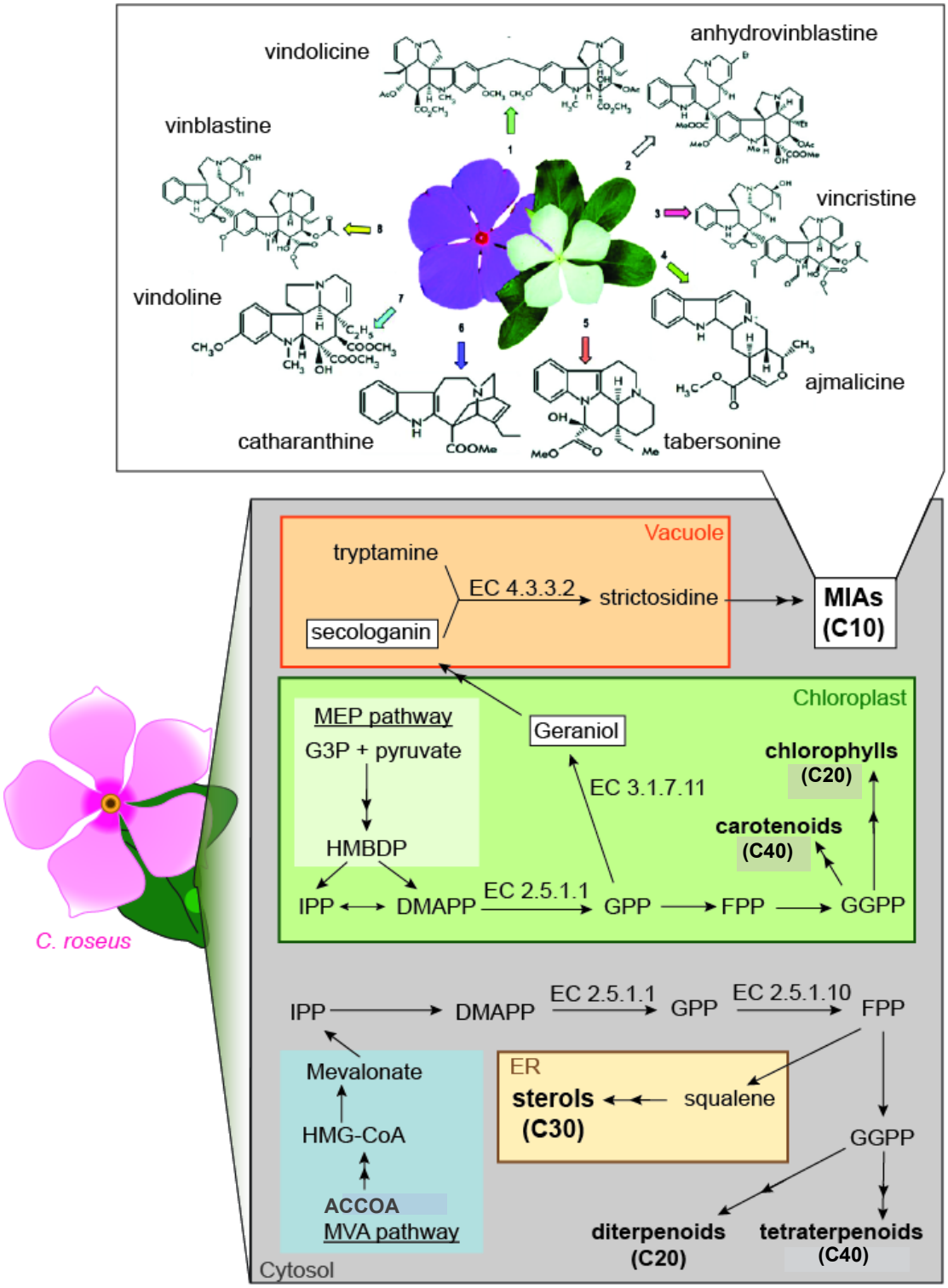


Figure 5. Schematic representation of an overview of the key metabolic pathways and enzymatic reactions (indicated by a single arrow and EC number) involved with geraniol and monoterpenoid-indole alkaloid (MIA) synthesis occurring in various cellular compartments and cell types of the medicinal plant, *Catharanthus roseus*. The universal terpenoid precursor molecules, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are synthesised via two independent metabolic pathways: the mevalonate (MVA) pathway (blue box) in the cytosol (grey

compartment) and the methylerythritol phosphate (MEP) pathway (light green box) in the chloroplast (dark green compartment). The MVA pathway uses acetyl-coenzyme A (ACCOA) to produce mevalonate, which is converted into IPP and DMAPP. In the chloroplast, the MEP pathway uses glyceraldehyde-3-phosphate (G-3-P) and pyruvate to produce (E)-4-Hydroxy-3-methyl-but-2-enyl pyrophosphate (HMBDP), which can be converted into IPP and DMAPP. DMAPP is converted to geranyl diphosphate (GPP) by GPP synthase (EC 2.5.1.1) in the plastid. Plastidial GPP is converted to the monoterpenoid geraniol by geraniol synthase (EC 3.1.7.11). Geraniol is the first monoterpenoid (white box) and produced in the chloroplasts of *C. roseus* internal phloem-associated parenchyma (IPAP) cells and vascular cells. It is then used for the production of another key monoterpenoid, secologanin, which is produced in the vacuoles (orange compartment) of epidermal cells. The secologanin monoterpenoid is then condensed with tryptamine, an indole produced in the shikimate pathway, to form the base molecule, strictosidine, in the Pictet–Spengler reaction. Strictosidine is the universal precursor of all MIAs, including key therapeutic molecules: vindolicine, anhydrovinblastine, vincristine, ajmalicine, tabersonine, catharanthine, vindoline and vinblastine. GPP is also used to make farnesyl diphosphate (FPP), a precursor for squalene and all sterols, which are produced in the endoplasmic recticulum (yellow compartment). Likewise, GPP is also used to make FPP in the plastid, where it acts as a precursor for production of photosynthetic and accessory pigments, such as chlorophylls and carotenoids, respectively.

Synthetic biology can enable heterologous production of MIAs in microorganisms

MIAs are in high demand due to their important therapeutic properties; however, their extraction from *C. roseus* is extremely expensive and unreliable due to the naturally occurring low concentrations in the plants tissues (0.0002% fresh weight) (Miettinen et al., 2014). This is because monoterpenoids and MIAs are often only synthetised in trace amounts in tissue-specific locations, and usually only in response to particular stresses and environmental stimuli (Zhou et al., 2009).

Therefore, industries have moved towards methods for terpenoid production that are not dependent on biosynthesis in these medicinal plant species. One alternative is the chemical synthesis of terpenoids. Chemical synthesis is widely used in industry, however, it is complex and expensive (Kawamura et al., 2016). Chemical synthesis is also only possible for a certain terpenoid molecules every possible terpenoid molecule,

as high chirality and many functional moieties in some compounds make synthesis too complex or expensive, with many purification steps (Schwab et al., 2015). This is particularly true for vincristine and vinblastine, which are particularly difficult to synthesise chemically. Alternatively, synthetic biology offers new possibilities to address the limitations associated with both plant extraction and chemical synthesis of monoterpenoids. The commercial feasibility of such microbial production systems is made evident by the growing number of companies using these technologies: from pharmaceuticals such as Arteminisin (Amyris, Sanofi) and Sitagliptin (Codexis, Merck) to food products such as Vanillin (Evolva, IFF) and Resveratrol (Evolva) to fine chemicals including plasticizers (BioAmber) and polyhydroxyalkanoates (Metabolix) (Jullesson et al., 2015). Here, the benefits of biological production of complex and expansive molecules associated with native plant biosynthesis can be combined with the high-yielding, fast growing traits of microbes. Given that MIA metabolic engineering is highly complex, many studies have explored the possibility of heterologous MIA biosynthesis by optimising geraniol production (Fischer et al, 2011; Jiang et al., 2017; Liu et al., 2016; Shah et al., 2013; Zhou et al, 2015; Brown et al., 2015).

Geraniol

Geraniol is a useful monoterpenoid to investigate the impact of heterologous monoterpenoid production in a microbial host for various reasons. First, geraniol is converted in a single enzymatic reaction, requiring expression of only a single plant gene geraniol synthase (GES, EC 3.1.7.1). Second, this makes it a useful direct readout tool to quantify the carbon flux directed towards GPP synthesis (Qian et al., 2019) in order to evaluate the impact of a particular rational design. This is because it is the first monoterpenoid in *C. roseus* MIA biosynthesis, and because *GES* expression will terminate the synthetic monoterpenoid pathway in heterologous host

microorganisms. Therefore, investigating strategies which optimise flux towards geraniol production will be useful for more complex, multi-step MIA synthetic pathways. Third, geraniol is itself also a valuable and commercially relevant monoterpenoid, used as a fragrance and insect repellent (Chen et al., 2010). Fourth, it is a volatile compound that can be directly captured in an organic solvent supplied to the cell culture (Liu et al., 2016). This solvent can then be analysed by gas chromatography–mass spectrometry (GCMS) for accurate geraniol quantification.

Terpenoid biosynthesis

Terpenoid metabolism is highly complex for various reasons. First, terpenoid pathways –such as those involved in MIA biosynthesis in plants– occur in different cellular compartments and in different cell types in plants (Figure 5). Second, terpenoid metabolism is comprised of branching metabolic pathways, such as the alkaloid and monoterpenoid branches which come together to form secologanin, a key intermediate in the production of MIAs (Figure 5). Concurrently, it also depends on flux through central carbon metabolism, such as glycolysis and Calvin-Benson cycle, and highly specialised terpenoid specific pathways, such as the strictosidine pathway for producing the universal MIA precursor, strictosidine. Being such an integral path of cellular metabolism, terpenoid pathways are also tightly controlled, for example through the action of regulatory enzymes and flux and metabolite feedback mechanisms. Finally, terpenoid biosynthesis occurs in all eukaryotic and prokaryotic cells and consequently, some aspects are highly evolutionarily conserved, whilst others are species-specific (Athanasakoglou & Kampranis, 2019). For these reasons, it is useful to consider terpenoid metabolism as conceptually separate modules (Vavitsas et al., 2018) or stages (Athanasakoglou & Kampranis, 2019). It is important

to note that these models are entirely hypothetical and are not indicative of any biologically present factions.

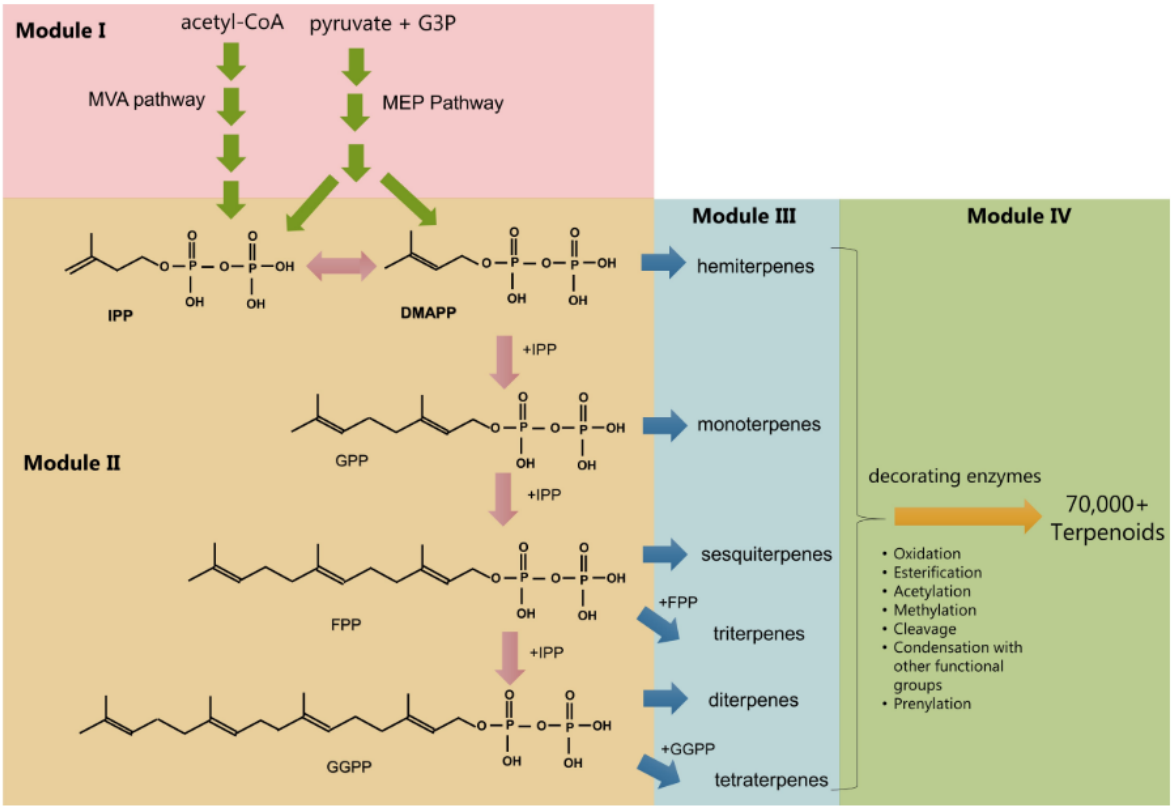


Figure 6. Biosynthesis of terpenoids. The pathways have been conceptually separated into four modules, which is representative of the modularised method many metabolic engineers use to approach isoprenoid pathway engineering. Mevalonate (MVA) and methyl-D-erythritol phosphate (MEP) pathways lead to IPP (isopentenyl pyrophosphate) and DMAPP (dimethylallyl pyrophosphate) (module I). Additions of IPP produce higher-order prenyl phosphates (module II), dephosphorylation (often coincident with or followed by bond rearrangement and/or cyclisation) to form specialized terpenoid backbones (module III), chemical decorations, and other modifications to yield end products. Note that not all end products undergo decorations of the carbon skeleton. From Vavitsas et al. (2018).

Module One: The production of IPP and DMAPP

The production of the five carbon prenylated phosphates IPP and DMAPP, the universal isoprenoid building blocks of all terpenoids, can be seen as the first module of terpenoid biosynthesis (Figure 6). There are two pathways which produce IPP and DMAPP: the mevalonate (MVA) pathway, which is mostly present in eukaryotes and

archea; and the methyl-D-erythritol (MEP) pathway, which is usually found in bacteria and in the chloroplasts of photosynthetic organisms due to its endosymbiotic origin (Figure 6 and 7). Plants and some microalgae, such as diatoms, are unique in that they contain both a cytosolic MVA pathway and a plastidial MEP pathway.

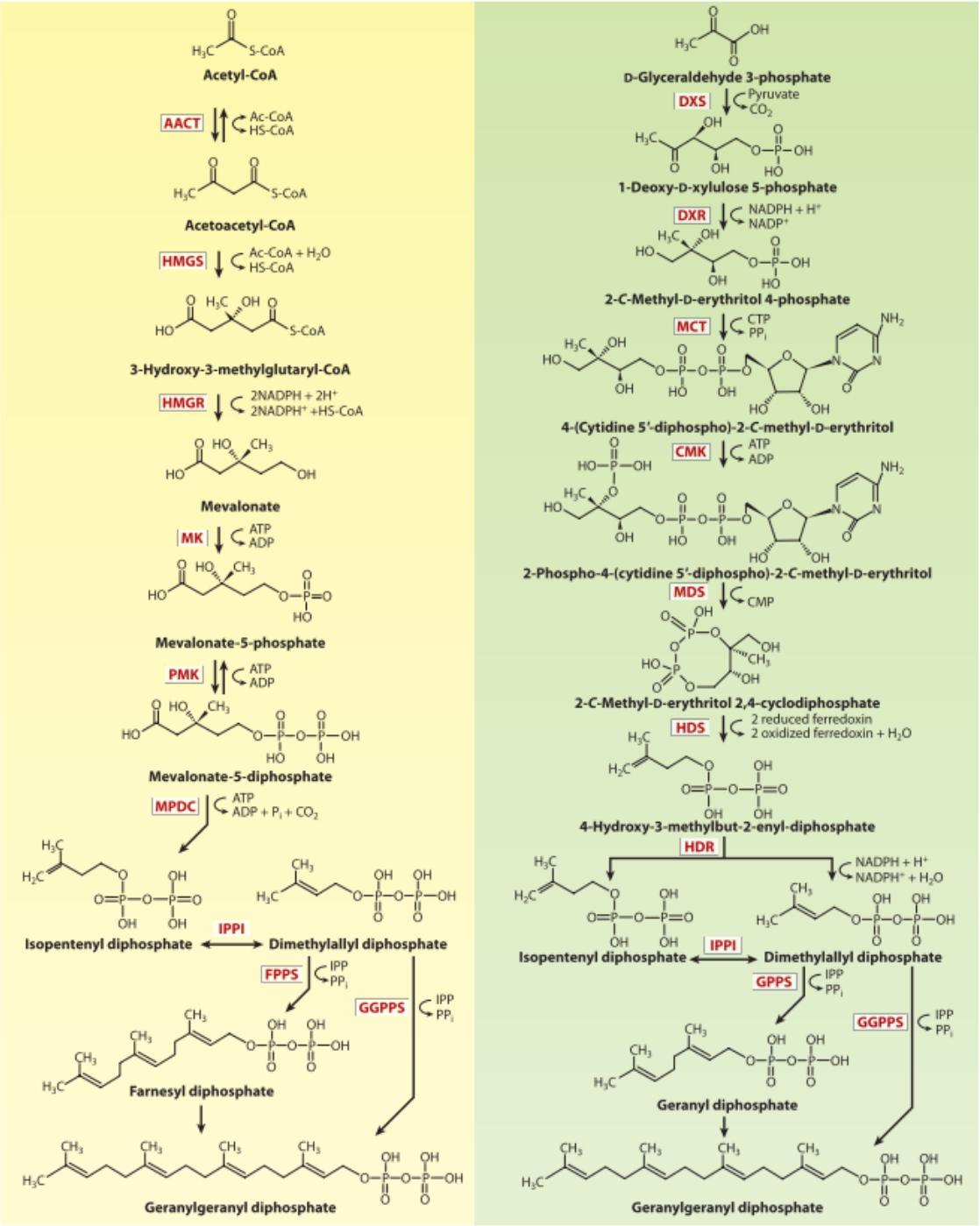


Figure 7. Enzymatic reactions in the MVA and MEP pathways in plants and synthesis of the short-chain prenyl diphosphates. The MVA pathway is shown in yellow; the MEP

pathway is shown in green. AACT; Acetyl-coenzyme A c-acetyltransferase, HMGS; hydroxy-methylglutaryl-coenzyme A synthase, HMGR; hydroxyl-methylglutaryl-coenzyme A reductase, MVK; mevalonate kinase, PMK; phospho-mevalonate kinase, MVD; mevalonate disphosphate decarboxylase, DXS; 1-deoxy-D- xylulose 5-phosphate synthase, DXR; 1-deoxy-D-xylulose 5-phosphate reductoisomerase, MCT; 2-C-methyl-D-erythritol-4-phosphate-cytidylyltransferase, CMK; 4- diphosphocytidyl-2c-methyl-d-erythritol kinase, MDS; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, HDS; (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase, HDR; 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase, IPPI; isopentenyl diphosphate isomerase, GPPS; geranyl diphosphate synthase, FPP; farnesyl diphosphate synthase, GGPP; geranylgeranyl diphosphate synthase. From Vranová et al. (2013).

This module includes any metabolic reactions which commit resources towards these primary metabolic pathways. For example, carbon in the form of acetyl-coenzyme A (ACCOA), pyruvate and glyceraldehyde-3-phosophate (GAP) are required for the isoprenoid backbone component of IPP and DMAPP. GAP, pyruvate, and ACCOA are all by-products of the breakdown of glucose to produce energy for the cell during glycolysis. The MEP pathway involves seven reactions initiated by the condensation of pyruvate and GAP by 1-deoxy-D- xylulose-5-phosphate synthase (DXS) (EC 2.2.1.7). DXS has been shown to be a crucial rate-limiting enzyme in *C. roseus* (Han et al., 2013) and other plant species (Enfissi et al., 2005; Morris et al., 2006) as well as the bacteria *E. coli* (Kim & Keasling, 2001), cyanobacteria *Synechocystis sp. PCC6803* (Kudoh et al., 2014) and recently in the microalga, *C. reinhardtii* (Wichmann et al., 2018). The second enzyme 1-deoxy-D- xylulose 5-phosphate synthase (DXR) and the final enzyme 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase (HDR), have also shown to be rate-limiting in various plant species (Botella-Pavía et al., 2004; Veau et al., 2000; Walter et al., 2000).

In the MVA pathway, two ACCOA molecules are condensed by acetyl-coenzyme A c-acetyltransferase (AACT; EC 2.3.1.9) to form acetoacetyl-coenzyme A. Three subsequent reactions catalysed by hydroxy-methylglutaryl-coenzyme A synthase

(HMGS) hydroxyl-methylglutaryl-coenzyme A reductase (HMGR), and mevalonate kinase (MVK), respectively result in the production of mevalonic acid. The final reaction converts mevalonate-5-diphosphate into IPP by MVA diphosphate decarboxylase (MPDC; EC 4.1.1.33).

IPP generated by both the MVA and MEP pathways is converted to DMAPP by isopentenyl diphosphate isomerase (IPPI or IDI; EC 5.3.3.2) in a reversible reaction (Vranová et al., 2013) . The balance between IPP:DMAPP metabolites is extremely important for terpenoid biosynthesis; for example in diatoms, sterols depend on a ratio of 2:1, whereas carotenoids require higher IPP in a 3:1 ratio (Athanasakoglou & Kampranis, 2019). Many metabolic engineering approaches have focused on overexpressing rate-limiting enzymes (Kim & Keasling, 2001; Liu et al., 2013; Zhao et al., 2016) involved in module one or even supplementing the native IPP and DMAPP biosynthesis pathway with the presence of the other, non-natively present pathway in *E. coli* (Qian et al., 2019; C. Yang et al., 2016; J. Yang et al., 2012) and *S. cerevisiae* (Carlsen et al., 2013; Partow et al., 2012).

Module Two: Synthesis of prenyl diphosphates

IPP and DMAPP are five carbon prenyl phosphates that forms the carbon backbone for hemiterpenoids (C5). The consecutive addition of a five carbon IPP functional group to a five carbon DMAPP generates three additional terpenoid moieties of increasing complexity: geranyl pyrophosphate (GPP), farnesyl pyrophosphate (FPP) and geranylgeranyl pyrophosphate (GGPP). This conversion of IPP and DMAPP carbon skeletons can be considered the second conceptual module of terpenoid biosynthesis.

IPP and DMAPP are condensed by GPP synthase (GPPS; EC 2.5.1.1) to produce the two isoprene unit-containing prenyl phosphate GPP. In plastids of numerous medicinal plants such as *C. roseus* and *Abies grandis*, geraniol synthase converts GPP into the monoterpenoid (C10) geraniol. Concurrently, cytosolic GPP is condensed with IPP by FPP synthase to produce FPP (EC 2.5.1.10), which is comprised of three isoprenes (C15) and is required for all sesquiterpenoids. When two FPP molecules are condensed by squalene synthase (EC 2.5.1.21), they give rise to a thirty carbon squalene metabolite, which is required for all triterpenoids (C30). An essential class of triterpenoids found in all eukaryotes are sterols, which make up a crucial component of cell membranes for maintaining fluidity and flexibility, as well as functioning as signalling molecules in cell growth and differentiation (Benveniste, 2004).

E. coli and most *S. cerevisiae*–with the exception of a few winemaking strains–do not possess any GPP synthase enzymes and consequently, do not natively produce monoterpenoids (Carrau et al., 2005). Instead, they use a farnesyl diphosphate synthase (ispA for *E.coli* and ERG20 for *S. cerevisiae*) that converts IPP and DMAPP into GPP (EC 2.5.1.1) and then immediately into FPP (EC 2.5.1.10). In this way, very little GPP escapes and diffuses away from the FPP synthase domain and consequently, significant efforts have gone into engineering both species to redirect flux away from FPP biosynthesis to allow for a sufficient enough free pool of GPP which heterologous metabolic pathways could draw from in order to produce monoterpenoids (Blanchard & Karst, 1993; Fischer et al., 2011; Oswald et al., 2007). While bacteria do not produce sterols, they too use FPP to produce sterol-like compounds called hopanoids (Dufourc, 2008).

Finally, GGPP is the prenyl phosphate precursor for the biosynthesis of all pigments in photosynthetic organisms, including chlorophylls and carotenoids. GGPP is

comprised of 4 isoprene subunits (C20) and is the base for diterpenoids, which are mostly found in plants, fungi and some marine microalgae. When two GGPP molecules are condensed together they form tetraterpenoids (C40), making up most carotenoids. These are mostly produced in the chloroplast and are essential as antioxidants (Sachindra et al., 2007) and UV attenuators for photoprotection (Kuczynska et al, 2015).

Module Three: Terpenoid synthesis and diversification

The production of hemi-, mono-, sesqui-, di-, tri- and tetra-terpenoids can be considered the third module of terpenoid biosynthesis, which builds upon and modifies the four major prenyl phosphate carbon skeletons, IPP, DMAPP, GPP, FPP, GGPP, using terpene synthases. Many of the enzymes involved are cytochrome p450s. Because of the vastness of terpenoids that exist in nature, this module is exponentially more complex than module one or two and comprises of many metabolic pathways in various cellular compartments. For example, module three of MIA biosynthesis includes the secoiridoid-, shikimate-, and strictosidine pathways.

In heterologous monoterpenoid systems, *S. cerevisiae* and *E. coli* have been extensively engineered to express plant enzymes involved in module three of MIA biosynthesis (Campbell et al., 2016; Fischer et al., 2011; Jiang et al., 2017; Oswald et al., 2007; Zhou et al., 2014). A noteworthy example in *S. cerevisiae* optimised aspects within module one, two and three for production of strictosidine (Brown et al., 2015). In module one, researchers overexpressed three endogenous genes involved in the MVA pathway for enhanced flux. In module two, they knocked out endogenous ERG20 and replaced it with *A. grandis* GPP synthase to increase the free pool of GPP. In module three, they introduced a synthetic strictosidine pathway, encoding 13

additional plant enzymes for converting GPP into strictosidine. Finally, they also knocked out two endogenous enzymes which compete with the synthetic strictosidine pathway for geraniol; alcohol acetyltransferase (ATF1) and NADPH oxidoreductase (OYE2), which are involved with the production of geranyl acetate and citronellol, respectively. This metabolic engineering feat was made possible by HR-driven TGI and resulted in an impressive 0.53 mg/L yield of strictosidine (Brown et al., 2015).

Module Four: Terpenoid modification and diversification

While module three comprises of the metabolic reactions involved in synthesising a range of terpenoid compounds, module four can be considered the last step of terpenoid biosynthesis before the final product is released to perform it biological role. Here, the terpenoid is further modified by a suite of 'decorating enzymes' capable of oxidation, esterification, acetylation and methylation, cleavage, condensation, or prenylation. In many heterologous terpenoid systems, the final products are the most complex to produce (Ward et al., 2018). In both *E. coli* and *S. cerevisiae*, heterologous geraniol accumulation has been associated with toxic effects that hinder its production (Shah et al., 2013; Zhao et al., 2016). Both module three and four contain an exceptional amount of diverse metabolic reactions required to generate the vast myriad of terpenoid compounds that exist in nature.

Heterologous terpenoid production in microalgae

Unlike yeasts and bacteria, microalgae (not including cyanobacteria) have only recently been explored for their capacity to produce plant terpenoids. Heterologous terpenoid production in microalgae has been demonstrated in only a handful of recent proof of concept publications in model species. In the chlorophyte *C. reinhardtii*, (E)-alpha-bisabolene was produced at 11 mg/L (Wichmann et al., 2018), patchoulol at

0.47 mg/L (Lauersen et al., 2016), and manoyl oxide at 80 mg/g dry weight (Lauersen et al., 2018). In the marine stramenopile diatom *Phaeodactylum tricornutum,* the triterpenoid betulinic acid was produced at up to 0.1 mg/L (D'Adamo et al., 2018). While these are foundational studies that validate the possibility of heterologous production of terpenoids in model microalgal species, they all relied on first generation genetic engineering approaches. Such engineering strategies are not amenable to more complex gene stacking and metabolic pathway engineering that would be required for taking these strains from proof of concepts to industrially relevant strains. Recently, we demonstrated that extrachromosomal expression (EE) can be used to efficiently express the fusion protein CrGES-mVenus in *P. tricornutum* cytosol to produce up to 0.309 mg/L (0.21 µg/10^7cells) geraniol following bacterial conjugation (Fabris et al., 2020). EE of transgenes is not subject to position effect (Karas et al., 2015) and could therefore provide highly reproducible, consistent, and controllable expression, which is a basic requisite for synthetic biology. In contrast, randomly integrated chromosomal expression (RICE) can result in genetically dissimilar transformants and consequently varied transgene expression among them. It is generally accepted that position effect causes these varied transgene expression phenotypes, and this has been shown by comparing an average of three to five *P. tricornutum* transformant cell lines. However, diatom phenotypes derived from EE and RICE have not previously been systematically parameterised. Because little is known regarding the mechanisms and effects following both EE and RICE of transgenes, it is unclear how these different engineering strategies will compare regarding the expression of *CrGES-mVenus*, and consequently, heterologous monoterpenoid production.

Microalgae are under-explored chassis organisms for monoterpene applications

Model species of microalgae are promising chassis organisms for heterologous terpenoid production–compared to the routinely used *S. cerevisiae* and *E. coli*–due to their photosynthetic capabilities. Photosynthetic production of sugar allows for lower culturing costs and more sustainable approaches commercially, as no external carbon sources are required to feed these microbes (Vickers *et al.*, 2014; Davies et al., 2015; Vavitsas et al., 2018). Phototrophs also require a naturally larger abundance and diversity of terpenoids in the form of photosynthetic pigments, such as chlorophylls, and accessory pigments, such as carotenoids (Lohr et al., 2012). For example, microalgae naturally produce β-carotene and lycopene, both of which protect the photosystem from light stress and offer antioxidant properties desired by the nutraceutical and cosmetic industries (Gille et al., 2016; Cicero & Colletti, 2017). It is hypothesised that microalgae have the potential to commit more energy and resources called flux, through naturally their occurring terpenoid biosynthetic pathways (Lohr et al., 2012). This includes the reducing agent NADPH, an essential electron carrying co-factor required by cytochrome P450s, and organic carbon, which is fixed during photosynthesis and required as the isoprene carbon backbone of terpenoids.

Diatoms offer a potential uniquely suited background for monoterpenoid production compared to not only yeast and bacteria, but other microalgae too. This is due to the diatoms peculiar metabolism (Ashworth et al., 2016; Longworth et al., 2016; Remmers et al., 2018*;* Smith et al., 2019; Allen et al., 2011; Fabris et al., 2014; Fabris et al., 2012; Kroth et al., 2008), suggested to have arisen from numerous endosymbiotic events and frequent horizontal gene transfers (HGT). This peculiar metabolism is important for two main processes in conceptual module one of terpenoid biosynthesis.

Like plants, diatoms contain two pathways for IPP and DMAPP production: a cytosolic MVA pathway and a plastidial MEP pathway. They also bear two additional glycolytic pathways, which are hypothesised to be contributors to the production of carbon precursors required by the MEP and MVA pathways (Fabris et al., 2012). The Entner-Douroroff pathway, more conventionally associated with prokaryotic metabolism and the phosphoketolase pathway, which is common in fungal metabolism, are both present in *P. tricornutum*. Pyruvate and G-3-P are produced via the Entner-Douroroff pathway and could be precursors for MEP pathway, whereas ACCOA is produced via phosphoketolase pathway and could be a precursor for MVA pathway (Fabris et al., 2012; Meadows et al., 2016).

Altogether, the presence of additional pathways allows for greater supply of IPP and DMAPP for native diatom terpenoid metabolism (such as pigments and sterols) as well as for heterologous monoterpenoid production. It also allows for many more engineer-able nodes –genes and enzymes to knock-out or overexpress– to modify flux through these pathways to both support native terpenoid production and drive heterologous production. In these ways, *P. tricornutum* is highly unique and theoretically well suited for metabolic engineering for heterologous terpenoid production. Additionally, research in our laboratory recently uncovered that *P. tricornutum* has a detectable free pool of geranyl pyrophosphate (GPP) (Fabris et al., 2020), the first precursor in *C. roseus* monoterpenoid biosynthesis pathways, such as the strictosidine pathways for producing vinblastine and vincristine.

C. reinhardtii and *P. tricornutum* are primed for synthetic biology

Although synthetic biology-based research is delayed in *C. reinhardtii* and *P. tricornutum*, developments in their engineering toolkits have the potential to change

this in the near future. Firstly, both offer high-growth rates and adaptation to growth in high density conditions required for biotechnologically useful chassis organisms. Second, there is fully sequenced genomic data available for both species (Merchant et al., 2007; Bowler et al., 2008) and increasingly emerging metabolomics, transcriptomics and proteomics datasets for informing more complex genetic modification (Keeling et al., 2014; Kleessen et al., 2015; Maheswari et al., 2009; Remmers et al., 2018). Third, there is a range of selectable markers, reporter genes and tags, versatile promoters available for genetic engineering (Scranton et al., 2015; Daboussi et al., 2014), as well as the modular systems; MoClo for *C. reinhardtii* (Werner et al., 2012) and Universal Loop (uLoop) assembly for diatoms including *P. tricornutum* (Pollak et al., 2019) for more complex gene-stacking designs. Finally, next generation genetic engineering strategies have been validated with programmable endonuclease-driven targeted engineering in both species (Baek et al., 2016; Greiner et al., 2017; Moosburner et al., 2020; Serif et al., 2018; Sharma et al., 2018; Shin et al., 2016). *P. tricornutum* has also been demonstrated for extrachromosomal expression via bacterial conjugation (Diner et al., 2016; Karas et al., 2015).

Altogether, these developments mark *C. reinhardtii* and *P. tricornutum* as high potential synthetic biology chassis organisms, particularly regarding heterologous production of monoterpenoids. However, before more complex multigene pathways can be constructed and testing in these species, it is crucial that new research is conducted to characterise both the native biochemistry of the microalga, as well as the engineering tools available. Such knowledge will be foundational in advancing synthetic biology in these microalgae, not only for monoterpenoid production, but other synthetic biology ventures in these under explored species.

AIMS AND OBJECTIVES

Aim One: CRISPR-Cas9 method development for *Chlamydomonas reinhardtii*

There is a need to shift genetic engineering practises towards next generation genome editing, particularly regarding the highly flexible system of CRISPR-Cas9 technology. Therefore, we aimed to:

- Generate a microalgal-specific pipeline for CRISPR-Cas9, informed by both algal and non-algal approaches.
- Demonstrate reproducibility of CRISPR-Cas9 based gene editing in *C. reinhardtii.*
- Investigate the impact of alternative, theoretically more suitable CRISPR-Cas9 delivery methods, never before tested in this species.

These aims were addressed in Chapter Two.

Aim Two: Engineer *Phaeodactylum tricornutum* for monoterpenoid production and large-scale phenotyping of transformant libraries

Although randomly integrated chromosomal expression (RICE) is widely used in microalgal genetics, the molecular mechanisms involved are poorly understood. It is also generally accepted that RICE results in transgene expression-related issues; however, this has never before been parametrised. The recent development of extrachromosomal expression (EE) and targeted genome editing for the marine diatom *P. tricornutum* could replace RICE altogether; however, a better understanding of these strategies at the phenotypic level is required for rigorous comparison and for informing these next generation approaches. Furthermore, *P. tricornutum* has been identified as high potential biofactory for heterologous production of monoterpenoids due to its native ability to generate a diverse range of terpenoids, as well as highly unique features of the diatom's terpenoid metabolism (D'Adamo et al., 2018; Fabris et

al., 2014, 2012; Keeling et al., 2014; Pollier et al., 2019; Cvejic & Rohmer, 2000; Athanasakoglou & Kampranis, 2019). One such relevant aspect of diatom metabolism–not seen in yeast or bacterial species–is it's use of both a naturally present MVA and MEP biosynthesis pathway. This is highly unusual even in microalgae, of which most species contain a plastidial MEP pathway, but lost a cytosolic MVA pathway through evolution (Lohr et al., 2012).

Therefore, we aimed to:

- Generate large libraries of *P. tricornutum* genetically engineered to express the *Catharanthus roseus* enzyme geraniol synthase, fused to a fluorescent reporter mVenus, (CrGES-mVenus) following both RICE and EE.
- Generate a high-throughput method to interrogate the phenotypes of the mutant cell libraries at both the intra- and inter-cellular level, in order to comprehensively parametrise these engineering approaches.
- Identify superior cell lines within the RICE and EE libraries that demonstrate high and stable mVenus fluorescence, a proxy for CrGES-mVenus expression.
- Validate the superior cell lines by quantifying the product of geraniol synthase, geraniol.

These aims were addressed in Chapter Three.

Aim Three: Transgenome interrogation of superior geraniol yielding strains by long-read whole-genome DNA sequencing

Prior to this research, there was virtually no knowledge of the implications of RICE on microalgal transgenomes. There were also no previously known safe harbour loci for any eukaryotic microalgae. Therefore, we hypothesised that long-read whole-genome sequencing could be applied to uncover the types of rearrangements and the frequency and size of integration events associated with superior geraniol yielding RICE cell lines. Additionally, it is not yet known if exogenous DNA is inadvertently integrated into the nuclear genome following EE. Such knowledge is crucial, as EE is

understood to not require integration and thus any integration events can be overlooked. Therefore, we aimed to:

- Reveal genome wide, sub-chromosomal details of superior geraniol yielding *P. tricornutum* cell lines engineered by RICE, namely the number and location of integration events, their genetic arrangements, and identify any putative safe harbour loci for targeted integration strategies.
- Determine if episomal DNA is inadvertently integrated into nuclear genome following bacterial conjugation.

These aims were addressed in Chapter Three.

Aim Four: Develop the first CRISPR-Cas9 mediated targeted genomic integration (TGI) strategy for heterologous geraniol production

Following the identification of the first putative safe harbour loci for *P. tricornutum*, and given the flexibility of CRISPR-Cas9 technology, we aimed to:

- Demonstrate reproducibility of the recently published protocol describing CRISPR-Cas9 genome editing in *P. tricornutum* via biolistic delivery of ribonucleoprotein (RNP).
- Adapt this approach for the first CRISPR-Cas9 driven targeted integration in this species.
- Validate the putative safe harbours identified in aim three.
- Test the appropriateness for the recently described endogenous marker, *uridine-5'-monophosphate synthase*.

These aims were addressed in Chapter Four.

Aim Five: To increase heterologous geraniol production in *P. tricornutum* and uncover the implication on sterol and pigment biosynthesis

There is a significant gap in knowledge regarding algal systems for heterologous monoterpenoid production. Therefore, we aimed to:

- Explore two rational design strategies to increase geraniol production using extrachromosomal expression.
- Determine the impact of geraniol accumulation on production of sterols and pigments.

These aims were addressed in Chapter Five.

REFERENCES

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., … Bowler, C. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, *473*(7346), 203–207. https://doi.org/10.1038/nature10074

Amin, S. A., Parker, M. S., & Armbrust, E. V. (2012). Interactions between Diatoms and Bacteria. *Microbiology and Molecular Biology Reviews*, *76*(3), 667–684. https://doi.org/10.1128/mmbr.00007-12

Andrew E. Allen, Julie LaRoche, Uma Maheswari, Markus Lommer, Nicolas Schauer, Pascal J. Lopez, Giovanni Finazzi, Alisdair R. Fernie, and C. B. (2008). Whole cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation_2008_Proc Nat Acad Sci 105_10438-10443.pdf. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 10438–10443.

Angstenberger, M., Krischer, J., Aktaş, O., & Büchel, C. (2019). Knock-Down of a ligIV Homologue Enables DNA Integration via Homologous Recombination in the Marine Diatom Phaeodactylum tricornutum. *ACS Synthetic Biology*, *8*(1), 57–69. https://doi.org/10.1021/acssynbio.8b00234

Apt, K. E., Kroth-Pancic, P. G., & Grossman, A. R. (1996). Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Molecular and General Genetics*, *252*(5), 572–579. https://doi.org/10.1007/s004380050264

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, *459*(7244), 185–192. https://doi.org/10.1038/nature08057

Ashworth, J., Turkarslan, S., Harris, M., Orellana, M. V., & Baliga, N. S. (2016). Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. *Marine Genomics*, *26*, 21–28. https://doi.org/10.1016/j.margen.2015.10.011

Athanasakoglou, A., & Kampranis, S. C. (2019). Diatom isoprenoids: Advances and biotechnological potential. *Biotechnology Advances*, *37*(8). https://doi.org/10.1016/j.biotechadv.2019.107417

Baek, K., Kim, D. H., Jeong, J., Sim, S. J., Melis, A., Kim, J.-S., … Bae, S. (2016). DNA-free two-gene knockout in Chlamydomonas reinhardtii via CRISPR-Cas9 ribonucleoproteins. *Scientific Reports*, *6*, 30620. https://doi.org/10.1038/srep30620

Bateman, J. M., & Purton, S. (2000). Tools for chloroplast transformation in Chlamydomonas: Expression vectors and a new dominant selectable marker. *Molecular and General Genetics*, *263*(3), 404–410. https://doi.org/10.1007/s004380051184

Bekker, A., Holland, H. D., Wang, P. L., Rumble, D., Stein, H. J., Hannah, J. L., … Beukes, N. J. (2004). Dating the rise of atmospheric oxygen. *Nature*, *427*(6970), 117–120. https://doi.org/10.1038/nature02260

Benoiston, A. S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 8–10. https://doi.org/10.1098/rstb.2016.0397

Bentley, F. K., Zurbriggen, A., & Melis, A. (2014). Heterologous expression of the mevalonic acid pathway in cyanobacteria enhances endogenous carbon partitioning to isoprene. *Molecular Plant*, *7*(1), 71–86. https://doi.org/10.1093/mp/sst134

Benveniste, P. (2004). Biosynthesis and Accumulation of Sterols. *Annual Review of Plant Biology*, *55*(1), 429–457. https://doi.org/10.1146/annurev.arplant.55.031903.141616

Blanchard, L., & Karst, F. (1993). Characterization of a lysine-to-glutamic acid mutation in a conservative sequence of farnesyl diphosphate synthase from Saccharomyces cerevisiae. *Gene*, *125*(2), 185–189. https://doi.org/10.1016/0378-1119(93)90326-X

Borowitzka, M. a. (2013). High-value products from microalgae-their development and commercialisation. *Journal of Applied Phycology*, *25*, 743–756. https://doi.org/10.1007/s10811-013-9983-9

Botella-Pavía, P., Besumbes, Ó., Phillips, M. A., Carretero-Paulet, L., Boronat, A., & Rodríguez-Concepción, M. (2004). Regulation of carotenoid biosynthesis in plants: Evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. *Plant Journal*, *40*(2), 188–199. https://doi.org/10.1111/j.1365-313X.2004.02198.x

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., … Grigoriev, I. V. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*(7219), 239–244. https://doi.org/10.1038/nature07410

Brown, S., Clastre, M., Courdavault, V., & O'Connor, S. E. (2015). De novo production of the plant-derived alkaloid strictosidine in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3205–3210. https://doi.org/10.1073/pnas.1423555112

Cahoon, A. B., & Timko, M. P. (2000). Yellow-in-the-dark mutants of Chlamydomonas lack the CHLL subunit of light-independent protochlorophyllide reductase. *Plant Cell*, *12*(4), 559–568. https://doi.org/10.1105/tpc.12.4.559

Campbell, A., Bauchart, P., Gold, N. D., Zhu, Y., De Luca, V., & Martin, V. J. J. (2016). Engineering of a Nepetalactol-Producing Platform Strain of Saccharomyces cerevisiae for the Production of Plant Seco-Iridoids. *ACS Synthetic Biology*, *5*(5), 405–414. https://doi.org/10.1021/acssynbio.5b00289

Cantos, C., Francisco, P., Trijatmiko, K. R., Slamet-Loedin, I., & Chadha-Mohanty, P. K. (2014). Identification of "safe harbor" loci in indica rice genome by harnessing the property of zinc-finger nucleases to induce DNA damage and repair. *Frontiers in Plant Science*, *5*(June), 302. https://doi.org/10.3389/fpls.2014.00302

Carlsen, S., Ajikumar, P. K., Formenti, L. R., Zhou, K., Phon, T. H., Nielsen, M. L., … Stephanopoulos, G. (2013). Heterologous expression and characterization of bacterial 2-C-methyl-d-erythritol-4-phosphate pathway in Saccharomyces cerevisiae. *Applied Microbiology and Biotechnology*, *97*(13), 5753–5769. https://doi.org/10.1007/s00253-013-4877-y

Carqueijeiro, I., Brown, S., Chung, K., Dang, T. T., Walia, M., Besseau, S., … Courdavault, V. (2018). Two tabersonine 6,7-epoxidases initiate lochnericine-derived alkaloid biosynthesis in catharanthus roseus. *Plant Physiology*, *177*(4), 1473–1486. https://doi.org/10.1104/pp.18.00549

Carrau, F. M., Medina, K., Boido, E., Farina, L., Gaggero, C., Dellacassa, E., … Henschke, P. A. (2005). De novo synthesis of monoterpenes by Saccharomyces cerevisiae wine yeasts. *FEMS Microbiology Letters*, *243*(1), 107–115. https://doi.org/10.1016/j.femsle.2004.11.050

Cerutti, H. (1997). Epigenetic Silencing of a Foreign Gene in Nuclear Transformants of Chlamydomonas. *The Plant Cell Online*, *9*(6), 925–945. https://doi.org/10.1105/tpc.9.6.925

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., … Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, *155*(7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001

Chen. (2010). Geraniol—a review of a commercially important fragrance material. *South African Journal of Botany Official Journal of the South African Association of Botanists* 76(4). https://doi.org/10.1016/j.sajb.2010.05.008

Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., … Sharp, P. A. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, *160*(6), 1246–1260. https://doi.org/10.1016/j.cell.2015.02.038

Cheng, R. Bin, Lin, X. Z., Wang, Z. K., Yang, S. J., Rong, H., & Ma, Y. (2011). Establishment of a transgene expression system for the marine microalga Schizochytrium by 18S rDNA-targeted homologous recombination. *World Journal of Microbiology and Biotechnology*, *27*(3), 737–741. https://doi.org/10.1007/s11274-010-0510-8

Choo, J. H., Han, C., Kim, J. Y., & Kang, H. A. (2014). Deletion of a KU80 homolog enhances homologous recombination in the thermotolerant yeast Kluyveromyces marxianus. *Biotechnology Letters*, *36*(10), 2059–2067. https://doi.org/10.1007/s10529-014-1576-4

Cicero, A. F. G., & Colletti, A. (2017). Effects of Carotenoids on Health: Are All the Same? Results from Clinical Trials. *Current Pharmaceutical Design*, *23*(February), 1–6. https://doi.org/10.2174/1381612823666170207

D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., … Matthijs, M. (2018). Engineering The Unicellular Alga Phaeodactylum tricornutum For High-Value Plant Triterpenoid Production. *Plant Biotechnology Journal*, 0–2. https://doi.org/10.1111/pbi.12948

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., … Duchateau, P. (2014). Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nature Communications*, *5*(May), 1–7. https://doi.org/10.1038/ncomms4831

Dai, J., Cui, X., Zhu, Z., & Hu, W. (2010). Non-homologous end joining plays a key role in transgene concatemer formation in transgenic zebrafish embryos. *International Journal of Biological Sciences*, *6*(7), 756–768. https://doi.org/10.7150/ijbs.6.756

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6640–6645. https://doi.org/10.1073/pnas.120163297

De Luca, V., Salim, V., Thamm, A., Masada, S. A., & Yu, F. (2014). Making iridoids/secoiridoids and monoterpenoid indole alkaloids: Progress on pathway elucidation. *Current Opinion in Plant Biology*, *19*, 35–42. https://doi.org/10.1016/j.pbi.2014.03.006

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., & Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research*, *37*(14). https://doi.org/10.1093/nar/gkp448

Demogines, A., East, A. M., Lee, J. H., Grossman, S. R., Sabeti, P. C., Paull, T. T., & Sawyer, S. L. (2010). Ancient and recent adaptive evolution of primate non-homologous end joining genes. *PLoS Genetics*, *6*(10), 1–12. https://doi.org/10.1371/journal.pgen.1001169

Deriano, L., & Roth, D. B. (2013). Modernizing the Nonhomologous End-Joining Repertoire: Alternative and Classical NHEJ Share the Stage. *Annual Review of Genetics*, *47*(1), 433–455. https://doi.org/10.1146/annurev-genet-110711-155540

Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E., & Weyman, P. D. (2016). Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods. *Frontiers in Bioengineering and Biotechnology*, *4*(August). https://doi.org/10.3389/fbioe.2016.00065

Dubini, A., Mus, F., Seibert, M., Grossman, A. R., & Posewitz, M. C. (2009). Flexibility in anaerobic metabolism as revealed in a mutant of chlamydomonas reinhardtii lacking hydrogenase activity. *Journal of Biological Chemistry*, *284*(11), 7201–7213. https://doi.org/10.1074/jbc.M803917200

Dudareva, N., Negre, F., Nagegowda, D. A., & Orlova, I. (2006). Plant volatiles: Recent advances and future perspectives. *Critical Reviews in Plant Sciences*, *25*(5), 417–440. https://doi.org/10.1080/07352680600899973

Dufourc, E. J. (2008). Sterols and membrane dynamics. *Journal of Chemical Biology*, *1*(1–4), 63–77. https://doi.org/10.1007/s12154-008-0010-6

Eaton-Rye, J. J. (2011). Construction of Gene Interruptions and Gene Deletions in the Cyanobacterium Synechocystis sp. Strain PCC 6803. *Photosynthesis Research Protocols*, *684*, 363–374. https://doi.org/10.1007/978-1-60761-925-3

Enfissi, E. M. A., Fraser, P. D., Lois, L. M., Boronat, A., Schuch, W., & Bramley, P. M. (2005). Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato. *Plant Biotechnology Journal*, *3*(1), 17–27. https://doi.org/10.1111/j.1467-7652.2004.00091.x

Evy, A. V. A. L. (1999). Stimulation of homologous recombination in plants by expression of the bacterial resolvase RuvC, *96*(June), 7398–7402.

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., … Ralph, P. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.9b00455

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., … Goossens, A. (2014). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *The New Phytologist*, 521–535. https://doi.org/10.1111/nph.12917

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., & Baart, G. J. E. (2012). The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant Journal*, *70*(6), 1004–1014. https://doi.org/10.1111/j.1365-313X.2012.04941.x

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, *281*(5374), 200–206. https://doi.org/10.1126/science.281.5374.200

Ferenczi, A., Pyott, D. E., Xipnitou, A., & Molnar, A. (2017). Efficient targeted DNA editing and replacement in *Chlamydomonas reinhardtii* using Cpf1 ribonucleoproteins and single-stranded DNA. *Proceedings of the National Academy of Sciences*, *114*(51), 201710597. https://doi.org/10.1073/pnas.1710597114

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, *281*(5374),

237–240. https://doi.org/10.1126/science.281.5374.237

Fischer, M. J. C., Meyer, S., Claudel, P., Bergdoll, M., & Karst, F. (2011). Metabolic engineering of monoterpene synthesis in yeast. *Biotechnology and Bioengineering*, *108*(8), 1883–1892. https://doi.org/10.1002/bit.23129

Garg, S. G., & Gould, S. B. (2016). The Role of Charge in Protein Targeting Evolution. *Trends in Cell Biology*, *26*(12), 894–905. https://doi.org/10.1016/j.tcb.2016.07.001

Gille, A., Trautmann, A., Posten, C., & Briviba, K. (2016). Bioaccessibility of carotenoids from Chlorella vulgaris and Chlamydomonas reinhardtii. *International Journal of Food Sciences and Nutrition*, *67*(5), 507–513. https://doi.org/10.1080/09637486.2016.1181158

Gold, D. A., Caron, A., Fournier, G. P., & Summons, R. E. (2017). Paleoproterozoic sterol biosynthesis and the rise of oxygen. *Nature*, *543*(7645), 420–423. https://doi.org/10.1038/nature21412

Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., & Hegemann, P. (2017). Targeting of Photoreceptor Genes in Chlamydomonas reinhardtii via Zinc- finger Nucleases and CRISPR/Cas9. *Plant Cell Advance Publication. Published on October*, *4*. https://doi.org/10.1105/tpc.17.00659

Grossman, A. R., Croft, M., Gladyshev, V. N., Merchant, S. S., Posewitz, M. C., Prochnik, S., & Spalding, M. H. (2007). Novel metabolism in Chlamydomonas through the lens of genomics. *Current Opinion in Plant Biology*, *10*(2), 190–198. https://doi.org/10.1016/j.pbi.2007.01.012

Guri Giaever1, Angela M. Chu2, LiNi3, Carla Connelly4, Linda Riles5, Steeve Ve´ronneau6, Sally Dow7, Ankuta Lucau-Danila8, Keith Anderson1, Bruno Andre´9, Adam P. Arkin10, Anna Astromoff2, Mohamed El Bakkoury11, Rhonda Bangham3, Rocio Benito12, Sophie Bra, 2 & Mark Johnston. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, (418), 387–391.

Hallmann, A., & Hallman, a. (2007). Algal transgenics and biotechnology. *Transgenic Plant J*, *1*(1), 81–98. https://doi.org/10.3390/ijms12010633

Han, M., Heppel, S. C., Su, T., Bogs, J., Zu, Y., An, Z., & Rausch, T. (2013). Enzyme Inhibitor Studies Reveal Complex Control of Methyl-D-Erythritol 4-Phosphate (MEP) Pathway Enzyme Expression in Catharanthus roseus. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0062467

Hebert Jair Barrales-Cureño, César Reyes Reyes, Irma Vásquez García, Luis Germán López Valdez, Adrián Gómez De Jesús, Juan Antonio Cortés Ruíz, Leticia Mónica Sánchez Herrera, María Carmina Calderón Caballero, Jesús Antonio Salazar Magallón, J. E. P. and J. M. M. (2012). Alkaloids of Pharmacological Importance in Catharanthus roseus. *Intech*, 13. https://doi.org/10.1016/j.colsurfa.2011.12.014

Heinemann, J. a et al. (1989). © 198 9 Nature Publishing Group. *Nature*, *342*, 189–192. https://doi.org/10.1038/340301a0

Hong, S. G., Yada, R. C., Choi, K., Carpentier, A., Liang, T. J., Merling, R. K., … Dunbar, C. E. (2017). Rhesus iPSC Safe Harbor Gene-Editing Platform for Stable Expression of Transgenes in Differentiated Cells of All Germ Layers. *Molecular Therapy*, *25*(1), 44–53. https://doi.org/10.1016/j.ymthe.2016.10.007

Huang, W., & Daboussi, F. (2017). Genetic and metabolic engineering in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160411. https://doi.org/10.1098/rstb.2016.0411

Jeon, S., Lim, J.-M., Lee, H.-G., Shin, S.-E., Kang, N. K., Park, Y.-I., … Chang, Y. K. (2017). Current status and perspectives of genome editing technology for microalgae. *Biotechnology for Biofuels*, *10*(1), 267. https://doi.org/10.1186/s13068-017-0957-z

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., … Yuan, Y. J. (2017). Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metabolic Engineering*, *41*(March), 57–66. https://doi.org/10.1016/j.ymben.2017.03.005

Jinkerson, R. E., & Jonikas, M. C. (2015). Molecular techniques to interrogate and edit the Chlamydomonas nuclear genome, 393–412. https://doi.org/10.1111/tpj.12801

Jullesson, D., David, F., Pfleger, B. and Nielsen, J. (2015). Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances*, *33*(7). http://dx.doi.org/10.1016/j.biotechadv.2015.02.011

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., … Ecker, J. R. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics*, *15*(1), 1–25. https://doi.org/10.1371/journal.pgen.1007819

Karas, B. J., Diner, R. E., Lefebvre, S. C., McQuaid, J., Phillips, A. P. R., Noddings, C. M., … Weyman, P. D. (2015). Designer diatom episomes delivered by bacterial conjugation. *Nature Communications*, *6*, 6925. https://doi.org/10.1038/ncomms7925

Katz, M. E., Finkel, Z. V., Grzebyk, D., Knoll, A. H., & Falkowski, P. G. (2004). Evolutionary trajectories and biogeochemical impacts of marine eukaryotic phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 523–556. https://doi.org/10.1146/annurev.ecolsys.35.112202.130137

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, *12*(6). https://doi.org/10.1371/journal.pbio.1001889

Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., … Weiss, R. (2014). CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nature Methods*, *11*(7), 723–726. https://doi.org/10.1038/nmeth.2969

Kilian, O., Benemann, C. S. E., Niyogi, K. K., & Vick, B. (2011). High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. *Proceedings of the National Academy of Sciences*, *108*(52), 21265–21269. https://doi.org/10.1073/pnas.1105861108

Kim, E. J., Ma, X., & Cerutti, H. (2015). Gene silencing in microalgae: Mechanisms and biological roles. *Bioresource Technology*, *184*, 23–32. https://doi.org/10.1016/j.biortech.2014.10.119

Kim, S. W., & Keasling, J. D. (2001). Metabolic engineering of the nonmevalonate isopentenyl diphosphate synthesis pathway in Escherichia coli enhances lycopene production. *Biotechnology and Bioengineering*, *72*(4), 408–415. https://doi.org/10.1002/1097-0290(20000220)72:4<408::AID-BIT1003>3.0.CO;2-H

Kindle, K. L., Schnell, R. A., Fernandez, E., & Lefebvre, P. A. (1989). Stable nuclear transformation of Chlamydomonas using the Chlamydomonas gene for nitrate reductase. *Journal of Cell Biology*, *109*(6 I), 2589–2601. https://doi.org/10.1083/jcb.109.6.2589

Kleessen, S., Irgang, S., Klie, S., Giavalisco, P., & Nikoloski, Z. (2015). Integration of

transcriptomics and metabolomics data specifies the metabolic response of Chlamydomonas to rapamycin treatment. *Plant Journal*, *81*(5), 822–835. https://doi.org/10.1111/tpj.12763

Kohli, A., González-Melendi, P., Abranches, R., Capell, T., Stoger, E., & Christou, P. (2006). The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signaling and Behavior*, *1*(4), 185–195. https://doi.org/10.4161/psb.1.4.3195

Kohli, A., Miro, B., & Twyman, R. M. (2010). *Transgene Integration , Expression and Stability in Plants : Strategies for Improvements*. https://doi.org/10.1007/978-3-642-04809-8

Krejci, L., Altmannova, V., Spirek, M., & Zhao, X. (2012). Homologous recombination and its regulation. *Nucleic Acids Research*, *40*(13), 5795–5818. https://doi.org/10.1093/nar/gks270

Kroth, P. G., Chiovitti, A., Gruber, A., Martin-Jezeque, V., Mock, T., Parker, M. S., … Bowler, C. (2008). A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS ONE*, *3*(1). https://doi.org/10.1371/journal.pone.0001426

Kuczynska, P., Jemiola-Rzeminska, M., & Strzalka, K. (2015). Photosynthetic pigments in diatoms. *Marine Drugs*, *13*(9), 5847–5881. https://doi.org/10.3390/md13095847

Kudoh, K., Kawano, Y., Hotta, S., Sekine, M., Watanabe, T., & Ihara, M. (2014). Prerequisite for highly efficient isoprenoid production by cyanobacteria discovered through the over-expression of 1-deoxy-D-xylulose 5-phosphate synthase and carbon allocation analysis. *Journal of Bioscience and Bioengineering*, *118*(1), 20–28. https://doi.org/10.1016/j.jbiosc.2013.12.018

Lan, E. I., Chuang, D. S., Shen, C. R., Lee, A. M., Ro, S. Y., & Liao, J. C. (2015). Metabolic engineering of cyanobacteria for photosynthetic 3-hydroxypropionic acid production from CO2 using Synechococcus elongatus PCC 7942. *Metabolic Engineering*, *31*, 163–170. https://doi.org/10.1016/j.ymben.2015.08.002

Lauersen, K. J., Baier, T., Wichmann, J., Wördenweber, R., Mussgnug, J. H., Hübner, W., … Kruse, O. (2016). Efficient phototrophic production of a high-value sesquiterpenoid from the eukaryotic microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *38*, 331–343. https://doi.org/10.1016/j.ymben.2016.07.013

Lauersen, K. J., Kruse, O., & Mussgnug, J. H. (2015). Targeted expression of nuclear transgenes in Chlamydomonas reinhardtii with a versatile, modular vector toolkit. *Applied Microbiology and Biotechnology*, *99*(8), 3491–3503. https://doi.org/10.1007/s00253-014-6354-7

Lauersen, K. J., Wichmann, J., Baier, T., Kampranis, S. C., Pateraki, I., Møller, B. L., & Kruse, O. (2018). Phototrophic production of heterologous diterpenoids and a hydroxy-functionalized derivative from Chlamydomonas reinhardtii. *Metabolic Engineering*, *49*. https://doi.org/10.1016/j.ymben.2018.07.005

Lee, S. H., Kim, S., & Hur, and J. K. (2018). CRISPR and Target-Specific DNA Endonucleases for Efficient DNA Knock-in in Eukaryotic Genomes. *Mol. Cells*, *41*(11), 943–952. https://doi.org/10.14348/MOLCELLS.2018.0408

León-Bañares, R., González-Ballester, D., Galván, A., & Fernández, E. (2004). Transgenic microalgae as green cell-factories. *Trends in Biotechnology*, *22*(1), 45–52. https://doi.org/10.1016/j.tibtech.2003.11.003

Li, H., Shen, C. R., Huang, C.-H., Sung, L.-Y., Wu, M.-Y., & Hu, Y.-C. (2016). CRISPR-Cas9

for the Genome Engineering of Cyanobacteria and Succinate Production. *Metabolic Engineering*, *38*(August), 293–302. https://doi.org/10.1016/j.ymben.2016.09.006

Li, J. X., Vaidya, M., White, C., Vainstein, A., Citovsky, V., & Tzfira, T. (2005). Involvement of KU80 in T-DNA integration in plant cells. *Proc Natl Acad Sci U S A*, *102*(52), 19231–19236. https://doi.org/10.1073/pnas.0506437103

Li Jianming, & Timko, M. P. (1996). The pc-1 phenotype of Chlamydomonas reinhardtii results from a deletion mutation in the nuclear gene for NADPH:protochlorophyllide oxidoreductase. *Plant Molecular Biology*, *30*(1), 15–37. https://doi.org/10.1007/bf00017800

Li, Z., & Bock, R. (2018). Replication of bacterial plasmids in the nucleus of the red alga Porphyridium purpureum. *Nature Communications*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-05651-1

Liu. (2018). Genome-scale sequence disruption following biolistic transformation in rice and maize.

Liu, Jidong, Zhang, W., Du, G., Chen, J., & Zhou, J. (2013). Overproduction of geraniol by enhanced precursor supply in Saccharomyces cerevisiae. *Journal of Biotechnology*, *168*(4), 446–451. https://doi.org/10.1016/j.jbiotec.2013.10.017

Liu, Jin, Gerken, H., Huang, J., & Chen, F. (2013). Engineering of an endogenous phytoene desaturase gene as a dominant selectable marker for Chlamydomonas reinhardtii transformation and enhanced biosynthesis of carotenoids. *Process Biochemistry*, *48*(5–6), 788–795. https://doi.org/10.1016/j.procbio.2013.04.020

Liu, W., Xu, X., Zhang, R., Cheng, T., Cao, Y., Li, X., … Xian, M. (2016). Engineering Escherichia coli for high-yield geraniol production with biotransformation of geranyl acetate to geraniol under fed-batch culture. *Biotechnology for Biofuels*, *9*(1), 1–8. https://doi.org/10.1186/s13068-016-0466-5

Lohr, M., Schwender, J., & Polle, J. E. W. (2012). Isoprenoid biosynthesis in eukaryotic phototrophs: A spotlight on algae. *Plant Science*, *185–186*, 9–22. https://doi.org/10.1016/j.plantsci.2011.07.018

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

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

Maheswari, U., Mock, T., Armbrust, E. V., & Bowler, C. (2009). Update of the Diatom EST Database: A new tool for digital transcriptomics. *Nucleic Acids Research*, *37*(SUPPL. 1), 1001–1005. https://doi.org/10.1093/nar/gkn905

Materna, A. C., Sturm, S., Kroth, P. G., & Lavaud, J. (2009). First induced plastid genome mutations in an alga with secondary plastids: psbA mutations in the diatom phaeodactylum tricornutum (bacillariophyceae) reveal consequences on the regulation of photosynthesis 1. *Journal of Phycology*, *45*(4), 838–846. https://doi.org/10.1111/j.1529-8817.2009.00711.x

Matsuura, K., Lefebvre, P. A., Kamiya, R., & Hirono, M. (2004). Bld10p, a novel protein essential for basal body assembly in Chlamydomonas: Localization to the cartwheel, the first ninefold symmetrical structure appearing during assembly. *Journal of Cell*

Biology, *165*(5), 663–671. https://doi.org/10.1083/jcb.200402022

Maury, J., Asadollahi, M. A., Møller, K., Schalk, M., Clark, A., Formenti, L. R., & Nielsen, J. (2008). Reconstruction of a bacterial isoprenoid biosynthetic pathway in Saccharomyces cerevisiae. *FEBS Letters*, *582*(29), 4032–4038. https://doi.org/10.1016/j.febslet.2008.10.045

Meadows, A. L., Hawkins, K. M., Tsegaye, Y., Antipov, E., Kim, Y., Raetz, L., … Tsong, A. E. (2016). Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature*, *537*(7622), 694–697. https://doi.org/10.1038/nature19769

Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., … Werck-Reichhart, D. (2014). The seco-iridoid pathway from Catharanthus roseus. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4606

Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T., & Tanaka, K. (2004). Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, *Cyanidioschyzon merolae* 10D. *Plant and Cell Physiology*, *45*(6), 667–671. https://doi.org/10.1093/pcp/pch087

Moosburner, M. A., Gholami, P., McCarthy, J. K., Tan, M., Bielinski, V. A., & Allen, A. E. (2020). Multiplexed Knockouts in the Model Diatom Phaeodactylum by Episomal Delivery of a Selectable Cas9. *Frontiers in Microbiology*, *11*(January), 1–13. https://doi.org/10.3389/fmicb.2020.00005

Morris, W. L., Ducreux, L. J. M., Hedden, P., Millam, S., & Taylor, M. A. (2006). Overexpression of a bacterial 1-deoxy-D-xylulose 5-phosphate synthase gene in potato tubers perturbs the isoprenoid metabolic network: Implications for the control of the tuber life cycle. *Journal of Experimental Botany*, *57*(12), 3007–3018. https://doi.org/10.1093/jxb/erl061

Nadia Sharif, Neelma Munir, Shagufta Naz, Rehana Iqbal, W. R. (n.d.). *Origin of Algae and Their Plastids*.

Naduthodi, M. I. S., Barbosa, M. J., & van der Oost, J. (2018). Progress of CRISPR-Cas based genome editing in Photosynthetic microbes. *Biotechnology Journal*, 1700591. https://doi.org/10.1002/biot.201700591

Nagegowda, D. A., & Gupta, P. (2020). Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant Science*, *294*(February), 110457. https://doi.org/10.1016/j.plantsci.2020.110457

Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A., & Quéguiner, B. (1995). Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, *9*(3), 359–372. https://doi.org/10.1029/95GB01070

Oswald, M., Fischer, M., Dirninger, N., & Karst, F. (2007). Monoterpenoid biosynthesis in Saccharomyces cerevisiae. *FEMS Yeast Research*, *7*(3), 413–421. https://doi.org/10.1111/j.1567-1364.2006.00172.x

Oudin, A., Courtois, M., Rideau, M., & Clastre, M. (2007). The iridoid pathway in Catharanthus roseus alkaloid biosynthesis. *Phytochemistry Reviews*, *6*(2–3), 259–276. https://doi.org/10.1007/s11101-006-9054-9

Pan, Q., Mustafa, N. R., Tang, K., Choi, Y. H., & Verpoorte, R. (2016). Monoterpenoid indole alkaloids biosynthesis and its regulation in Catharanthus roseus: a literature review from genes to metabolites. *Phytochemistry Reviews*, *15*(2), 221–250. https://doi.org/10.1007/s11101-015-9406-4

Papapetrou, E. P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L. M. S., … Sadelain, M. (2011). Genomic safe harbors permit high ??-globin transgene expression in thalassemia induced pluripotent stem cells. *Nature Biotechnology*, *29*(1), 73–81. https://doi.org/10.1038/nbt.1717

Park, S. Y., Vaghchhipawala, Z., Vasudevan, B., Lee, L. Y., Shen, Y., Singer, K., … Gelvin, S. B. (2015). Agrobacterium T-DNA integration into the plant genome can occur without the activity of key non-homologous end-joining proteins. *Plant Journal*, *81*(6), 934–946. https://doi.org/10.1111/tpj.12779

Partow, S., Siewers, V., Daviet, L., Schalk, M., & Nielsen, J. (2012). Reconstruction and Evaluation of the Synthetic Bacterial MEP Pathway in Saccharomyces cerevisiae. *PLoS ONE*, *7*(12), 1–12. https://doi.org/10.1371/journal.pone.0052498

Pazour, G. J., Agrin, N., Walker, B. L., & Witman, G. B. (2006). Identification of predicted human outer dynein arm genes: Candidates for primary ciliary dyskinesia genes. *Journal of Medical Genetics*, *43*(1), 62–73. https://doi.org/10.1136/jmg.2005.033001

Pflueger, C., Tan, D., Swain, T., Nguyen, T., Pflueger, J., Nefzger, C., … Lister, R. (2018). A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Research*, *28*(8), 1193–1206. https://doi.org/10.1101/gr.233049.117

Pichersky, E., & Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytologist*, *220*(3), 692–702. https://doi.org/10.1111/nph.14178

Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., … Tamagnini, P. (2015). Improving a Synechocystis-based photoautotrophic chassis through systematic genome mapping and validation of neutral sites. *DNA Research*, *22*(6), 425–437. https://doi.org/10.1093/dnares/dsv024

Pollak, B., Matute, T., Nunez, I., Cerda, A., Lopez, C., Kan, A., … Roscoff, S. B. De. (2019). Universal Loop assembly (uLoop): open, efficient, and species- agnostic DNA fabrication.

Qian, S., Clomburg, J. M., & Gonzalez, R. (2019). Engineering Escherichia coli as a platform for the in vivo synthesis of prenylated aromatics. *Biotechnology and Bioengineering*, *116*(5), 1116–1127. https://doi.org/10.1002/bit.26932

R. F. CHANDLER, 2 S. N. HOOPER, 2 AND M. J. HARVEY. (1982). Ethnobotany and Phytochemistry of Yarrow, Achillea millefolium, Compositae. *New York*, *36*(September 1981), 203–223.

Radakovits, R., Jinkerson, R. E., Darzins, A., & Posewitz, M. C. (2010). Genetic engineering of algae for enhanced biofuel production. *Eukaryotic Cell*, *9*(4), 486–501. https://doi.org/10.1128/EC.00364-09

Rasala, B. A., Barrera, D. J., Ng, J., Plucinak, T. M., Rosenberg, J. N., Weeks, D. P., … Mayfield, S. P. (2013). Expanding the spectral palette of fluorescent proteins for the green microalga Chlamydomonas reinhardtii. *Plant Journal*, *74*(4), 545–556. https://doi.org/10.1111/tpj.12165

Remacle, C., Cardol, P., Coosemans, N., Gaisne, M., & Bonnefoy, N. (2006). High-efficiency biolistic transformation of Chlamydomonas mitochondria can be used to insert mutations in complex I genes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4771–4776. https://doi.org/10.1073/pnas.0509501103

Remmers, I. M., D'Adamo, S., Martens, D. E., de Vos, R. C. H., Mumm, R., America, A. H.

P., … Lamers, P. P. (2018). Orchestration of transcriptome, proteome and metabolome in the diatom Phaeodactylum tricornutum during nitrogen limitation. *Algal Research*, *35*(August), 33–49. https://doi.org/10.1016/j.algal.2018.08.012

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M. A., … Gunawardane, R. N. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular Biology of the Cell*, *28*(21), 2854–2874. https://doi.org/10.1091/mbc.E17-03-0209

Rohr, J., Sarkar, N., Balenger, S., Jeong, B. R., & Cerutti, H. (2004). Tandem inverted repeat system for selection of effective transgenic RNAi strains in Chlamydomonas. *Plant Journal*, *40*(4), 611–621. https://doi.org/10.1111/j.1365-313X.2004.02227.x

Rosales-Mendoza, S., Paz-Maldonado, L. M. T., & Soria-Guerra, R. E. (2012). Chlamydomonas reinhardtii as a viable platform for the production of recombinant proteins: Current status and perspectives. *Plant Cell Reports*, *31*(3), 479–494. https://doi.org/10.1007/s00299-011-1186-8

S. Merchant, Simon E. Prochnik, Olivier Vallon, Elizabeth H. Harris, Steven A. Sanderfoot, Martin H. Spalding, Vladimir V. Kapitonov, Qinghu Ren, Patrick Laurence Maréchal-Drouard, Wallace F. Marshall, Liang-Hu Qu, David R. Nels, A. (2007). The Chlamydomonas Genome Reveals the Evolution of Key. *National Institutes of Health*, *318*(5848), 245–250.

Sachindra, N. M., Sato, E., Maeda, H., Hosokawa, M., Niwano, Y., Kohno, M., & Miyashita, K. (2007). Radical scavenging and singlet oxygen quenching activity of marine carotenoid fucoxanthin and its metabolites. *Journal of Agricultural and Food Chemistry*, *55*(21), 8516–8522. https://doi.org/10.1021/jf071848a

Sadelain, M., Papapetrou, E. P., & Bushman, F. D. (2012). Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer*, *12*(1), 51–58. https://doi.org/10.1038/nrc3179

Saika, H., Nishizawa-Yokoi, A., & Toki, S. (2014). The non-homologous end-joining pathway is involved in stable transformation in rice. *Frontiers in Plant Science*, *5*(October), 560. https://doi.org/10.3389/fpls.2014.00560

San Filippo, J., Sung, P., & Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. *Annual Review of Biochemistry*, *77*(1), 229–257. https://doi.org/10.1146/annurev.biochem.77.061306.125255

Scaife, M. a., & Smith, a. G. (2016). Towards developing algal synthetic biology. *Biochemical Society Transactions*, *44*, 716–722. https://doi.org/10.1042/BST20160061

Schwab, W., Fischer, T., & Wüst, M. (2015). Terpene glucoside production: Improved biocatalytic processes using glycosyltransferases. *Engineering in Life Sciences*, *15*(4), 376–386. https://doi.org/10.1002/elsc.201400156

Scranton, M. A., Ostrand, J. T., Fields, F. J., & Mayfield, S. P. (2015). Chlamydomonas as a model for biofuels and bio-products production. *Plant Journal*, *82*(3), 523–531. https://doi.org/10.1111/tpj.12780

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., & Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-06378-9

Shah, A. A., Wang, C., Chung, Y. R., Kim, J. Y., Choi, E. S., & Kim, S. W. (2013). Enhancement of geraniol resistance of Escherichia coli by MarA overexpression.

Journal of Bioscience and Bioengineering, *115*(3), 253–258. https://doi.org/10.1016/j.jbiosc.2012.10.009

Shahar, N., Landman, S., Weiner, I., Elman, T., Dafni, E., Feldman, Y., … Yacoby, I. (2020). The Integration of Multiple Nuclear-Encoded Transgenes in the Green Alga Chlamydomonas reinhardtii Results in Higher Transcription Levels. *Frontiers in Plant Science*, *10*(February), 1–9. https://doi.org/10.3389/fpls.2019.01784

Shalem. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells, *343*(January), 84–88.

Sharma, A. K., Nymark, M., Sparstad, T., Bones, A. M., & Winge, P. (2018). Transgene-free genome editing in marine algae by bacterial conjugation – comparison with biolistic CRISPR/Cas9 transformation. *Scientific Reports*, *8*(1), 14401. https://doi.org/10.1038/s41598-018-32342-0

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Jeong, B. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. *Scientific Reports*, *6*(April), 27810. https://doi.org/10.1038/srep27810

Shuhei Kawamura, Hang Chu, Jakob Felding, and P. S. B. (2016). Nineteen-Step Total Synthesis of (+)-Phorbol. *Nature*, *532*(7597), 90–93. https://doi.org/10.1038/nature17153

Sikorski, R. S., Michaud, W., Levin, H. L., Boeke, J. D., & Hieter, P. (1990). Trans-kingdom promiscuity [5]. *Nature*, *345*(6276), 581–582. https://doi.org/10.1038/345581b0

Simkin, A. J., Miettinen, K., Claudel, P., Burlat, V., Guirimand, G., Courdavault, V., … Clastre, M. (2013). Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. *Phytochemistry*, *85*, 36–43. https://doi.org/10.1016/j.phytochem.2012.09.014

Sims, P. A., Mann, D. G., & Medlin, L. K. (2006). Evolution of the diatoms: Insights from fossil, biological and molecular data. *Phycologia*, *45*(4), 361–402. https://doi.org/10.2216/05-22.1

Smith, S. R., Dupont, C. L., McCarthy, J. K., Broddrick, J. T., Oborník, M., Horák, A., … Allen, A. E. (2019). Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12407-y

Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H., … Moritz, T. (2016). Transcriptional Orchestration of the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron Limitation. *PLOS Genetics*, *12*(12), e1006490. https://doi.org/10.1371/journal.pgen.1006490

Spicer, A., & Molnar, A. (2018). Gene Editing of Microalgae: Scientific Progress and Regulatory Challenges in Europe. *Biology*, *7*(1), 21. https://doi.org/10.3390/biology7010021

Steinert, J., Schiml, S., & Puchta, H. (2016). Homology-based double-strand break-induced genome engineering in plants. *Plant Cell Reports*, *35*(7), 1429–1438. https://doi.org/10.1007/s00299-016-1981-3

Sternberg, S. H., & Doudna, J. A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, *58*(4), 568–574. https://doi.org/10.1016/j.molcel.2015.02.032

Taparia, Y., Zarka, A., Leu, S., Zarivach, R., Boussiba, S., & Khozin-Goldberg, I. (2019). A novel endogenous selection marker for the diatom Phaeodactylum tricornutum based

on a unique mutation in phytoene desaturase 1. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-44710-5

Thabet, I., Grégory Guirimand, G., Guihur, A., Lanoue, A., Courdavault, V., Papon, N., … Clastre, M. (2012). Characterization and subcellular localization of geranylgeranyl diphosphate synthase from Catharanthus roseus. *Molecular Biology Reports*, *39*(3), 3235–3243. https://doi.org/10.10.1007/s11033-011-1091-9

Thakore, P. I., Song, L., Safi, A., Shivakumar, K., Kabadi, A. M., Reddy, T. E., … Gersbach, C. A. (2015). Highly Specific Epigenome Editing by CRISPR/Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nature Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630.Highly

Tuggle, C. K., & Waters, W. R. (2015). Tuberculosis-resistant transgenic cattle. *Proceedings of the National Academy of Sciences*, *112*(13), 3854–3855. https://doi.org/10.1073/pnas.1502972112

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G. J. E., Rombauts, S., Hasnain, G., … Goossens, A. (2013). CathaCyc, a metabolic pathway database built from catharanthus roseus RNA-seq data. *Plant and Cell Physiology*, *54*(5), 673–685. https://doi.org/10.1093/pcp/pct039

Vavitsas, K., Fabris, M., & Vickers, C. E. (2018). Terpenoid Metabolic Engineering in, (Figure 1). https://doi.org/10.3390/genes9110520

Veau, B., Courtois, M., Oudin, A., Chénieux, J. C., Rideau, M., & Clastre, M. (2000). Cloning and expression of cDNAs encoding two enzymes of the MEP pathway in Catharanthus roseus. *Biochimica et Biophysica Acta - Gene Structure and Expression*, *1517*(1), 159–163. https://doi.org/10.1016/S0167-4781(00)00240-2

Vickers, C. E., Bongers, M., Liu, Q., Delatte, T., & Bouwmeester, H. (2014). Metabolic engineering of volatile isoprenoids in plants and microbes. *Plant, Cell and Environment*, *37*(8), 1753–1775. https://doi.org/10.1111/pce.12316

Vranová, E., Coman, D., & Gruissem, W. (2013). Network Analysis of the MVA and MEP Pathways for Isoprenoid Synthesis. *Annual Review of Plant Biology*, *64*(1), 665–700. https://doi.org/10.1146/annurev-arplant-050312-120116

Walter, M. H., Fester, T., & Strack, D. (2000). Arbuscular mycorrhizal fungi induce the non-mevalonate methylerythritol phosphate pathway of isoprenoid biosynthesis correlated with accumulation of the “yellow pigment” and other apocarotenoids. *Plant Journal*, *21*(6), 571–578. https://doi.org/10.1046/j.1365-313X.2000.00708.x

Wang, C., Kim, J.-H., & Kim, S.-W. (2014). Synthetic Biology and Metabolic Engineering for Marine Carotenoids: New Opportunities and Future Prospects. *Marine Drugs*, *12*(9), 4810–4832. https://doi.org/10.3390/md12094810

Wang, C. T., Liu, H., Gao, X. S., & Zhang, H. X. (2010). Overexpression of G10H and ORCA3 in the hairy roots of Catharanthus roseus improves catharanthine production. *Plant Cell Reports*, *29*(8), 887–894. https://doi.org/10.1007/s00299-010-0874-0

Ward, V. C. A., Chatzivasileiou, A. O., & Stephanopoulos, G. (2018). Metabolic engineering of Escherichia coli for the production of isoprenoids. *FEMS Microbiology Letters*, *365*(10), 1–9. https://doi.org/10.1093/femsle/fny079

Waters, V. L. (2001). Conjugation between bacterial and mammalian cells. *Nature Genetics*, *29*(4), 375–376. https://doi.org/10.1038/ng779

Werner, S., Engler, C., Weber, E., Gruetzner, R., & Marillonnet, S. (2012). Fast track assembly of multigene constructs using golden gate cloning and the MoClo system.

Bioengineered Bugs, *3*(1), 38–43. https://doi.org/10.4161/bbug.3.1.18223

Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., Mccarthy, J. K., Heuberger, A. L., … Dupont, C. L. (2015). Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnology Journal*, *13*(4), 460–470. https://doi.org/10.1111/pbi.12254

White, R. A., Wolfe, G. R., Komine, Y., & Hoober, J. K. (1996). Localization of light-harvesting complex apoproteins in the chloroplast and cytoplasm during greening of Chlamydomonas reinhardtii at 38 °C. *Photosynthesis Research*, *47*(3), 267–280. https://doi.org/10.1007/BF02184287

Wichmann, J., Baier, T., Wentnagel, E., Lauersen, K. J., & Kruse, O. (2018). Tailored carbon partitioning for phototrophic production of (E)-α-bisabolene from the green microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *45*(October 2017), 211–222. https://doi.org/10.1016/j.ymben.2017.12.010

Yamamoto, K., Takahashi, K., Mizuno, H., Anegawa, A., Ishizaki, K., Fukaki, H., … Mimura, T. (2016). Cell-specific localization of alkaloids in Catharanthus roseus stem tissue measured with Imaging MS and Single-cell MS. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(14), 3891–3896. https://doi.org/10.1073/pnas.1521959113

Yang, C., Gao, X., Jiang, Y., Sun, B., Gao, F., & Yang, S. (2016). Synergy between methylerythritol phosphate pathway and mevalonate pathway for isoprene production in Escherichia coli. *Metabolic Engineering*, *37*, 79–91. https://doi.org/10.1016/j.ymben.2016.05.003

Yang, J., Xian, M., Su, S., Zhao, G., Nie, Q., Jiang, X., … Liu, W. (2012). Enhancing production of bio-isoprene using hybrid MVA pathway and isoprene synthase in E. coli. *PLoS ONE*, *7*(4), 1–8. https://doi.org/10.1371/journal.pone.0033509

Yool, A., & Tyrrell, T. (2003). Role of diatoms in regulating the ocean's silicon cycle. *Global Biogeochemical Cycles*, *17*(4), n/a-n/a. https://doi.org/10.1029/2002gb002018

Zebec, Z., Wilkes, J., Jervis, A. J., Scrutton, N. S., Takano, E., & Breitling, R. (2016). Towards synthesis of monoterpenes and derivatives using synthetic biology. *Current Opinion in Chemical Biology*, *34*, 37–43. https://doi.org/10.1016/j.cbpa.2016.06.002

Zelensky, A. N., Schimmel, J., Kool, H., Kanaar, R., & Tijsterman, M. (2017). Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/s41467-017-00124-3

Zhao, J., Bao, X., Li, C., Shen, Y., & Hou, J. (2016). Improving monoterpene geraniol production through geranyl diphosphate synthesis regulation in Saccharomyces cerevisiae. *Applied Microbiology and Biotechnology*, *100*(10), 4561–4571. https://doi.org/10.1007/s00253-016-7375-1

Zhou, J., Wang, C., Yang, L., Choi, E. S., & Kim, S. W. (2015). Geranyl diphosphate synthase: An important regulation point in balancing a recombinant monoterpene pathway in Escherichia coli. *Enzyme and Microbial Technology*, *68*, 50–55. https://doi.org/10.1016/j.enzmictec.2014.10.005

Zhou, J., Wang, C., Yoon, S. H., Jang, H. J., Choi, E. S., & Kim, S. W. (2014). Engineering Escherichia coli for selective geraniol production with minimized endogenous dehydrogenation. *Journal of Biotechnology*, *169*(1), 42–50. https://doi.org/10.1016/j.jbiotec.2013.11.009

Zhou, M.-L., Shao, J.-R., & Tang, Y.-X. (2009). Production and metabolic engineering of

terpenoid indole alkaloids in cell cultures of the medicinal plant Catharanthus roseus (L.) G. Don (Madagascar periwinkle). *Biotechnology and Applied Biochemistry*, *52*(4), 313. https://doi.org/10.1042/ba20080239

Zorin, B., Lu, Y., Sizova, I., & Hegemann, P. (2009). Nuclear gene targeting in Chlamydomonas as exemplified by disruption of the PHOT gene. *Gene*, *432*(1–2), 91–96. https://doi.org/10.1016/j.gene.2008.11.028

Method optimisation of CRISPR-Cas9 ribonucleoprotein delivery for genome editing in *Chlamydomonas reinhardtii*

ABSTRACT

Chlamydomonas reinhardtii is a widely genetically engineered model microalgal species that has been explored for its usefulness in both basic and applied research. However, strategies to modify *C. reinhardtii* nuclear genome have relied solely on random integration approaches, resulting in instability, detrimental “position-effects” such as transgene silencing, integration in transcriptionally-inactive regions, and disruption of endogenous sequences. The advent of CRISPR-Cas9 technology –a bacterially derived system for editing genomic DNA *in vivo* in a targeted fashion– has revolutionised molecular biology across many model species including human cell lines, animal models and plants. However, microalgae remain an exception to this, as only a small selection of low efficiency protocols are currently available. Herein, we investigated existing electroporation approaches for CRISPR-Cas9 ribonucleoprotein (RNP) delivery in *C. reinhardtii* and showed that we were unable to achieve a CRISPR-based edit. In order to address the low efficiency issues associated with these electroporation-based protocols, we developed large (>100 kDa) protein delivery systems previously unexplored in microalgae, based on biolistic bombardment and a lipofection-mediated delivery based on endocytosis. Our results demonstrated a 0.2% efficiency of a fluorescently labelled antibody proxy which was co-delivered with CRISPR-Cas9 RNP following proteolistic bombardment, but no detectable CRISPR-based genome editing. Similarly, we detected about 50% of lipofected cells with a phenotype associated with CRISPR-Cas9 editing, however, we did not confirm any edits from Sanger sequencing. Altogether, these results demonstrated that CRISPR-Cas9 delivery systems in *C. reinhardtii* are limited by unreliable endogenous reporter markers, a poor knowledge of large protein delivery in microalgal cells, and limited knowledge of CRISPR-Cas9 binding and endonuclease activity in vivo in microalgae.

The work presented here is important in highlighting challenges and limitations of CRISPR-Cas9 based genome editing approaches in *C. reinhardtii*, and to underscore that more work needs to be done to efficiently appropriate CRISPR into this microalgal molecular toolkit.

INTRODUCTION

CRISPR revolution

CRISPR-Cas9 is a bacterial ribonucleoprotein that was first described for its genome editing capacity in 2012 (Jinek et al., 2012). Since then, CRISPR has been integrated into molecular research of most model species, accelerating discoveries in industrial biotechnology, plant breeding, and disease investigation and treatment (Sternberg & Doudna, 2015; Stovicek et al., 2017). While CRISPR is not the first nuclease to be employed for genome editing, its widespread and rapid success is attributed to its components being cheaper and quicker to synthesise than pre-existing genome editing tools and its ability to target virtually any region in the genome. Pre-CRISPR nucleases, such as zinc finger nucleases (ZFN) and transcription activator-like effector nuclease (TALENs), depend on complex DNA-protein interactions and consequently, require more time and capital to design and test, and are not necessarily suitable for any location (Gaj et al., 2013; Noman et al., 2016). On the contrary, CRISPR-Cas9 depends on simple Watson-Crick base pairing between its RNA component and the target genomic DNA locus of interest (Figure 1). Once based paired, the endonuclease component of the CRISPR-Cas9 complex is able to generate a double stranded break in the genomic DNA at this precise location (Figure 1). The break site can be repaired to result in either a gene knock-out or knock-in modification (Figure 1).

Beyond genomic DNA editing, Cas9 can be modified for epigenetic editing (Pflueger et al., 2018; Thakore et al., 2015), gene localisation (Chen et al., 2013; Roberts et al., 2017), genome-wide screening (Chen et al., 2015; Shalem, 2014), regulating gene circuits in synthetic biology (Kiani et al., 2014), and potential for *in vivo* gene therapy (Xue et al., 2016), making it a revolutionary discovery.

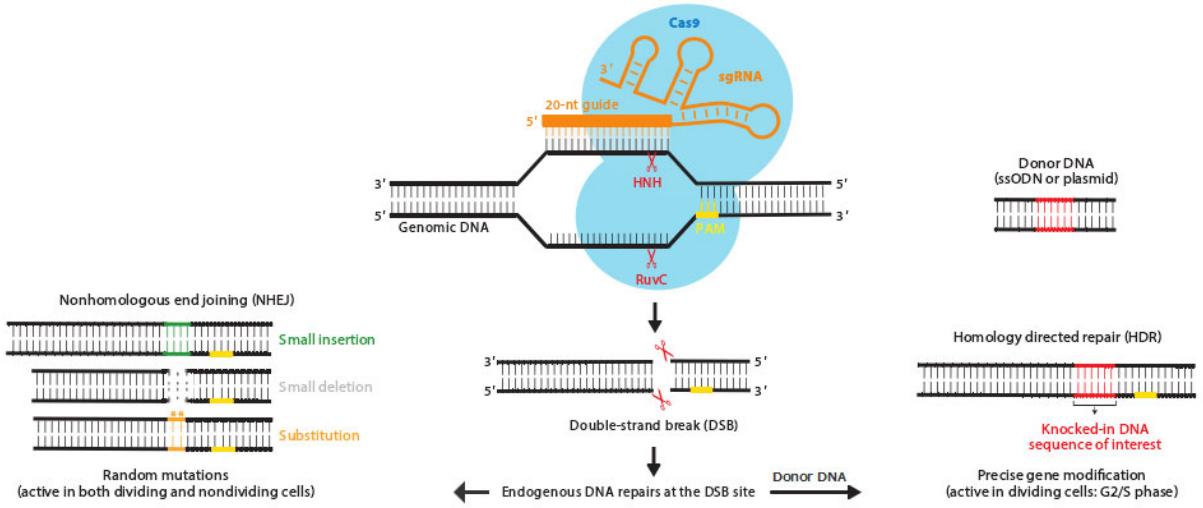


Figure 1. Schematic representation of CRISPR-Cas9 editing technology. The orange guide RNA component of the CRISPR-Cas9 complex is made up of a 20 nucleotide long guide region and a longer regulatory region that helps the guide to assemble with the Cas9 enzyme. The blue Cas9 enzyme contains a HNH and RuvC endonuclease domain for double stranded DNA digestion. After delivery into the target species, the CRISPR-Cas9 ribonucleoprotein complex scans the host genome to find the region that aligns to the guide RNA, which is always upstream of a protospacer adjacent motif (yellow). The double stranded break can be repaired by non-homologous end joining (NHEJ) resulting in various possibilities of INDELs (insertions, deletions, or a combination). Alternatively, if donor DNA or single stranded donor oligonucleotide (ssODN) containing homologous regions to the cut site is supplied to the cell, the repair can be mended by homology driven (or directed) repair (HDR), also commonly referred to as homologous recombination. Numerous reports have also demonstrated that donor DNA without homologous flanks can be integrated by NHEJ. Adapted from https://www.sinobiological.com/crispr-cas9-system.html

CRISPR-Cas9 editing technology has had widespread success across kingdoms including microorganisms, plants, and vertebrates (Stovicek et al., 2017). The first reports of CRISPR-Cas9 editing in mammalian cells (Cong et al., 2013; Mali et al., 2013), yeast (Dicarlo et al., 2013) and plant systems (Feng et al., 2013) were all published in 2013, closely followed by mouse models (Mashiko et al., 2013; Wang et al., 2013) and the microalga, *C. reinhardtii*, in 2014 (Jiang et al. , 2014). Since then, applications of CRISPR-Cas9 technology in these models has boomed –with the exception of microalgae– which is still limited by low editing efficiencies. Reviews

covering genome editing in microalgae have documented the progress thus far (Jeon et al., 2017; Naduthodi et al., 2018; Spicer & Molnar, 2018; Patel et al., 2019) but it is still unclear why the editing efficiency (usually calculated as the percentage of mutated colonies out of the total number of colonies) in these photosynthetic eukaryotic microbes seems to be lower than in other eukaryotic systems.

CRISPR mechanism of action

While CRISPR-Cas9 is often described as a genome editing technology, it is in fact a simple DNA double stranded break (DSB) inducer, originating from a bacterial immune response to identify and digest invading viral DNA (Barrangou et al., 2007). The editing function arises from the host cell's natural DNA repair mechanisms, classed as either non-homologous end joining (NHEJ), the random repair of a DSB, or homologous recombination (HR), the specific repair of a DSB using a template–either exogenously supplied recombinant DNA or an identical sister chromatid (Figure 1). NHEJ occurs when CRISPR-Cas9 is delivered into the cell without any template or donor DNA, often resulting in a frame shift or functionally disruptive mutation and a loss of function, knock-out (KO) organism (Sander & Joung, 2014) (Figure 1). Conversely, when donor DNA that bearing flanks homologous to the break site is delivered to the cell, the cell's native HR mechanisms can incorporate the donor DNA at the target site (Figure 1). Approaches which couple CRISPR-Cas9 with donor DNA delivery are used for the next generation genetic engineering strategy of targeted genomic integration (TGI).

While these descriptions of NHEJ and HR mediated genome editing are important, they can be reductive and therefore inaccurate. For example, NHEJ is usually referred to as error-prone and random. However, there is increasing evidence that this is an inaccurate, overly simplistic definition (Bétermier et al., 2014; van Overbeek et al.,

2016). It was previously thought that HR was the only mechanism by which to insert a DNA fragment at a specific DSB, resulting in precise TGI. However, this too has shown to be facile, as more and more reports showing homology-independent knock in at target sites, also referred to as imprecise TGI (Bachu et al., 2015; He et al., 2016; Shin et al., 2016). In *C. reinhardtii* specifically, Shin et al. (2016) observed that 80% of their edited colonies had integrated the antibiotic resistance plasmid DNA, which had been co-delivered with the CRISPR-Cas9 expression plasmid, instead of at a random site in the genome. Like random integration, imprecise TGI, is not well understood, but one possible mechanism of action is via microhomologies (2 - 25bp) between the exogenous DNA fragment and the free ends of the DSB, known as microhomology-mediated end joining (MMEJ) (Mcvey et al., 2017). These types of findings again highlight how little we know about the molecular processes driving exogenous DNA integration, or the chronology of the mechanisms involved (Kohli et al., 2006; Kohli et al., 2010).

Off-target mutation in CRISPR-Cas9 genome editing

Although CRISPR-Cas9 tools have received significant hype and 'gone viral' in molecular research (Pennisi, 2013; Sternberg & Doudna, 2015), an important consideration of the technology is the occurrence of off-target mutations. These occur when any region(s) of the host genome other than the selected target locus is edited by the genome editing endonuclease. Theoretically, this should not be possible, as any sequence longer than 16 nucleotides is specific enough to statistically not occur anywhere else in a eukaryotic genome of 430 megabases or smaller (Kim & Kim, 2014). However, the mechanisms of off-targeting are complex, not yet fully understood and do not occur in the same way or at the same rate in different organisms (Haeussler et al., 2016). Some important factors already identified to cause off-target mutation

include ‘bulging’, whereby off-target sites with high sequence similarity to the guide RNA and only a few mismatches will still anneal (Lin et al., 2014); chromatin structure at target sites (Kuscu et al., 2014; Singh et al., 2015); and high dose or exposure of CRISPR-Cas9 components, resulting in toxicity from DNA damage (Morgens et al., 2017; Wendt et al., 2016). Consequently, researchers are continuously developing new algorithms to predict the off-targeting score of a potential guide RNA molecule to improve guide RNA design (Doench et al., 2016; Haeussler et al., 2016; Park et al., 2015; Tsai et al., 2015) as well as detect and validate any that might occur (Cradick et al., 2014; Wang et al., 2015).

DNA-independent CRISPR-Cas9 genome editing

In order to manage off-targeting and reduce toxicity, many researchers have turned to directly delivering the CRISPR-Cas9 assembled ribonucleoprotein (RNP) to the cells, instead of the more conventional approach of recombinant DNA technology and expression of the CRISPR-Cas9 components from plasmid DNA (Figure 2A). Consequently, recombinant Cas9 protein and customised guide RNA molecules are commercially available from various refutable suppliers, such as IDT. Unlike hereditary DNA, RNP degrades quickly *in vivo*, significantly reducing off-targeting and possible toxicity issues, while still accomplishing editing. In human cells, Kim et al. (2014) showed that Cas9 protein was degraded within 24 hours post-transfection, while achieving editing frequencies of up to 79%.

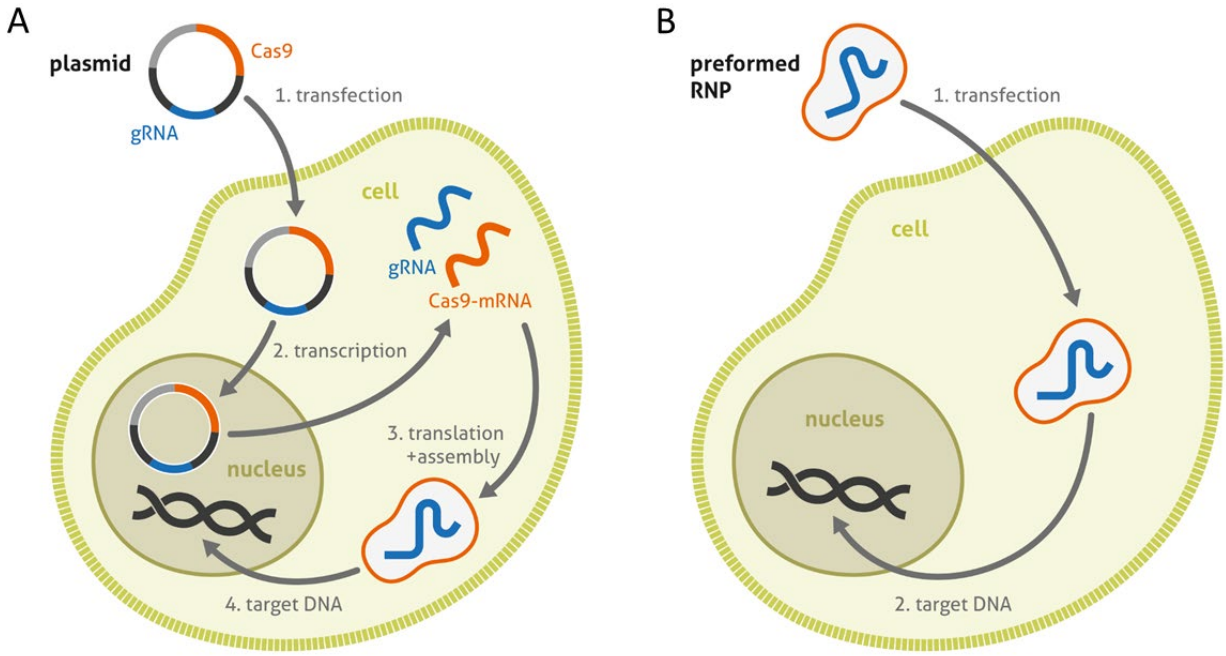


Figure 2. Schematic representation of CRISPR-Cas9 technology based on either (A) conventional plasmid delivery system or (B) ribonucleoprotein (RNP) delivery system. From https://viromer-transfection.com/crispr.

It is surprising that while CRISPR-Cas9 driven off-targeting is so widely highlighted as a concern if this technology, the implications of randomly integrating CRISPR-Cas9 exogenous DNA into the host genome is often overlooked. Recombinant technology often depends on randomly integrated chromosomal expression (RICE), including in microalgae. RICE is associated with numerous positional effects and it seems counter intuitive to apply a precision technology and focus on minimising off-targeting, while allowing random disruptions to the genome through RICE. Therefore, RNP delivery covers another important benefit over RICE mediated recombinant DNA technology.

A final benefit to RNP delivery is that it is possible to perform multiple edits within the same strain with no genetic trace left, as demonstrated by Baek et al. (2016), who created a zeaxanthin epoxidase (Cre02.g082550) CpFTSY (Cre05.g241450) double knock-out (KO) mutant *C. reinhardtii* strain. Only two years later, this strain was trialled

as a chicken feed supplement for producing macular-pigment enriched eggs, a testament to the benefit of generating transgene-free mutants by RNP editing (Baek et al., 2018).

CRISPR-Cas9 genome editing *C. reinhardtii*

C. reinhardtii is theoretically an ideal microalgal candidate for CRISPR-Cas9 genome editing via RNP delivery for a few reasons. First, *C. reinhardtii* has a haploid nuclear genome, requiring only a single allele edit for total knock-out and has no sister chromatid for HR mechanisms to repair an edit back to its original state. This would avoid any problems of heterozygous mutant genotypes. Second, unlike model plants and other widely studied eukaryotic animals such as mice and zebrafish, *C. reinhardtii* is unicellular, avoiding issues of tissue mosaicism seen these multicellular organisms. Third, *C. reinhardtii* is one of only a few microalgae that has been studied extensively in molecular biology, for which a fully sequenced genome is available for effective guide RNA design and extensive bioinformatics and -omics data for targeted genome engineering projects, such as a Pathway/Genome Databases (PGDB) (http://pmn.plantcyc.org/CHLAMY/) and ChlamyNET (http://viridiplantae.ibvf.csic.es/ChlamyNet/index.html). Finally, various endogenous selectable marker genes have been validated in *C. reinhardtii* (Crespo et al., 2005; Liu et al., 2013; Shin et al., 2016). Endogenous marker genes are native genes that, when knocked out, result in a distinct phenotype that can be used for selection or screening of mutants, such as a colour change phenotype associated with magnesium protoporphyrin O-methyltransferase (ChlM; Cre12.g498550) knock-out (Meinecke et al., 2010) or antibiotic resistance associated with FKB12/rapamycin complex (FKB12; Cre07.g347250) knock-out (Crespo et al., 2005). Endogenous markers are crucial for developing CRISPR-based knock-out approaches, as they result in a mutation that is

directly related the CRISPR editing capacity. This is in contrast to cases in which endogenous markers are not available, whereby researchers rely on integration and expression of an exogenous selectable marker gene, such as an antibiotic resistance phenotype. Exogenous selectable marker genes require random integration which makes them susceptible to position effects, whereby transgenes are integrated into important endogenous regions, or transcriptionally repressed regions, or loci that induce epigenetic silencing of the transgene (Jupe et al., 2019).

Given this context, it is unsurprising that the first CRISPR-Cas9 publication describing microalgal genome editing was demonstrated in *C. reinhardtii* using already widely established recombinant DNA technology and random integration (Jiang et al., 2014). Since then, three more approaches have been published, all of which have focused solely on electroporation-mediated RNP delivery. Shin et al., (2016) showed first DNA-facilitated RNP delivery by electroporation. Baek et al. (2016) showed first DNA-free RNP delivery via electroporation. Greiner et al. (2017) developed an optimised electroporation protocol that did deliver higher editing efficiencies of about 16%; however, this requires a very specific electroporator for delivering a poring pulse, a feature which is not available using most standard electroporators. Greiner et al. (2017) also published a modified DNA encoded system and in both their RNP delivery and DNA encoded approaches, they suggested that synchronising the cells and administering a heat shock treatment prior to electroporation increased the CRISPR-Cas9 activity *in vivo*. Each of these strategies above described targeted knock-out approaches via electroporation, including insertional knock-out, in which antibiotic resistance encoding selection plasmid DNA is integrated at a targeted location in order to disrupt the gene at that locus. While all of these approaches have been successful, the reported overall efficiency is still relatively low (below 16% editing efficiency),

suggesting that generating transient pores in the cell membrane via electroporation may not be an effective way of delivering the large 160 kDa globular protein across the tough cell walls of microalgae (Greiner et al., 2017).

What is a CRISPR-Cas9 workflow?

Method development depends on a strong understanding of the required workflow, as this allows collation of diverse strategies to inform design choices, enables appropriate identification of aspects to be modified, and facilitates correct analysis of optimisation strategies. Therefore, we constructed a generic CRISPR-Cas9 pipeline summarising tools and assays required for a microalgae-specific CRISPR-Cas9 RNP workflow (Figure 3). Some aspects of the pipeline, namely guide RNA design and *in vitro* analysis (Step 1 and 2), sub-cloning (Step 3), and edit detection and verification (Step 5 and 6) are already well described in the literature (Bortesi & Fischer, 2015; Lin et al., 2018). For example, *in vitro* guide RNA analysis and edit detection protocols are easily reproducible, including commercially available assays such as the T7E1 assay (Woo et al., 2015; Kim et al., 2014). This allowed us to identify and categorise aspects of CRISPR-based approaches for microalgae that are less well defined and therefore open to optimisation; namely, the process of delivery, enrichment and phenotype screening (Step 3 and 4).

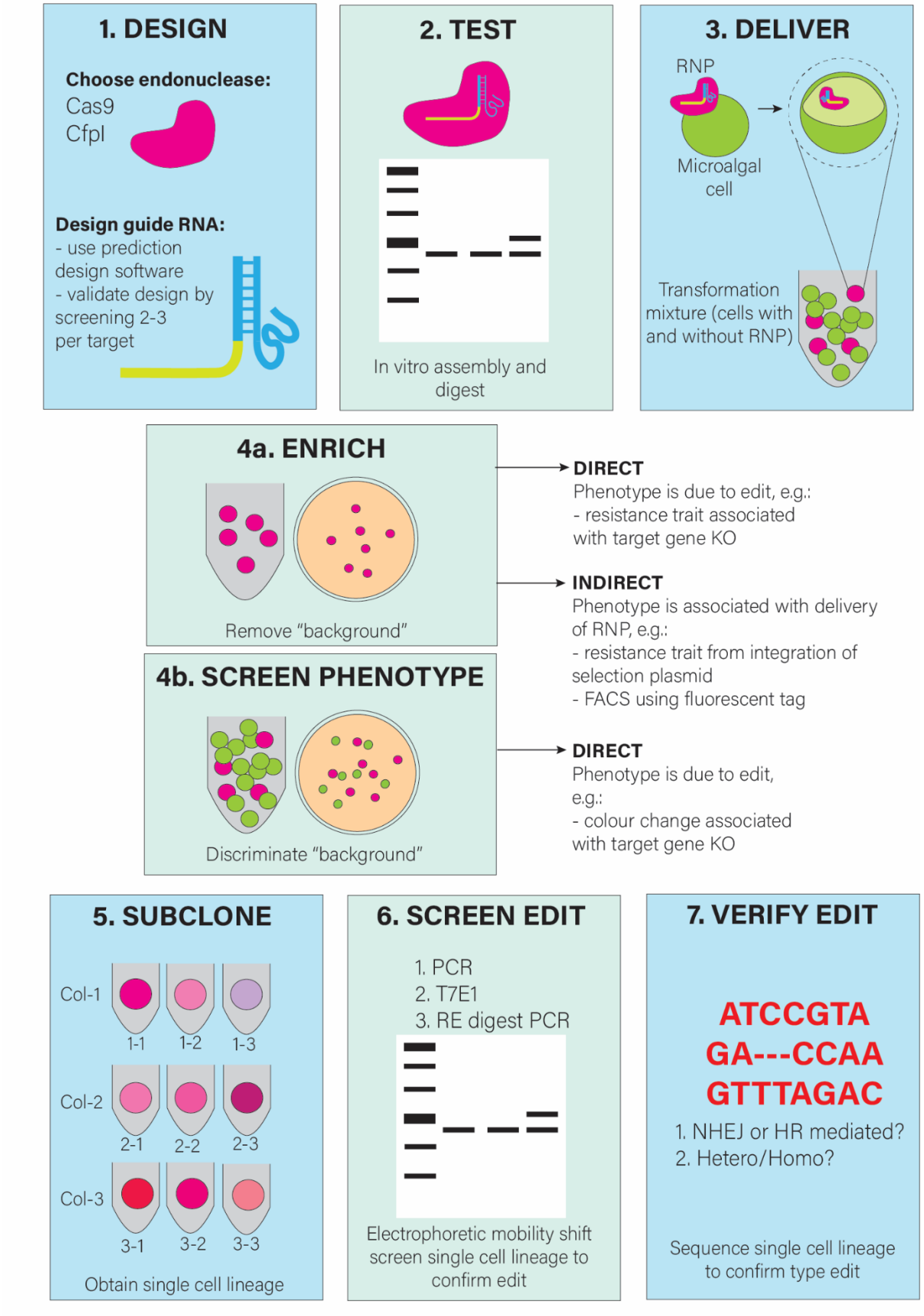

1. DESIGN
Choose endonuclease:
Cas9
CfpI
Design guide RNA:
- use prediction design software
- validate design by screening 2-3 per target
2. TEST
In vitro assembly and digest
3. DELIVER
RNP
Microalgal cell
Transformation mixture (cells with and without RNP)
4a. ENRICH
Remove "background"
DIRECT
Phenotype is due to edit, e.g.:
- resistance trait associated with target gene KO
INDIRECT
Phenotype is associated with delivery of RNP, e.g.:
- resistance trait from integration of selection plasmid
- FACS using fluorescent tag
4b. SCREEN PHENOTYPE
Discriminate "background"
DIRECT
Phenotype is due to edit, e.g.:
- colour change associated with target gene KO
5. SUBCLONE
Col-1
1-1
1-2
1-3
Col-2
2-1
2-2
2-3
Col-3
3-1
3-2
3-3
Obtain single cell lineage
6. SCREEN EDIT
1. PCR
2. T7E1
3. RE digest PCR
Electrophoretic mobility shift screen single cell lineage to confirm edit
7. VERIFY EDIT
ATCCGTA
GA---CCAA
GTTTAGAC
1. NHEJ or HR mediated?
2. Hetero/Homo?
Sequence single cell lineage to confirm type edit

Figure 3. Schematic representation of the various workflow strategies involved with generating a CRISPR-Cas9 mutant. Step 1. Selection of the type of endonuclease such as Cas9 or CfpI, and design of the guide RNA targets, which requires the use of software such as CRISPOR. It is also recommended to test 2-3 guide RNAs for those which have not previously been validated. Step 2. In vitro assembly of the guide RNA and endonuclease to form an active ribonucleoprotein (RNP). To confirm correct assembly and activity of the RNP, an in vitro digestion should be performed, whereby a PCR amplicon containing the target site is incubated with the RNP and the product is screened by electrophoretic mobility shift assay. Step 3. Intracellular delivery of the tested RNP into microalgal cells results in a mixture of transformed mutants and untransformed (background) cells. Step 4a. Enrichment involves removing background cells and can occur via direct method, in which the phenotype used for enrichment is the result of the CRISPR-Cas9 edit itself. Indirect methods rely on a phenotype that is not the direct result of the CRISPR-Cas9 edit, but instead relies on a proxy such as use of fluorescently tagged CRISPR-Cas9 RNP for FACS, or expression of a selectable marker which has been integrated during co-transformation with RNP. In cases where enrichment is not appropriate or not possible, screening can be used. Step 4b. Here, background cells are not removed but instead can be distinguished from the RNP edited cell lines. Screening should be directly related to CRISPR-Cas9 edit, such as colour change following knock-out of ChlM gene. Screening can also be used in conjunction with indirect enrichment, as indirect enrichment can result in a mixture of both edited and non-edited cell lines. Step 4 can be performed either using step 4a, or step 4b, or both. Step 5. Because the intracellular presence of CRISPR-Cas9 can cause some daughter cells to be repaired by NHEJ causing INDELs and others to be repaired faithfully with no edits, mutants may present as colony mosaics and consequently must be subcloned in order to generate single cell lineages (Huang & Daboussi, 2017). Step 6. Subclones can be screened for CRISPR-Cas9 induced mutation via electrophoretic mobility shift assay following various analyses of the target DNA. For example, if the experiment was designed for targeted integration or if two endonucleases were used for cutting a fragment out of a target region, conventional PCR can be used to screen for the size change mutants. Alternatively, where targeted knock-out via NHEJ-mediated repair was expected, the T7E1 assay or restriction digestion analysis can be used. Step 7. Mutants identified as positive for CRISPR-Cas9 mutation in the screen must be sequenced in order to determine the sequence and confirm the nature of the CRISPR edit.

Alternative CRISPR-Cas9 RNP delivery methods

Although RNP-based CRISPR-Cas9 offers important advantages over integrating DNA encoded CRISPR-Cas9, delivering the large 160 kDa globular RNP demonstrates new challenges. In non-microalgal organisms, numerous methods for RNP delivery have been investigated and optimised, namely via physical or endocytotic methods.

Physical methods involve direct transfer of RNPs intracellularly, for example, using microinjection or delivery via biolistic bombardment (proteolistic). At the time of this research, there were no proteolistic protocols for delivering CRISPR RNPs into any microalgae. Protocols had depended entirely on electroporation, until the recent publication in 2018 regarding *Phaeodactylum tricornutum* (Serif et al., 2018). Proteolistic delivery of CRISPR RNPs has been used in plants for generating 4 - 5% independently edited plant lines (Liang et al., 2017; Liang et al., 2018). On the other hand, microinjection is commonly used in multicellular organisms such as zebrafish and mouse embryos (Horii et al., 2014; Xin & Duan, 2018), but is not suitable for most microalga, as it requires a large cell size (around 100 μm) to be effective.

Unlike physical methods of delivery, endocytotic methods rely on fusion of RNP cargo with the cell membrane and is usually induced by a chemical condition. Therefore, unicellular organisms with cell walls are not appropriate for chemical delivery, as the RNP cargo is unlikely to make sufficient direct contact with the cell membrane. However, cell wall deficient microalgae, or protoplasts, could be good candidates for endocytotic delivery. Indeed, cell wall deficient strains of *C. reinhardtii* do exist, such as cc-503. This may be very useful, given the current limitations on *C. reinhardtii* strategies, but has not been reported before. Lipofection is the most commonly used endocytotic delivery method for RNP in various mammalian cell lines, generated editing efficiencies of 5 – 51%. Lipofection is particularly attractive because it requires a simple, quick, and inexpensive protocol with no expensive or specialised equipment and very low concentrations (< 1 μg/reaction; compared to 50 - 200 μg/reaction using proteolistic delivery) of the more expensive component, the Cas9 protein (Liang et al., 2015; Zuris et al., 2014). Here, the lipofection cargo material is complexed with a lipid solution to create vehicles contained inside phospholipid bilayer which is able to

conjugate with the cell membrane (Figure 2B). A similar lipid-based protocol for plants called polyethylene glycol (PEG)-mediated delivery is used in plant protoplast cells (Liang et al., 2018). Because many plants are difficult and slow to generate from protoplasts (Lin et al., 2018), this technique is often used as a quick *in vivo* assay for screening guide RNA targets. Such endocytotic delivery protocols could be optimised for microalgal protoplasts or cell-wall deficient strains.

Other endocytotic delivery methods have harnessed nanotechnology for precise donor DNA delivery. For example, CRISPR-Gold uses gold nanoparticles conjugated donor DNA and RNP secured inside a polymer coating (Lee et al., 2017). This was shown to be efficient for various cell types and successful in replacing the disease-associated gene in a mouse model for Duchenne muscular dystrophy (Lee et al., 2017).

Enrichment and screening in *C. reinhardtii*

From our microalgae-specific RNP workflow, we identified that the process of identifying CRISPR-Cas9 putative mutant *C. reinhardtii* strains from non-transformed background, classified as Step 4 of the workflow (Figure 3), can occur in three possible ways; by enrichment (Step 4a), or screening (Step 4b), or both. Enrichment describes the process of reducing the background –cells present in the transformation mixture but which did not become transformed– and can be achieved directly or indirectly (Step 4a; Figure 3). Direct approaches enrich cells based on successful CRISPR-Cas9 RNP delivery, binding and editing activity, and resultant mutation associated phenotype. In *C. reinhardtii*, examples include an antibiotic resistance phenotype associated with endogenous gene knock-out of FK506-binding protein (FKB12; Crespo et al., 2005) and an auxotrophic growth phenotype associated with

endogenous gene knock-out of argininosuccinate lyase (ARG7; Debuchy et al., 1989) (Step 4a; Figure 3).

In contrast, indirect enrichment of cells selects for those that have acquired a selectable trait that is not related to the CRISPR-Cas9 mutation event itself; such as expression of a co-delivered antibiotic resistance plasmid or a fluorescently tagged CRISPR-Cas9 RNP (Step 4a; Figure 3). In *C. reinhardtii*, Shin et al. (2016) and Greiner et al. (2017) both used indirect enrichment by co-delivering a hygromycin and paromomycin resistance plasmid with RNP, respectively. This is less desirable, as it assumes that cells which receive, integrate and express the selectable marker plasmid DNA will also have received the RNP. In reality, cells that receive the RNP might not integrate and express the plasmid DNA to confer resistance and vice versa and consequently, co-delivery via indirect enrichment has a very low frequency of success. Alternatively, when a fluorescently tagged CRISPR-Cas9 RNP is delivered into cells, these cells can be sorted by fluorescence activated cell sorting (FACS), as demonstrated in mammalian cells (Seki & Rutz, 2018). Although this is also indirect, as cells which receive the protein are not necessarily edited, it is a more reliable method than co-delivery of an antibiotic resistance plasmid and can therefore be considered ‘semi-indirect’. Overall, direct enrichment is preferential over indirect enrichment, as it both removes background for decreased screening load, and is a marker of CRISPR-Cas9 activity.

Unlike enrichment, which removes background, screening depends on the ability to discriminate between background and mutants edited by CRISPR-Cas9 RNP (Step 4b; Figure 3). Instead of introducing an exogenous selectable trait encoded by recombinant DNA, direct screening depends on a detectable phenotype change that is associated with the CRISPR-Cas9 RNP edit, such as the colour change phenotype

is associated with magnesium protoporphyrin O-methyltransferase (ChlM; Cre12.g498550) disruption in *C. reinhardtii* (Shin et al., 2016).

Shin et al. (2016) combined indirect enrichment with direct screening by co-delivering RNP targeting the ChlM gene with a hygromycin resistance plasmid. Interestingly, they reported NHEJ-mediated insertion of the hygromycin plasmid DNA into the CRISPR-Cas9 induced break site. Direct screening is particularly useful over enrichment when a 'clean' recombinant DNA-free mutant is desired. Baek et al. (2016) used direct screening to detect a colour change phenotype with no enrichment (DNA-free) and reported extremely low editing efficiency of 0.56%. This highlights the low efficiency of RNP delivery via electroporation and demonstrates that an enrichment step is essential where gene knock-out does not generate a detectable phenotype.

CRISPR-Cas9 in *C. reinhardtii* requires more efficient RNP strategies

CRISPR-Cas9 technology holds great potential for metabolic engineering and synthetic biology, including integrating heterologous terpenoid pathway associated plant genes into the model green microalga, *C. reinhardtii*. However, this depends heavily on the availability of robust and reproducible protocols that should both provide both efficient genome editing and low off-target mutation events. Although CRISPR-Cas9 genome editing can be achieved via either recombinant DNA technology, ribonucleoprotein delivery can help to overcome off-target mutations, as it avoids long-term exposure to the CRISPR-Cas9 components. CRISPR-Cas9 RNP delivery systems have been described as quick, efficient and cheap and indeed, numerous publications have described CRISPR-Cas9 genome editing in *C. reinhardtii* via RNP delivery (Baek et al., 2016, 2018; Shin et al., 2016; Greiner et al., 2017). However, all of these publications demonstrated electroporation-driven delivery, which depends on

the passive migration of the large 160 kDa RNP into the cell via transiently created cell membrane pores. Given that these studies lay the groundwork for more complex CRISPR-Cas9 editing in *C. reinhardtii*, we aimed to both investigate the available electroporation strategies currently available, as well as develop novel RNP delivery approaches via proteolistic bombardment and a microalgal-compatible lipofection strategy never before reported in any microalga.

RESULTS AND DISCUSSION

Electroporation for RNP delivery in *C. reinhardtii*

In order to access the suitability of an indirect enrichment approach, we designed two guide RNA molecules to target the first exon Mg-protoporphyrin IX S-adenosyl methionine O-methyl transferase *(ChlM;* Cre12.g498550; GenBank accession XM_001702328) (Figure 4A, B; Suppl. Table 1). *ChlM* is a useful endogenous marker gene, as the ChlM enzyme converts magnesium-protoporphyrin IX into magnesium-protoporphyrin IX 13-monomethyl esterchlorophyll (EC 2.1.1.11) as one of the first steps in the chlorophyll-a biosynthesis pathway (Meinecke et al., 2010) (Figure 4A). Knocking out *ChlM* generates a low chlorophyll-a mutant, giving rise to a light green phenotype under low light conditions, easily identifiable for screening (Meinecke et al., 2010; Shin et al., 2016). We confirmed the correct assembly and activity of both RNP-*ChlM*-1 and RNP-*ChlM*-2 by *in vitro* digest (Figure 4C). Here, a 320 bp region of the *ChlM* gene which harboured *ChlM*-1 and *ChlM*-2 target sites was amplified by PCR from isolated genomic *C. reinhardtii* DNA. The PCR products were digested with either the assembled RNP complex or the Cas9 protein alone (no guide RNA). The digest confirms that RNP-*CrChlM*-1 and RNP-*CrChlM*-2 digest the *ChlM* gene, suggesting these guide RNAs recognize the target site specifically (Figure 4C). Because we required extensive amounts of Cas9 throughout the optimisation experiments, we chose to work with only RNP-*ChlM*-2.

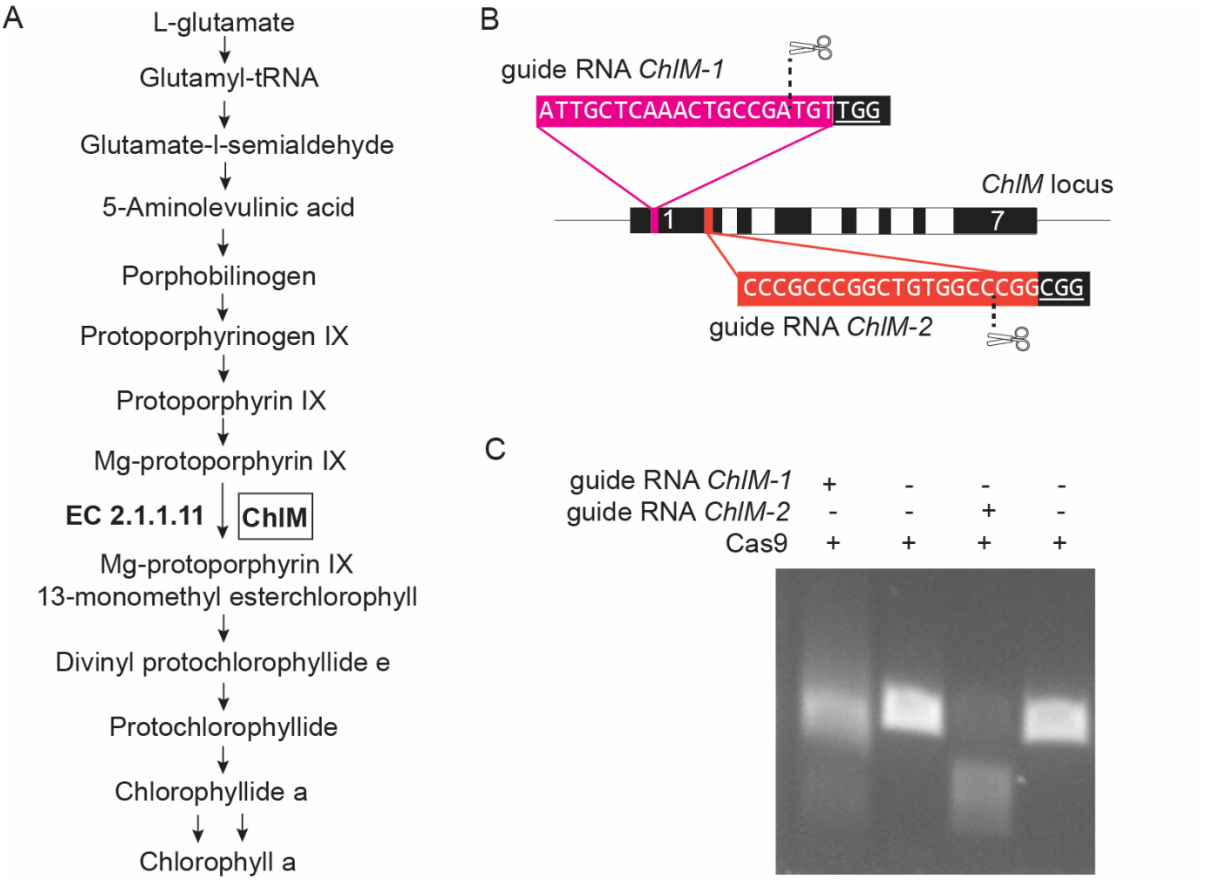


Figure 4. CRISPR-Cas9 RNP design and validation targeting *C. reinhardtii ChlM* gene. (A) Scheme of the tetrapyrrole biosynthetic pathway in *C. reinhardtii*, in which ChlM catalyses the conversion of Mg-protoporphyrin IX into Mg-protoporphyrin IX 13-monomethyl esterchlorophyll for the eventual production of chlorophyll a. Adapted from Meinecke et al., 2010. (B) Graphic representation of the genetic sequences for guide RNA *ChlM*-1 and guide RNA *ChlM*-2 with their respective 20 nucleotide recognition sites (pink and red, respectively) alongside protospacer adjacent motifs (PAMs) (black). Both guide RNAs target the first exon of *C. reinhardtii ChlM* gene. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break location, three base pairs upstream of the PAM site. (C) Electrophoretic mobility shift assay following *in vitro* digest of RNP-*ChlM*-1 and RNP-*ChlM*-2 activity to validate that RNP components were correctly assembled and are able to recognise and cut their target regions. The 320 bp amplicon showed no shift when incubated with Cas9 protein alone, but the presence of smaller bands when incubated with fully assembled CRISPR-Cas9 RNPs.

Wild type *C. reinhardtii* cells were electroporated as described by Shin et al. (2016) using RNP-*ChlM*-2 co-delivered with a selection plasmid pChlamy4 for zeocin resistance enrichment. Cells co-transformed with pChlamy4 and RNP-*ChlM*-2 did not result in any single colonies following zeocin selection, but instead grew as a thick

mat, suggesting problems at the selection stage. However, in the pChlamy4 only control treatment, single colonies were obtained, suggesting that there were no problems with the delivery, integration and expression of the pChlamy4 resistance cassette into the *C. reinhardtii* genome (Supp. Figure 1).

Because we were unable to generate single colonies using co-delivery, we explored a DNA-free RNP delivery strategy (Baek et al., 2016) which has the benefit of avoiding random integration associated with RICE, but lacks an enrichment step to remove background. DNA-free strategies rely either on extremely high editing efficiencies not yet possible with microalgae, or on direct screening by visual phenotype associated with the CRISPR edit (Figure 3). Given that Baek et al. (2016) reported an editing efficiency of 0.56%, this should result in approximately 11 edited mutant colonies per 2,000 cells per plate of solid media. After electroporating wild type *C. reinhardtii* cells with RNP-*ChlM-2*, we successfully obtained single colonies by diluting transformation mixtures to 10^4 cells/mL and plating 500 µL. However, we did not detect any colonies with a colour change phenotype (photographs not shown). These results were confirmed by chlorophyll fluorescence intensity detection analysis by pulse amplitude modulation (PAM) fluorometry, an assay measures quantum yields of photosystem II in response to a light pulse. Low quantum yields should be identified by dark red colonies compared to higher quantum yields associated with yellow to green colonies; however, no single colony showed visible colour change compared to one another (Supp. Figure 2A). This was confirmed by flow cytometry, where chlorophyll fluorescence was measured and no significant differences between any three given colonies per plate was detected (Supp. Figure 2B).

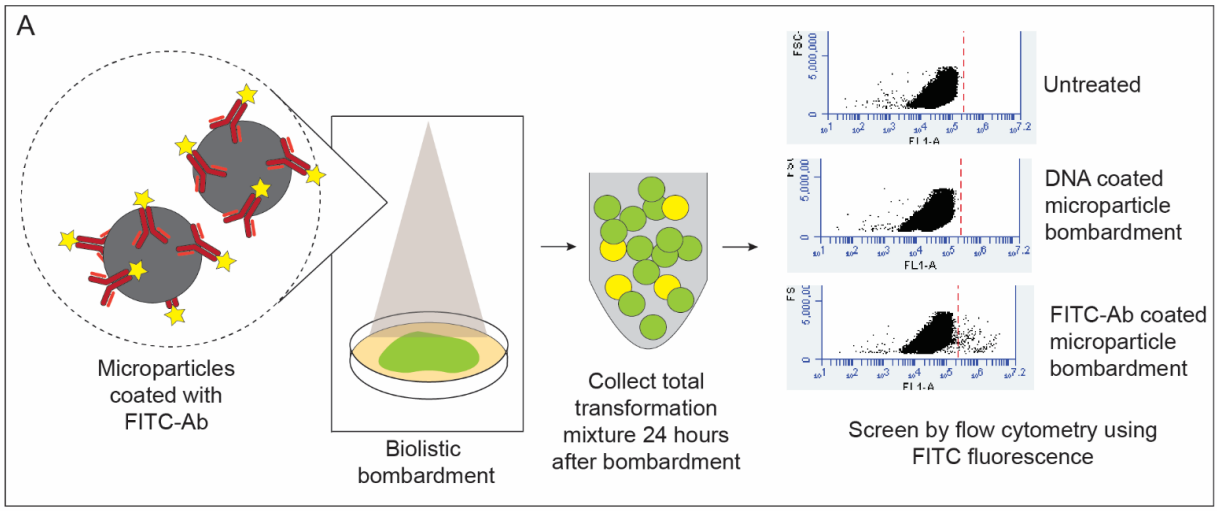
During the course of this research, Greiner et al. (2017) published a protocol using a poring pulse electroporator for higher efficiency CRISPR RNP editing. Unfortunately,

this feature is not available using most standard electroporators and consequently, we were unable to replicate it in our laboratory using the exact same conditions. Because of the problems we faced with replicating the co-electroporation (Shin et al., 2016) and the DNA-free electroporation (Baek et al., 2016) strategies, and because these are still relatively low efficiency protocols, we decided to develop a new approach for reliable, high efficiency RNP editing in *C. reinhardtii*. Indeed, various other delivery methods for delivering large proteins, including RNPs, have been explored and validated in other non-microalgal species, such as lipofection and microinjection (Horii et al., 2014; Yu et al., 2016). While electroporation methods for delivering DNA molecules intracellularly are well developed and widely used in *C. reinhardtii*, these may very well be inappropriate for delivering large 160 kDa proteins intracellularly, especially given the presence of a cell wall in most algal cells.

Development of a large protein proteolistics delivery method for *C. reinhardtii*

We identified biolistic delivery as a potentially useful strategy for large protein delivery in microalgae because this method results in physical disruption to the cell and is not hindered by the presence of a cell wall. At the time of this research there were no proteolistic delivery studies published for any microalgae, including *C. reinhardtii*. Therefore, we designed a unique hybrid approach combining different studies developed for different uses. In 2014, Martin-Ortigosa & Wang described the co-delivery of a variety of proteins with DNA to mouse ear pinnae tissue and onion epidermis via biolistic bombardment. They demonstrated that each protein (including enzymes and fluorescent proteins) retained its functionality post-delivery intracellularly. This study was appropriated for the preparation of the CRISPR-Cas9 RNP cargo with the tungsten microprojectiles. Particularly, this included an air-drying step to allow for buffer evaporation and favour the adhesion of RNP to the

microprojectiles in absence of other chemicals, as conventional chemical binding of DNA cargo depends on a calcium chloride and spermidine treatment which could impact the activity of the RNP. Given that this study was designed for multicellular tissues which are very different to unicellular microalgal cells, we consulted Barrera et al. (2014) for the preparation of the *C. reinhardtii* sample. Herein, Barrera et al. demonstrate the efficient transformation of *C. reinhardtii* chloroplast via biolistic bombardment with DNA cargo. A combination of approaches described by both Martin-Ortigosa & Wang. (2014) and Barrera et al. (2014) were used to determine parameters for the biolistic particle delivery system instrument set-up.



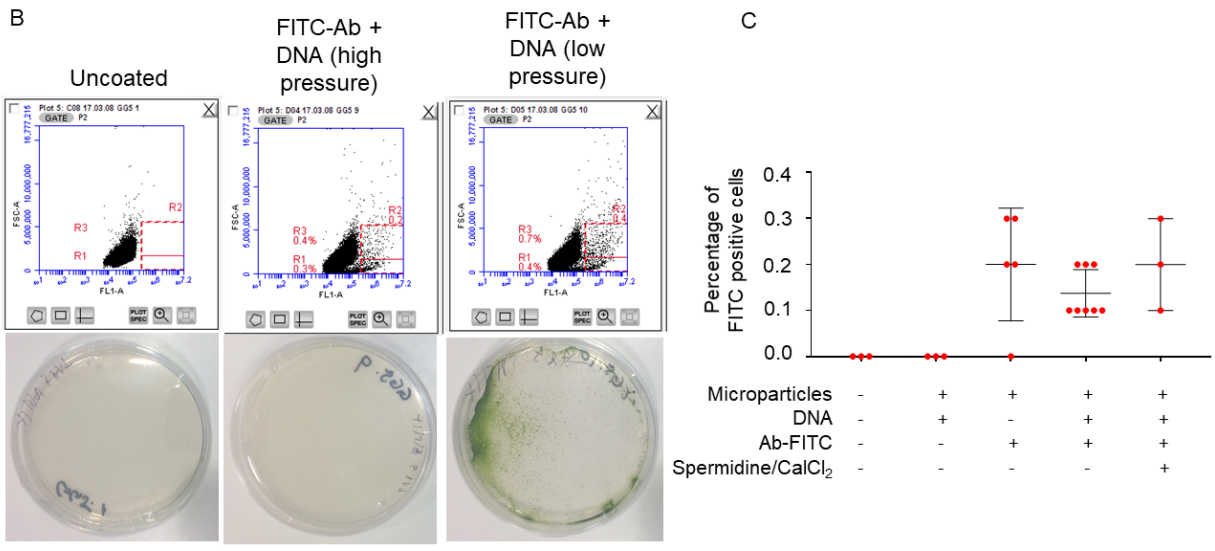


Figure 5. Method development for proteolistic bombardment of *C. reinhardtii* with CRISPR-Cas9 targeting *ChlM* gene. (A) Schematic representation of experimental design FITC-labelled antibody (FITC-Ab) was coated onto 1.1 µm tungsten microparticles and bombarded into cells. The cell biomass was scraped off bombarded agar plate a 24 hours post treatment and analysed by flow cytometry to determine the delivery efficiency based on FITC fluorescence. (B) In order to determine that proteolistic treatment did not prevent the cells from being able to recover on selectable agent, the cells were co-bombarded with the FITC-Ab and resistance DNA cassette at high pressure (1350 psi) or low pressure (650 psi). Cells treated with FITC-Ab and donor DNA and plated on TAP supplemented with zeocin only survived only when using a low pressure rupture disc (650 psi) and not when using a high pressure rupture disc (1350 psi), demonstrating that the delivery method allowed for cells to recover even after selective treatment, but only at lower rupture disc pressure. (C) Flow cytometry analysis demonstrated the percentage of FITC positive cells following various treatments.

After creating a hybrid protocol (detailed in methods), we optimised each component of the experimental setup one at a time. Because of the extensive number of experiments, and the complexities associated with enriching, screening and validating a CRISPR-Cas9 induced edit, we used a 150 kDa commercial FITC-conjugated IgG antibody as a proxy delivery protein for indirect enrichment (Figure 5A). The FITC fluorophore is easily detected by flow cytometry for quick, single cell level screening of the entire transformation mixture. This proxy protein offered a semi-indirect screening approach; it is a direct measure of protein delivery into the cell (unlike co-delivery approaches using antibiotic resistance plasmids), but is not a direct measure of CRISPR-Cas9 protein activity. The proxy design implemented sped up each analysis, as fluorescence was detected within 24 hours of the transformation and allowed us to bypass an enrichment step, which usually required a minimum of two weeks before single, antibiotic resistant colonies can be screened (Figure 5A). It should be noted that the tungsten microparticle size (1.1 µm) is about 10 times smaller than that of the single *C. reinhardtii* algal cell size (10 µm), making it highly unlikely to mistaken a coated particle for an algal cell.

The first hybrid protocol we established used 160 µg protein air-dried for 1.5 hours to gold microparticles. We observed 0.02% FITC fluorescent cells (data not shown), demonstrating the first unoptimised approach was successful. Furthermore, it is important to note that, differently from previous experiments, the efficiency reported here refers to the total number of cells bombarded and does not have any enrichment stage. In most protocols, a selection step is included, which removes background cells, thus enriching the population of cells receiving RNP successfully. These enriched efficiencies, such as those reported by Shin et al. (2016) should not be compared to non-enriched efficiencies, such as those reported by Baek et al. (2016) and herein.

We determined that a high enough delivery velocity required for successful cargo deposition intracellularly must be balanced against a low enough delivery velocity to prevent cell death. For example, we showed that at 650 psi rupture disc pressure, we were able to obtain both FITC positive fluorescent cells and single cells using zeocin selection following co-transformation (Figure 5B). However, 1350 psi pressure condition with the same settings, did not yield zeocin resistant colonies, but did see FITC positive fluorescing cells (Figure 5B). In this way, co-transformation was an important and useful condition to determine this trade-off during method optimisation, despite not being used in the final genome editing strategy (Figure 5B).

Following extensive optimisation, we conducted the finalised FITC-conjugated antibody proxy protocol in replicate. We showed that no cells were present in FITC positive gate for untreated wild type *C. reinhardtii* cells or cells bombarded with unloaded tungsten microparticles (Figure 5C). Cells bombarded with tungsten loaded with FITC antibody showed average of 0.2% cells in FITC positive gate and cells co-bombarded with tungsten and that this was consistent across both chemically bound and air-dried samples. As expected, cells bombarded with tungsten co-loaded with

FITC antibody and selection plasmid pChlamy4 in a 50:50 ratio showed average of 0.13% cells in FITC positive gate (Figure 5C).

Altogether, these results demonstrate a protocol suitable for delivering large proteins (>100 kDa) into *C. reinhardtii* cells, which has not previously been shown. We successfully improved delivery from initial 0.02% to about 0.1 – 0.2% (maximum of 10-fold increase) while reducing the total protein required from 150 µg per shot to 100 µg per shot. Others have reported the delivery of dyes and BSA protein (~50 kDa) into *C. reinhardtii* (Bothwell et al., 2006; Azencott et al., 2007 and Hyman et al., 2012) and other small proteins (<70 kDa) have been delivered to other tissue types at about 800 cells per sample while retaining their bioactivity intracellularly (Martin-Ortigosa et al., 2014). Furthermore, we detected a delivery efficiency in the same range (0.2%) of that described by Baek et al. (2016), who reported a 0.56% efficiency. We concluded that while our protocol works to deliver large proteins intracellularly, a 0.2% efficiency was too low to feasibly use without an enrichment step and consequently we investigated the potential of co-delivery for improving this novel approach.

FACS enriched DNA-free CRISPR-Cas9 RNP delivery in *C. reinhardtii*

In order to improve the proteolistics approach described above, we included an enrichment step by co-delivering RNP in conjunction the FITC-Ab. Herein, we hypothesised that cells which received FITC-Ab would be highly likely to also receive RNP. Following co-delivery, cells were allowed to recover briefly before using fluorescence activated cell sorting (FACS) to isolate cells which received FITC-Ab (Figure 6A). Sorted cell populations were left to recover and multiply for two days to increase biomass. Majority of this biomass was used for genomic DNA extraction and screening via T7E1 analysis (Figure 6A). The remainder of the biomass was used for

dilution plating and single cell lineage isolation and sequencing analysis (Figure 6A).

We chose to analyse the total sorted population mixture and single cell lineages, which

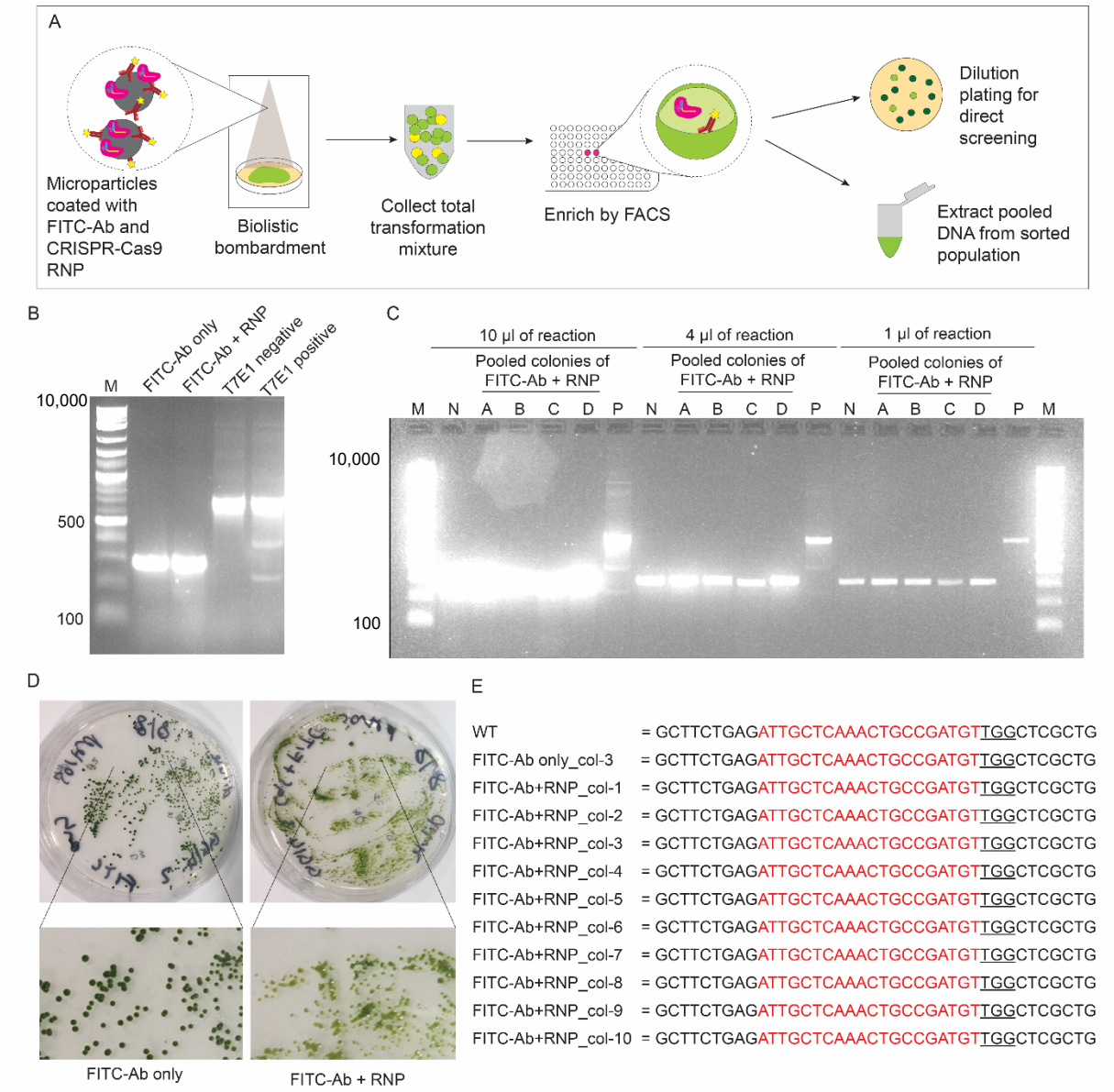


Figure 6. (A) Schematic flow diagram illustrating the proteolistic delivery of RNP-*ChlM*-2 and editing analysis. After co-delivery with RNP-*ChlM*-2 and FITC-Ab, cell biomass was scraped off bombarded agar plate and sorted for FITC fluorescence. Once sorted cells have grown for increased biomass, each sorted population was split and either plated in low cell density dilution for isolating single colonies, or used for genomic DNA extraction on pooled population for T7E1 analysis. (B) T7E1 analysis of pooled proteolistic control (bombarded with FITC-Ab only) and pooled RNP-*ChlM*-2 sample. The T7E1 commercially supplied positive and negative controls indicate that the analysis was successful and that no editing occurred. (C) Photographs of diluted cell plating following proteolistic delivery. The images suggest that co-delivery sample has the expected light green phenotype associated with *ChlM* gene knockout, compared to FITC-Ab only control. (D) Sanger sequencing alignment results of the *ChlM* amplicon amplified from wild type *C. reinhardtii* CC-125 mt+, a single colony from

the FITC-Ab only control proteolistic bombardment and ten colonies from RNP-*ChlM*-2 co-delivery bombardment showing no mismatches and therefore unsuccessful editing. Red indicates target sequence, underlined text indicates PAM motif.

were pooled into groups containing 5 colonies each. The T7E1 analysis uses the T7 endonuclease which cleaves double-stranded DNA at positions of mismatches. Because NHEJ repair of RNP-induced breaks results in a varying INDELs, there will be a variety of different mutations present in both the sorted population mixture and the pooled single cell lineages. In order to perform the T7E1 analysis, the target region of the samples must be denatured and renatured to allow for mismatches at the target site, and a positive result can be visualized by electrophoretic mobility shift assay. Most T7E1 commercial kits are supplied with a positive control, showing that the denaturing, renaturing and cleaving aspects work correctly; and likewise, a negative control in which there are no mismatches and therefore no capacity for the T7 endonuclease to digest the amplicon.

We were able to successfully isolate genomic DNA, amplify the 320 bp *ChlM* amplicon and perform the T7E1 assay on both the sorted mixture and pooled single cell lineages. However, the results show no detectable edit in *ChlM* amplicon (Figure 6B and C), despite colonies in FITC-Ab and RNP co-delivery treatment showed generally lighter green phenotype compared to FITC-Ab only control samples (Figure 6D). Therefore, we directly analysed single cell lineages. We screened ten colonies from the co-delivery treatment, one from the FITC-Ab only treatment and one wild type by Sanger sequencing. These results confirmed that none of the isolates had been edited (Figure 6E). We hypothesised that the light green phenotype may be due to non-genetic factors, for example as a product of chemicals present in the buffer, which commercial Cas9 is supplied in, may be toxic to the cells. It is possible that the sorted

cells represent those that had FITC antibodies and RNP-*ChlM*-2 associated to their outer membranes, thus never able to enter the cell intracellularly.

The sequencing results confirm that this was not possible, but altogether, this work reveals that visual screening and indirect enrichment strategies for RNP delivery in *C. reinhardtii* can be inconsistent and unreliable. It highlights the complexities of developing a reliable, robust CRISPR-Cas9 RNP delivery system and the importance of a selectable growth phenotype as a marker of CRISPR-Cas9 activity.

Lipofection delivery of CRISPR-Cas9 RNP in *C. reinhardtii*

Because lipofection is commonly used in mammalian cell systems with high CRISPR-Cas9 editing efficiencies (60-90%) and also uses dramatically less RNP compared to other methods (about 100-1000 times less) (Liang et al., 2015; Yu et al., 2016), we investigated the feasibility of lipofection-mediated RNP delivery in *C. reinhardtii*. Lipofection is a chemical delivery method which depends on lipid particles encapsulating the RNP and fusing with the cell membrane. Therefore, lipofection can only be achieved with cell wall deficient cells. We used a cell wall deficient strain of *C. reinhardtii*, CC-503, and optimised a protocol for lipofection based on supplier's recommended protocol and IDT protocol for hard-to-transfect cells (Figure 7A). Because lipofection uses so little RNP per reaction (approximately 250 ng/reaction), we were able to optimise this protocol using RNP directly instead of a proxy protein. Once again, it was critical to use a fast, easily detectable screen to evaluate our optimisation experiments. Therefore, we used RNP-*ChlM*-2 for direct screening and assessed chlorophyll content per cell using flow cytometry (Figure 7A) instead of via visual screening. We created a low chlorophyll gate (P5) to identify cells that may have been edited at *ChlM* gene (Figure 7A). This approach allowed us to optimise the

lipofection protocol and parameters rapidly before performing a full CRISPR experiment.

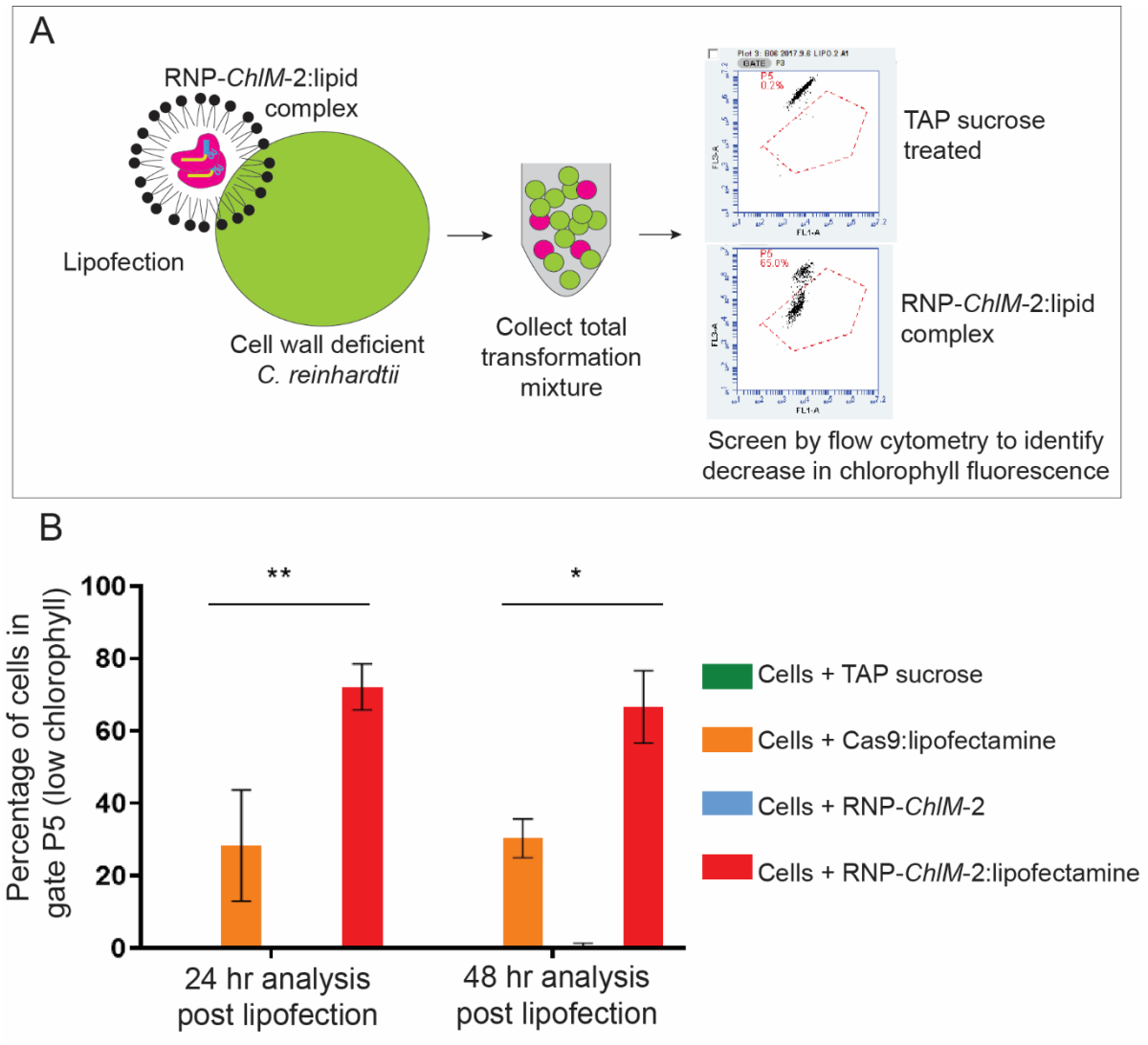


Figure 7. Method development for lipofection of cell wall-deficient *C. reinhardtii* cc-503 with CRISPR-Cas9 targeting *ChlM* gene. (A) Schematic representation of experimental design where RNP-*ChlM*-2 is complexed with lipofectamine to make a RNP-*ChlM*-2:lipid complex. When the complex comes into contact with the cell membrane of the cc-503 cell (green), it fused and enters the cell, releasing the CRISPR-Cas9 cargo intracellularly. Cells which successfully receive this cargo are edited resulting in a *ChlM* knock out (pink). Daughter cells of the *ChlM* knocked out mutants have a characteristic light green phenotype due to reduced chlorophyll content which can be detected by flow cytometry. The mixture of edited and non-edited cells is analysed by flow cytometry to detect the percentage of cells with reduced chlorophyll 24 and 48 hours after lipofection. Flow cytometry gate P5 indicated cells which have lost chlorophyll fluorescence. (B) Flow cytometry analysis of lipofection method optimisation. Cells receiving RNP-*ChlM*-2:lipid complex show significantly increased percentage of cells in low chlorophyll gate P5 compared to cells exposed to

the inactive Cas9 enzyme:lipid complex. N = 3, error bars indicate SEM, significance analysis by two way ANOVA where * indicates $p<0.05$ and ** indicates $p<0.01$.

Although the CRISPR driven *ChlM* knock-out is associated with decreased chlorophyll accumulation, this is not necessarily evident in the parent cell that is edited, as there is already chlorophyll present in the cell from before the mutation occurred. Therefore, it is important to give the treated cells time to divide and allow for analysis of daughter cells. Given that *C. reinhardtii* doubles every 5-8 hours (Moses et al., 2017; Vítová et al., 2011), we analysed the lipofected cell mixture 24 hr and 48 hr post treatment. The 24 hr treatment allowed for at least two cell division events under normal growth conditions. In case of reduce growth rate due to treatment-induced stress, a second 48 hour treatment was included.

In order to confirm that the CRISPR-Cas9 components could not cause the mutation without the delivery via lipofection, we included a cargo-only treatment in which cells were mixed with RNP-*ChlM*-2 without the lipofection mix. As expected, these cell mixtures showed no cells in low chlorophyll gate (Figure 7B). In order to confirm that the lipofection reagents did have any cytotoxic effect that may have been detected as a mutated, low chlorophyll fluorescence phenotype, we included a control in which cells were exposed to lipofectamine loaded with only Cas9 enzyme and no guide RNA. Here we detected approximately 30% of the cell population with a decrease in chlorophyll fluorescence, indicating that these reagents may have had some toxic effect on the cells or might have interfered with fluorescence detection (Figure 7B). However, an average of ~70% cells were present in low chlorophyll gate following lipofection with RNP-*ChlM*-2, compared to about 30% when lipofected with Cas9 only (Figure 7B). This suggests that ~40% of the treated cells were mutated at the *ChlM*

gene by the lipofection-delivered CRISPR-Cas9 cargo, a significant improvement over editing efficiencies in *C. reinhardtii* to date.

Although the colour change phenotype has been shown to be useful in previous studies (Shin et al., 2016), we were only able to detect this change using flow cytometry. Therefore, it would be more beneficial to make use of an auxotrophic phenotype for isolating CRISPR-edited colonies in order to validate this high potential lipofection-based CRISPR-Cas9 approach. Therefore, we designed new guide RNAs which would be compatible with both the T7E1 assay, as well as the restriction fragment length polymorphism (RFLP) assay. Here, guide RNAs which straddle a restriction enzyme site are used, allowing screening by detection of the loss of a restriction enzyme recognition site following CRISPR-Cas9 driven mutation. This analysis is much cheaper than T7E1, although equally labour intensive. We designed a guide RNA targeting the locus *Cre01.g021251,* which encodes an *a*rgininosuccinate lyase/Omega-N-(L-arginino)succinate arginine-lyase (*ARG7*). *ARG7* is responsible for the conversion of L-arginino-succinate into L-arginine (EC 4.3.2.1), which is required for translation of proteins. When knocked out, *ARG7* loss-of-function mutants result in arginine auxotrophic phenotype (Mages et al., 2007) and must be supplied arginine artificially to survive (Figure 8A; Suppl. Table 2).

Guide RNA *ARG7*-2 was designed to straddle the *HaeIII* enzyme site (Figure 8A). Upon confirming correct assembly and activity of RNP-*ARG7*-2 *in vitro* (Figure 8B), we lipofected *C. reinhardtii* CC-503 cells with RNP-*ARG7*-2 following the optimised method described above; however, both the electrophoretic mobility shift assay following both T7E1 and RFLP analysis showed no CRISPR-Cas9 edited mutants (Figure 8C). The electrophoretic assay was performed at three dilutions and intentionally overexposed in order to detect any faint smaller bands. These results

demonstrated clearly that the lipofection approach did not work. The low chlorophyll fluorescence phenotype detected in the optimisation experiments were not the result of a CRISPR-driven mutation, but rather an artifact of the experiment, similar to what was demonstrated in the proteolistics optimisation approach.

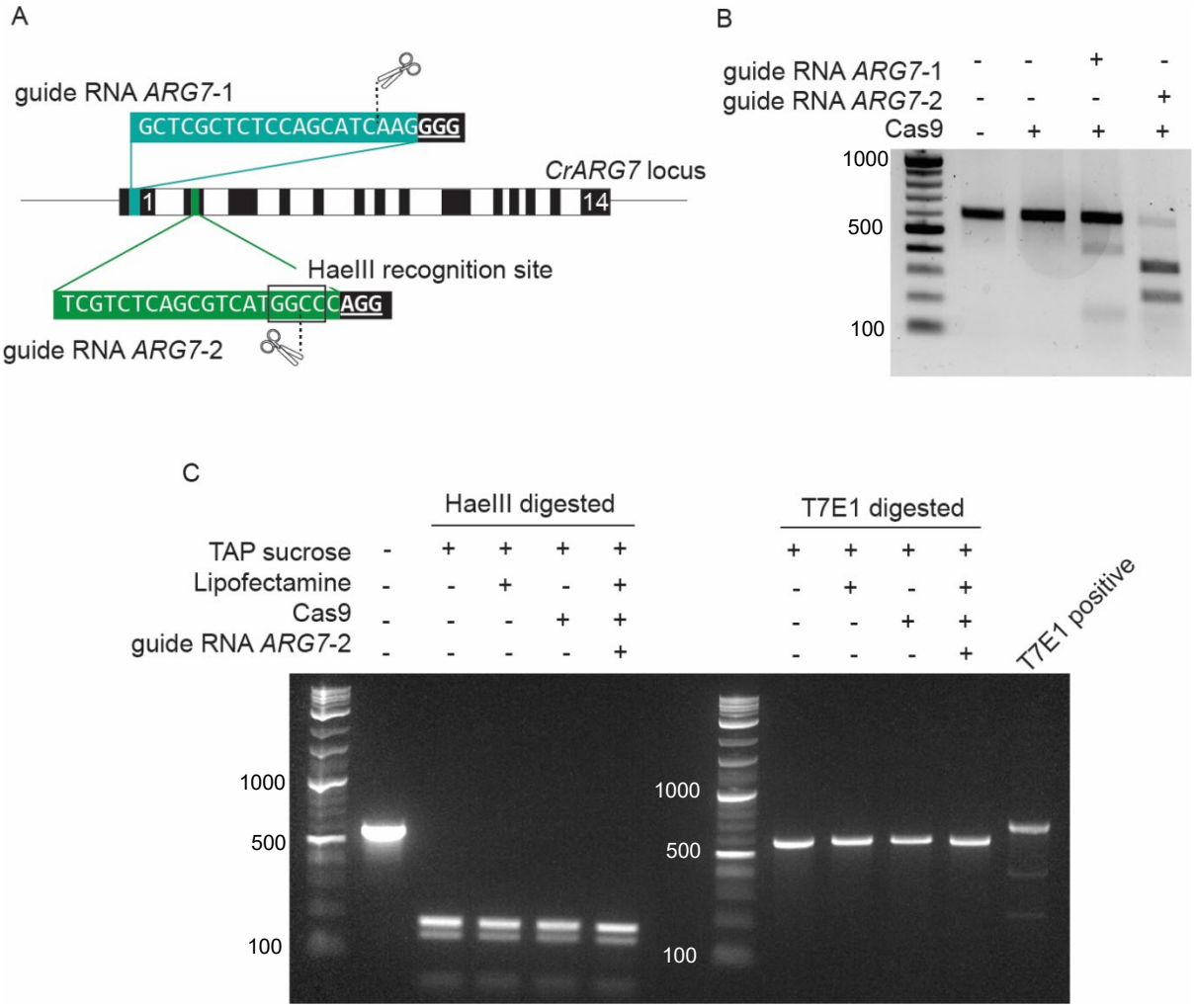


Figure 8. CRISPR-Cas9 RNP design and validation targeting *C. reinhardtii ARG7* gene. (A) Graphic representation of the genetic sequences for RNP-*ARG7*-1 and RNP-*ARG7*-2 guide RNAs showing two 20 nt recognition sites (teal and green, respectively) alongside protospacer adjacent motifs (PAMs) (black, underlined). RNP-*ARG7*-1 targeted the first exon of *C. reinhardtii ChlM* gene and RNP-*ARG7*-2 targeted the second of the 14 exon *ARG7* gene. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break location, which is always three base pairs upstream of the PAM site. (B) *In vitro* digest analysis of RNP-*ARG7*-1 and RNP-*ARG7*-2 activity to validate that RNP components were correctly assembled in vitro and are able to recognise and cut their target regions. The 600 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the presence of smaller bands when incubated with fully assembled CRISPR-Cas9 RNPs

indicating correct assembly. (C) The electrophoretic mobility shift assay following both T7E1 and RFLP analysis showed no CRISPR-Cas9 edited mutants. The electrophoretic assay was performed at three dilutions and intentionally overexposed in order to detect any faint smaller bands.

CONCLUSION

The success of CRISPR-Cas9 editing technology across organisms and fields of molecular biology research has resulted in widespread hype and acceptance that CRISPR technology is ‘quick, easy and cheap’. While this somewhat true, it is only quick, easy and cheap when it can be robustly replicated in numerous laboratories. Given that this is not being demonstrated yet, we worked to examine and categorise the CRISPR-Cas9 ribonucleoprotein (RNP) work flow for editing *C. reinhardtii* to better understand the complexities that have not been discussed in the literature to date. This work flow informed the design of two newly developed RNP-based CRISPR-Cas9 editing approaches for *C. reinhardtii*, as well as two previously published strategies using electroporation. The novel strategies we developed, based on proteolistic delivery, which has been validated in plant models, and lipofection, which is widely used in mammalian cell research, invested a noteworthy amount of time and resources optimising these RNP delivery strategies.

From our work, it is clear that enrichment and screening are crucial nodes in the CRISPR-Cas9 work flow and current studies may not have addressed these issues. For example, even though we targeted an endogenous marker gene that should infer a phenotype change which can be used for both direct screening and direct enrichment; we found that this *ChlM* gene in *C. reinhardtii*, resulted in false positives. This light green, low chlorophyll fluorescence phenotype was detected both visually and by flow cytometry analysis, respectively but T7E1 analysis, RFLP analysis and Sanger sequencing confirmed these were all false positives. Without reliable enrichment or screening, the work-load to optimise CRISPR-Cas9 RNP protocols solely by genotyping becomes unmanageable, both regarding cost and labour intensity.

Even though we faced issues with false positives in both *C. reinhardtii*, we were very surprised to not detect a single *C. reinhardtii* mutant, as this species is theoretically an ideal CRISPR-Cas9 candidate cell line. *C. reinhardtii* is haploid, which overcomes issues regarding heterogeneity in edits, has cell wall deficient mutants amenable for chemical delivery strategies such as lipofection, and is unicellular, which means it does not experience tissue mosaicism associated with more complex eukaryotes such as zebrafish, mice and plants. Instead, only one cell has to be edited for whole organism to be edited. Unfortunately, our results do not help to elucidate at which point the problem is occurring: be it at the intracellular delivery level, the CRISPR-Cas9 double stranded break induction level, or the DNA repair and editing level. This is because of the lack of a reliable endogenous target gene to knock-out. However, this work underscored that there is an urgent need to develop CRISPR-Cas9 protocols for genome editing in microalgae that are reproducible in numerous laboratories. It also suggests and confirms that developing such protocols is more complex than other species.

METHODS

Microalgal strain and culture conditions

C. reinhardtii CC-125 mt+ cells were obtained from GeneArt Chlamydomonas Protein Expression Kit (Invitrogen, A24244) and *C. reinhardtii* CC-503 cw92 mt+ cell-wall deficient cells were obtained from the Chlamy Collection (https://www.chlamycollection.org). *C. reinhardtii* CC-125 mt+ cells were used for all experiments except transformation by lipofection. All *C. reinhardtii* cell lines were cultured mixotrophically in Tris-Acetate-Phosphate (TAP) medium (Harris, 1989) and maintained on 1.5% TAP agar plates. All cultures were incubated at 25°C under continuous illumination of 50 μE m^{-2} s^{-1} with agitation.

CRISPR-Cas9 guide RNA design

Guide RNAs used in this study were designed using Cas-Designer www.rgenome.net/cas-designer (Park et al., 2015) and CRISPOR http://crispor.tefor.net/ (Hsu et al., 2013) according to the following criteria: (i) the target site is located in the first exon; (ii) greater than 60% out-of-frame score (indicating likelihood of cut resulting in a frameshift); (iii) no mismatches and (iv) GC content 25 - 75% (Shin et al., 2016). Once designed, the crispr RNAs were custom synthesized by IDT.

CRISPR-Cas9 *in vitro* assembly and digestion of target DNA

All of the CRISPR components required were obtained from the Alt-R CRISPR Cas9 suite by IDT. This included the *Streptococcus pyogenes* Cas9 protein, containing three nuclear localization sequences (NLS) and a his tag, the tracrRNA and custom synthesized target specific crRNAs, both containing chemical modifications for increased resistance to cellular RNases. The tracrRNA and crRNA were mixed in

equimolar concentrations, heated to 94°C for 2 minutes and allowed to cool to room temperature in order to form the duplex RNA. The duplex RNA was mixed with the Cas9 protein in a 1:1.3 molar ratio and incubated at 37°C for 30 min to allow the formation of the RNP. To confirm the correct assembly and targeting efficiency of the complex, an in vitro digestion of the target DNA by the Cas9 RNP was performed. In order to obtain the target DNA, a region of genomic DNA flanking either end of the target site was amplified by PCR. One hundred nanograms of target DNA or non-target DNA was incubated with 600 ng of Cas9 RNP or Cas9 protein (sans duplex RNA) at 37°C for 2 hours. One microliter of RNase1 was added to the reaction and incubated at 37°C for 20 minutes. DNA loading buffer was added to stop the reaction. The reaction was analysed by electrophoresis.

Co-transformation plasmid for selection

Where appropriate, CRISPR-Cas9 RNP was co-delivered with pChlamy4 plasmid, which contains the *Streptoalloteichus hindustanus bleomycin resistance (shble)* gene conferring zeocin resistance and codon optimised for *C. reinhardtii* (Invitrogen, A24244). The vector was transformed, maintained and cultured in Top10 Chemically Competent *E. coli* (Thermo Fisher, C404010) according to manufacturer's instructions. pChlamy4 was isolated using the Zyppy Plasmid Maxiprep Kit (Zyppy Research, D4028).

Preparation of *C. reinhardtii* cells for transformation

C. reinhardtii cells were cultured for 1-2 days until reaching early to mid log-phase growth, indicated by reaching a cell density of about 1.6 x 10^6 cells/mL (cells were only transformed if they reached a density between 0.8 – 2 x 10^6 cells/mL) in 100 mL aliquots in 250 mL conical flasks. Cell density was determined in two ways; by using

a hemacytometer and by measuring absorbance at 750 nm using the equation $\frac{OD-0.088}{9x10\hat{}-8}$ cells/mL. Cells were harvested by centrifugation at 3,500 rpm for 4 minutes and resuspended according to the appropriate transformation protocol. Cells were transformed with CRISPR-Cas9 RNP in one of three ways: electroporation (co-delivered with pChlamy4 or DNA-free), biolistic bombardment or lipofection.

Electroporation (co-transformation)

Harvested cells were washed in 2 mL of MAX Efficiency Transformation Reagent for Algae (Thermo Fisher Scientific) twice before being resuspended in the appropriate amount of MEB for 75 x 10^6 cells in a 250 µL reaction (300 x 10^6 cells/mL). Cells were added to ice cold 0.2 cm gap cuvette (165-2086, Biorad) along with 10 µg of Cas9 complexed with duplex RNA and 1 µg of undigested pChlamy4 empty vector and incubated at 4°C for 5 minutes. Electroporation was performed using GenePulser with the following conditions: voltage 250 V, capacitance 50 µF, resistance 800 Ω, duration of pulse 15 ms, number of pulses 1. Electroporated cells were recovered in TAP media supplemented with 40 mM sucrose for 16 hours before being spread on TAP plates containing 5 mg/L zeocin (Invivogen, ant-zn-1). This protocol was based on Shin et al. protocol (2016).

Electroporation (DNA-free)

Electroporation was performed as described in Baek et al. (2016) with 10 µg Cas9 RNP-1 and -2 and not the full 200 µg, because this lower concentration still showed successful targeting, as seen in the supplementary material. Electroporated cells were resuspended in TAP media supplemented with 40 mM sucrose and then immediately plated onto 1.5% agar TAP plates without antibiotics. The full reaction volume was

plated across 10 plates (50 µL per plate) so that single colonies could grow without the presence of a selectable marker.

Biolistic delivery (co-transformation)

In order to trial numerous settings for optimal delivery method, we used a fluorescently labeled proxy for the RNP complex, a rabbit anti-bovine IgG-FITC conjugated antibody (Sigma Aldrich, F7887). This allowed us to affordably trial large quantities of protein (hundreds of micrograms per shot) at different conditions, while permitting same-day screening for intracellular protein delivery using flow cytometry. Furthermore, Cas9 protein (163.7 kDa) is very similar in size to an antibody (150 kDa).

Plasmid DNA pChlamy4 (1.5 mg/mL) was concentrated by ethanol precipitation. Tungsten microparticles (Biorad) of 1.1 µm diameter were prepared to 100 µg/mL according to Fabris et al. (2014) and sonicated for 1 hour. For proxy experiments, 100 µg of IgG-FITC conjugated (10 mg/mL) and 10 µg of plasmid DNA (where appropriate) were added to 3 mg of freshly sonicated tungsten microparticles. The DNA and protein was bound to the microparticles by adding 20 µL of fresh 0.1 M Spermidine (Sigma Aldrich, S0226) and 50 µL of filter sterilized $CaCl_2$ (Fabris et al., 2014). Coated microparticles were resuspended in 45 µL of 200 proof molecular grade ethanol (Sigma Aldrich, E7023) and the total suspension was pipetted onto the macrocarriers (BioRad), which were previously sterilized in 100% isopropanol. For chemical-free binding, the microparticles and transformation mixture was placed directly onto macrocarriers and air-dried for two hours. For CRISPR-Cas9 experiments, 100 µg of IgG-FITC conjugated and 100 µg of CRISPR-Cas9 RNP were air-dried to 3 mg of freshly sonicated tungsten microparticles on the macrocarriers.

Once dry, the microparticle coated macrocarriers were assembled into the Gene Gun according to Barrera et al. (2014) described protocol, with the following changes: (i) 650 psi rupture disc was used instead of the recommended 1350 psi (ii) plate holder was placed in second slot from the top.

After bombardment, the cells were left to recover for 1.5 hours. Plates were then scraped and the cells were resuspended in 1 mL TAP media. The cells were washed once in fresh TAP media by centrifugation at 3,500 rpm for 5 minutes. The cell pellet was then resuspended in 250 µL TAP media. For antibody proxy experiments, 125 µL transformed cells were fixed in 2% glutaraldehyde for FACS and flow cytometry analysis. The remaining 125 µL was plated into TAP plates containing 50 mg/L zeocin (TAP-Z) and left to grow for 9 days for single colonies. For CRISPR-Cas9 experiments, cells were sorted by FACS and left to recover for 2 days before analysis.

Lipofection

C. reinhardtii CC-503 cell wall deficient cells were used for lipofection experiments. Here, cells were diluted to a final concentration of 80,000 cells/mL in TAP supplemented with 40 mM sucrose. One hundred microlitres of cells were treated with 250 ng CRISPR-Cas9 RNP complexed with 3 µL of Lipofectamine 3000 (ThermoFisher Scientific, L3000001) following the manufacturer's instructions for a final reaction volume of 50 µL in TAP supplemented with 40 mM sucrose. After 1 hour incubation, 10 µL cells were diluted and plated on freshly prepared TAP plates for single colonies. The remaining transformation mixture was left to increase in biomass for 24 or 48 hours. The biomass was used for flow cytometry analysis and genomic DNA extraction for T7E1 and/or restriction enzyme digest analysis.

Flow cytometry and FACS

Flow cytometry was performed using a BD CytoFLEX S flow cytometer (BD Biosciences). The bombarded, fixed cells were diluted 1:10, and immediately analysed at medium speed until 100,000 events were counted. FITC fluorescence was excited with a 488 nm laser and emission was acquired using a 525/40 nm optical filter. The distribution of two bombardment experiments –wild type cells bombarded with uncoated microparticles and microparticles coated with DNA only– were used as controls to set the gate (R3) for the FITC fluorescent events and (P5) for low chlorophyll events.

PCR based analyses

Genomic DNA was extracted and purified using Monarch PCR & DNA Cleanup Kit (New England Biolabs, Hitchin, UK). PCR screening was performed using GoTaq Flexi DNA polymerase (Promega, Wisconsin, United States) according to the manufacturer's instructions. Forward primer 5'-CTGTCGCTTTATATTTAGGACC-3' and reverse primer 5'-ATCTGCATTAAGATCTGAGG-3' were used to amplify a fragment of *ChlM* which contained both RNP-*ChlM*-1 and -2 target loci. Similarly, 5'-GCACACGTGACCAAATTTATGC-3' and reverse primer 5'-TTGATGTCCTCAGCCCACAG-3' were used to amplify a fragment of *ARG7* which contained both RNP-*ARG7*-1 and -2 target loci. For RFLP was performed as described by Shin et al. (2016) and T7E1 was performed Alt-R® Genome Editing Detection Kit (IDT, 1075931) as per manufacturer's instructions.

SUPPLEMENTARY TABLES

Suppl. Table 1. Details of the guide RNAs designed to target *Chlamydomonas reinhardtii Magnesium protoporphyrin O-methyltransferase (ChlM; XM_001702328)*

Guide RNA name	5' – 3' Sequence (including PAM)	Specificity score*	GC %	Mismatches (0-1-2-3-4)	Exon target	Guide RNA functionality validation
ChlM-1	CCCGCCCGGCTGTGGCCCG G**CGG**	78	90	0 - 0 - 0 - 1 - 0	1 of 7	Shin et al., 2016 and *in vitro* digest
ChlM-2	ATTGCTCAAACTGCCGATGT **TGG**	100	45	0 - 0 - 0 - 0 - 0	1 of 7	*In vitro* digest

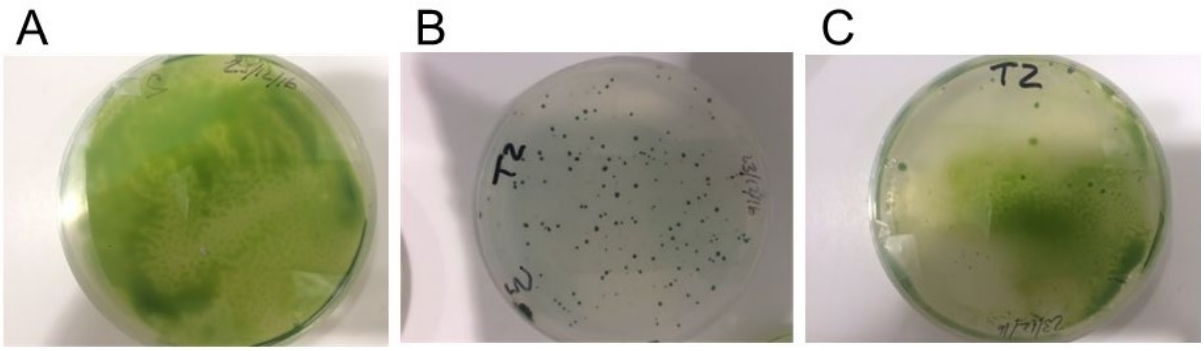
*Specificity score is a measure of uniqueness. The higher the specificity score, the lower are off-target effects in the genome. The specificity score ranges from 0-100 and measures the uniqueness of a guide in the genome (Hsu et al., 2013).

Suppl. Table 2. Details of the guide RNAs designed to target *Chlamydomonas reinhardtii Argininosuccinate lyase / Omega-N-(L-arginino)succinate arginine-lyase(ARG7; X16619.1)*

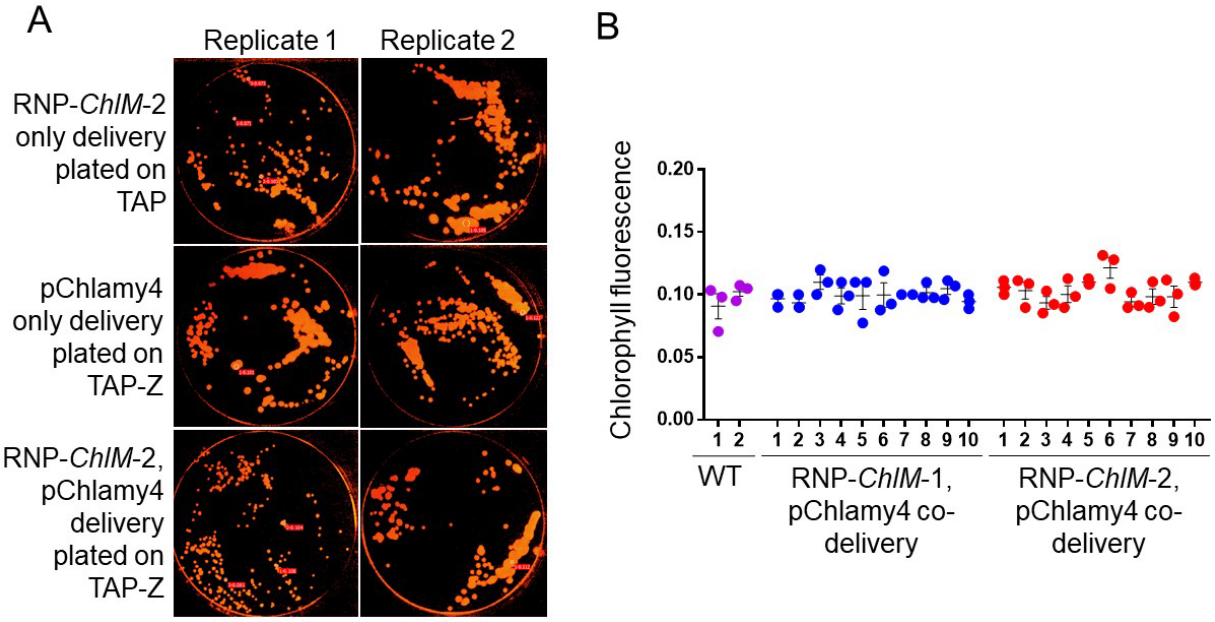
Guide RNA name	5' – 3' Sequence (including PAM)	Specificity score*	GC %	Mismatches (0-1-2-3-4)	Exon target	Guide RNA functionality validation
ARG7-1	GCTCGCTCTCCAGCATCAA G**GGG**	99	50	0 - 0 - 0 - 0 - 0	1 of 14	*In vitro* digest
ARG7-2	TCGTCTCAGCGTCAT*GGCC C***AGG**	98	52.8	0 - 0 - 0 - 0 - 0	2 of 14	*In vitro* digest

*Specificity score is a measure of uniqueness. The higher the specificity score, the lower are off-target effects in the genome. The specificity score ranges from 0-100 and measures the uniqueness of a guide in the genome (Hsu et al., 2013).

SUPPLEMENTARY FIGURES



Suppl. Figure 1. Indirect enrichment and screening of *C. reinhardtii* following electroporation delivery of RNP-*ChlM*-2. (A) Cells electroporated with RNP-*ChlM*-2 only plated on TAP agar plates show normal growth indicating cells were able to survive and grow normally without inhibition of antibiotic treatment. (B) Cells electroporated with pChlamy4 DNA only plated on TAP agar plates supplemented with 50 mg/L zeocin (TAP-Z) show single colony growth as expected. (C) Cells electroporated with RNP-*ChlM*-2 and pChalmy4 DNA plated on TAP-Z agar plates showing no single colonies but instead a thick mat of cells.



Suppl. Figure 2. Fluorometry analysis of *C. reinhardtii* single colonies following co-delivery of RNP-*ChlM*-1 or RNP-*ChlM*-2 with pChlamy4 DNA, where colonies were selected on TAP supplemented with 50 mg/L zeocin (TAP-Z). (A) Photographs of selection plates excited by 450nm light during pulse amplitude modulation (PAM) fluorometry. (B) Chlorophyll fluorescence of three randomly selected colonies per selection plate measured by flow cytometry (n=2 selection plates for wild type (WT) controls; n=10 selection plates for bombarded samples).

REFERENCES

Bachu, R., Bergareche, I., & Chasin, L. A. (2015). CRISPR-Cas targeted plasmid integration into mammalian cells via non-homologous end joining. *Biotechnology and Bioengineering*, *112*(10), 2154–2162. https://doi.org/10.1002/bit.25629

Baek, K., Kim, D. H., Jeong, J., Sim, S. J., Melis, A., Kim, J.-S., … Bae, S. (2016). DNA-free two-gene knockout in Chlamydomonas reinhardtii via CRISPR-Cas9 ribonucleoproteins. *Scientific Reports*, *6*, 30620. https://doi.org/10.1038/srep30620

Baek, K., Yu, J., Jeong, J., Sim, S. J., Bae, S., & Jin, E. S. (2018). Photoautotrophic production of macular pigment in a Chlamydomonas reinhardtii strain generated by using DNA-free CRISPR-Cas9 RNP-mediated mutagenesis. *Biotechnology and Bioengineering*, *115*(3), 719–728. https://doi.org/10.1002/bit.26499

Barrera, D., Gimpel, J., & Mayfield, S. (2014). Rapid Screening for the Robust Expression of Recombinant Proteins in Algal Plastids. *Methods in Molecular Biology (Clifton, N.J.)*, *1132*(March), 401–411. https://doi.org/10.1007/978-1-62703-995-6

Bétermier, M., Bertrand, P., & Lopez, B. S. (2014). Is Non-Homologous End-Joining Really an Inherently Error-Prone Process? *PLoS Genetics*, *10*(1). https://doi.org/10.1371/journal.pgen.1004086

Bortesi, L., & Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances*, *33*(1), 41–52. https://doi.org/10.1016/j.biotechadv.2014.12.006

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., … Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, *155*(7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001

Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., … Sharp, P. A. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, *160*(6), 1246–1260. https://doi.org/10.1016/j.cell.2015.02.038

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., … Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems, (February), 819–824. https://doi.org/10.1126/science.1229223

Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J., & Bao, G. (2014). COSMID: A web-based tool for identifying and validating CRISPR/Cas off-target sites. *Molecular Therapy - Nucleic Acids*, *3*(12), e214. https://doi.org/10.1038/mtna.2014.64

Crespo, J. L., Díaz-Troya, S., & Florencio, F. J. (2005). Inhibition of target of rapamycin signaling by rapamycin in the unicellular green alga Chlamydomonas reinhardtii. *Plant Physiology*, *139*(4), 1736–1749. https://doi.org/10.1104/pp.105.070847

Debuchy, R., Purton, S., & Rochaix, J. D. (1989). The argininosuccinate lyase gene of Chlamydomonas reinhardtii: an important tool for nuclear transformation and for correlating the genetic and molecular maps of the ARG7 locus. *The EMBO Journal*, *8*(10), 2803–2809.

Dicarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J., & Church, G. M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Research*, *41*(7), 4336–4343. https://doi.org/10.1093/nar/gkt135

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., … Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184–191.

https://doi.org/10.1038/nbt.3437

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., … Goossens, A. (2014). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *The New Phytologist*, 521–535. https://doi.org/10.1111/nph.12917

Feng, Z., Zhang, B., Ding, W., Liu, X., Yang, D.-L., Wei, P., … Zhu, J.-K. (2013). Efficient genome editing in plants using a CRISPR/Cas system. *Cell Research*, *23*(10), 1229–1232. https://doi.org/10.1038/cr.2013.114

Gaj, T., Gersbach, C. A., & Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, *31*(7), 397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., & Hegemann, P. (2017). Targeting of Photoreceptor Genes in Chlamydomonas reinhardtii via Zinc- finger Nucleases and CRISPR/Cas9. *Plant Cell Advance Publication. Published on October*, *4*. https://doi.org/10.1105/tpc.17.00659

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J. B., … Concordet, J. P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, *17*(1), 1–12. https://doi.org/10.1186/s13059-016-1012-2

He, X., Tan, C., Wang, F., Wang, Y., Zhou, R., Cui, D., … Feng, B. (2016). Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. *Nucleic Acids Research*, *44*(9), 1–14. https://doi.org/10.1093/nar/gkw064

Horii, T., Arai, Y., Yamazaki, M., Morita, S., Kimura, M., Itoh, M., … Hatada, I. (2014). Validation of microinjection methods for generating knockout mice by CRISPR/Cas-mediated genome engineering. *Scientific Reports*, *4*(2), 1–6. https://doi.org/10.1038/srep04513

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., … Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832. https://doi.org/10.1038/nbt.2647

Huang, W., & Daboussi, F. (2017). Genetic and metabolic engineering in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160411. https://doi.org/10.1098/rstb.2016.0411

Jeon, S., Lim, J.-M., Lee, H.-G., Shin, S.-E., Kang, N. K., Park, Y.-I., … Chang, Y. K. (2017). Current status and perspectives of genome editing technology for microalgae. *Biotechnology for Biofuels*, *10*(1), 267. https://doi.org/10.1186/s13068-017-0957-z

Jiang, W., Brueggeman, A. J., Horken, K. M., Plucinak, T. M., & Weeks, D. P. (2014). Successful Transient Expression of Cas9 and Single Guide RNA Genes in Chlamydomonas reinhardtii, *13*(11), 1465–1469. https://doi.org/10.1128/EC.00213-14

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA – Guided, *337*(August), 816–822.

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., … Ecker, J. R. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics*, *15*(1), 1–25. https://doi.org/10.1371/journal.pgen.1007819

Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., … Weiss, R. (2014). CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nature Methods*, *11*(7), 723–726. https://doi.org/10.1038/nmeth.2969

Kim, H., & Kim, J. S. (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, *15*(5), 321–334. https://doi.org/10.1038/nrg3686

Kim, S., Kim, D., Cho, S. W., Kim, J., & Kim, J.-S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Research*, *128*, 1–32. https://doi.org/10.1101/gr.171322.113.Freely

Kohli, A., González-Melendi, P., Abranches, R., Capell, T., Stoger, E., & Christou, P. (2006). The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signaling and Behavior*, *1*(4), 185–195. https://doi.org/10.4161/psb.1.4.3195

Kohli, A., Miro, B., & Twyman, R. M. (2010). *Transgene Integration , Expression and Stability in Plants : Strategies for Improvements*. https://doi.org/10.1007/978-3-642-04809-8

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., & Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature Biotechnology*, *32*(7), 677–683. https://doi.org/10.1038/nbt.2916

Lee, K., Conboy, M., Park, H. M., Jiang, F., Kim, H. J., Dewitt, M. A., … Murthy, N. (2017). Nanoparticle delivery of Cas9 ribonucleoprotein and donor DNA in vivo induces homology-directed DNA repair. *Nature Biomedical Engineering*, *1*(11), 889–901. https://doi.org/10.1038/s41551-017-0137-2

Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., … Chesnut, J. D. (2015). Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *Journal of Biotechnology*, *208*, 44–53. https://doi.org/10.1016/j.jbiotec.2015.04.024

Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., … Gao, C. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nature Communications*, *8*(January), 1–5. https://doi.org/10.1038/ncomms14261

Liang, Z., Chen, K., Zhang, Y., Liu, J., Yin, K., Qiu, J.-L., & Gao, C. (2018). Genome editing of bread wheat using biolistic delivery of CRISPR/Cas9 in vitro transcripts or ribonucleoproteins. *Nature Protocols*, *13*(3), 413–430. https://doi.org/10.1038/nprot.2017.145

Lin, C. S., Hsu, C. T., Yang, L. H., Lee, L. Y., Fu, J. Y., Cheng, Q. W., … Shih, M. C. (2018). Application of protoplast technology to CRISPR/Cas9 mutagenesis: from single-cell mutation detection to mutant plant regeneration. *Plant Biotechnology Journal*, *16*(7), 1295–1310. https://doi.org/10.1111/pbi.12870

Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., … Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research*, *42*(11), 7473–7485. https://doi.org/10.1093/nar/gku402

Liu, J., Gerken, H., Huang, J., & Chen, F. (2013). Engineering of an endogenous phytoene desaturase gene as a dominant selectable marker for Chlamydomonas reinhardtii transformation and enhanced biosynthesis of carotenoids. *Process Biochemistry*, *48*(5–6), 788–795. https://doi.org/10.1016/j.procbio.2013.04.020

Mages, W., Heinrich, O., Treuner, G., Vlcek, D., Daubnerova, I., & Slaninova, M. (2007). Complementation of the Chlamydomonas reinhardtii arg7-8 (arg2) Point Mutation by Recombination with a Truncated Nonfunctional ARG7 Gene. *Protist*, *158*(4), 435–446. https://doi.org/10.1016/j.protis.2007.05.001

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., … Church, G. M. (2013).

RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826. https://doi.org/10.1126/science.1232033

Martin-Ortigosa, S., & Wang, K. (2014). Proteolistics: a biolistic method for intracellular delivery of proteins. *Transgenic Research*, *23*(5), 743–756. https://doi.org/10.1007/s11248-014-9807-y

Mashiko, D., Fujihara, Y., Satouh, Y., Miyata, H., Isotani, A., & Ikawa, M. (2013). Generation of mutant mice by pronuclear injection of circular plasmid expressing Cas9 and single guided RNA. *Scientific Reports*, *3*, 1–6. https://doi.org/10.1038/srep03355

Mcvey, M., Lee, S. E., Avenue, H., & Antonio, S. (2017). MMEJ repair of double-strand breaks: deleted sequences and alternative endings. *Trends Genet.*, *24*(11), 529–538. https://doi.org/10.1016/j.tig.2008.08.007.MMEJ

Meinecke, L., Alawady, A., Schroda, M., Willows, R., Kobayashi, M. C., Niyogi, K. K., … Beck, C. F. (2010). Chlorophyll-deficient mutants of Chlamydomonas reinhardtii that accumulate magnesium protoporphyrin IX. *Plant Molecular Biology*, *72*(6), 643–658. https://doi.org/10.1007/s11103-010-9604-9

Morgens, D. W., Wainberg, M., Boyle, E. A., Ursu, O., Araya, C. L., Kimberly Tsui, C., … Bassik, M. C. (2017). Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nature Communications*, *8*(May), 1–8. https://doi.org/10.1038/ncomms15178

Moses, T., Mehrshahi, P., Smith, A. G., & Goossens, A. (2017). Synthetic biology approaches for the production of plant metabolites in unicellular organisms. *Journal of Experimental Botany*, *68*(15), 4057–4074. https://doi.org/10.1093/jxb/erx119

Naduthodi, M. I. S., Barbosa, M. J., & van der Oost, J. (2018). Progress of CRISPR-Cas based genome editing in Photosynthetic microbes. *Biotechnology Journal*, 1700591. https://doi.org/10.1002/biot.201700591

Noman, A., Aqeel, M., & He, S. (2016). CRISPR-Cas9: Tool for Qualitative and Quantitative Plant Genome Editing. *Frontiers in Plant Science*, *7*(November), 1–17. https://doi.org/10.3389/fpls.2016.01740

Park, J., Bae, S., & Kim, J. (2015). Sequence analysis Cas-Designer : A web-based tool for choice of CRISPR-Cas9 tar- get sites. *Bioinformatics*, *31*(July), 1–3. https://doi.org/10.1101/005074.Bae

Patel, V. K., Soni, N., Prasad, V., Sapre, A., Dasgupta, S., & Bhadra, B. (2019). CRISPR–Cas9 System for Genome Engineering of Photosynthetic Microalgae. *Molecular Biotechnology*, *61*(8), 541–561. https://doi.org/10.1007/s12033-019-00185-3

Pennisi, E. (2013). The CRISPR craze. *Science*, *341*(6148), 833–836. https://doi.org/10.1126/science.341.6148.833

Pflueger, C., Tan, D., Swain, T., Nguyen, T., Pflueger, J., Nefzger, C., … Lister, R. (2018). A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Research*, *28*(8), 1193–1206. https://doi.org/10.1101/gr.233049.117

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M. A., … Gunawardane, R. N. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular Biology of the Cell*, *28*(21), 2854–2874. https://doi.org/10.1091/mbc.E17-03-0209

Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, M. R., & Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, P. H. (2007). CRISPR Provides Acquired

Resistance Against Viruses in Prokaryotes. *Science*, *315*(March), 1709–1712. https://doi.org/10.1126/science.1138140

Sander, J. D., & Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, *32*(4), 347–355. https://doi.org/10.1038/nbt.2842

Seki, A., & Rutz, S. (2018). Optimized RNP transfection for highly efficient CRISPR / Cas9-mediated gene knockout in primary T cells. *Journal of Experimental Medecine*, 1–13. https://doi.org/10.1084/jem.20171626

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., & Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-06378-9

Shalem. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells, *343*(January), 84–88.

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Jeong, B. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii SUPP. *Scientific Reports*, *6*(April), 27810. https://doi.org/10.1038/srep27810

Singh, R., Kuscu, C., Quinlan, A., Qi, Y., & Adli, M. (2015). Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Research*, *43*(18), 1–8. https://doi.org/10.1093/nar/gkv575

Sizova, I., Greiner, A., Awasthi, M., Kateriya, S., & Hegemann, P. (2013). Nuclear gene targeting in Chlamydomonas using engineered zinc-finger nucleases. *Plant Journal*, *73*(5), 873–882. https://doi.org/10.1111/tpj.12066

Spicer, A., & Molnar, A. (2018). Gene Editing of Microalgae: Scientific Progress and Regulatory Challenges in Europe. *Biology*, *7*(1), 21. https://doi.org/10.3390/biology7010021

Sternberg, S. H., & Doudna, J. A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, *58*(4), 568–574. https://doi.org/10.1016/j.molcel.2015.02.032

Stovicek, V., Holkenbrink, C., & Borodina, I. (2017). CRISPR/Cas system for yeast genome engineering: advances and applications. *FEMS Yeast Research*, *17*(5), 1–16. https://doi.org/10.1093/femsyr/fox030

Thakore, P. I., Song, L., Safi, A., Shivakumar, K., Kabadi, A. M., Reddy, T. E., … Gersbach, C. A. (2015). Highly Specific Epigenome Editing by CRISPR/Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nature Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630.Highly

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., … Joung, J. K. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, *33*(2), 187–198. https://doi.org/10.1038/nbt.3117

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., … May, A. P. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, *63*(4), 633–646. https://doi.org/10.1016/j.molcel.2016.06.037

Vítová, M., Bišová, K., Umysová, D., Hlavová, M., Kawano, S., Zachleder, V., & Čížková, M. (2011). Chlamydomonas reinhardtii: Duration of its cell cycle and phases at growth rates affected by light intensity. *Planta*, *233*(1), 75–86. https://doi.org/10.1007/s00425-010-1282-y

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., & Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/cas-mediated genome engineering. *Cell*, *153*(4), 910–918. https://doi.org/10.1016/j.cell.2013.04.025

Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., … Yee, J. K. (2015). Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nature Biotechnology*, *33*(2), 175–179. https://doi.org/10.1038/nbt.3127

Wendt, K. E., Ungerer, J., Cobb, R. E., Zhao, H., & Pakrasi, H. B. (2016). CRISPR/Cas9 mediated targeted mutagenesis of the fast growing cyanobacterium Synechococcus elongatus UTEX 2973. *Microbial Cell Factories*, *15*(1), 1–8. https://doi.org/10.1186/s12934-016-0514-7

Woo, J. W., Kim, J., Kwon, S. Il, Corvalán, C., Cho, S. W., Kim, H., … Kim, J. (2015). DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins, *33*(11), 1162–1165. https://doi.org/10.1038/nbt.3389

Xue, H. Y., Zhang, X., Wang, Y., Xiaojie, L., Dai, W. J., & Xu, Y. (2016). In vivo gene therapy potentials of CRISPR-Cas9. *Gene Therapy*, *23*(7), 557–559. https://doi.org/10.1038/gt.2016.25

Yi Xin and Cunming Duan. (2018). Microinjection of Antisense Morpholinos, CRISPR/Cas9 RNP, and RNA/DNA into Zebrafish Embryos. *Sex and the Brain*, (January). https://doi.org/10.7551/mitpress/7458.003.0026

Yu, X., Liang, X., Xie, H., Kumar, S., Ravinder, N., Potter, J., … Chesnut, J. D. (2016). Improved delivery of Cas9 protein/gRNA complexes using lipofectamine CRISPRMAX. *Biotechnology Letters*, *38*(6), 919–929. https://doi.org/10.1007/s10529-016-2064-9

Zuris, J. A., Thompson, D. B., Shu, Y., Guilinger, J. P., Bessen, J. L., Hu, J. H., … Liu, D. R. (2014). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature Biotechnology*, *33*(1), 73–80. https://doi.org/10.1038/nbt.3081

Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology

Published in *Frontiers Biotechnology and Bioengineering* 5 June 2020

George, J., Kahlke, T., Abbriano, R. M., Kuzhiumparambil, U., Ralph, P. J., & Fabris, M. (2020). Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology. *Frontiers in Bioengineering and Biotechnology*, 8 (June), 1–19. https://doi.org/10.3389/fbioe.2020.00513

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 June 2020
doi: 10.3389/fbioe.2020.00513

Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology

Jestin George[1], *Tim Kahlke*[1], *Raffaela M. Abbriano*[1], *Unnikrishnan Kuzhiumparambil*[1], *Peter J. Ralph*[1] *and Michele Fabris*[1,2]*

[1] University of Technology Sydney, Climate Change Cluster, Faculty of Science, Ultimo, NSW, Australia, [2] CSIRO Synthetic Biology Future Science Platform, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Maurycy Daroch, Peking University, China

Reviewed by:
Konstantinos Vavitsas, National and Kapodistrian University of Athens, Greece
Manuel Serif, Norwegian University of Science and Technology, Norway

***Correspondence:**
Michele Fabris
michele.fabris@uts.edu.au

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *16 December 2019*
Accepted: *30 April 2020*
Published: *05 June 2020*

Citation:
George J, Kahlke T, Abbriano RM, Kuzhiumparambil U, Ralph PJ and Fabris M (2020) Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology. Front. Bioeng. Biotechnol. 8:513. doi: 10.3389/fbioe.2020.00513

Diatoms are photosynthetic microeukaryotes that dominate phytoplankton populations and have increasing applicability in biotechnology. Uncovering their complex biology and elevating strains to commercial standards depends heavily on robust genetic engineering tools. However, engineering microalgal genomes predominantly relies on random integration of transgenes into nuclear DNA, often resulting in detrimental "position-effects" such as transgene silencing, integration into transcriptionally-inactive regions, and endogenous sequence disruption. With the recent development of extrachromosomal transgene expression via independent episomes, it is timely to investigate both strategies at the phenotypic and genomic level. Here, we engineered the model diatom *Phaeodactylum tricornutum* to produce the high-value heterologous monoterpenoid geraniol, which, besides applications as fragrance and insect repellent, is a key intermediate of high-value pharmaceuticals. Using high-throughput phenotyping we confirmed the suitability of episomes for synthetic biology applications and identified superior geraniol-yielding strains following random integration. We used third generation long-read sequencing technology to generate a complete analysis of all transgene integration events including their genomic locations and arrangements associated with high-performing strains at a genome-wide scale with subchromosomal detail, never before reported in any microalga. This revealed very large, highly concatenated insertion islands, offering profound implications on diatom functional genetics and next generation genome editing technologies, and is key for developing more precise genome engineering approaches in diatoms, including possible genomic safe harbour locations to support high transgene expression for targeted integration approaches. Furthermore, we have demonstrated that exogenous DNA is not integrated inadvertently into the nuclear genome of extrachromosomal-expression clones, an important characterisation of this novel engineering approach that paves the road to synthetic biology applications.

Keywords: microalgae, *Phaeodactylum tricornutum*, extrachromosomal expression, random integration, long-read sequencing, integration islands, heterologous monoterpenoids, synthetic biology

INTRODUCTION

Diatoms are a diverse group of unicellular Stramenopile microalgae that have received substantial attention for their ecological importance (Armbrust, 2009) and biotechnological potential (Huang and Daboussi, 2017). Newly developed genetic resources hold much promise for diatom functional genetics studies and have propelled the model pennate diatom *Phaeodactylum tricornutum* into the field of synthetic biology. Firstly, next-generation genetic engineering tools have now been established in *P. tricornutum*. This includes targeted genetic engineering via transcription activator-like effector nucleases (TALENs) (Daboussi et al., 2014; Weyman et al., 2015; Serif et al., 2017) and CRISPR-Cas9 (Nymark et al., 2016; Serif et al., 2018; Sharma et al., 2018) techniques with wide applications for biotechnology and basic research. Next-generation engineering strategies also include the recently demonstrated extrachromosomal approach. Here, potentially large episomes that contain various DNA parts can be maintained and expressed without requiring genomic integration (Karas et al., 2015a). Consequently, extrachromosomal transformation is anticipated to become increasingly widely-used in diatom genetic engineering (Huang and Daboussi, 2017). These next-generation engineering strategies are central to diatom genetics and synthetic biology primarily because they allow multi-gene stacking approaches (Goyal et al., 2009; Ainley et al., 2013). Secondly, the recently developed Universal Loop (uLoop) assembly kit provides a collection of useful parts for modular DNA assembly and high-throughput testing as well as more complex gene-stacking designs for *P. tricornutum* and other diatoms (Pollak et al., 2019). Altogether, these resources offer unparalleled potential of diatoms such as *P. tricornutum* compared to other model algal chassis.

Further to these developments, intrinsic desirable biological traits have elevated this microbe as a promising alternative to well-established chassis, *Escherichia coli* and *Saccharomyces cerevisiae*. Such traits include its robustness and scalability for industrial-scale growth (Hamilton et al., 2015); and—unlike bacteria and yeast species—its ability to fix carbon via photosynthesis for cheaper culture conditions. There is also an increased availability of transcriptomic, metabolomic, and proteomic datasets for uncovering previously unknown traits in diatoms (Ashworth et al., 2016; Longworth et al., 2016; Remmers et al., 2018; Smith et al., 2019) including unique aspects with potential biotechnological relevance (Kroth et al., 2008; Allen et al., 2011; Fabris et al., 2012, 2014).

Phaeodactylum tricornutum is poised to become a widely-used, reliable chassis organism and has been validated in various high interest biotechnological applications, including in the production of bioplastic precursor compounds (Hempel et al., 2011a), therapeutic antibodies (Hempel et al., 2011b; Hempel and Maier, 2012), biofuels (Yao et al., 2014) and nutritional supplements (Hamilton et al., 2014). However, these approaches have all relied on the first generation genetic engineering strategy of randomly integrated chromosomal expression (RICE) of exogenous DNA. While RICE has been crucial for generating a myriad of transgenic *P. tricornutum* strains, both for basic (Lavaud et al., 2012; Liu et al., 2016) and applied research (Hempel and Maier, 2012); it is beset with silencing issues resulting in low transgene expression and stability (Cerutti et al., 2011). This is largely attributed to the "position effect" phenomenon, whereby transgenes integrate into regions in the genome that are unfavourable for transgene expression (Gangl et al., 2015; Doron et al., 2016; Huang and Daboussi, 2017), such as transcriptionally repressed regions (Elgin, 1996). In microalgal research, there is virtually no information regarding the mechanisms driving and regulating RICE. Uncovering this knowledge is important for better understanding of a strategy that is still widely used today, including for CRISPR-Cas9 targeted genome editing (Hopes et al., 2016; Nymark et al., 2016; Greiner et al., 2017).

In order for next-generation engineering tools to deliver a variety of synthetic biology applications in *P. tricornutum* and to replace RICE, they need to be better characterised. For example, even though targeted integration has been demonstrated in this species (Weyman et al., 2015), there are no known safe harbour loci–regions in the nuclear genome that facilitate stable, high transgene expression–identified in *P. tricornutum* or any other eukaryotic photosynthetic microbe. Similarly, non-integrative episomes are extremely appealing for synthetic biology applications as backbones of self-maintaining mini-chromosomes. However, there is little knowledge available regarding mechanisms of episomal maintenance (Diner et al., 2016), including the level of transgene expression that can be achieved by extrachromosomal expression (EE), and whether fragments of episomal DNA are inadvertently integrated into the nuclear genome. This is because EE technology has only recently been described in diatoms with limited examples of its use to express transgenes. Resolving these knowledge gaps will enable understanding of how different genetic engineering strategies may alter the diatom's biology and DNA integration patterns for developing more reliable next-generation engineering approaches required for complex synthetic biology.

Given the developments for diatom synthetic biology, *P. tricornutum* is now being explored for its potential for heterologous terpenoid production (Vavitsas et al., 2018). *Phaeodactylum tricornutum* is a promising alternative chassis compared to the more widely used bacteria and yeast species, which require extensive engineering to increase relatively low flux to isoprenoid biosynthesis (Zurbriggen et al., 2012; Paddon et al., 2013; Bian et al., 2017; Wang et al., 2017). Until recently, *P. tricornutum* had only been metabolically engineered to produce heterologous terpenoids betulin and lupeol (D'Adamo et al., 2018) by RICE. However, we have since demonstrated the first use of EE for metabolic engineering in *P. tricornutum* (Fabris et al., 2020) by expressing a geraniol synthase from the medicinal plant *Catharanthus roseus* (EC 3.1.7.11, *CrGES*) for the heterologous production of geraniol. Geraniol is a commercially relevant monoterpenoid with a variety of applications as flavourant, fragrances, and insect repellent (Chen and Viljoen, 2010). Geraniol is also the first intermediate in the monoterpenoid

Abbreviations: EE, Episomal expression; RICE, Randomly integrated chromosomal expression.

indole alkaloids (MIAs) biosynthesis pathway, which in *C. roseus* leads to the synthesis of the very high-value products with pharmaceutical applications (Van Moerkercke et al., 2013; Caputi et al., 2018).

Herein, with the aim of laying the basis for more sophisticated synthetic biology and metabolic engineering strategies, we generated and thoroughly profiled libraries of transgenic *P. tricornutum* cell lines engineered to produce geraniol via either EE or RICE. By adopting high-throughput phenotyping, we unveiled important intrinsic differences both between and within EE and RICE cell lines. This revealed a small selection of high-performing RICE strains, as well as a highly consistent transgene expression phenotype across EE strains. We used long-read DNA sequencing to interrogate the genomes of selected EE and RICE lines. Our results provide a complete analysis of all integration events, genomic locations, and transgene arrangements associated with high-performing RICE strains at both genome-wide and subchromosomal scale, never reported before in any microalga, and confirmed the non-integrative nature of EE.

Our findings are key for understanding the underexplored dynamics of EE in diatoms, to be used as the basis for gene-stacking based synthetic biology applications such as metabolic pathways and complex genetic circuit assembly and expression. They also highlight the importance of moving to next-generation genetic engineering strategies in *P. tricornutum* and lay the groundwork for identifying putative genomic safe harbour locations that support high transgene expression for targeted integration strategies.

METHODS

Microbial Strains and Growth Conditions

Phaeodactylum tricornutum CCAP1055/1 was grown in liquid ESAW (Berges et al., 2001) supplemented with 50 μg/mL zeocin (Invivogen, San Diego, CA, USA) where appropriate, under 100 μE m^{-2} s^{-1} light in 21 °C shaking at 95 rpm. *Phaeodactylum tricornutum* induction media was prepared following ESAW protocol but without any addition of phosphate. *Escherichia coli* was grown in Luria broth supplemented with 100 μg/mL ampicillin.

Cloning and Genetic Construct Assembly

Plasmids and episomes were constructed using Gibson assembly cloning kit (New England Biolabs, Hitchin, UK). Plasmids were propagated in *E. coli* strain Top10 and purified by Monarch Plasmid Miniprep Kit (New England Biolabs, Hitchin, UK). PCR amplification was performed using Q5 high fidelity polymerase (New England Biolabs, Hitchin, UK) and PCR screening was performed using GoTaq Flexi DNA polymerase (Promega, Wisconsin, United States) according to the manufacturer's instructions. Plasmid coding sequences were validated by Sanger sequencing (Macrogen Korea, Seoul, Korea). Episomes *pPtPBR11_AP1p_CrGES-mVenus* (mVenus NCBI Accesssion: AAZ65844.1) and a control *pPtPBR11_AP1p_mVenus* are described in Fabris et al. (2020). Expression plasmid for chromosomal integration, *pUC19_AP1p_CrGES-mVenus,* was ligated by Gibson assembly into the pUC19 cloning vector linearised with BamHI. Both the *AP1p_CrGES-mVenus_FCBPt* expression cassette and the *FCBPp_ShBle_FCBPt* zeocin resistant cassette were amplified from *pPtPBR11_AP1p_CrGES-mVenus* using 5′-tcgagctcggtacccgggCTAACAGGATTAGTGC AATTC-3′ forward primer and 5′-aggtcgactAGACGAGCTA GTGTTATTC-3′ reverse primer; and 5′-agctcgtctAGTCGACC TGCACATATG-3′ forward primer and 5′-tgcaggtcgactctaga gAGACGAGCTAGTGTTATTC-3′ reverse primer, respectively. Similarly, *pUC19_AP1p_mVenus* expression plasmid for genomic integration was ligated by Gibson assembly into the pUC19 cloning vector linearised with BamHI. The *AP1p_mVenus_FCBPt* expression cassette was amplified using the same primers for amplifying *AP1p_CrGES-mVenus_FCBPt* expression cassette from *pPtPBR11_AP1p_mVenus* episome. Sequences of the vectors used in this work are provided in **Supplementary File 2**.

Diatom Transformation and Conjugation

Phaeodactylum tricornutum was transformed by biolistic bombardment using PDS-1000/He System with Hepta adapter (Bio-Rad) for nuclear genomic integration of plasmid DNA (Kroth, 2007). Afterward, the cell mixture was left to recover 12 h before being scraped and plated onto fresh ½ ESAW 50 μg/mL zeocin agar plates and left for 4–5 weeks for single colonies to appear. *E. coli* containing *pTA-Mob* (Karas et al., 2015a) and *pPtPBR11_AP1p_CrGES-mVenus* or *pPtPBR11_AP1p_mVenus* plasmids (Fabris et al., 2020) were used for conjugation with *P. tricornutum* according to protocol described by Diner et al. (2016). The cell mixture was scraped and plated onto 3–5 fresh ½ ESAW zeocin agar plates and left for 10–15 days when single colonies appeared. Single colonies generated by biolistic bombardment and conjugation were picked and inoculated into individual wells of 96-well round bottom plates containing 200 μl of ESAW supplemented with 50 μg/mL zeocin. The EE and RICE generated cell lines were incubated at 21°C with 100 μE m^{-2} s^{-1} light for 1 week to adjust to liquid growth, after which they were subcultured every 4 days. For high-throughput screening, cell lines were subcultured over 3 weeks in ESAW supplemented with 50 μg/mL zeocin (or without supplementation for stability analysis), and induced with phosphate-free ESAW for 24 h before flow cytometry.

Flow Cytometry and Fluorescence-Activated Cell Sorting (FACS)

Once antibiotic resistant single colonies were established in liquid culture, they were used to inoculate fresh 200 μL of ESAW with and without zeocin. Cell lines were subcultured 1:7 (v:v) every 4 days for 3 weeks under these conditions. On day 4 of culture, plates were centrifuged at 2,300 g for 3 min to pellet cells. The supernatant was removed and cells were washed with 150 μL induction media twice before being resuspended in 200 μL of induction media and induced for 24 hrs. Induced cells were screened by flow cytometry using CytoFLEX S (Beckman Coulter). Fluorescence was excited using

a 488 nm laser. mVenus fluorescence was detected using a 525/40 nm filter and chlorophyll fluorescence was detected using 690/50 nm filter. Compensation of chlorophyll channel was set to 0.3. The distribution of eight single colonies of wild type *P. tricornutum* cultured and induced in the same way as transgenic cell lines were used as controls to determine the mVenus auto fluorescence for screening. Only chorophyll positive cells were included in the analysis to account for cell debris and other background events.

The per cell mVenus fluorescence of 20,000 events was log normalised and violin plots were created in Python using the seaborn data visualisation library (v. 0.9.0) (Allen et al., 2018). Cell lines selected for geraniol production analysis were sorted using BD Influx FACS (BD Biosciences). All cell lines were cultured and induced as described above prior to FACS. Yellow mVenus fluorescence was detected using a 488 nm laser for excitation and a 530/40 nm filter. Wild type *P. tricornutum* cultured and induced in the same way as the transformants, was used as a control to determine the mVenus auto fluorescence distribution. A preliminary screen of a pooled sample of the top eight RICE_mV transformants was used to define the mVenus positive gate which did not overlap with the wild type control. One thousand cells from each transformant cell line that fell into this gate were collected into 96-well round bottom plate wells containing 200 μL ESAW media. After sorting, cells were incubated for 1–2 weeks in 200 μL ESAW supplemented with zeocin 50 μg/mL in 96-well round bottom plate, after which they were subcultured as described previously.

Geraniol Capture and Analysis

Cell lines analysed for geraniol production were scaled up to 50 mL preculture in ESAW supplemented with zeocin 50 μg/mL. Precultures were used to inoculate 50 mL fresh ESAW media without zeocin in 250 mL shake flasks at 10,000 cells/mL density on Day 0. Cultures in late-exponential phase were induced in the presence of isopropyl myristate ($C_{17}H_{34}O_2$) to capture volatile monoterpenoids (Jiang et al., 2017). On Day 4 the total culture was collected and centrifuged at 3,000 g for 4 min to pellet cells. Cells were resuspended in 4 mL induction media and washed twice in induction media before being resuspended in 30 mL induction media in fresh 250 mL shake flasks with 1.6 mL isopropyl myristate, which was harvested 72 h after induction and stored −80°C until being analysed by GCMS. Geraniol was captured, sampled and analysed as described in Fabris et al. (2020).

High Molecular Weight Genomic DNA (gDNA) Extraction

Genomic DNA from *P. tricornutum* transformants was extracted using 7×10^7 cells in 4–6 extractions to obtain ∼7.5 μg high molecular weight, purified gDNA (dx.doi.org/10.17504/protocols.io.qzudx6w). The DNA was resuspended in 25–45 μL ultrapure water overnight at room temperature.

MinION Sequencing

MinION sequencing libraries were prepared according to the 1D Genomic DNA by ligation (SQK-LSK108) protocol supplied by the MinION manufacturer (Oxford Nanopore Technologies) with modifications. Briefly, the DNA fragmentation step was replaced with two bead-cleaning steps. An initial 1:0.1 bead clean of gDNA to GC Biotech CleanNGS (CNGS-0005) beads was performed and the sample was gently mixed by flicking, and was slowly repeatedly inverted for 5 min, and pelleted on a magnetic rack to collect the supernatant. A second 1:1 bead clean was performed using the supernatant and beads. The sample was gently mixed by flicking, incubated by slowly rotating by hand for 5 min, and pelleted on a magnetic rack. The supernatant was removed and DNA bound to the beads was washed with 200 μL freshly prepared 70% ethanol twice without removing the tube off the magnet or disturbing the pellet. The beads and DNA were resuspended in 46 μL ultrapure water, incubated at room temperature for 5 min, the beads were pelleted on the magnet and the supernatant containing the DNA was collected and used according to the manufacturer protocol. Samples were sequenced until a coverage of seven to ten times was achieved. Raw reads were base called and quality filtered using albacore v2.2.6 with default settings. All the sequencing data has been deposited in NCBI BioProject under the ID PRJNA593624.

Identifying Transgene Integration Locations in Nuclear Genome

To identify reads which contain RICE plasmid DNA and remove reads made up of genomic DNA only, all reads were aligned against *pUC19_AP1p_CrGES-mVenus* RICE plasmid using BLAST. Reads that did align to RICE plasmid were defined as initial hits. To account for regions in *pUC19_AP1p_CrGES-mVenus* RICE plasmid that contain native *P. tricornutum* genomic regions, such as promoters or terminators, the *pUC19_AP1p_CrGES-mVenus* RICE plasmid sequence was aligned against *P. tricornutum* genome using BLAST (EnsemblProtists, ASM15095v2). Initial hits that only aligned to those regions of *pUC19_AP1p_CrGES-mVenus* that are native to *P. tricornutum* were filtered out as false-positive hits using custom awk commands. The resulting true-positive hits were manually checked to identify the chromosomes of the integration sites. For detailed analyses all reads were mapped to each of the matching chromosomes using bwa v0.7.15 (Li and Durbin, 2009). The resulting sam files were sorted, converted to bam-format and indexed using samtools 1.3.1 (Li et al., 2009) and potential integration sites were manually checked using the Integrative Genomics Viewer v2.4.16 (Robinson et al., 2011). Further analysis for ambiguous hits was performed using the Artemis Comparison Tool (ACT) v13.0.0 (Carver et al., 2005).

RESULTS AND DISCUSSION

Terpenoid engineering in *P. tricornutum* has only recently been reported (D'Adamo et al., 2018; Fabris et al., 2020) and consequently, there is limited prior knowledge to inform metabolic engineering strategies in diatoms to obtain

elevated terpenoid production. Recently, we demonstrated that extrachromosomal expression (EE) can be used to efficiently express the fusion protein CrGES-mVenus in *P. tricornutum* cytosol to produce up to 0.21 $\mu g/10^7$ cells (0.309 mg/L) geraniol following bacterial conjugation (Fabris et al., 2020). EE of transgenes is not subject to position effect (Karas et al., 2015a) and could therefore provide highly reproducible, consistent, and controllable expression, which is a basic requisite for synthetic biology. In contrast, randomly integrated chromosomal expression (RICE) can result in genetically dissimilar transformants and consequently varied transgene expression among them. However, diatom phenotypes derived from EE and RICE have not been systematically parameterised. Because little is known regarding the mechanisms and effects following EE and RICE of transgenes, it is unclear how these different engineering strategies will compare regarding the expression of *CrGES-mVenus*, and consequently, heterologous geraniol production. Therefore, we analysed the phenotypes of EE and RICE of *Ap1_CrGES-mVenus* in *P. tricornutum* cell lines both at the expression level and in terms of geraniol yield.

Two identical DNA expression cassettes of *AP1p_CrGES-mVenus* and *AP1p_mVenus* as control (Fabris et al., 2020) were cloned either into an pPtPBR11 episome (Diner et al., 2016) or a pUC19 plasmid (Norrander et al., 1983), and delivered either by bacterial conjugation or DNA-coated particle bombardment, respectively, in order to create EE and RICE *P. tricornutum* transformant libraries. Both *AP1p_CrGES-mVenus* constructs contained the *CrGES* gene fused at the carboxyl-terminus to a mVenus yellow fluorescent protein (YFP) (Kremers et al., 2006) for rapidly screening the cell lines by flow cytometry. The *CrGES-mVenus* fusion gene was driven by the *P. tricornutum* native alkaline phosphatase (*AP1*, *Phatr3_J49678*) promoter (hereafter *AP1p*), which is induced in low phosphate conditions for controllable expression (Lin et al., 2017).

Upon transformation of *P. tricornutum* with the episomes and plasmids described above, the resulting antibiotic resistant cell lines were used to create four transgenic diatom libraries. Cell lines transformed with *pPtPBR11_AP1p_CrGES-mVenus* and *pPtPBR11_AP1p_mVenus* for EE were denoted as EE_GmV and EE_mV, respectively. Likewise, cell lines transformed with *pUC19_AP1p_CrGES-mVenus* and *pUC19_AP1p_mVenus* for RICE were denoted as RICE_GmV and RICE_mV, respectively.

EE Transformants Demonstrate Consistent mVenus Fluorescence, While RICE Transformants Demonstrate Higher, but More Variable Signals

In order to compare transgene expression across all four transgenic libraries (EE_GmV, EE_mV, RICE_GmV, and RICE_mV), we required a high-throughput screening strategy to quantify the relative heterologous protein production. We used flow cytometry to rapidly evaluate the CrGES-mVenus expression in an unprecedented number of transformants to identify unique features among and within cell lines generated by EE and RICE. This strategy enabled us to confirm the correct expression of the fusion protein, as well as quantify its relative abundance, offering a proxy for identifying high-expressing transformant variants from the low expressing or silenced variants (Sheff and Thorn, 2004; Delvigne et al., 2014). It is generally accepted that RICE transformants will exhibit different levels of heterologous protein production from each other (Hallmann, 2007; Jeon et al., 2017; Tanwar et al., 2018). Conversely, EE exconjugants theoretically offer more consistent levels of expression (Karas et al., 2015a). However, this has never been shown over a large scale of transformants or through direct comparison. Therefore, it is not known to what extent the expression of heterologous protein can vary across EE exconjugants or RICE transformants.

Single colonies from each of the four libraries were screened according to mVenus fluorescence to evaluate differences in transgene expression. Our results demonstrate that, as predicted and previously reported on a smaller sample (Fabris et al., 2020), EE results in consistent transgene expression among exconjugants. The mean mVenus fluorescence fold change of all EE_GmV lines and all EE_mV lines was 58.50- and 57.73-fold, respectively, when compared to wild type (WT) auto-fluorescence and were not significantly different from each other ($p > 0.9999$) (**Figure 1A**). This suggested that construct size and complexity—at least of this degree—does not affect expression, and confirms that EE might not be affected by variable transgene silencing typically associated with position effect. For EE_GmV strains, the DNA construct contained the fusion gene geraniol synthase and mVenus (total construct size of 10,849 bp); whereas EE_mV strains were transformed with the mVenus containing DNA (total construct size 9,082 bp). Within each EE library, there was also little variation among EE_GmV lines (SEM = 3.54) and EE_mV lines (SEM = 2.95) (**Figure 1A**). These results indicate that EE transformants are highly similar. Interestingly, every EE transformant analysed, across both EE_GmV and EE_mV libraries, showed a mean mVenus fluorescence greater than WT auto-fluorescence (**Figure 1B**). This shows that EE is highly efficient and reliable in generating cell lines that express transgenes and does not require extensive screening, as is required for RICE engineering strategies.

Unlike EE, RICE is subject to position effects and is therefore expected to result in transformants with variable transgene expression. We confirmed this in both RICE_GmV and RICE_mV libraries, in which transformants showed highly variable mVenus fluorescence intensities (**Figure 1A**). For example, the mean mVenus fluorescence in RICE_GmV lines ranged between 0.04 and 719.20- fold change (SEM = 21.37) and RICE_mV lines ranged between 0.40 and 1695.00- fold change (SEM = 43.22). Furthermore, the RICE_GmV and RICE_mV libraries were significantly different from each other ($p < 0.0001$), demonstrating mean mVenus fold changes of 137.80- and 368.60-fold, respectively. Together these results suggested that when transgenes are integrated randomly in to the genome, features of the transgene, such as size and complexity, may affect its expression. It is plausible that gene silencing plays a role, as mVenus is present as a large fusion protein in the RICE_GmV library, whereas it is a smaller, free fluorescent protein in the RICE_mV library.

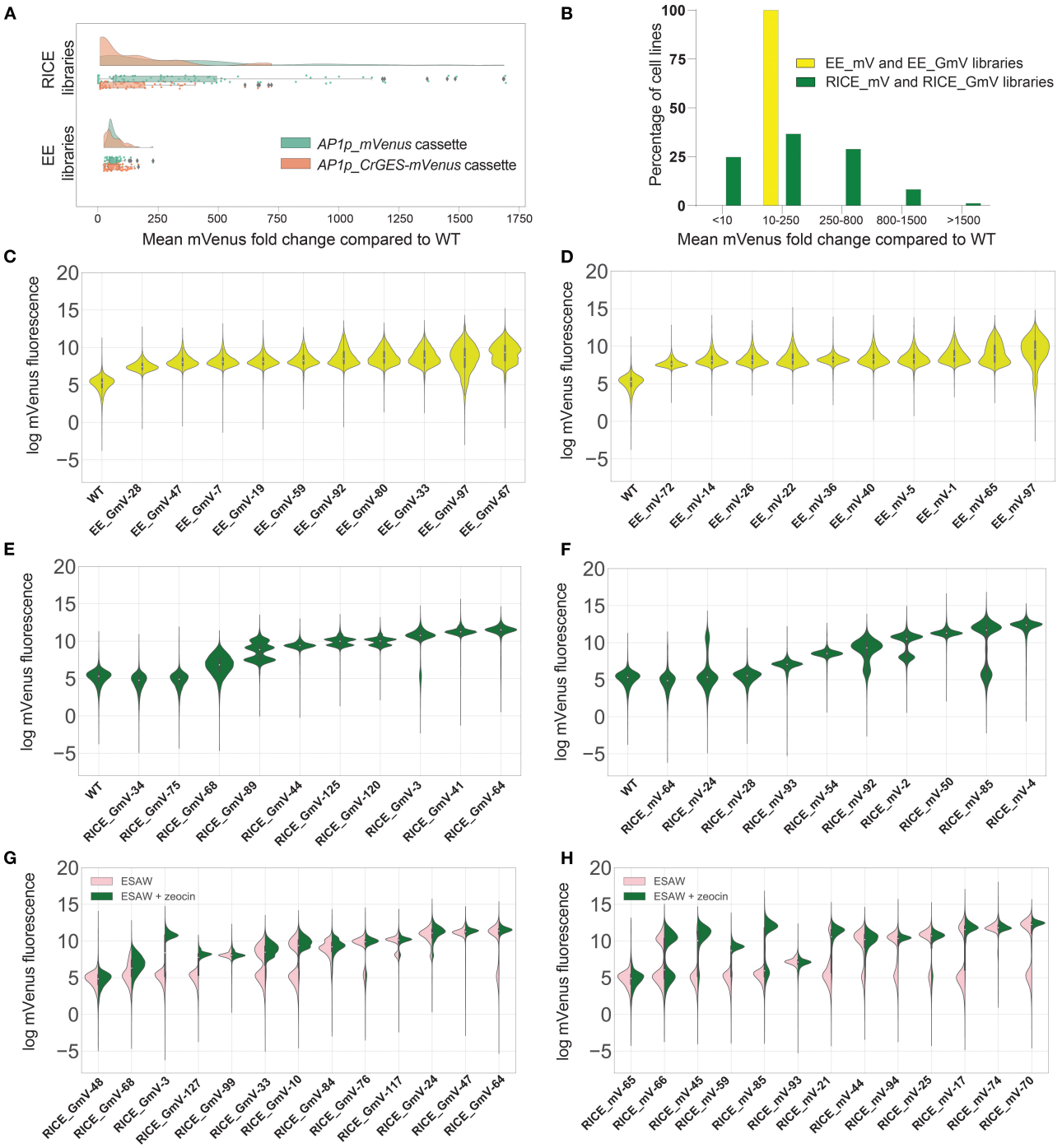


FIGURE 1 | mVenus fluorescence intensities of transgenic *P. tricornutum* extrachromosomal expression (EE) and randomly integrated chromosomal expression (RICE) libraries. **(A)** Fold change of mean mVenus fluorescence of RICE_GmV and RICE_mV transformant libraries and EE_GmV and EE_mV libraries compared to wild type auto-fluorecence. Peach indicates CrGES-mVenus transgenic cell lines and teal indicates mVenus transgenic cell lines. Statistical comparisons were made using Kruskal-Wallis non-parametric ANOVA and Dunn's *post-hoc* test. For EE_GmV library $n = 96$ cell lines total, EE_mV library $n = 96$ cell lines total, RICE_GmV library $n = 74$ cell lines total and RICE_mV library $n = 95$ cell lines total. **(B)** Percentage of pooled RICE libraries (green) compared to percentage of pooled EE libraries (yellow) binned according to mean mVenus fluorescence fold change. **(C–F)** Violin plots indicate the per cell mVenus fluorescence intensity of ten representative cell lines for each library **(C)** EE_GmV **(D)** EE_mV **(E)** RICE_GmV **(F)** RICE_mV, ranked from lowest to mean mVenus expression ($n = 20{,}000$ cells for each cell line). **(G,H)** Representative cell lines from transgene stability analysis for **(G)** RICE_GmV and **(H)** RICE_mV libraries. Pink indicates selection free growth conditions, green indicates zeocin selection growth conditions, cell lines are ranked by mean mVenus intensity ($n = 20{,}000$ cells for each cell line).

To further evaluate the heterogeneity of mVenus fluorescence across RICE libraries, we arbitrarily binned transformants based on the vast spread of mean mVenus fluorescence profiles recorded. We generated five groups comprising of <10-, 10- to 250-, 250- to 800-, 800- to 1,500-, and >1,500-fold change in mVenus fluorescence compared to wild type auto-fluorescence. In the RICE_GmV and RICE_mV libraries (**Figure 1B**), ∼25% of transformants showed <10-fold mean mVenus than wild type auto-fluorescence (**Figure 1B**). Biolistic bombardment is expected to result in random fragmentation of plasmid DNA (Hopes et al., 2016). This could theoretically result in antibiotic resistant transformants that contain the selection cassette without the intact *AP1p_CrGES-mVenus* transcriptional unit, possibly resulting in antibiotic resistant transformants with fluorescence profiles indistinguishable from WT auto-fluorescence. Additionally, it is plausible that transformants associated with such low mean mVenus signals might have integrated the *CrGES-mVenus* expression cassette at transcriptionally repressed genomic loci, or in arrangements that may have triggered gene silencing (Kim et al., 2015). About 37% of RICE transformants and 100% of EE exconjugants demonstrated mean mVenus fluorescence 10–250-fold greater than WT auto-fluorescence. A further 28% of RICE transformants showed a 250–800-fold increase. Only 7% showed an 800- to 1,500-fold increase, and 1% (corresponding to 2 cell lines both from RICE_mV library) reached a remarkable 1,500-fold increase in fluorescence compared to WT auto-fluorescence.

Together, these results demonstrate that transformants generated by RICE require extensive screening at the protein expression level, as up to a quarter can show no to low expression. Interestingly, RICE transformants were able to demonstrate exceptionally higher maximum transgene expression compared to EE using the same transgene cassette design. This warrants further investigation into these transgenic genomes, as there may be aspects of chromosome-integration that could be useful, particularly with regard to multi-generation transgene stability (Kohli and Christou, 2008). Although this high RICE-related expression could be advantageous for simple transgenic constructs, our results show that it would not be suitable for testing larger, more complex assemblies especially without reporter genes. Instead, the high, virtually size-independent consistency of phenotypes associated with EE offer a more suitable platform for applications involving multi-gene constructs, with the advantage of not requiring large scale screening.

Clonal Variegation Is Broadly Distributed in EE but Discretely Defined in RICE

After having determined marked differences in the expression profile between EE and RICE libraries, we exploited the resolution of high-throughput flow cytometry to investigate the population composition within each cell line of EE and RICE libraries. In doing so, we identified relevant variations in the distributions of mVenus fluorescence within individual cell lines, known as cell mosaicism or variegation (Kaufman et al., 2008). Most EE transformants showed a relatively homogenous distribution of mVenus abundance within each cell line, such as that of EE_GmV-28,−47,−7 and−19, and EE_mV-72,−14 and−36 (**Figures 1C,D**). However, some showed increasingly diverse mVenus distribution profiles within individual cell lines, such as EE_GmV-92,−80,−33,−97, and−67 and EE_mV-22,−40,−5,−1,−65, and−97. Intriguingly, these cell lines tended to show higher mean mVenus abundance (**Supplementary Figure 1**). This suggested that EE transformants which exhibited higher mean mVenus signals were composed of cells that were highly dissimilar from each other in a more continuous, non-discrete manner. This observation raises important questions about the dynamics of EE in diatoms; namely, how are episomal copies maintained within each cell, and how dynamic is episomal copy number and segregation across individual cells at different stages of the life cycle? Such mechanisms have been uncovered in other eukaryotic, non-microalgal species. For example, maintenance of viral episomes in mammalian cells and plasmids in *S. cerevisiae* have been attributed to chromosome tethering and hitchhiking mechanisms (Ghosh et al., 2007; Liu et al., 2008; McBride, 2008; Sau et al., 2019). In yeast synthetic biology research, episomal DNA sequences have been characterised based on traits including transformation efficiency, copy number, transgene expression and plasmid stability of clonal populations (Bouton and Smith, 1986; Nakamura et al., 2018; Gu et al., 2019). For example, Nakamura et al. (2018) tested various episome-regulation sequences in *Pichia pastoris* transgenic strains expressing EGFP extrachromosomally. They reported that strains containing the autonomously replicating sequence (ARS) without a centromeric region (CEN) showed broad fluorescence profiles similar to those that we report here, whereby cells within a single clonal population show a wide spread of EGFP fluorescence. However, when combined with centromeric region (CEN2), they reported a more discrete distribution of high EGFP fluorescence profiles that were more consistent with our RICE transformants. This was likely due to a strong bias for the mother cell over the daughter cells during cell division (Gehlen et al., 2011). While the pPtPBR11 plasmid used in this study contains CEN region (Diner et al., 2016), it is still not yet known how such features contribute to episome expression in diatoms (Karas et al., 2015a). Such factors could influence this cell-to-cell phenotypic heterogeneity, but for unknown reasons, this seemed to become more prominent at higher mean mVenus fluorescence.

Overall, RICE transformants demonstrated homogeneous fluorescence distribution profiles within individual cell lines, such as RICE_GmV-44,−41, and−64 and RICE_mV-93,−54,−50, and−4 (**Figures 1E,F**). Other RICE transformants also demonstrated heterogeneous mean mVenus profiles, but as numerous discrete populations within single cell lines (**Figures 1E,F**). For example, RICE_GmV-125,−120 and−3 and RICE_mV-24,−92,−2, and−85 all were composed of two unique populations of mVenus fluorescence distribution (**Figures 1E,F**). Transformant RICE_GmV-89 even showed three populations (**Figure 1E**). These results demonstrate that individual cells within a clonal transformant RICE cell line, generally assumed to have identical phenotypes, can be highly heterogeneous

with regard to transgene expression, but that this heterogeneity can be distributed into unique, discrete populations. This is a major difference with the highly heterogeneous EE cell lines, which were instead characterized by a wide distribution of heterogeneity within the population, although RICE_GmV-68 transformant also followed this distribution.

RICE Cell Lines Exhibit Dramatically Varied Stability That Does Not Correlate to *CrGES-mVenus* Expression

Given the extremely high outliers, we investigated the stability of the RICE libraries and how this related to expression level. Random chromosomal integration can result in stable maintenance and expression of transgenes, even in absence of selective pressure. Transformants that looked indistinguishable from wild type auto-fluorescence in the selective treatment did not change when selective pressure was removed (RICE_GmV-48 and RICE_mV-65, **Figures 1G,H**). We also identified RICE transformants that did not retain their mVenus fluorescence in absence of selective pressure (**Figures 1G,H**). For example, RICE_GmV-3 and−127 and RICE_mV-45,−59,−85 (**Figures 1G,H**, respectively) demonstrated a complete reduction in mVenus fluorescence when cultured in the absence of zeocin that was indistinguishable from wild type auto-fluorescence. Without selective pressure, cells that have silenced their resistance transgene (and by proxy, the transgene of interest) can outcompete and take over the culture due to the disadvantages of reduced energy and resource investment associated with transgene expression.

Interestingly, transgene expression level did not correlate to transgene stability, as seen in RICE_GmV-127 and−99, which showed similar expression levels with selection, but only−127 lost mVenus fluorescence when selection was removed. Likewise, RICE_GmV-3 and−127 cell lines show dissimilar mVenus signals in presence of selection but lost signal completely in selection-free conditions (**Figures 1G,H**). This suggested that there may be transgene integration events or arrangements that facilitate stable transgene expression and highlight the importance of designing screening procedures based on stability not only on transgene expression. Potential mechanisms of action include progressive transcriptional silencing via DNA methylation or histone modification, including *de novo* DNA methylation triggered by transgene recognition (Kohli et al., 2010); and posttranscriptional silencing known as RNA interference (Meyer, 1995; León-Bañares et al., 2004; Cerutti et al., 2011; Doron et al., 2016). Epigenetic silencing of nuclear-integrated exogenous DNA have been attributed to defense mechanisms against viruses and transposable elements in plants (Rajeevkumar et al., 2015) and mammalian cells (Alhaji et al., 2019) alike. Silencing mechanisms, and indeed transgene regulation mechanisms yet to be identified, can influence daughter cells from the same original clonal population differently (Kaufman et al., 2008). In fact, it is not known how stable randomly integrated exogenous DNA fragments are once they have been integrated, or how these insertions are genetically maintained over time.

In other RICE transformants, such as RICE_GmV-68,−33 and−10 and RICE_mV-17 and−70, we detected a reduction in mVenus abundance in absence of selection. Here, a distinct population of cells within each transformant showed signals similar to those in presence of selection, as well as a secondary population of noticeably lower mVenus fluorescence (**Figures 1G,H**). Other lines showed only a very small reduction in expression, such as RICE_GmV-76,−117, and−64 and RICE_mV-44,−94, and−25. Finally, we were able to identify some RICE transformants that maintained mVenus signals both in presence and absence of selective pressure, namely transformants RICE_GmV-99,−84,−24, and−47 and RICE_mV-66,−93, and−74 (**Figures 1G,H**). Some of these transformants also demonstrated comparatively high mVenus fluorescence abundance, particularly RICE_GmV-41,−47 and RICE_mV-74 (**Supplementary Figure 1**). This suggested that they might contain integration events or arrangements that bypass silencing mechanisms. These transformants could provide empirical evidence for putative safe-harbour loci, which have been previously verified in various other organisms including mammalian cell lines (Lee et al., 2015; Cheng et al., 2016; Papapetrou and Schambach, 2016; Salsman and Dellaire, 2016), rice (Cantos et al., 2014) and cyanobacteria (Bentley et al., 2014; Pinto et al., 2015).

Together, these results once again highlight that RICE is not suitable for more complex synthetic biology and that efforts to move toward next-generation genetic engineering strategies is crucial. High expressing RICE transformants can be unstable, as well as contain unknown genomic disturbances and mutations due to damage to the genome itself, as demonstrated in rice and maize (Liu et al., 2018). However, a better understanding of high transgene expression in RICE transformants may reveal aspects of exogenous DNA integration that would be useful for targeted insertion strategies.

Long-Read Whole-Genome Sequencing Reveals No Chromosomal Integration of Episomal DNA, Whereas Biolistic Bombardment Caused Exogenous DNA to Integrate at Unique Chromosomal Loci

To date, transgenic genomic research has been restricted by prohibitive costs of whole-genome sequencing and limited techniques that only reveal certain aspects of random integration events (Scaife and Smith, 2016; Jeon et al., 2017). In diatoms, such strategies include Southern blotting, which showed 1–10 transgene copies of foreign DNA integrated into the genome (Falciatore et al., 1999); quantitative polymerase chain reaction (qPCR), which revealed that copy number was relatively consistent between transformants variants (average of three) (D'Adamo et al., 2018); and thermal asymmetric interlaced PCR (TAIL-PCR), which revealed integration loci of exogenous DNA (Johansson et al., 2019). Similarly, the recently developed method of EE has been shown to require no transgene integration via episome recovery experiments (Karas et al., 2015a; Diner et al., 2016). However, it is not yet known if integration does occur alongside extrachromosomal maintenance of episomes.

TABLE 1 | Summarised details of MinION sequencing of EE and RICE transformants.

	EE_ GmV-97	RICE_GmV-41_A	RICE_GmV-41_B	RICE_GmV-47_A	RICE_GmV-47_B
Nucleotides (total)	203,640,859	254,953,905	279,294,842	257,611,430	325,110,058
Reads (total)	26,455	37,373	31,581	18,525	20,153
Average read length	7,697.63	6,822	8,843.60	13,906	16,132.09
Coverage estimated	∼5.8x	∼ 7.2x	∼7.9x	∼ 7.3x	∼9.2x
Total reads aligning to RICE plasmid	26	248	210	157	151
Reads aligning to RICE plasmid and genome on both borders	0	1	0	0	0
Reads aligning to RICE plasmid and genome on either border	0	18	14	19	27
Reads aligning to RICE plasmid only	26	230	196	138	124

For RICE_GmV cell lines, the analysis was performed on two biological replicates (A and B).

Consequently, there is no knowledge of RICE or EE transgenic microalgal genomes with regard to exogenous DNA integration arrangements, the frequency of integration events throughout the genome, or any precise genomic integration loci. Answering some of these knowledge gaps is required for advancing synthetic biology design, progressing next-generation engineering tools –such as possible safe harbour loci to target–, and providing better understanding of the transgenic genome architecture, particularly regions associated with high transgene expression.

Developments in long-read sequencing technologies, namely Oxford Nanopore and PacBio, have allowed more continuous genome assemblies which can be done in real-time in the lab (Jain et al., 2018). To date, these technologies are mostly used for metagenomics analyses (Robertsen et al., 2016; Pinder et al., 2019) or sequencing new, non-model species (Davis et al., 2016; Fournier et al., 2017). Herein, we applied Oxford Nanopore sequencing to interrogate bacteria-conjugated and biolistic-bombarded transgenic *P. tricornutum* cell lines to explore integrated transgene arrangements, integration locations, and associated genetic architecture, as has been recently done in *Arabidopsis thaliana* and mouse models (Jupe et al., 2019; Nicholls et al., 2019). Given the phenotypic consistency between EE lines, we analysed a single EE line, EE_GmV-97, and assessed all the reads for alignment to the *pPtPBR11_AP1p_CrGES-mVenus* episome DNA. For RICE lines, we analysed two lines that showed high stability and mVenus fluorescence, RICE_GmV-41 and−47. These cell lines showed very similar transgene expression profiles and stabilities. Therefore, we aimed to identify differences or similarities regarding their transgenes at the genome-wide scale. These RICE reads were assessed for alignment to the RICE plasmid *pUC19_AP1p_CrGES-mVenus*. All EE and RICE reads with hits to their respective exogenous DNA constructs were then aligned to the wild type *P. tricornutum* genome in order to identify genomic integration events. The results of the Oxford Nanopore sequencing analysis are summarised in **Table 1**. For each cell line, we sequenced over 250 million nucleotides, resulting in genome coverage of five to nine times, with a probability >99% that the complete genome of each cell line was covered (Clarke and Carbon, 1976).

Our results strongly suggest that no traces of episome were integrated into EE_GmV-97 (**Tables 1**, **2**). Although we identified 26 reads that aligned to the episome, these reads contained no regions which aligned to the *P. tricornutum* genome (**Table 1** and **Figure 2A**), strongly suggesting that these DNA fragments did not get integrated into the nuclear chromosomes. This is the first demonstration that no episomal exogenous DNA is inadvertently integrated into the nuclear genome following bacterial conjugation. Such knowledge is important for identifying any genetic disturbances that may go undetected in exconjugants, progressing knowledge for a better understanding of episomal regulation mechanisms, and for synthetic biology applications with *P. tricornutum* more broadly.

In the RICE lines, we identified only two independent integration loci in both RICE_GmV-41 and RICE_GmV-47, respectively (**Tables 1**, **2**; **Figures 2B,C**). These four sites were detected and confirmed in both biological replicates (**Table 2**; **Supplementary Figure 2**). The integration events associated with these four independent loci consisted of concatenations of various fragments of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid (**Figures 2B,C**). The genomic features and details associated with each of the four integration events are summarised in **Table 2**.

DNA extracted from each cell line does not come from a single cell, but instead a clonal population, and it is plausible that endogenous genomic regions around the insertion site are not stable (Kohli et al., 1998, 2006). Therefore, it was not possible to resolve every integration event to the single nucleotide level, but only at < 60 bp range. For example, Kohli demonstrated that endogenous DNA concatenations can be assembled prior to or during integration in rice crop species, and suggested recombination could occur even after integration (Kohli et al., 1998, 2006). It is also possible that insertions, deletions, or a combination of both (INDELs) can occur at the borders of an exogenous DNA integration event, driven by non-homologous integration (Shin et al., 2016). Such INDELs would cause reads at this small border region to show no alignment to wild type genome, as seen in RICE_GmV-47 integration event 47-10 (**Supplementary Figure 2**). Finally, the high sequencing

TABLE 2 | Summarised details of integration events of EE and RICE transformants.

Clone	Insertion site ID	Estimated chromosomal location (bp)	*In silico* assembly of island complete?	Size island (Kbp)	Genetic feature at integration site	Putative annotation of feature at integration site	AA size	Upstream features within 1 Kbp	Downstream features within 1 Kbp
EE_GmV-97	None				NA				
RICE_GmV-41	41-1	ch1: 2,477,260	Incomplete.	>43	Intergenic	None	NA	Phatr3_J8770 (protein coding)	Phatr3_J54066 (protein coding)
	41-11	ch11: 316,959–317,016	Complete. (Single read spanned island).	~10	Intergenic	None	NA	Phatr3_J46733 (protein coding)	Phatr3_EG00809 (protein coding)
RICE_GmV-47	47-9	ch9: 865,083–865,119	Incomplete.	>124	Phatr3_J46300 (single exon protein coding gene).	CM000612 Genomic DNA Translation: EEC47937.1	402 aa	NA	Phatr3_J46301 (protein coding)
	47-10	ch10: 609,260–609,276	Incomplete.	>87	Phatr3_J46528 (single exon protein coding gene).	CM000613 Genomic DNA Translation: EEC47687.1	429 aa	NA	NA

NA, Not applicable.

error rate associated with Nanopore sequencing (15%) can also influence integration site determination.

In RICE_GmV-41, fragments of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid were inserted at two unique genomic loci, ch1: 2,477,260 (integration island 41-1) and ch11: 316,959–317,016 (integration island 41-11) (**Table 2**). Both integration island 41-1 and 41-11 occur at intergenic regions in the genome; however, they are both flanked by predicted protein coding genes (**Table 2**; **Figure 2B**). Integration island 41-1 is situated 199 bp downstream of the 3′ end of *Phatr3_J8770* and 479 bp upstream of the 5′ start of *Phatr3_J54066* (**Table 2**; **Figure 2B**). *Phatr3_J8770* contains dynamin domains and *Phatr3_J54066* is putatively involved in vesicle trafficking functions according to HMMER (Finn et al., 2011) searches. Integration island 41-11 occurs ∼900 bp downstream of the 3′ end of *Phatr3_J46733* and ∼100 bp upstream of the 5′ start of *Phatr3_EG00809* (**Table 2**; **Figure 2B**). *Phatr3_J46733* contains a transmembrane feature at its C terminus (Uniprot) and a VAD1 Analog of StAR-related lipid transfer domain (VASt) according to HMMER (Finn et al., 2011). *Phatr3_EG00809* showed no predicted functional annotations, nor similarity to known protein domains (Finn et al., 2011).

Neither of these islands disrupted the protein coding regions of these neighbouring genes and we did not detect any growth defective phenotypes for these cell lines. However, the close proximity of the islands to these neighbouring genes means that the integration events may have affected their associated endogenous regulatory regions (**Table 2**, **Figure 2B**).

In transformant RICE_GmV-47, two integration events were localised to ch9: 865,083–865,119 (integration island 47-9) and ch10: 609,260–609,276 (integration island 47-10) (**Table 2**, **Figure 2C**). Both of these loci harbour predicted single-exon protein coding regions *Phatr3_J46300* and *Phatr3_J46528*, respectively, with no predicted functional annotations, nor similarity to known protein domains (Finn et al., 2011).

Interestingly, all four integration events were contained within unique sites across the entire genome of both cell lines, instead of occurring in a more scattered arrangement at a high number of locations, as has been demonstrated following biolistic bombardment in the plants *Oryza sativa* and *Zea mays* (Liu et al., 2018).

Biolistic Bombardment Results in Extremely Large Integration Islands Containing Highly Repetitive Arrangements of Exogenous DNA

Due to the size of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid (6.5 Kbp) and the length of the longer reads we obtained (up to 193.5 Kbp in length), we expected that single reads sequenced using this technology could span the entire integration site. We found this to be true for integration island 41-11, as demonstrated with left-right border read (LRB-R), 28.6 Kbp in length (**Figure 2B**). This read revealed ∼10 Kbp aligned to the RICE plasmid and 4 Kbp upstream and 14 Kbp downstream of the "integration island" aligning to adjacent loci in the reference genome (**Figure 3A**). This is the first visualisation of a complete

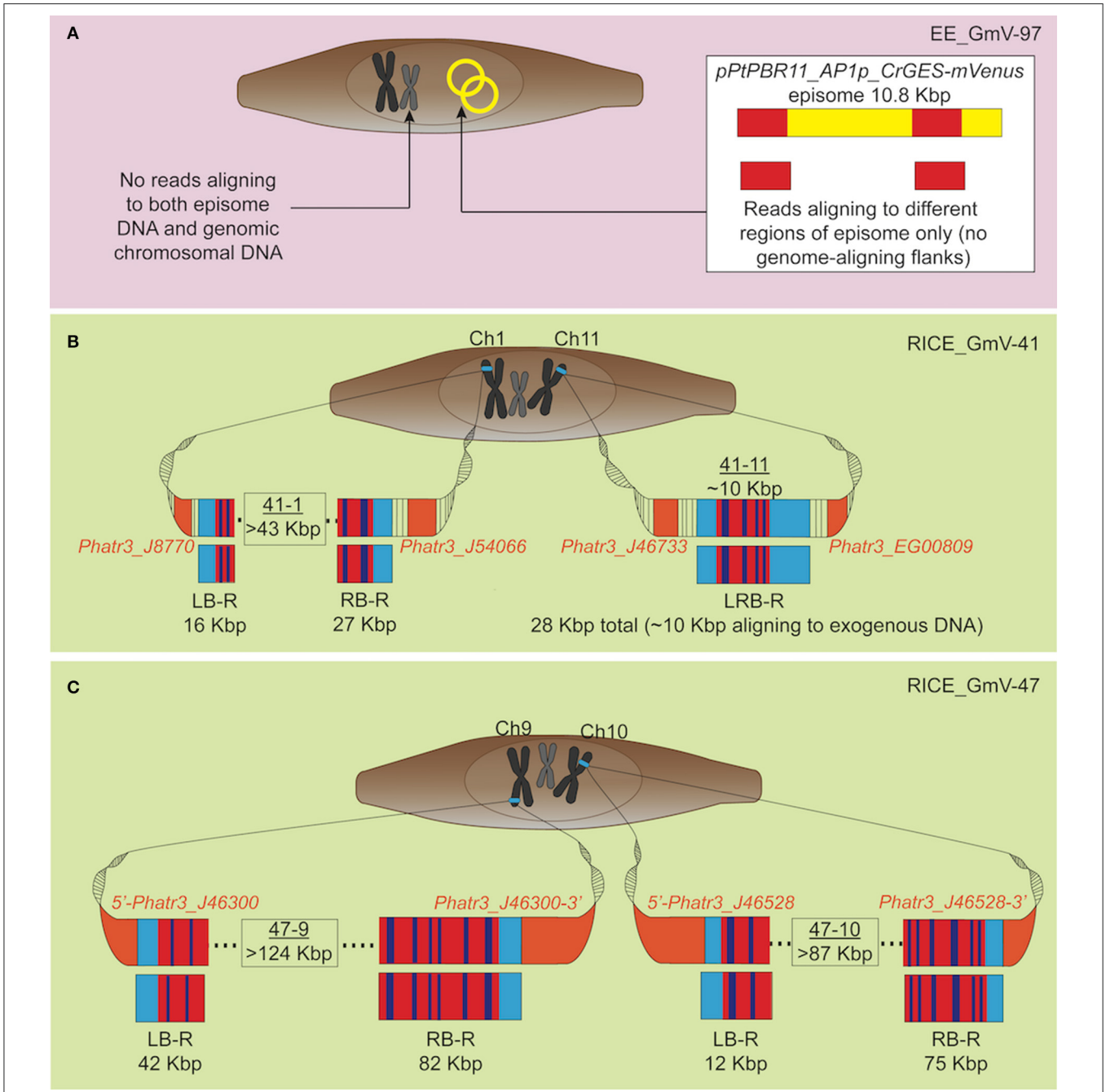


FIGURE 2 | Graphic representation of exogenous DNA constructs in extrachromosomal and chromosomal DNA of the transgenic cell lines, based on long-read sequencing. Only reads aligning to both exogenous DNA and wild type *P. tricornutum* genome, and not those aligning to the genome alone are depicted. **(A)** EE_GmV-97 transformant showed no reads which aligned to both exogenous episomal DNA *pPtPBR11_AP1p_CrGES-mVenus* (yellow), and the reference *P. tricornutum* genome, indicating that no exogenous DNA was integrated into the genome. Instead, some reads showed alignment (red) only to episome DNA, suggesting these reads came from episomal DNA which was extracted and analysed with genomic DNA. **(B)** Transformant RICE_GmV-41 generated by biolistic bombardment showed reads which aligned to both exogenous RICE plasmid, *pUC19_AP1p_CrGES-mVenus*, and the reference *P. tricornutum* genome, indicating a frequency of only two integration islands, 41-1 and 41-11, occurring throughout the whole genome. Island 41-1 occurred on chromosome 1 where the longest left border read (LB-R) and right border read (RB-R) collectively indicated that this island was a minimum of 43 Kbp in size. Island 41-11 occurred on chromosome 11 and was spanned by a single read, left-right border read (LRB-R), which aligned to the reference genome at both left and right borders (light blue), as well as the exogenous RICE plasmid. Red indicates alignment in sense orientation and dark blue indicates alignment in antisense orientation, representing the highly concatenated integration events observed. **(C)** RICE_GmV-47 transformant showed reads which aligned to both exogenous RICE plasmid, *pUC19_AP1p_CrGES-mVenus*, and the reference *P. tricornutum* genome, indicating a frequency of only two integration islands, 47-9 and 47-10, occurring throughout the whole genome. Island 47-9 occurred on chromosome 9 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 124 kB in size. Island 47-10 occurred on chromosome 10 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 87 Kbp in size.

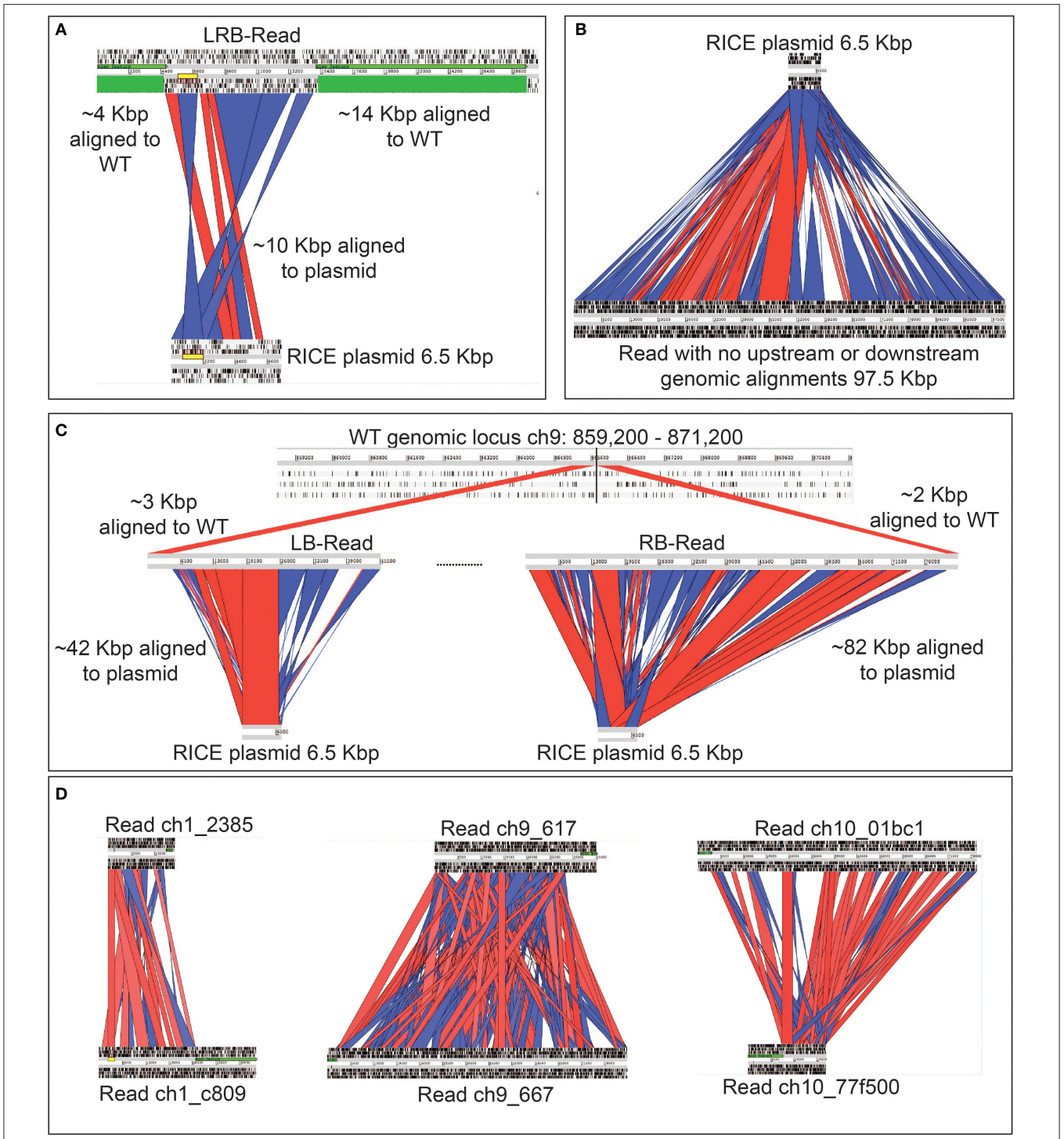


FIGURE 3 | Graphic representation of rearrangements of exogenous DNA in *P. tricornutum* chromosomes, based on long-read sequencing. Red channels show alignment in sense orientation and blue channels show alignment in antisense orientation. Regions that are not highlighted did align to the plasmid, but with below-threshold for hit length of percent identity used for the visualisation, which was performed manually. **(A)** Alignment of a left-right border read (LRB-Read) (top) from integration event 41-11 to RICE plasmid *pUC19_AP1p_CrGES-mVenus* (bottom) and to the wild type *P. tricornutum* genome (green). **(B)** A single 97.5 Kbp read (bottom) with no regions of similarity to the *P. tricornutum* wild type reference genome aligned to the RICE plasmid *pUC19_AP1p_CrGES-mVenus* (top). **(C)** Integration island 47-9 made up by two reads; the left border read (LB-Read) (middle) contains approximately 42 Kbp aligned to the RICE plasmid (bottom) and 3 Kbp aligned to the *P. tricornutum* wild type reference genome (top). The right border read (RB-Read) (middle) contains approximately 82 Kbp of aligned to the RICE plasmid and 2 Kbp aligned to the *P. tricornutum* wild type reference genome. **(D)** Alignments of left and right border reads to each other for integration island 41-1, 47-9, and 47-10. These reads do not align to each other to "close" the integration island, suggesting that some "filler" reads are missing.

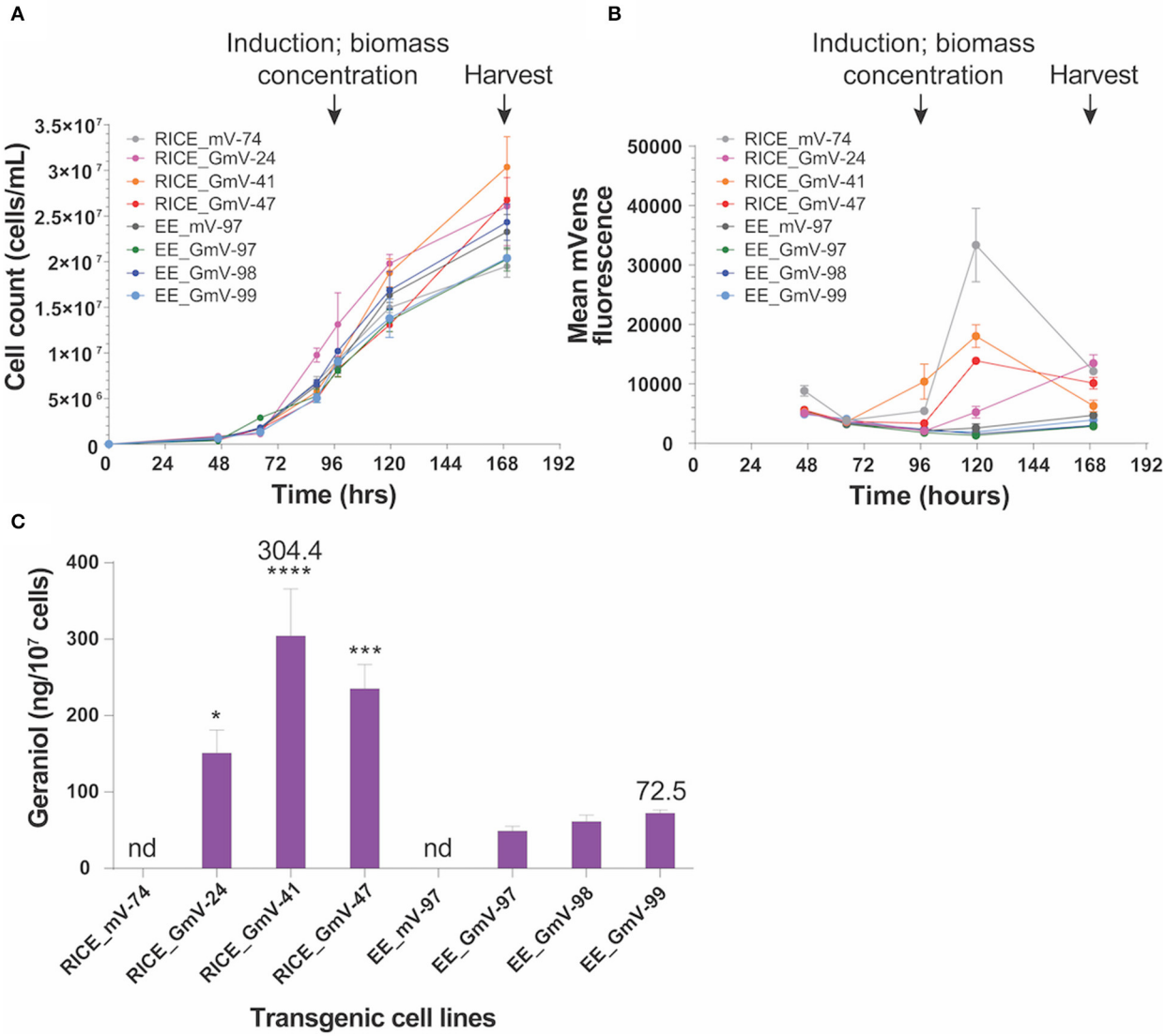


FIGURE 4 | Geraniol production in three selected RICE_GmV and EE_GmV cell lines. $N = 3$, error bars represent SEM, statistical comparisons were made using one-way ANOVA and Tukey's multiple comparisons *post-hoc* test. Significance is demonstrated with asterisks, where $^{****}p \leq 0.0001$; $^{***}p \leq 0.001$; and $^{*}p \leq 0.05$. **(A)** Growth curve for all cell lines. **(B)** mVenus fluorescence intensity 24 h after induction. **(C)** geraniol produced after 72 h induction.

integration event in any microalgae and clearly shows both the integration location in the genome as well as the integration event arrangement.

However, this was the only integration event which was detected on a single read. We did not expect that every other integration event would be so large that it would not be detectable within a single long read, but instead only align to one border of the integration island. This is demonstrated by representative reads 47-9_LB-R and 47-9_RB-R that contained a short 5′ flank aligning to ch9: 860,407–865,083 and a 3′ flank aligning to ch9: 865,119–867,673, respectively (**Figure 3C**). The majority portion of these two reads (42 Kbp for 47-9_RB-R and 82 Kbp for 47-9_RB-R) in fact aligned to RICE plasmid and were found to contain a high frequency of adjacent concatenations of *pUC19_AP1p_CrGES-mVenus* RICE plasmid, in both sense and antisense orientations (**Figure 3C**). This type of "bordered" configuration was detected around the integration islands 41-1, 47-9, and 47-10.

The highly repetitive, shuffled structure of the integration islands may be responsible for the high *CrGES-mVenus* expression associated with these cell lines compared to others in this library. Alternatively or in addition to this, it is plausible that either high intact transcriptional unit copy number, or highly repetitive, chimeric promoter and/or terminator arrangements of the *CrGES-mVenus* cassette might act as enhancers for the expression of this construct. Given that we were able to detect mVenus fluorescence and geraniol production (**Figures 4B,C**), we can infer that some of these fragments contained functional

transgene cassettes. However, the ∼15% error rate of Nanopore sequencing technology (Jain et al., 2018) and the inability to detect single nucleotide polymorphisms make it near impossible to infer the exact number of functional copies present. Furthermore, it is unclear how stable these large integration islands are, if they are prone to recombination events, or what molecular mechanisms occurred to generate them.

Together, the left and right border reads for integration islands 41-1, 47-9, and 47-9 indicate that these islands are a minimum of 43, 124, and 87 Kbp, respectively. We then looked to assess whether these left and right border reads for these three integration islands aligned to each other to "close" the integration island, but were unable to due to their repetitive nature (**Figure 3D**).

Interestingly, over 80% of the reads we identified that aligned to the *pUC19_AP1p_CrGES-mVenus* RICE plasmid did so in their entirety (**Table 1**); i.e., these reads contained no flanks which aligned to the genome at all (**Figure 3B**). Theoretically, these large reads with no genome-aligning flanks could sit between the left and right border reads of the integration islands. Although we cannot align these highly concatenated reads to each other to confirm this, we speculate that islands 41-1, 47-9, and 47-10 could certainly be hundreds of kilobase pairs in size. Given that RICE_GmV-41 transformant has only two integration events and we knew the size of integration island 41-11 (∼10 Kbp) and the sizes of all the "RICE plasmid only" aligning reads together (1,808,244 nt), we can roughly estimate the size of integration island 41-1 through

$$\frac{ni - (l_{11} * c)}{c}$$

where *ni* is the number of nucleotides in "RICE plasmid only" aligned reads, l_{11} is the length of integration island 41-11, and *c* is the estimated coverage. Following this the estimated length of the second integration island 41-1 is ∼250 Kpbs. This correlates to 38 hypothetical back-to-back integrations of the full *pUC19_AP1p_CrGES-mVenus* RICE plasmid. Such large hypothetical islands are supported by similar results in transgenic rice and maize lines obtained following biolistic bombardment, in which large integration islands up to 1.6 Mbp in size were reported (Liu et al., 2018). Jupe et al. (2019) also used whole-genome sequencing to elucidate transgene integration structure following *Agrobacterium*-mediated transformation in *Arabidopsis thaliana*. They reported between one and seven integration islands of between 20 and 230 Kbp per strain. While these results are from higher plant species, they confirm that these huge, highly concatenated islands are not specific to biolistic transformation, nor to diatoms.

As previously described, mechanisms involved with RICE are not well-understood but have been investigated in plants (Kohli et al., 2006). Our results highlight a need to explore mechanisms driving the assembly and maintenance of these islands in diatoms, especially given the range of sizes possible that suggest numerous strategies may be at play. This is the first insight into how nuclear integration occurs in diatoms with widespread implications for existing understanding and future studies. Previous short-read sequencing techniques such as targeted gene-walking (Parker et al., 1991), and inverse PCR or TAIL-PCR (Huang et al., 2000; Liu and Chen, 2007; Johansson et al., 2019) have been useful for identifying integration loci, but would not have been able to detect such large integration events of hundreds of kilobase pairs in size, nor would it be able to detect the complex transgene rearrangements (Nicholls et al., 2019) that we unveiled through long-read sequencing. Furthermore, highly concatenated integrations contain many repeated sequences (**Figure 3**), which nested primers used in TAIL-PCR are able to anneal to. Consequently, PCR strategies would result in many non-specific amplicons and difficulty in determining an unknown integration location. Whilst Southern blotting is a useful approach for determining gene copy number, it does not provide information about the integration site. Hence, our results demonstrate that long-read whole-genome sequencing is an ideal, rapid and affordable approach for determining highly complex transgene integration events.

RICE *CrGES-mVenus* Transformants Are Associated With Higher Expression and Higher Geraniol Yield

We previously demonstrated that wild type *P. tricornutum* does not naturally produce geraniol, but that it can be efficiently engineered to extrachromosomally express *CrGES-mV* to produce it heterologously (Fabris et al., 2020). In order to determine whether extrachromosomal or chromosomal-integrated expression of the fusion construct *CrGES-mVenus* affected the heterologous production of monoterpenoids, we quantified the amount of geraniol produced in three independent RICE_GmV transformant lines and three EE_GmV exconjugant lines. Using mVenus fluorescence as a proxy for GES expression, we selected six transformants (RICE_GmV-24,−41,−24; and EE_GmV-97,−98 and−99, respectively) based on their mean fluorescence intensity and stability (**Supplementary Figure 1**). With the aim of enriching the clonal populations associated with higher mean mVenus fluorescence, RICE_mV-74, RICE_GmV-24,−41 and−47 and EE_mV-97, EE_GmV-97,−98 and−99 were induced and sorted based on mVenus fluorescence using fluorescence activated cell sorting (FACS). A preliminary screen of a pooled sample of the top eight RICE_mV transformants was used to define the mVenus positive gate which did not overlap with the wild type control. During FACS, one thousand cells from each transformant that fell into this gate were collected and scaled up. We concluded that cell sorting did not enrich phenotypic populations, and in this specific case did not improve mean mVenus intensity (**Supplementary Figure 3**).

Geraniol production in transgenic cell lines was evaluated in a bi-phasic, batch fermentation experiment (Fabris et al., 2020). All *CrGES-mVenus* transformants and exconjugants were induced by resuspension in lower volumes (30 ml) of phosphate-free media and showed similar growth to *mVenus* transformant and exconjugant controls prior to induction, indicating no identifiable loss of fitness in *CrGES-mVenus* expressing lines (**Figure 4A**).

We tracked mVenus fluorescence daily and quantified the accumulated geraniol 72 h after inducing the expression of the *CrGES-mVenus* fusion gene. The RICE_GmV transformants, which showed increased mVenus fluorescence 24 h after induction, produced more geraniol than the episomal exconjugant equivalents (**Figures 4B,C**). Chromosome-integrated transformant production yields were 150.9 ng/10^7 cells (0.37 mg/L), 304.4 ng/10^7 cells (0.89 mg/L) and 235.3 ng/10^7 cells (0.61 mg/L) for RICE_GmV-24, RICE_GmV-41 and RICE_GmV-47, respectively (**Figure 4C**). Conversely, non-integrated exconjugants demonstrated consistently lower yields that were 49.1 ng/10^7 cells (0.10 mg/L), 61.4 ng/10^7 cells (0.15 mg/L) and 72.5 ng/10^7 cells (0.15 mg/L) for EE_GmV-97, EE_GmV-98, and EE_GmV-99, respectively (**Figure 4C**). Neither EE nor RICE mVenus controls showed any detectable geraniol (**Figure 4C**). These results indicate that mVenus fluorescence is a reliable proxy for geraniol production, as mVenus fluorescence correlated with geraniol yields. Also, the high geraniol yields achieved support the hypothesis that *P. tricornutum* may have an available free pool of cytosolic geranyl diphosphate (GPP), the prenylphosphate precursor that CrGES converts into geraniol (Fabris et al., 2020), and that the heterologous synthesis of this monoterpenoid might not be limited by substrate availability in these settings. In light of these results, strategies involving promoter optimisation and targeted integration—both currently being evaluated in our laboratory—could further increase the geraniol production. Together, our results warrant the development of *P. tricornutum* for enhanced monoterpenoid production and suggest that it would be possible to improve production levels further by optimising *CrGES* expression at the genetic level.

CONCLUSIONS

Within the emerging application of monoterpenoid engineering in diatoms, we set out to generate specific knowledge to inform genetic optimisation strategies, including multi-gene approaches for synthetic biology and pathway engineering. We provided for the first time a comprehensive comparative analysis of two main types of transgene expression in diatoms, the conventional RICE and in the newly developed EE, using large-scale, high-throughput phenotyping, which allowed us to uncover details of these genetic resources between and within cell lines. The genetic differences between EE and RICE were reflected by the varied yields of the relevant monoterpenoid geraniol in a selection of transgenic diatom cell lines expressing a CrGES-mVenus fusion enzyme. The geraniol yield was more than 4-fold higher in the best RICE transformant, reaching the titre of 304.4 ng/10^7 cells (0.89 mg/L), compared to the best EE exconjugants, reaching 72.47 ng/10^7 cells (0.15 mg/L). Thus, this work evaluated how previously unexplored genetic strategies can improve heterologous production of geraniol in *P. tricornutum*, in addition to more conventional strategies such as metabolic engineering or bioprocessing. While diatom engineering for terpenoid production has only recently been demonstrated, our results show that *P. tricornutum* is a promising photosynthetic microbial factory (D'Adamo et al., 2018; Fabris et al., 2020).

We report profound differences in the phenotypes of RICE and EE *P. tricornutum* cell lines in terms of expression levels, phenotypic consistency and sub-clonal population composition. Non-integrative episomes are associated with much more consistent phenotypes in the scenarios we tested regarding overexpression and EE exconjugants do not seem to require extensive screening. Furthermore, bacterial conjugation tends to result in more clones in a shorter amount of time than biolistic bombardment. Altogether, these results indicate that EE will be an invaluable resource for genetic parts validation and modular assembly, and even automation of the design-build-test-learn cycle. These aspects of more complex synthetic biology strategies are crucial for heterologous production of high-value products such as monoterpenoids. Our results highlighted the particularly limited knowledge available on key aspects of EE in diatoms. This included stability, copy number and segregation patterns, and episome re-arrangement, which we did not observe but has been reported (Slattery et al., 2018). Such characterisations still need to be addressed to fully exploit EE as a synthetic biology platform in diatoms.

In contrast, RICE cell lines are associated with high variability and overall higher expression levels, and by using large-scale screening it is possible to isolate particularly high expressing lines. These superior diatom cell lines bear highly concatenated arrangements of exogenous DNA, present as vast islands within or nearby predicted protein-coding genes. This raises a concern about this widespread method of generating transgenic diatom cell lines, as disrupting numerous protein coding regions can introduce unknown changes to *P. tricornutum* physiology that may not be easily detected. This is a particularly relevant issue in functional genetics studies involving overexpression, knock-down or knock-out constructs, which are traditionally delivered by biolistics, and randomly integrated in the genome of diatoms. On the contrary, it is not yet known if such large, highly concatenated integration events might be a factor in transgene stability and expression. In such scenario, RICE via biolistic bombardment, might be preferable over EE for obtaining high expressing cell lines. Finally, although it has been shown that high copy number and transgene tandem repeats can cause transcriptional silencing of transgene cassettes in other organisms (Kaufman et al., 2008; Moritz et al., 2016), our findings highlight the need to explore copy number and transgene arrangement optimisation in more detail, as this may well not be the case in *P. tricornutum*.

Whilst it is generally accepted that exogenous DNA delivered by biolistic bombardment randomly integrates in diatom chromosomes, the implications of this may have previously been overlooked, particularly at a time when CRISPR-Cas9 technology is being developed. While there is a general concern in CRISPR research to monitor and prevent off-target cutting by CRISPR-Cas9 itself, our results demonstrate that off-target effects from random integration of exogenous constructs such as vector backbone and DNA-encoded CRISPR-Cas9 components, could be just as much cause for concern. In this way, generating a precise knock-in or knock-out genotype by

randomly integrating CRISPR-Cas9 components is suboptimal. As suggested by other works (Sharma et al., 2018; Stukenberg et al., 2018), our findings clearly demonstrate the need to move toward non-integrative alternatives, such as episomal expression (Slattery et al., 2018) and ribonucleoprotein delivery (Serif et al., 2018).

This research also identified putative safe-harbour or neutral loci that could be tested for targeted integration in *P. tricornutum*. Neutral sites in cyanobacteria have been used for targeted integration in metabolic engineering for multigene pathway assembly (Bentley et al., 2014) and dual knock-in knock-out modifications (Li et al., 2016). Synthetic "landing pads" are useful for gene stacking via "domino cloning," but depend on the knowledge of robust, reliable safe harbour loci prior to being feasibly applied to diatoms (Karas et al., 2015b). While some of the loci we identified harbour predicted protein coding regions, this is not unusual for safe harbours, as seen in human cell lines (e.g., CCR5 and ROSA26 loci) and mouse cell lines (e.g., Rosa26 locus), which all occur within protein coding regions. Furthermore, regions that may currently appear to be intergenic or non-functional may be re-categorised in the future, as more information about "junk DNA," transcripts without function (TUFs) and unannotated regulatory regions are discovered (Gingeras, 2007).

Finally, to the best of our knowledge, this work reports for the first time the suitability and utility of third generation long-read whole-genome sequencing to reveal the previously unknown nature of chromosomal integration sites, that would not have been feasible with conventional short-read sequencing. Future work investigating trans-genomes, such as low expression RICE cell lines or epigenetic modifications including DNA methylation patterns (Jain et al., 2018; Jupe et al., 2019), would build upon this knowledge to help uncover mechanisms driving transgene integration in diatoms. Such knowledge is important for developing better functional genomics tools including targeted genome editing. Our research primarily aimed at tracking specific, known transgenic constructs in EE and RICE transgenic diatoms cell lines. Long-read whole-genome sequencing technology can also be used to identify changes to the genome independent of an integration event, such as large translocations (Jupe et al., 2019) and deletions (Nicholls et al., 2019), purely due to the disruptive nature of the DNA delivery method. Our work lays the basis for future research efforts specifically focused on these relevant aspects, to investigate the impact of biolistic bombardment itself on genome integrity.

In conclusion, advancing synthetic pathway construction in *P. tricornutum* would ideally combine the reproducibility of EE with the high expression achievable through RICE, which could be achieved by targeted chromosomal integration. This work lays key groundwork for these developments which are crucial for extending knowledge on diatom biology and elevating model species such as *P. tricornutum* as a widely used synthetic biology chassis organism in a broad array of biotechnological applications.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in NCBI BioProject repository under the ID PRJNA593624.

AUTHOR CONTRIBUTIONS

JG designed the study, conducted the experiments, analysed data and wrote the manuscript. TK performed DNA extraction and sequencing, analysed data, and contributed writing the manuscript. RA analysed data and contributed writing the manuscript. UK performed GC-MS analyses and contributed writing the manuscript. PR revised and edited the manuscript. MF designed the study, conducted and supervised the experiments, and contributed writing the manuscript.

FUNDING

This work was funded by the University of Technology Sydney and the CSIRO Synthetic Biology Future Science Platform. JG was supported by a UTS Doctoral Scholarship. MF was supported by a CSIRO Synthetic Biology Future Science Platform Fellowship co-funded by CSIRO and the University of Technology Sydney.

ACKNOWLEDGMENTS

The authors would like to thank Kun Xiao and Taya Lapshina for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00513/full#supplementary-material

REFERENCES

Ainley, W. M., Sastry-Dent, L., Welter, M. E., Murray, M. G., Zeitler, B., Amora, R., et al. (2013). Trait stacking via targeted genome editing. *Plant Biotechnol. J.* 11, 1126–1134. doi: 10.1111/pbi.12107

Alhaji, S. Y., Ngai, S. C., and Abdullah, S. (2019). Silencing of transgene expression in mammalian cells by DNA methylation and histone modifications in gene therapy perspective. *Biotechnol. Genet. Eng. Rev.* 35, 1–25. doi: 10.1080/02648725.2018.1551594

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., et al. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473, 203–207. doi: 10.1038/nature10074

Allen, M., Poggiali, D., Whitaker, K., and Marshall, T. R. (2018). Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* 4:63. doi: 10.7287/peerj.preprints.27137

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature* 459, 185–192. doi: 10.1038/nature08057

Ashworth, J., Turkarslan, S., Harris, M., Orellana, M. V., and Baliga, N. S. (2016). Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Mar. Genomics* 26, 21–28. doi: 10.1016/j.margen.2015.10.011

Bentley, F. K., Zurbriggen, A., and Melis, A. (2014). Heterologous expression of the mevalonic acid pathway in cyanobacteria enhances endogenous carbon partitioning to isoprene. *Mol. Plant* 7, 71–86. doi: 10.1093/mp/sst134

Berges, J. A., Franklin, D. J., and Harrison, P. J. (2001). Evolution of an artificial seawater medium: improvements in enriched seawater, artificial water over the last two decades. *J. Phycol.* 37, 1138–1145. doi: 10.1046/j.1529-8817.2001.01052.x

Bian, G., Deng, Z., and Liu, T. (2017). Strategies for terpenoid overproduction and new terpenoid discovery. *Curr. Opin. Biotechnol.* 48, 234–241. doi: 10.1016/j.copbio.2017.07.002

Bouton, A. H., and Smith, M. M. (1986). Fine-structure analysis of the DNA sequence requirements for autonomous replication of *Saccharomyces cerevisiae* plasmids. *Mol. Cell. Biol.* 6, 2354–2363. doi: 10.1128/MCB.6.7.2354

Cantos, C., Francisco, P., Trijatmiko, K. R., Slamet-Loedin, I., and Chadha-Mohanty, P. K. (2014). Identification of "safe harbor" loci in indica rice genome by harnessing the property of zinc-finger nucleases to induce DNA damage and repair. *Front. Plant Sci.* 5:302. doi: 10.3389/fpls.2014.00302

Caputi, L., Franke, J., Farrow, S. C., Chung, K., Payne, R. M. E., Nguyen, T. D., et al. (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in *Madagascar periwinkle*. *Science* 360, 1235–1239. doi: 10.1126/science.aat4100

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., and Parkhill, J. (2005). ACT: the artemis comparison tool. *Bioinformatics* 21, 3422–3423. doi: 10.1093/bioinformatics/bti553

Cerutti, H., Ma, X., Msanne, J., and Repas, T. (2011). RNA-mediated silencing in algae: biological roles and tools for analysis of gene function. *Eukaryotic Cell* 10, 1164–1172. doi: 10.1128/EC.05106-11

Chen, W., and Viljoen, A. M. (2010). Geraniol - a review of a commercially important fragrance material. *South Afr. J. Bot.* 76, 643–651. doi: 10.1016/j.sajb.2010.05.008

Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., and Alper, H. S. (2016). Identifying and retargeting transcriptional hot spots in the human genome. *Biotechnol. J.* 11, 1100–9. doi: 10.1002/biot.201600015

Clarke, L., and Carbon, J. (1976). A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. *Cell* 9, 91–99. doi: 10.1016/0092-8674(76)90055-6

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., et al. (2014). Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. *Nat. Commun.* 5:3831. doi: 10.1038/ncomms4831

D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., et al. (2018). Engineering the unicellular alga phaeodactylum tricornutum for high-value plant triterpenoid production. *Plant Biotechnol. J.* 17, 75–87. doi: 10.1111/pbi.12948

Davis, A. M., Iovinella, M., James, S., Robshaw, T., Dodson, R., Herrero-davila, L., et al. (2016). Using MinION nanopore sequencing to generate. *bioRxiv*. doi: 10.1101/076208

Delvigne, F., Zune, Q., Lara, A. R., Al-Soud, W., and Sørensen, S. J. (2014). Metabolic variability in bioprocessing: implications of microbial phenotypic heterogeneity. *Trends Biotechnol.* 32, 608–616. doi: 10.1016/j.tibtech.2014.10.002

Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E., and Weyman, P. D. (2016). Refinement of the diatom episome maintenance sequence and improvement of conjugation-based DNA delivery methods. *Front. Bioeng. Biotechnol.* 4:65. doi: 10.3389/fbioe.2016.00065

Doron, L., Segal, N., and Shapira, M. (2016). Transgene expression in microalgae—from tools to applications. *Front. Plant Sci.* 7:505. doi: 10.3389/fpls.2016.00505

Elgin, S. C. R. (1996). Heterochromatin and gene regulation in *Drosophila*. *Curr. Opin. Genet. Dev.* 6, 193–202. doi: 10.1016/S0959-437X(96)80050-5

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., et al. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synth. Biol.* 9, 598–612. doi: 10.1021/acssynbio.9b00455

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., et al. (2014). Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*. *N. Phytol.* 204, 521–535. doi: 10.1111/nph.12917

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., and Baart, G. J. E. (2012). The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic entner-doudoroff glycolytic pathway. *Plant J.* 70, 1004–1014. doi: 10.1111/j.1365-313X.2012.04941.x

Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., and Bowler, C. (1999). Transformation of nonselectable reporter genes in marine diatoms. *Mar. Biotechnol.* 1, 239–251. doi: 10.1007/PL00011773

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, 29–37. doi: 10.1093/nar/gkr367

Fournier, T., Gounot, J. S., Freel, K., Cruaud, C., Lemainque, A., Aury, J. M., et al. (2017). High-quality *de novo* genome assembly of the dekkera bruxellensis yeast using nanopore MinION sequencing. *G3 (Bethesda)* 7, 3243–3250. doi: 10.1534/g3.117.300128

Gangl, D., Zedler, J. A. Z., Rajakumar, P. D., Martinez, E. M. R., Riseley, A., Wodarczyk, A., et al. (2015). Biotechnological exploitation of microalgae. *J. Exp. Bot.* 66, 6975–6990. doi: 10.1093/jxb/erv426

Gehlen, L. R., Nagai, S., Shimada, K., Meister, P., Taddei, A., and Gasser, S. M. (2011). Nuclear geometry and rapid mitosis ensure asymmetric episome segregation in yeast. *Curr. Biol.* 21, 25–33. doi: 10.1016/j.cub.2010.12.016

Ghosh, S. K., Hajra, S., and Jayaram, M. (2007). Faithful segregation of the multicopy yeast plasmid through cohesin-mediated recognition of sisters. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13034–13039. doi: 10.1073/pnas.0702996104

Gingeras, T. (2007). Origin of phenotypes: genes and transcripts. *Genome Res.* 17, 682–690. doi: 10.1101/gr.6525007

Goyal, A., Bhowmik, P. K., and Basu, S. K. (2009). Minichromosomes : the second generation genetic engineering tool. *Plant Omics J.* 2, 1–8. Available online at: https://www.pomics.com/Saikat_January2009_2_1_1_8.pdf

Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., and Hegemann, P. (2017). Targeting of photoreceptor genes in *Chlamydomonas reinhardtii* via zinc- finger nucleases and CRISPR/Cas9. *Plant Cell* 29, 2498–2518. doi: 10.1105/tpc.17.00659

Gu, Y., Gao, J., Cao, M., Dong, C., Lian, J., Huang, L., et al. (2019). Construction of a series of episomal plasmids and their application in the development of an efficient CRISPR/Cas9 system in *Pichia pastoris*. *World J. Microbiol. Biotechnol.* 35:79. doi: 10.1007/s11274-019-2654-5

Hallmann, A. (2007). Algal transgenics and biotechnology. *Transgenic Plant J*, 1, 81–98. Available online at: http://www.globalsciencebooks.info/Online/GSBOnline/images/0706/TPJ_1(1)/TPJ_1(1)81-98o.pdf

Hamilton, M. L., Haslam, R. P., Napier, J. A., and Sayanova, O. (2014). Metabolic engineering of *Phaeodactylum tricornutum* for the enhanced accumulation of omega-3 long chain polyunsaturated fatty acids. *Metab. Eng.* 22, 3–9. doi: 10.1016/j.ymben.2013.12.003

Hamilton, M. L., Warwick, J., Terry, A., Allen, M. J., Napier, J. A., and Sayanova, O. (2015). Towards the industrial production of omega-3 long chain polyunsaturated fatty acids from a genetically modified diatom *Phaeodactylum tricornutum*. *PLoS ONE* 10:e0144054. doi: 10.1371/journal.pone.0144054

Hempel, F., Bozarth, A. S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., et al. (2011a). Microalgae as bioreactors for bioplastic production. *Microb. Cell Fact.* 10:81. doi: 10.1186/1475-2859-10-81

Hempel, F., Lau, J., Klingl, A., and Maier, U. G. (2011b). Algae as protein factories: expression of a human antibody and the respective antigen in the diatom *Phaeodactylum tricornutum*. *PLoS ONE* 6:e28424. doi: 10.1371/journal.pone.0028424

Hempel, F., and Maier, U. G. (2012). An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. *Microb. Cell Fact.* 11:126. doi: 10.1186/1475-2859-11-126

Hopes, A., Nekrasov, V., Kamoun, S., and Mock, T. (2016). Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods* 12:49. doi: 10.1186/s13007-016-0148-0

Huang, G., Zhang, L., and Birch, R. G. (2000). Rapid amplification and cloning of Tn5 flanking fragments by inverse PCR. *Lett. Appl. Microbiol.* 31, 149–153. doi: 10.1046/j.1365-2672.2000.00781.x

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

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi: 10.1038/nbt.4060

Jeon, S., Lim, J-M., Lee, H-G., Shin, S-E., Kang, N. K., Park, Y-I., et al. (2017). Current status and perspectives of genome editing technology for microalgae. *Biotechnol. Biofuels* 10:267. doi: 10.1186/s13068-017-0957-z

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., et al. (2017). Manipulation of GES and ERG20 for geraniol overproduction in *Saccharomyces cerevisiae*. *Metab. Eng.* 41, 57–66. doi: 10.1016/j.ymben.2017.03.005

Johansson, O. N., Töpel, M., Pinder, M. I. M., Kourtchenko, O., Blomberg, A., Godhe, A., et al. (2019). Skeletonema marinoi as a new genetic model for marine chain-forming diatoms. *Sci. Rep.* 9:5392. doi: 10.1038/s41598-019-41085-5

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., et al. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics* 15:e1007819. doi: 10.1371/journal.pgen.1007819

Karas, B. J., Diner, R. E., Lefebvre, S. C., McQuaid, J., Phillips, A. P. R., Noddings, C. M., et al. (2015a). Designer diatom episomes delivered by bacterial conjugation. *Nat. Commun.* 6:6925. doi: 10.1038/ncomms7925

Karas, B. J., Suzuki, Y., and Weyman, P. D. (2015b). Strategies for cloning and manipulating natural and synthetic chromosomes. *Chromosome Res.* 23, 57–68. doi: 10.1007/s10577-014-9455-3

Kaufman, W. L., Kocman, I., Agrawal, V., Rahn, H. P., Besser, D., and Gossen, M. (2008). Homogeneity and persistence of transgene expression by omitting antibiotic selection in cell line isolation. *Nucleic Acids Research*, 36:e111. doi: 10.1093/nar/gkn508

Kim, E. J., Ma, X., and Cerutti, H. (2015). Gene silencing in microalgae: mechanisms and biological roles. *Bioresour. Technol.* 184, 23–32. doi: 10.1016/j.biortech.2014.10.119

Kohli, A., and Christou, P. (2008). Stable transgenes bear fruit. *Nat. Biotechnol.* 26, 653–654. doi: 10.1038/nbt0608-653

Kohli, A., González-Melendi, P., Abranches, R., Capell, T., Stoger, E., and Christou, P. (2006). The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signal. Behav.* 1, 185–195. doi: 10.4161/psb.1.4.3195

Kohli, A., Leech, M., Vain, P., Laurie, D. A., and Christou, P. (1998). Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots. *Proc. Natl. Acad. Sci. U.S.A.* 95, 7203–7208. doi: 10.1073/pnas.95.12.7203

Kohli, A., Miro, B., and Twyman, R. M. (2010). "Transgene integration, expression and stability in plants : strategies for improvements," in *Transgenic Crop Plants,* eds C. Kole, C. H. Michler, A. G. Abbott, T. C. Hall (Berlin, Heidelberg: Springer), 201–237. doi: 10.1007/978-3-642-04809-8_7

Kremers, G. J., Goedhart, J., Van Munster, E. B., and Gadella, T. W. J. (2006). Cyan and yellow super fluorescent proteins with improved brightness, protein folding, and FRET förster radius. *Biochemistry* 45, 6570–6580. doi: 10.1021/bi0516273

Kroth, P. G. (2007). Genetic transformation: a tool to study protein targeting in diatoms. *Methods Mol. Biol.* 390, 257–267. doi: 10.1385/1-59745-466-4:257

Kroth, P. G., Chiovitti, A., Gruber, A., Martin-Jezeque, V., Mock, T., Parker, M. S., et al. (2008). A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PLoS ONE* 3:e1426. doi: 10.1371/journal.pone.0001426

Lavaud, J., Materna, A. C., Sturm, S., Vugrinec, S., and Kroth, P. G. (2012). Silencing of the violaxanthin de-epoxidase gene in the diatom *Phaeodactylum tricornutum* reduces diatoxanthin synthesis and non-photochemical quenching. *PLoS ONE* 7:e36806. doi: 10.1371/journal.pone.0036806

Lee, J. S., Kallehauge, T. B., Pedersen, L. E., and Kildegaard, H. F. (2015). Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep.* 5:8572. doi: 10.1038/srep08572

León-Bañares, R., González-Ballester, D., Galván, A., and Fernández, E. (2004). Transgenic microalgae as green cell-factories. *Trends Biotechnol.* 22, 45–52. doi: 10.1016/j.tibtech.2003.11.003

Li, H., and Durbin, R. (2009). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H., Shen, C. R., Huang, C-H., Sung, L-Y., Wu, M-Y., and Hu, Y-C. (2016). CRISPR-Cas9 for the genome engineering of cyanobacteria and succinate production. *Metab. Eng.* 38, 293–302. doi: 10.1016/j.ymben.2016.09.006

Lin, H. Y., Yen, S. C., Kuo, P. C., Chung, C. Y., Yeh, K. L., Huang, C. H., et al. (2017). Alkaline phosphatase promoter as an efficient driving element for exogenic recombinant in the marine diatom *Phaeodactylum tricornutum*. *Algal Res.* 23, 58–65. doi: 10.1016/j.algal.2017.01.007

Liu, J., Nannas, N. J., Fu, F., Shi, J., Aspinwall, B., Parrott, W. A., et al. (2018). Genome-scale sequence disruption following biolistic transformation in rice and maize. *Plant Cell* 31, 368–383. doi: 10.1101/379644

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

Liu, Y. G., and Chen, Y. (2007). High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. *BioTechniques* 43, 649–656. doi: 10.2144/000112601

Liu, Y-T., Sau1, S., Ma1, C-H., Kachroo1, A. H., Rowley1, P. A., Chang, K-M., et al. (2008). The partitioning and copy number control systems of the selfish yeast plasmid: an optimized molecular design for stable persistence in host cells. *Microbiol. Spectr.* 23, 1–7. doi: 10.1128/microbiolspec.PLAS-0003-2013

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

McBride, A. A. (2008). Chapter 4 replication and partitioning of papillomavirus genomes. *Adv. Virus Res.* 72, 155–205. doi: 10.1016/S0065-3527(08)00404-1

Meyer, P. (1995). Understanding and controlling transgene expression. *Trends Biotechnol.* 13, 332–337. doi: 10.1016/S0167-7799(00)88977-5

Moritz, B., Woltering, L., Becker, P. B., and Göpfert, U. (2016). High levels of histone H3 acetylation at the CMV promoter are predictive of stable expression in Chinese hamster ovary cells. *Biotechnol. Prog.* 32, 776–786. doi: 10.1002/btpr.2271

Nakamura, Y., Teruyuki, N., Noguchi, R., Ito, Y., Watanabe, T., Nishiyama, T., et al. (2018). A stable, autonomously replicating plasmid vector containing pichia pastoris centromeric *DNA*. *Appl. Environ. Microbiol.* 84:e02882–17. doi: 10.1128/AEM.02882-17

Nicholls, P. K., Bellott, D. W., Cho, T. J., Pyntikova, T., and Page, D. C. (2019). Locating and characterizing a transgene integration site by nanopore sequencing. *G3 (Bethesda)* 9, 1481–1486. doi: 10.1534/g3.119.300582

Norrander, J., Kempe, T., and Messing, J. (1983). Construction of improved Ml3 vectors using oligodeoxynucleotidedirected mutagenesis (Ml3 cloning; synthetic DNA primers; gene synthesis; phosphoramidites; DNA polymerase I Klenow frag-ment). *Bren Road East* 26, 935–7335.

Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M., and Winge, P. (2016). A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* 6:24951. doi: 10.1038/srep24951

Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., et al. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* 496, 528–532. doi: 10.1038/nature12051

Papapetrou, E. P., and Schambach, A. (2016). Gene insertion into genomic safe harbors for human gene therapy. *Mol. Ther.* 24, 678–684. doi: 10.1038/mt.2016.38

Parker, J. D., Rabinovitch, P. S., and Burmer, G. C. (1991). Walking polymerase chain reaction. *Gene* 19, 3055–3060. doi: 10.1093/nar/19.11.3055

Pinder, M. I. M., Johansson, O. N., Almstedt, A., Kourtchenko, O., Clarke, A. K., Godhe, A., et al. (2019). Genome sequence of kordia sp. strain SMS9 identified in a non-axenic culture of the diatom *Skeletonema marinoi*. *J. Genomics* 7, 46–49. doi: 10.7150/jgen.35061

Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., et al. (2015). Improving a Synechocystis-based photoautotrophic chassis through

systematic genome mapping and validation of neutral sites. *DNA Res.* 22, 425–437. doi: 10.1093/dnares/dsv024

Pollak, B., Matute, T., Nunez, I., Cerda, A., Lopez, C., Kan, A., et al. (2019). Universal Loop assembly (uLoop): open, efficient, and species- agnostic DNA fabrication. *Biorxiv.* ' doi: 10.1101/744854

Rajeevkumar, S., Anunanthini, P., and Sathishkumar, R. (2015). Epigenetic silencing in transgenic plants. *Front. Plant Sci.* 6:693. doi: 10.3389/fpls.2015.00693

Remmers, I. M., D'Adamo, S., Martens, D. E., de Vos, R. C. H., Mumm, R., America, A. H. P., et al. (2018). Orchestration of transcriptome, proteome and metabolome in the diatom *Phaeodactylum tricornutum* during nitrogen limitation. *Algal Res.* 35, 33–49. doi: 10.1016/j.algal.2018.08.012

Robertsen, E. M., Kahlke, T., Raknes, I. A., Pedersen, E., Semb, E. K., Ernstsen, M., et al. (2016). META-pipe - pipeline annotation, analysis and visualization of marine metagenomic sequence data. 1–22. arXiv:1604.04103.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *OECD Observ.* 29, 23–24. doi: 10.1038/nbt.1754

Salsman, J., and Dellaire, G. (2016). Precision Genome Editing in the CRISPR Era. *Biochem. Cell Biol.* 95, 187–201. doi: 10.1139/bcb-2016-0137

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

Scaife, M. A., and Smith, A. G. (2016). Towards developing algal synthetic biology. *Biochem. Soc. Transac.* 44, 716–722. doi: 10.1042/BST20160061

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., and Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom *Phaeodactylum tricornutum* by DNA-free genome editing. *Nat. Commun.* 9:3924. doi: 10.1038/s41467-018-06378-9

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

Sharma, A. K., Nymark, M., Sparstad, T., Bones, A. M., and Winge, P. (2018). Transgene-free genome editing in marine algae by bacterial conjugation – comparison with biolistic CRISPR/Cas9 transformation. *Sci. Rep.* 8:14401. doi: 10.1038/s41598-018-32342-0

Sheff, M. A., and Thorn, K. S. (2004). Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* 21, 661–670. doi: 10.1002/yea.1130

Shin, S-E., Lim, J-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., et al. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in *Chlamydomonas reinhardtii* SUPP. *Sci. Rep.* 6:27810. doi: 10.1038/srep27810

Slattery, S. S., Diamond, A., Wang, H., Therrien, J. A., Lant, J. T., Jazey, T., et al. (2018). An expanded plasmid-based genetic toolbox enables Cas9 genome editing and stable maintenance of synthetic pathways in *Phaeodactylum tricornutum*. *ACS Synth. Biol.* 7, 328–338. doi: 10.1021/acssynbio.7b00191

Smith, S. R., Dupont, C. L., McCarthy, J. K., Broddrick, J. T., Oborník, M., Horák, A., et al. (2019). Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nat. Commun.* 10:4552. doi: 10.1038/s41467-019-12407-y

Stukenberg, D., Zauner, S., Aquila, G. D., and Maier, U. G. (2018). Optimizing CRISPR/Cas9 for the diatom *Phaeodactylum tricornutum*. *Front. Plant Sci.* 9:740. doi: 10.3389/fpls.2018.00740

Tanwar, A., Sharma, S., and Kumar, S. (2018). Targeted genome editing in algae using CRISPR/Cas9. *Indian J. Plant Physiol.* 23, 653–669. doi: 10.1007/s40502-018-0423-3

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G. J. E., Rombauts, S., Hasnain, G., et al. (2013). CathaCyc, a metabolic pathway database built from catharanthus roseus RNA-seq data. *Plant Cell Physiol.* 54, 673–685. doi: 10.1093/pcp/pct039

Vavitsas, K., Fabris, M., and Vickers, C. E. (2018). Terpenoid metabolic engineering in photosynthetic microorganisms. *Genes* 9:520. doi: 10.3390/genes9110520

Wang, C., Zada, B., Wei, G., and Kim, S. W. (2017). Metabolic engineering and synthetic biology approaches driving isoprenoid production in *Escherichia coli*. *Bioresour. Technol.* 241, 430–438. doi: 10.1016/j.biortech.2017.05.168

Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., Mccarthy, J. K., Heuberger, A. L., et al. (2015). Inactivation of *Phaeodactylum tricornutum* urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnol. J.* 13, 460–470. doi: 10.1111/pbi.12254

Yao, Y., Lu, Y., Peng, K. T., Huang, T., Niu, Y. F., Xie, W. H., et al. (2014). Glycerol and neutral lipid production in the oleaginous marine diatom *Phaeodactylum tricornutum* promoted by overexpression of glycerol-3-phosphate dehydrogenase. *Biotechnol. Biofuels* 7:110. doi: 10.1186/1754-6834-7-110

Zurbriggen, A., Kirst, H., and Melis, A. (2012). Isoprene production via the mevalonic acid pathway in *Escherichia coli* (Bacteria). *Bioenergy Res.* 5, 814–828. doi: 10.1007/s12155-012-9192-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CRISPR-Cas9 mediated targeted chromosomal integration in *Phaeodactylum tricornutum* using the endogenous selectable marker *uridine-5'-monophosphate synthase*

In preparation to submit for publication.

ABSTRACT

Targeted genomic integration (TGI) stands to be a useful next generation genetic engineering approach for controlled and predictable integration and transgene expression in the model diatom, *Phaeodactylum tricornutum*. This is because targeting technologies, such as CRISPR-Cas9, facilitate integration into regions known to support transgene expression, referred to as neutral sites or safe harbours. This is a vast improvement from the first-generation strategy of randomly integrated chromosomal expression (RICE) in *P. tricornutum*, as well as many other algal and non-algal model species, which results in highly dissimilar transformant cell lines regarding their genetic configuration and transgene expression phenotypes. A recent publication described high efficiency CRISPR-driven gene knockout in *P. tricornutum* using an endogenous selectable marker gene, *uridine-5'-monophosphate synthase* (*UMPS),* which resulted in a uracil auxotrophic, 5-FOA resistant phenotype useful for direct enrichment in CRISPR-based screening. This study is foundational for CRISPR-based knock-out in *P. tricornutum*; however, CRISPR-driven TGI is yet to be described in highly studied diatom. Previously, we described four putative safe harbour loci that could be useful for TGI engineering strategies. Therefore, we aimed to validate the first putatively identified safe harbour site locus described in the previous chapter for the first CRISPR-based TGI approach in any diatom. We targeted geraniol synthase encoding donor DNA to within 5 bp of *ch1:2,477,260*, one of the putative sites previously identified, and showed that this protocol was too inefficient to warrant its use for TGI. We also demonstrated for the first time the highly mutagenic effect of 5-FOA, a widely used selectable agent in yeast synthetic biology, but only recently described in *P. tricornutum*. This work is very timely as CRISPR-based editing

becomes widely used in *P. tricornutum*, robust and reliable endogenous markers will be essential.

INTRODUCTION

Randomly integrated chromosomal expression (RICE) is not appropriate for metabolic engineering

It is generally accepted that RICE causes position effect, whereby transgenes are integrated into important endogenous regions, or transcriptionally repressed regions, or loci that induce epigenetic silencing of the transgene (Jupe et al., 2019). However, this has not been well-defined and exemplified until recently (George et al., 2020). In Chapter 3 we showed that RICE is associated with exogenous DNA fragmentation and concatenation, as well as unpredictable transgene stability that is not associated with the level of transgene expression. For example, some cell lines identified with high exogenously supplied mVenus expression maintained high levels of fluorescence over three weeks of culturing without antibiotic pressure, whilst others demonstrated a dramatic reduction in expression. These characteristics are exponentially problematic when multi-gene knock-in is required, as would be the case for more complex metabolic engineering approaches. For example, we showed that a *ShBle* transcriptional unit was successfully expressed for zeocin resistance trait in 100% of RICE cells lines, however, only 24% of these cell lines demonstrated low to no expression of a second transcriptional unit, *mVenus*.

Second, metabolic engineering depends on precise and controlled expression of exogenous DNA to not only uncover more details of the organism's biology, but also to inform and test rational designs. This is already in practise in yeast synthetic biology, where automation pipelines (Li & Borodina, 2015; Linshiz et al., 2016; Si et al., 2017) have been built to test thousands of constructs simultaneously (Szita et al, 2010). We demonstrated that in *P. tricornutum*, RICE generates unstable expression–not only

between cell lines developed–but within them, too (George et al., 2020). For example, we demonstrated that some RICE engineered cell lines showed up to three sub-populations with differing transgene expression profiles within a single clone. This is particularly problematic, as the highly dissimilar RICE cell lines make it near impossible to reliably compare different genetic circuits and designs.

Finally, a major milestone for metabolic engineering–and synthetic biology more broadly–is the exceptional complexity of cellular biology and the homeostatic mechanisms regulating metabolism (Mijakovic et al., 2005). This sets metabolic engineering apart from recombinant protein genetic engineering. While recombinant gene expression requires optimisation to express transgenes as highly as possible for increased titres of the protein of interest, enzymes in a metabolic pathway–or proteins regulating them–require differing levels of expression and consequently, metabolic engineering requires the capability to fine-tune genetic parts (Mijakovic et al., 2005). For example, the expression of an enzyme might be limited by: the availability of chaperones required for its correct folding; or a negative feedback loop; or the rate of cellular synthesis, to name a few (Mijakovic et al., 2005). Our recent publication demonstrates that RICE is not suitable for fine-tuning in *P. tricornutum*, as transgene cassettes are subject to significant rearrangement which could impact promoter sequences (George et al., 2020).

Recently, we demonstrated that *P. tricornutum* can be engineered to produce up to 0.309 mg/L (0.21 µg/10^7cells) geraniol following extrachromosomal expression (EE) of the fusion protein CrGES-mVenus via bacterial conjugation (Fabris et al., 2020). This work represents the first successful metabolic engineering study in this diatom that was not reliant upon RICE. Previously, we showed that unlike RICE, EE resulted in mVenus fluorescence higher than that of wild type in 100% of transformants and

was not prone to any inadvertent integration of recombinant DNA into the chromosomal genome (George et al., 2020). However, the maximum geraniol yield following EE was 4.2 times lower than that of the highest geraniol yielding RICE transformant. Furthermore, episomes could be purged if selective pressure is not maintained (Slattery et al., 2018). Altogether, we hypothesise that it might be possible to combine the benefits of high, stable expression associated with chromosomal integration with the advantages of reliable and consistent expression associated with extrachromosomal expression through targeted chromosomal expression.

Targeted genomic integration (TGI) in model species

Targeted genomic integration (TGI) depends on the natural DNA repair process of homologous recombination (HR) which involves repairing a DNA lesion by recombining two DNA strands that are similar or identical (San Filippo et al., 2008). HR is a universal eukaryotic mechanism which plays key roles in repairing double stranded breaks that can occur spontaneously or following exposure to DNA-damage inducing agents (San Filippo et al., 2008; Symington, 2002). HR is also required in meiotic cells for maintaining karyotype stability and driving genetic diversity by recombining related alleles (San Filippo et al., 2008; Symington, 2002). Although HR only occurs in the S and G2 phase of the cell cycle, some species perform HR at a high rate and consequently, this mechanism can be exploited for integrating recombinant DNA into a precise location (Lieber, 2010). This occurs when 'donor DNA' designed to contain flanks that are homologous to the target locus is delivered into the cell. Such HR-driven TGI has been demonstrated in *S. cerevisiae* (Giaever et al., 2002; Gietz & Woods, 2002), *E. coli* (Datsenko & Wanner, 2000), certain cyanobacterial strains (Eaton-Rye, 2011; Lan et al., 2015), the thraustochytrid *Schizochytrium* (Cheng et al., 2011), the heterokont *Nannochloropsis oceanica strain*

W2J3B (Kilian et al., 2011) and the red alga *Cyanidioschyzon merolae 10D* (Minoda et al., 2004).

TGI can be used to generate engineered cell lines with stable knock-in of recombinant genetic cassettes, allowing for continual expression of transgenes without selective pressure (Gaidukov et al., 2018). This is often attributed to the integration location in the genome, often denoted as a ‘safe harbour’ locus or ‘neutral site’, which is a region that supports stable transgene expression. Neutral sites in cyanobacteria have been used for targeted integration in metabolic engineering for multigene pathway assembly (Bentley et al., 2014) and dual knock-in knock-out modifications (Li et al., 2016). Synthetic “landing pads” are useful for gene stacking via “domino cloning”, but depend on the knowledge of robust, reliable safe harbour loci prior to being feasibly applied to diatoms (Karas et al., 2015). Numerous synthetic biology strategies have taken this further to develop synthetic landing pads at these sites for a ‘plug-and-play’ approach to targeted engineering, which allows rapid and reliable targeted integration and expression. This has been demonstrated in Chinese hamster ovary (CHO) cells (Gaidukov et al., 2018), in maize (Ainley et al., 2013) and tobacco plants (Hou et al., 2014), and in microorganisms *E. coli* (Kuhlman & Cox, 2010) and *S. cerevisiae* (Bourgeois et al., 2018).

Increasing HR-mediated TGI via programmable endonucleases

Programmable endonucleases are enzymes capable of cleaving the phosphodiester bond of a double stranded DNA molecule at a precise location of interest (Gaj et al., 2013). Examples include transcription activator-like effector nucleases (TALENs), zinc finger nucleases (ZFNs), meganucleases and clustered regularly interspaced short palindromic repeats (CRISPR)-based systems. Research in species ranging from

maize (D'Halluin et al, 2008) to yeast (Dicarlo et al., 2013) to mammalian cells (Donoho et al., 1998) have all demonstrated that double stranded breaks caused by programmable endonucleases can drastically increase the likelihood of TGI at the site of the break. In the diatom *P. tricornutum*, TGI was unable to occur when donor DNA was supplied without the CRISPR-Cas9 nuclease (Daboussi et al., 2014).

TGI requires the delivery of donor DNA, as well as the programmable endonuclease– either as DNA encoding the endonuclease or the protein itself. Such endonuclease-driven TGI strategies have resulted large DNA integration events (over 100 Kbp) encoding 21 transcriptional units in Chinese Hamster Ovary (CHO) cells (Gaidukov et al., 2018) and complex metabolic engineering strategies; for example, the introduction of the carotenoid pathway using 15 DNA parts integrated at three targeted chromosomal locations using CRISPR Cas9 in *S. cerevisiae*, as well as a strain producing tyrosine using 10 parts integrated at two loci (Jakočiunas et al., 2015).

In *P. tricornutum*, TGI was first demonstrated following meganuclease-induced double stranded break at the target region (Daboussi et al., 2014), which was the first validation that the HR pathway is functional in this species (https://www.genome.jp/kegg-bin/show_pathway?org_name=pti&mapno=03440&scale=&orgs=&auto_image=&nocolor=&show_description=hide). Soon after, Weyman et al. (2015) demonstrated TALEN-induced TGI of a selectable exogenous cassette into the urease gene (Pt_29702; EC 3.5.1.5) to generate an insertional knock-out. Insertional knock-out publications are useful proof-of-concepts for TGI, and have been demonstrated in other microalgal species, such as *C. reinhardtii* (Ferenczi et al., 2017; Greiner et al., 2017; Shin et al., 2016). However, the loci chosen were selected for feasible screening and were not validated safe harbours, but instead chosen for the unique phenotypes

associated with them, such as the inability to grow with urea as a sole nitrogen source in *P. tricornutum* (Weyman et al., 2015) or a light green phenotype in *C. reinhardtii* (Greiner et al., 2017; Shin et al., 2016). These works are foundational and very important for demonstrating that endonuclease-driven integration is possible in both of these model microalgae. Despite these important efforts, TGI in microalgae has not yet been able to replace first generation RICE engineering. Furthermore, the microalgal research field is still without any ideal candidate safe harbour loci for either *P. tricornutum* or *C. reinhardtii*, which would be invaluable for HR-mediated TGI.

In order for researchers to move away from RICE engineering strategies, we need to develop reliable, easy and cheap TGI and EE protocols. This is already well on its way for EE, with a refined high-throughput method published in 2017 –allowing up to 12 transformations at a time– being reproduced for novel research in numerous different labs (Diner et al., 2017; Fabris et al., 2020; Pollak et al., 2019; Pollier et al., 2019; Sharma et al, 2018; Slattery et al., 2018; Turnšek et al., 2019), often showing up to 100 times higher transformation efficiencies than RICE (Sharma et al., 2018).

However, TGI is more complex in its requirements and involves endonuclease reagents and reliable protocols, from transformation through to screening and monoclonal cell selection; appropriate genomic information to guide integration location choices, such as safe harbours; surety that TGI is not coupled with extensive random integration at regions unrelated to targeted site; and robust evidence that TGI outweighs RICE in *P. tricornutum*. When assessing these characteristics, it is clear that TGI in *P. tricornutum* is only in its infancy, with only two publications to date demonstrating this engineering strategy (Daboussi et al., 2014; Weyman et al., 2015).

CRISPR revolution

While TALENs, ZFNs and meganucleases have been useful, they all depend on protein-DNA interactions requiring extensive protein engineering and limited target location options, as not every region on the genome is suitable for binding (Pennisi, 2013). On the contrary, CRISPR-based technologies have made endonuclease-driven TGI more cost and time effective in various species, as reviewed by Sternberg & Doudna (2015). This is because the CRISPR system depends on simple Watson-Crick base pairing instead of complex protein-DNA interactions. Moreover, CRISPR can be used for much more than just TGI and has been successfully used for multiplex gene knock-out including in *P. tricornutum* (Moosburner et al., 2020). In non-algal species, CRISPR has been adapted for epigenetic modification (Pflueger et al., 2018; Thakore et al., 2015), gene localisation (Chen et al., 2013; Roberts et al., 2017), genome-wide screening (Chen et al., 2015; Shalem, 2014), regulating gene circuits in synthetic biology (Kiani et al., 2014). Altogether, these highlight the importance and need to shift genetic engineering practises towards CRISPR-based technologies.

CRISPR-Cas9 driven TGI depends on reliable endogenous selectable markers

Our previous work provided the first putative safe harbours for *P. tricornutum* which should be investigated for their capacity to support stable, high transgene expression. Of the four sites provided, two were not associated with disruption of any protein coding genes (Chapter 3). In addition to this, the recently published RNP strategy for targeted knock out in *P. tricornutum* is very promising (Serif et al., 2018). This work reported the highest editing rate of 60% using a CRISPR ribonucleoprotein system and two endogenous selectable markers; native genes which generate a useful phenotype for screening when knocked-out.

One such marker, uridine-5'-monophosphate synthase (*UMPS*; Phatr3_J11740; EC 2.4.2.10; EC 4.1.1.23), is involved in *de novo* pyrimidine biosynthesis and encodes the URA3 homolog protein-coding gene. *UMPS* knock-out mutants are tolerant to 5-fluoroorotic acid (5-FOA) and are uracil auxotrophic (Sakaguchi et al., 2011; Serif et al., 2018). 5-FOA is an analogue of oroate, which is converted by UMPS to form orotidine-5'-phosphate (EC 2.4.2.10), which is then converted by UMPS again to form uridine-5'-monophosphate (EC 4.1.1.23), a precursor of uracil (Figure 1). In the presence of 5-FOA and uracil, the UMPS enzyme catabolises the 5-FOA analogue compound to generate 5-fluorouracil (5-FU), a toxic molecule that causes cell death (Figure 1). When *UMPS* is knocked out, the resultant mutant requires uracil supplementation in the medium, but is also unable to catabolise 5-FOA into toxic 5-FU**,** making this a useful selectable marker gene for screening on 5-FOA selection plates (Figure 1). Endogenous markers are a crucial aspect of RNP technology, as they allow for direct enrichment, whereby colonies obtained are the result of efficient CRISPR RNP delivery and editing activity (compared to imprecise random integration and expression of resistance genes).

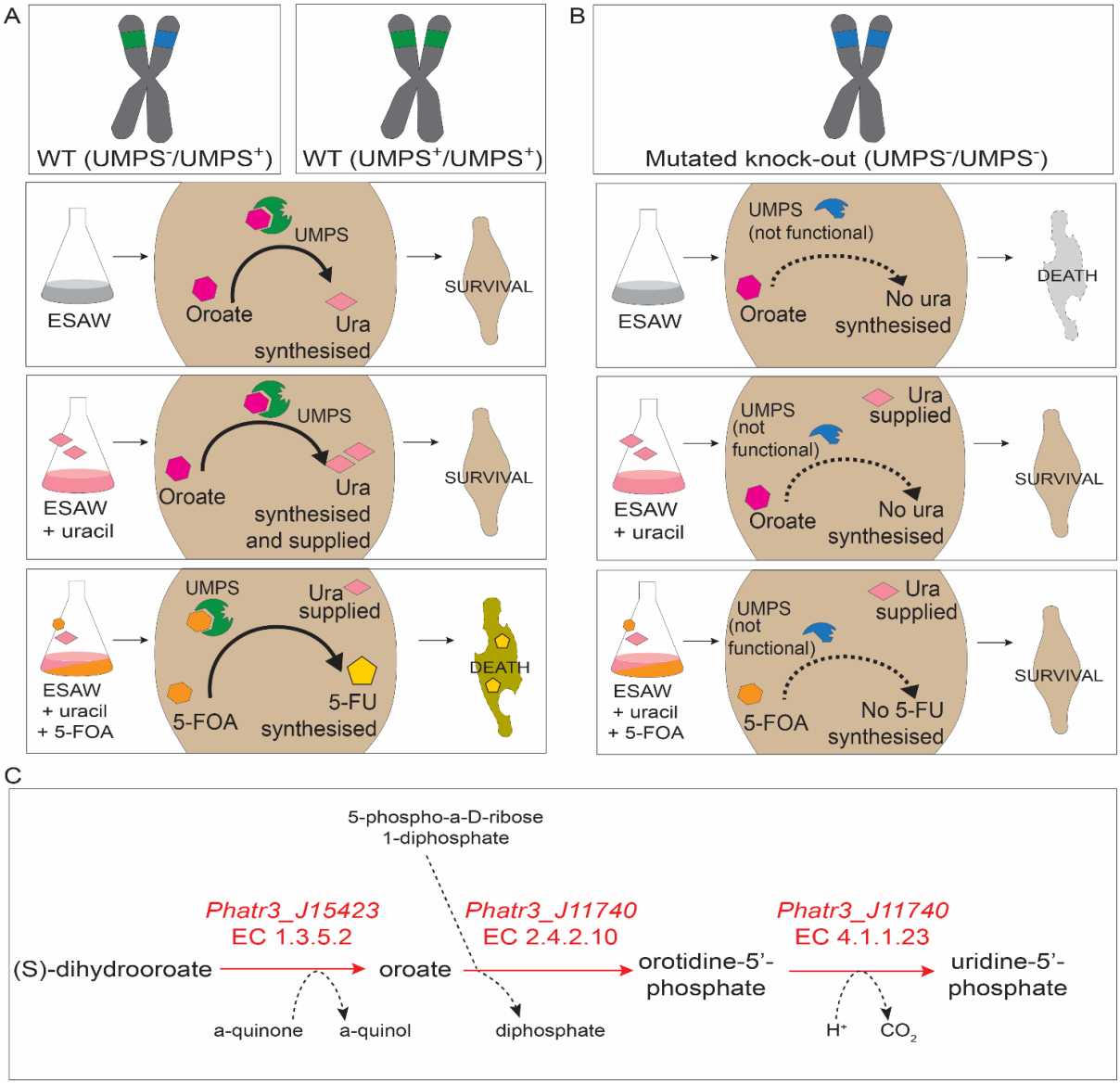


Figure 1. Graphic representation of uridine-5'-monophosphate synthase (*UMPS*; Phatr3_J11740) involved in de novo pyrimidine biosynthesis. (A) The heterozygous *UMPS* genotype from *P. tricornutum* UTEX LB 642 wild type strain sequenced by Sakaguchi et al. (2011) confirmed the presence of a functional and non-functional *UMPS* allele. Both the sequenced heterozygous and theoretical homozygous phenotypes will contribute to a functional UMPS enzyme able of converting oroate (dark pink hexagon) into uracil (light pink triangle) and survival of these strains in ESAW media with and without uracil supplementation. However, when supplemented with fluoroorotic acid (5-FOA) (orange hexagon), an analogue of oroate, *UMPS* will produce toxic 5-fluorouracil (5-FU) (yellow pentagon), which kills the cells. (B) On the contrary, a homozygous strain bearing both non-functional *UMPS* alleles–or an *UMPS* knock-out genotype–is not able to survive in ESAW unless uracil is supplemented. Furthermore, this mutant is able to survive in the presence of 5-FOA, as it is unable to use this substrate. (C) The biosynthesis of uridine-5'-phosphate from oroate in *P. tricornutum* is catabolised by the UMPS enzyme (Phatr3_J11740) in two subsequent reactions; the conversion of oroate to orotidine-5'-phosphate (EC 2.4.2.10) and then to uridine-5'-phosphate (EC 4.1.1.23).

CRISPR-Cas9 driven TGI in *P. tricornutum* for metabolic engineering

We previously demonstrated the heterologous production of geraniol in *P. tricornutum* following both extracellular and chromosome-integrated expression of *Catharanthus roseus geraniol synthase* fused to *mVenus* fluorescence reporter gene *(GES-mV).* This work demonstrated that while RICE was useful for generating stable, high mVenus fluorescing cell lines, these were very infrequent (occurring in less approximately 1% of the transformant cell lines following 13 transformation events) and harboured numerous large (hundreds of kilobases), highly concatenated integration islands (Chapter 3). We also demonstrated that EE resulted in transformants yielding approximately four times lower geraniol concentrations than the superior RICE cell lines; but that the EE cell lines were highly reliable and consistent, where 100% of exconjugants demonstrated an approximately 250-fold increased mVenus fluorescence compared to wild type auto fluorescence (Chapter 3).

Given the stability in high expression demonstrated in certain RICE cell lines (after growth without antibiotic supplementation for 3 weeks), we hypothesised that chromosomal integration in *P. tricornutum* could offer unique features that facilitate enhanced transgene expression or stability compared to extrachromosomal maintenance and expression. However, it is still not yet apparent whether this increase was due to (1) stable integration at favourable regions in the genome, known as safe harbours (Cantos et al, 2014; Hong et al., 2017; Papapetrou et al., 2011; Pinto et al., 2015; Sadelain et al., 2012); (2) other benefits related to the surrounding genomic architecture of the integration location, such as favourable chromosomal packing or epigenetic markers (Jain et al., 2018; Jupe et al., 2019); (3) the highly concatenated arrangement of the transgenes identified, which could result in enhanced chimeric regulatory regions; (4) a combination of these; or (5) yet to be determined factors.

Should the integration location or local genomic architecture be a significant contributor, it stands that targeting these sites in *P. tricornutum* would be highly desirable for TGI or synthetic landing pads aiding TGI approaches.

Although CRISPR-induced TGI has not yet been demonstrated in diatoms, Serif et al. (2018) recently described a highly efficient proteolistic RNP-based editing using the UMPS endogenous selectable marker in *P. tricornutum*. Therefore, we hypothesised that CRISPR-Cas9 RNP proteolistics and the *UMPS* gene marker could be used to drive targeted genomic integration (TGI) into one of the putative safe harbour loci identified in Chapter 3, combining the benefits of RICE and EE by taking advantage of stable, high expression associated with chromosomal integration, whilst generating more consistent phenotypes across cell lines.

RESULTS AND DISCUSSION

We previously demonstrated that Cr*GES-mVenus* can be extrachromosomally expressed for heterologous geraniol production in *P. tricornutum* generating yields up to of 0.21 µg/10^7 cells (Fabris et al., 2020). Therefore, we used donor DNA containing the same *P. tricornutum* functional zeocin resistance cassette and constitutive *Phatr3_49202p*-Cr*GES-mVenus* cassette (henceforth *GmV_ShBle* donor DNA) to attempt the first CRISPR-driven TGI in *P. tricornutum*. Conventional DNA transformations via biolistic bombardment usually require a chemical binding step to adhere the DNA molecules to the microprojectiles, using a combination of spermidine and calcium chloride (Apt et al., 1996; Falciatore et al., 1999). However, proteolistic delivery with an active CRISPR RNP cannot be treated with these chemicals and instead, air-drying techniques have been described (Martin-Ortigosa & Wang, 2014; Serif et al., 2018).

In order to demonstrate the first CRISPR-Cas9 mediated TGI event at a putative safe harbour locus, we co-delivered RNPs targeting *UMPS* endogenous marker, which generates a 5-FOA resistance uracil auxotrophic trait when knocked out, together with RNPs targeting 5-7 bp upstream of the putative safe harbor locus *ch1:2,477,260* identified in Chapter 3, as per Serif et al. (2018) air-drying method. For each site (referred to as *UMPS* and *ch1:247*, respectively) we targeted two loci using two RNPs in order to both improve chances of an edit occurring. This strategy was validated by Serif et al. (2018) who demonstrated that up to four guide RNAs, targeting the *UMPS* selection locus and locus of interest, can be used simultaneously.

We co-delivered these RNPs with *GmV_ShBle* donor DNA with the intention of demonstrating targeted gene integration (TGI) at *ch1:247* via NHEJ mediated mechanisms. This approach theoretically allows for donor DNA to be integrated at

either *UMPS* disrupted location, or *ch1:247* disrupted loci, or a combination of both. It also does not prevent possible off-target random integration. However, given the low frequency of only two random integration islands detected previously following biolistic bombardment with exogenous DNA (Chapter 3), it is possible that off-target integrations in this experimental set-up may be low.

Biolistic co-delivery of *GmV_ShBle* donor DNA and CRISPR RNPs using the *PtUMPS* endogenous selectable marker does not yield TGI clones

Two guide RNAs *UMPS*-1 and *UMPS*-3, were designed to target exon 1 and 2 of *UMPS*, respectively (Figure 2A and Suppl. Table 1). Similarly, two guide RNAs *ch1:247*-A and *ch1:247*-B, targeted the putative safe harbour locus, *ch1:247* (Figure 2A and Suppl. Table 2). All four guide RNAs were assembled with Cas9 protein to form RNP complexes, which were validated *in vitro* for their nuclease activity using PCR amplicons containing their recognition DNA sites (Figure 2B).

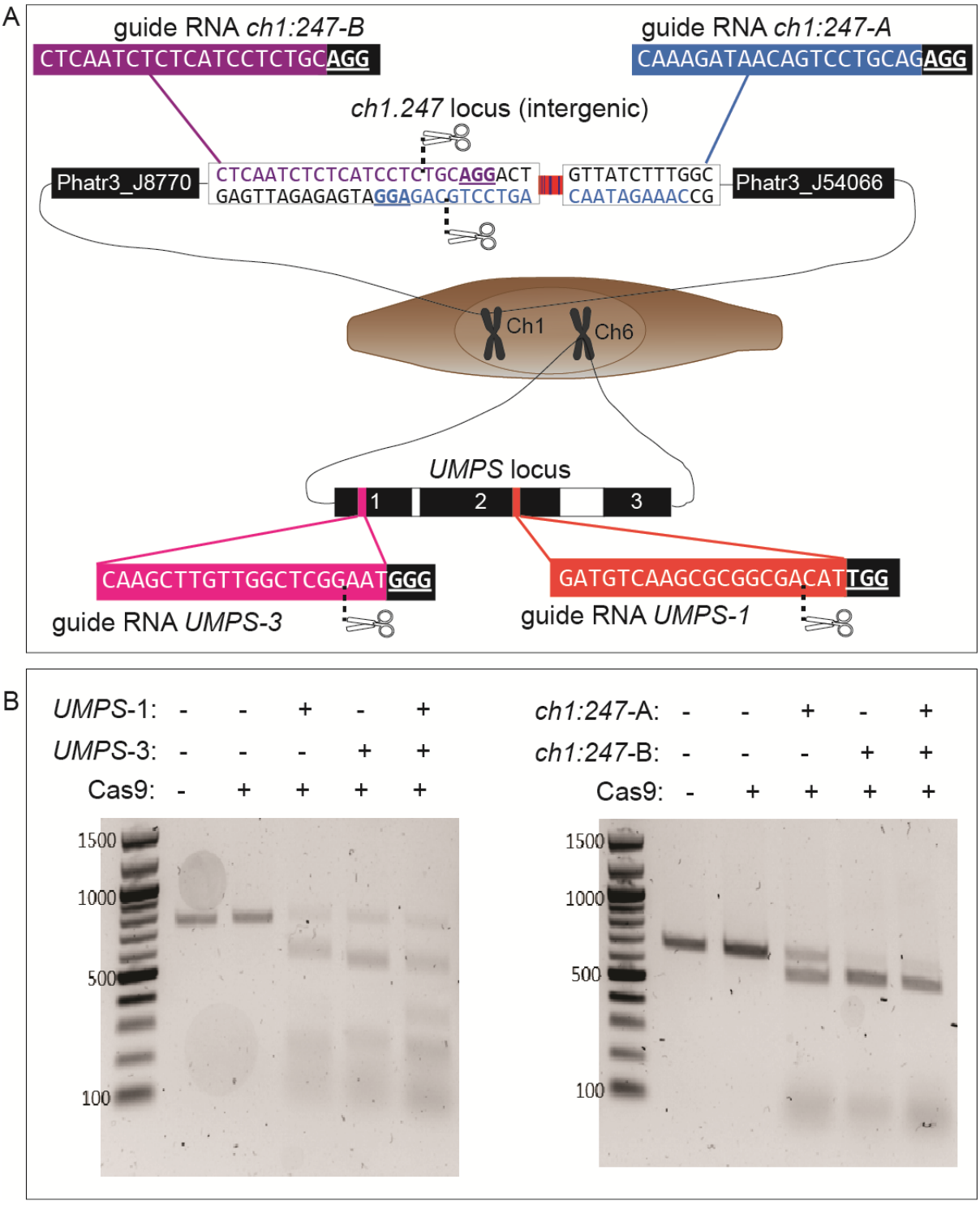


Figure 2. CRISPR-Cas9 RNP design and validation targeting *P. tricornutum UMPS* gene and putative safe harbour locus, the intergenic region *ch1:2,477,260*. (A) Graphic representation of the genetic sequences for RNP-UMPS-1 (red) and RNP-UMPS-3 (pink) guide RNAs targeting *UMPS*, and RNP-ch1:247-A (blue) and RNP-ch1:247-B (purple) guide RNAs targeting near to the putative safe harbour locus, *ch1:2,477,260*. All guide RNA recognition sites are 20 nt in length occurring alongside protospacer adjacent motifs (PAMs) (black, underlined). The ch1:247 locus occurs between predicted protein coding gene *Phatr3_J8770* and *Phatr3_J54066*. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break

location, which is always three base pairs upstream of the PAM site. (B) *In vitro* digest analysis of RNP-*UMPS*-1 and RNP-*UMPS*-3 activity to validate that RNP components were correctly assembled *in vitro* and are able to recognise and cut their target regions. The 740 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the expected 543 bp and 197 bp bands following incubation with fully assembled CRISPR-Cas9 RNP-*UMPS*-1. Likewise, the expected 516 bp and 224 bp bands occurred after incubation with RNP-*UMPS*-3. In the double digest with both RNP-*UMPS*-1 and RNP-*UMPS*-3, the expected 319 bp, 224 bp, and 197 bp bands are present, as well as the 516 bp band from RNP-*UMPS*-3 digest, indicating that some fragments were not fully digested. (C) *In vitro* digest analysis of RNP-*ch1:247*-B and RNP-*ch1:247*-B activity to validate that RNP components were correctly assembled *in vitro* and are able to recognise and cut their target regions. The 669 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the expected 565 bp and 104 bp bands following incubation with fully assembled RNP-*ch1:247*-A. The same expected pattern is seen following incubation with fully assembled RNP-*ch1:247*-B, however it appears RNP-*ch1:247*-B is more efficient than RNP-*ch1:247*-A, due to the faint undigested 669 bp band present in RNP-*ch1:247*-A. The double digest reaction would not allow for both RNP-*ch1:247*-A and RNP-*ch1:247*-B to digest the amplicon, as the activity of one would destroy the recognition site required for the other. However, we see that one of the two is able to digest the amplicon, which is useful to ensure *in vivo* digest, should one RNP be inactive.

We bombarded wild type *P. tricornutum* with microparticles coated with all four RNPs and *GmV_ShBle* donor DNA to generate targeted integration transformants, (henceforth Pt_TGI-Dual transformants). Targeting one gene with multiple guide RNAs has been shown to greatly increase the mutation frequency and the recovery of homozygous mutants in rice (Wang et al., 2016; Xie et al., 2015; Zhang et al., 2014). Pt_TGI-Dual transformants should result in a 5-FOA and zeocin double resistant phenotype which requires uracil supplementation and expresses the fusion protein, Cr*GES-mVenus*.

Although recent publications describing 5-FOA selection in *P. tricornutum CCAP1055/1* reported concentrations of 100 µg/mL 5-FOA (Serif et al., 2018; Slattery et al., 2020), the Daboussi laboratory provided us with a detailed, routinely protocol recommending 300 µg/mL 5-FOA. Furthermore, we conducted sensitivity tests by plating 150 $x10^6$ cells on plates containing varying concentrations of 5-FOA and

observed a mat of *P. tricornutum* cells at 100 µg/mL 5-FOA, a maximum of three colonies per plate at 300 µg/mL 5-FOA and complete cell death at 600 µg/mL 5-FOA. Given these results alongside the Daboussi laboratory protocol and Sakaguchi et al. (2011) report using concentrations of 100 – 300 µg/mL 5-FOA in *P. tricornutum*, we opted to use 300 µg/mL 5-FOA.

We performed three biolistic shots over two experiments, resulting in a total of six transformations, each containing 1.5 x 10^8 cells. Surprisingly, we did not identify a single Pt_TGI-Dual transformant colony from any of the transformations (Figure 3A). To confirm none of the 'tiny spots' present on the selection plate were very small possible colonies, we inoculated these specks in ESAW medium supplemented with 50 µg/mL uracil without 5-FOA; however, none of these specks grew, indicating they were never true diatom colonies (Figure 3B). In control transformations, microprojectiles were coated in RNPs targeting only *UMPS* and the *GmV_ShBle* donor DNA in order to generate Pt_TGI-UMPS transformants. However, we did not identify a single Pt_TGI-UMPS transformant colony from any of the transformations (Figure 3A and B).

To confirm that the *GmV_ShBle* donor DNA was in fact able to be delivered, integrated and expressed successfully via RICE, and that any lack of TGI colonies was not due to problems with *GmV_ShBle* donor DNA integration or zeocin resistance, we performed six transformations using only *GmV_ShBle* donor DNA. Following selection with zeocin, we identified only one single transformant in this condition, referred to as Pt_ *GmV_ShBle* (Figure 3). This suggested that the lack of colonies obtained following co-delivery with donor DNA was due to an exceptionally low transformation efficiency occurring at a rate of 1 in 9 x 10^8 cells in the presence of 6 µg donor DNA. We also confirmed that the donor DNA used was able to confer zeocin resistance, as the single

donor DNA colony was inoculated into ESAW supplemented with zeocin (Ez50) and was able to grow (Figure 3B, C).

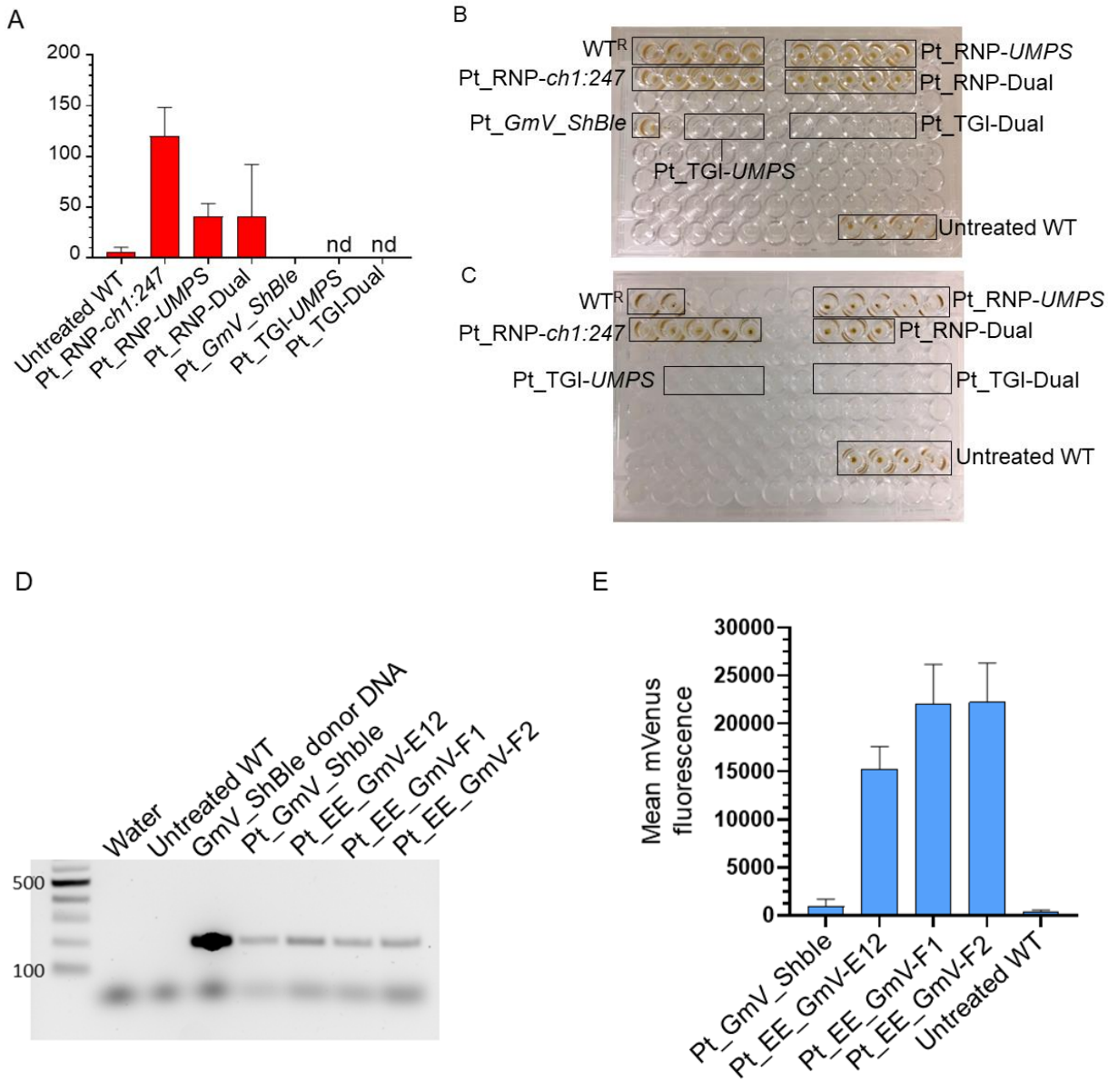


Figure 3. CRISPR-Cas9 RNP proteolistic bombardment does not mediate targeted genomic integration (TGI). (A) Number colonies appearing on selection plates (n = 2 for control transformations and n = 6 for CRISPR-Cas9 proteolistic transformations, error indicates SEM). Small specs which could be tiny colonies or could be non-living background were the only 'colony-like' spots on the TGI integration plates. (B, C) Growth of picked colonies in liquid culture confirming that TGI tiny specs were not colonies but just background from two independent experiments, where experiment 1 is shown in B and experiment 2 is shown in C). (D) PCR verifying that the donor DNA retained in mutant populations following both RICE and EE. (E) mVenus fluorescence of mutants generated by RICE and EE of donor DNA was able retained in mutant populations following both RICE and EE expressed by RICE and EE.

Furthermore, we detected the presence of *mVenus* from genomic DNA extracted from Pt_ *GmV_ShBle* (Figure 3D) and detected mVenus fluorescence in this strain (Figure 3E) compared to wild type auto-fluorescence control, albeit at a very low abundance compared to higher fluorescence associated with extrachromosomally expressing cell lines (Fabris et al., 2020). This is expected, given that 25% RICE clones show low to no transgene expression as demonstrated in Chapter 3. Together, these results suggested that co-delivery of *GmV_ShBle* donor DNA and RNP via the air-drying method is not appropriate for TGI due to low efficiency, most likely due to poor binding of DNA to the microprojectiles.

To confirm that the 5-FOA resistance phenotype was uniquely associated with *UMPS* gene disruption and not randomly associated with *ch1:247* disruption, we bombarded microparticles coated in RNP-*ch1:247*-A and RNP-*ch1:247*-B only and plated these on ESAW agar plates supplemented with 300 µg/mL 5-FOA and 50 µg/mL uracil (henceforth Pt_RNP-*ch1:247* mutants). To our surprise, we obtained an average of 110 Pt_RNP-*ch1:247* mutant colonies per transformation (Figure 4A), suggesting that the 5-FOA resistance phenotype was not uniquely associated to *UMPS* knockout. This result was confirmed by a second negative control, in which untreated wild type cells were also plated onto ESAW agar plates supplemented with 300 µg/mL 5-FOA and 50 µg/mL uracil, resulting in a total of ~9 wild type resistant (henceforth WT^R) strains per transformation (Figure 4A). Together, the presence of Pt_RNP-*ch1:247* and WT^R strains were very concerning, as neither of these conditions should result in any 5-FOA resistant cell lines. These results suggested that either the concentration of 5-FOA used for selection was too low for UMPS enzyme to favour 5-FOA substrate over oroate, resulting in a lack of toxic 5-FU build up and consequent cell death; or that 5-

FOA may be mutagenic to *P. tricornutum*, inducing mutations that allow for a resistance phenotype.

To test the unlikely event that the RNP we designed would cause off-target mutations leading to cell death, we also included two additional positive control conditions. To confirm that *UMPS* mutated cell lines can survive in the presence of 300 µg/mL 5-FOA and 50 µg/mL uracil, we coated microprojectiles in RNP-*UMPS*-1 and RNP-*UMPS*-3 only, which resulted in an average of 30 Pt_RNP-*UMPS* colonies per transformation following bombardment (Figure 4A). To confirm that dual *UMPS* and *ch1:247* mutated cell lines can survive in the presence of 300 µg/mL 5-FOA and 50 µg/mL uracil, we coated microprojectiles in RNPs targeting both *UMPS* and *ch1:247* and obtained an average 30 Pt_RNP-Dual mutant colonies per transformation following bombardment (Figure 4A). While these conditions are positive controls, the presence of colonies obtained in the negative controls suggested these may be false positives generated by a chemically induced mutation in wild type diatoms. However, it is possible that some of the Pt_RNP-*UMPS* and Pt_RNP-Dual mutants were generated by CRISPR-Cas9 driven mutations.

To determine the presence of genuine CRISPR-induced knock-out mutants, we screened sub-colonies via PCR amplifying the *UMPS* locus. In order to obtain sub-colonies, primary colonies were streaked on fresh 300 µg/mL 5-FOA and 50 µg/mL uracil ESAW agar plates to allow for sub-colony growth. We chose to analyse sub-colonies instead of primary colonies because it is possible that bombarded cells can divide before the RNP editing event has occurred, which would result in a mixed colony known as colony mosaicism (Greiner et al., 2017; Huang & Daboussi, 2017; Serif et al., 2018). Direct delivery of RNP protein can greatly reduce the likelihood of colony mosaicism (Serif et al., 2018). However, we chose to still screen sub-colonies because

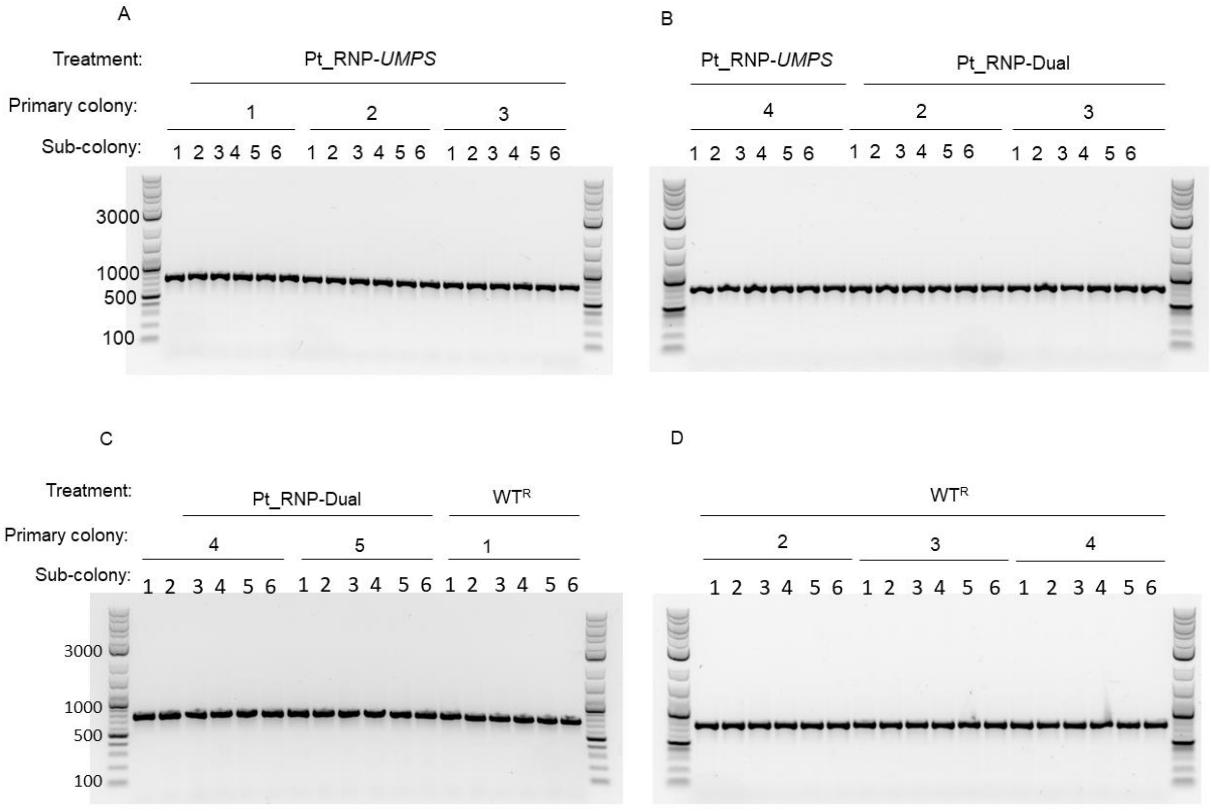
this approach is so under studied in diatoms and because of the high rate of false positives which we obtained.

In the case of the wild type control or hypothetically non-mutated colonies, the PCR should result in amplification of a single 740 bp band. The dual action of RNP-*PtUMPS*-1 and RNP-*PtUMPS*-3 should result in a 421 bp deletion; however, given that NHEJ driven repair causes variable INDELs, it is possible that the edited amplicon could be larger, smaller or equal to the unedited amplicon size of 740 bp. Therefore, this assay offers a useful quick, affordable, high throughout screen.

Six single sub-colonies from four primary Pt_RNP-*UMPS* colonies (total of 24 sub-colonies), and six single sub-colonies from four Pt_RNP-Dual colonies (total of 24 sub-colonies) were screened, as well as 30 WT^R sub-colonies and three control WT colonies that were not bombarded, nor exposed to 5-FOA (referred to as untreated WT). PCR amplification of the *UMPS* amplicon harbouring the recognition sites of RNP-*UMPS*-1 and RNP-*UMPS-3* showed no shift in the *UMPS* amplicon in any of the 81 colonies (Figure 4), which suggested that there was no CRISPR-Cas9 induced edit in any of these strains at the *UMPS-1* and *UMPS-3* loci.

These results do not confirm that there was no single detectable successful CRISPR edit in any of these screened cell lines. For example, it is possible that amplification of a heterologous knock-out mutant may favour the unedited allele over an edited allele. However, it is unlikely that this is true for all 48 RNP sub-colonies analysed, or that none of the 48 we screened contained a homozygous knock-out genotype. It is also possible that the edits that occurred were too small to detect (tens of base pairs in size) by mobility shift assay. However, this is again highly unlikely, given the use of two RNPs which would result in a 421 bp deletion. Furthermore, Serif et al. (2018) reported a 60% editing efficiency rate, which would equate to approximately 20 sub-

colonies showing an edit in our screen. Moreover, the scope of this work was to develop and validate a method for TGI in *P. tricornutum* using the recently described endogenous gene marker, *UMPS,* for direct enrichment using 5-FOA selection. Therefore, we did not progress to more labour-intensive, costly screening by T7E1 or sequencing to confirm the lack of CRISPR-Cas9 induced mutation. Instead, these findings were important for highlighting that *UMPS* knock-out may not be appropriate for genetic studies in *P. tricornutum* strain CCAP 1055/1 due to the putative mutagenic effects of 5-FOA on this species, which have never before been reported. Consequently, we set out to characterise these 5-FOA resistant cell lines.



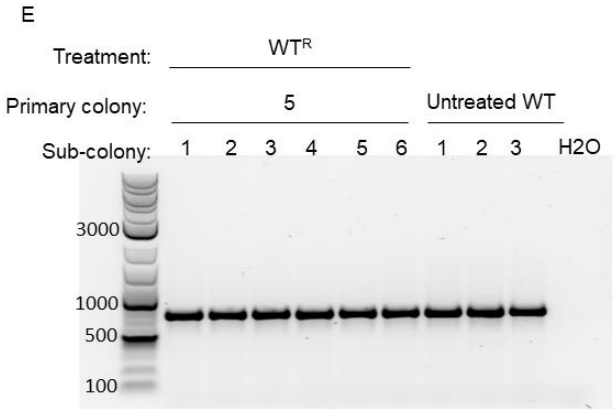


Figure 4. Electrophoretic mobility shift assay to screen sub-colonies for CRISPR-Cas9 driven knock-out of *UMPS* gene following proteolistic bombardment with two RNPs targeting *UMPS* (2% agarose). Primers annealing to regions approximately 200 bp up and downstream of the recognition sites of RNP-*UMPS*-1 and RNP-*UMPS*-3 were used to detect an edit in these regions. Edited cell lines are expected to have variable sized amplicons between these regions due to INDELs caused by NHEJ mediated repair of CRISPR-Cas9 driven double stranded breaks. The unedited wild type amplicon is expected to be 740 bp.

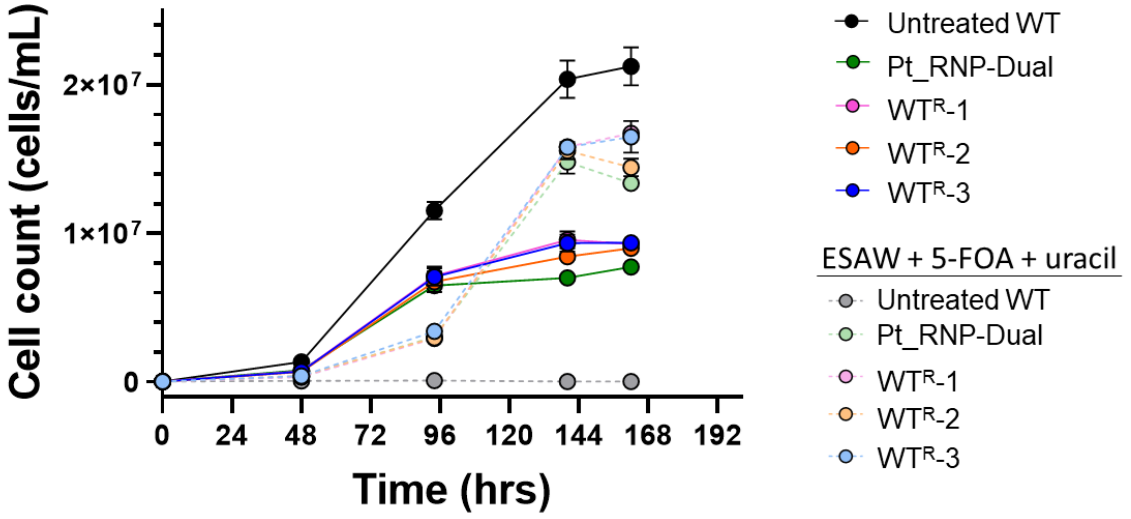
Wild type *P. tricornutum* exposed to 5-FOA results in a stable RURF phenotype

To date, there are no reports of 5-FOA induced mutation in *P. tricornutum*. However, Sakaguchi et al. (2011) reported that wild type *P. tricornutum* could become relying on uracil and resistant to 5-FOA (RURF) following chemical mutagenesis using N-ethyl-N-nitrosourea (ENU) and selection using 100 - 300 µg/mL of 5-FOA (Figure 2A). We hypothesised that the WT^R strains generated in this study had the same RURF phenotype described by Sakaguchi et al. (2011).

Three WT^R primary colonies and three untreated WT colonies were grown in the absence and presence of 300 µg/mL 5-FOA and 50 µg/mL uracil over seven days to compare their growth rates. Because we were unable to detect any CRISPR-Cas9 induced edits, we speculated that the colonies bombarded with RNP were in fact false positives, and consequently also represented WT^R mutants. Therefore, we also included one Pt_RNP-Dual colony.

As expected, untreated wild type grew normally in ESAW, indicating a functional UMPS enzyme and normal uracil biosynthesis, whereas untreated wild type cultured in 5-FOA and uracil supplemented media showed no cell growth due to the functional UMPS enzyme metabolising 5-FOA into toxic 5-FU (Figure 2A). In ESAW, the WT^R and Pt_RNP-Dual strains grew at a reduced rate in the early exponential phase (from 48 hrs to 96 hours (Figure 5), most likely due to the declining presence of an intracellular uracil pool available to the cells for RNA biosynthesis. Once depleted, cell growth was drastically slowed (from 96 hrs to 144 hrs) and eventually arrested at approximately 144 hrs, due to the inability to synthesise uracil without the functional UMPS enzyme (Figure 5A). In contrast, these strains were all able to grow in the presence of 5-FOA and uracil, confirming that they are unable to metabolise 5-FOA into toxic 5-FU, but are able to use supplemented uracil for RNA biosynthesis.

A



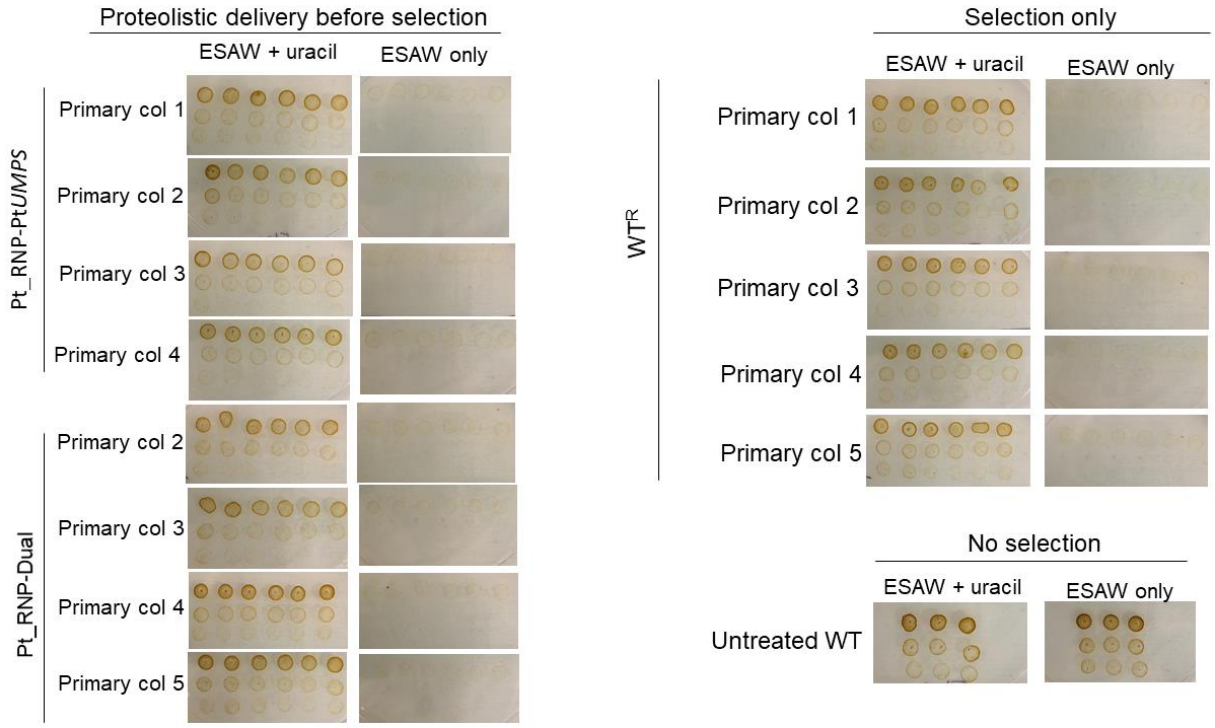


Figure 5. RURF phenotype in WT^R and Pt_RNP-Dual strains. (A) Growth analysis of untreated wild type, WT^R and Pt_RNP-Dual cultured in ESAW media supplemented with 5-FOA and uracil (dotted lines) compared to ESAW media without supplementation (solid lines). (B) Uracil auxotrophy analysis of 81 sub-colonies starved of uracil prior to dilution plating on ESAW agar plates with and without uracil.

These results confirmed that the WT^R and Pt_RNP-Dual strains screened were indeed 5-FOA resistant; however, it is not clear whether or not they are uracil auxotrophic. Therefore, we selected six single sub-colonies from four primary colonies (total of 24 sub-colonies per treatment) from Pt_RNP-*UMPS* and Pt_RNP-Dual bombarded colonies. We also included 30 WT^R sub-colonies and three WT colonies for a total of 81 colonies. We cultured the 81 sub-colonies without uracil for one week to deplete their endogenous cellular uracil pools and then dilution plated them onto ESAW or ESAW supplemented with uracil. We demonstrated that all 5-FOA exposed cell lines, including WT^R and RNP bombarded strains, were uracil auxotrophic and did not grow on ESAW plates at all, whereas the control wild type was able to grow without uracil supplementation.

Together, these results confirm that 5-FOA exposure results in chemical mutagenesis which can cause a RURF phenotype in WT^R strains. Furthermore, we have demonstrated that the false positives obtained have the perfect CRISPR-edited phenotype (5-FOA resistance and uracil auxotrophy) associated with knock-out of *UMPS* endogenous marker gene. This demonstrates that *UMPS* is not an appropriate endogenous marker for CRISPR screening.

While this work highlights the issue of false positives arising from using *UMPS* as a knock-out selectable marker, it may be possible to use one of the resultant WT^R strains for *UMPS* knock-in selection. Here, the uracil dependent WT^R strain would be transformed with donor DNA containing a functional *UMPS* gene copy, and transformation reactions could be screened on media without uracil. Only transformants that express the exogenously supplied *UMPS* gene would survive, which would be particularly useful for antibiotic free selection systems. Therefore, we next set out to confirm the genotype associated with 5-FOA induced chemical mutation and to determine if this differed between WT^R strains.

Genotyping WT^R strains revealed large mutations consistent with other 5-FOA sensitive organisms

In 2008, Sakaguchi et al. reported that wild type *P. tricornutum* UTEX LB 642 has a heterozygous *UMPS* genotype made up of one functional allele (allele 1), encoding the full UMPS enzyme, and one non-functional allele (allele 2), which contained 16 single nucleotide polymorphisms (SNPs) (Figure 1A). The SNPs present in the non-functional allele 2 resulted in 5 amino acid changes in the UMPS protein, none of which occurred in either of the two active domains, orotidine-5'-phosphate de- carboxylase (OCD) and purine/pyrimidine phosphoribosyl transferase (PPRT). Following treatment with *N*-ethyl-*N*-nitrosourea (ENU) and 5-FOA selection, Sakaguchi et al. identified

RURF mutant strains that required uracil and were 5-FOA resistant. Sanger sequencing revealed that the RURF mutants were homozygous for the non-functional allele 2, suggesting that the mutagenesis caused the loss of the functional allele 1 (Figure 1A).

In order to determine the genotypes of the *P. tricornutum* CCAP 1055/1 untreated WT strain and WT^R mutants used in this study, we too amplified *UMPS* and performed Sanger sequencing. We screened one untreated WT and three WR^R mutants WT^R-1-1, WT^R-2-1, and WT^R-2-2. Given the likelihood that the Pt_RNP-Dual colonies represent false positives, we included one Pt_RNP-Dual cell line.

While we were able to amplify the full *UMPS* locus in the untreated WT sample, we were not able to amplify it from any of the WT^R strains or the Pt_RNP-Dual strain. Therefore, we tested three different regions of *UMPS* to amplify, both within and outside the *UMPS* coding region, using six different primer combinations: (i) a region within the *UMPS* gene (ii) the full *UMPS* gene from start to stop codon and (iii) a region spanning 3' end of the upstream neighbouring gene, *Phatr3_45195,* including the full *UMPS* gene through to the 5' end of the downstream neighbouring gene, *Phatr3_45193* (Figure 6A and B).

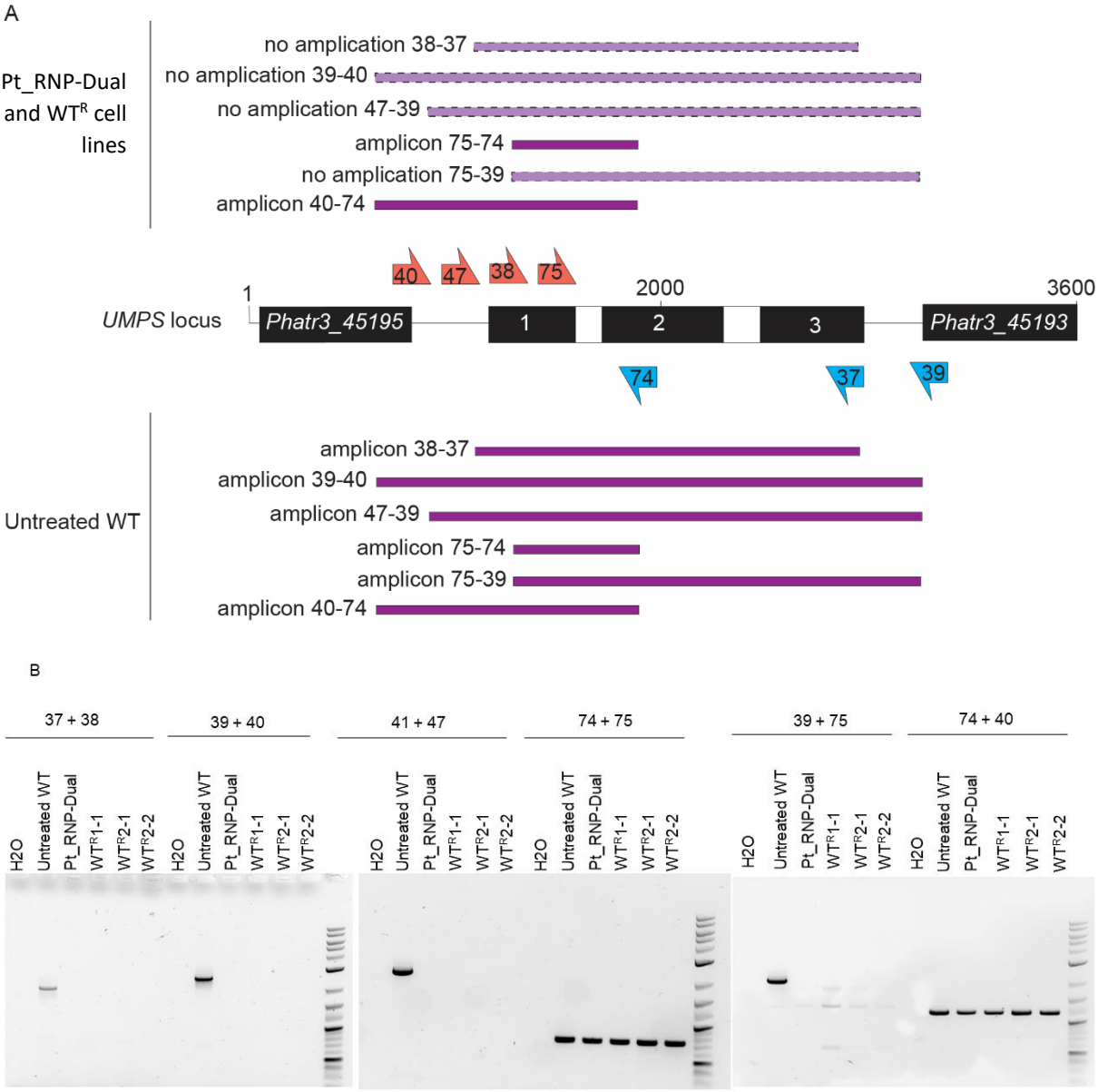


Figure 6. Analysis of *UMPS* genotype in untreated wild type *P. tricornutum* CCAP 1055/1 strain, WT^R strains and a Pt_RNP-Dual strain. (A) Graphic representation of the *UMPS* locus and neighbouring predicted protein coding regions *Phatr3_J45195* and *Phatr3_J45193*. Red markers indicate left primers and blue markers indicate right primers. Amplicons obtained from WT^R and Pt_RNP-Dual strains are depicted above the *UMPS* locus in purple fragments, where the light purple fragment indicates the region which was not amplifiable. Amplicons obtained from untreated wild type are depicted below the *UMPS* locus. Amplicons are named based on the forward and reverse primer ID codes used to amplify those products. (B) Amplicons from all six PCR amplifications.

We obtained correctly amplified PCR products from untreated WT samples using all six primer combinations (Figure 6A and B). However, we were only able to amplify a

small region within the *UMPS* gene for WT^R-1-1, WT^R-2-1, WT^R-2-2 and the Pt_RNP-Dual RNP mutants (Figure 6A and B). We were unable to amplify the full gene from start to stop codon, nor the full *UMPS* gene with 5' and 3' flanks of bordering neighbour genes for these cell lines, all of which were exposed to 5-FOA selection.

Sanger sequencing of the full *UMPS* gene from the untreated WT sample revealed a heterozygous genotype consisting of the same 16 SNPs identified by Sakaguchi et al. (SNP-1–16), as well as an additional three SNPs (SNP-A–C) (Suppl. Figure 1). These were identified by the presence of peak doublets in the sequencing chromatogram (Figure 7A). This suggested that untreated WT *P. tricornutum* CCAP 1055/1 contained one functional *UMPS* allele (allele A) and a second non-functional allele (allele B). Therefore, we hypothesised that WT^R mutants contained only the non-functional allele B.

Given that we were only able to amplify the 5' end of *UMPS* in the WT^R mutants, we sequenced amplicon 40-74 from WT^R-1-1, WT^R-2-1, WT^R-2-2 (Figure 6). As expected, sequencing confirmed a homozygous genotype in all three WT^R strains. However, the results confirmed both copies of *UMPS* contained functional allele A, evident by the loss of all the SNPs present in *UMPS*-frag. Interestingly, SNP-14 of all three WT^R strains had reverted to a guanine instead of an adenine, as was predicted by the *UMPS* sequence *Phatr3_J11740*. This was confirmed across three separately sequenced reads (read 1, 2 and 3) for WT^R-1-1, WT^R-2-1, WT^R-2-2, evident by the loss of the doublet peak in the chromatograms (Figure 7B). Consequently, the WT^R genotype is homozygous with a single point mutation and stunted coding region. This translated to a truncated protein (254 amino acids instead of the full 518 amino acids) and the loss of the PPRT active domain, as well as a single point mutation causing an amino acid change from isoleucine to methionine (Supp. Figure 2).

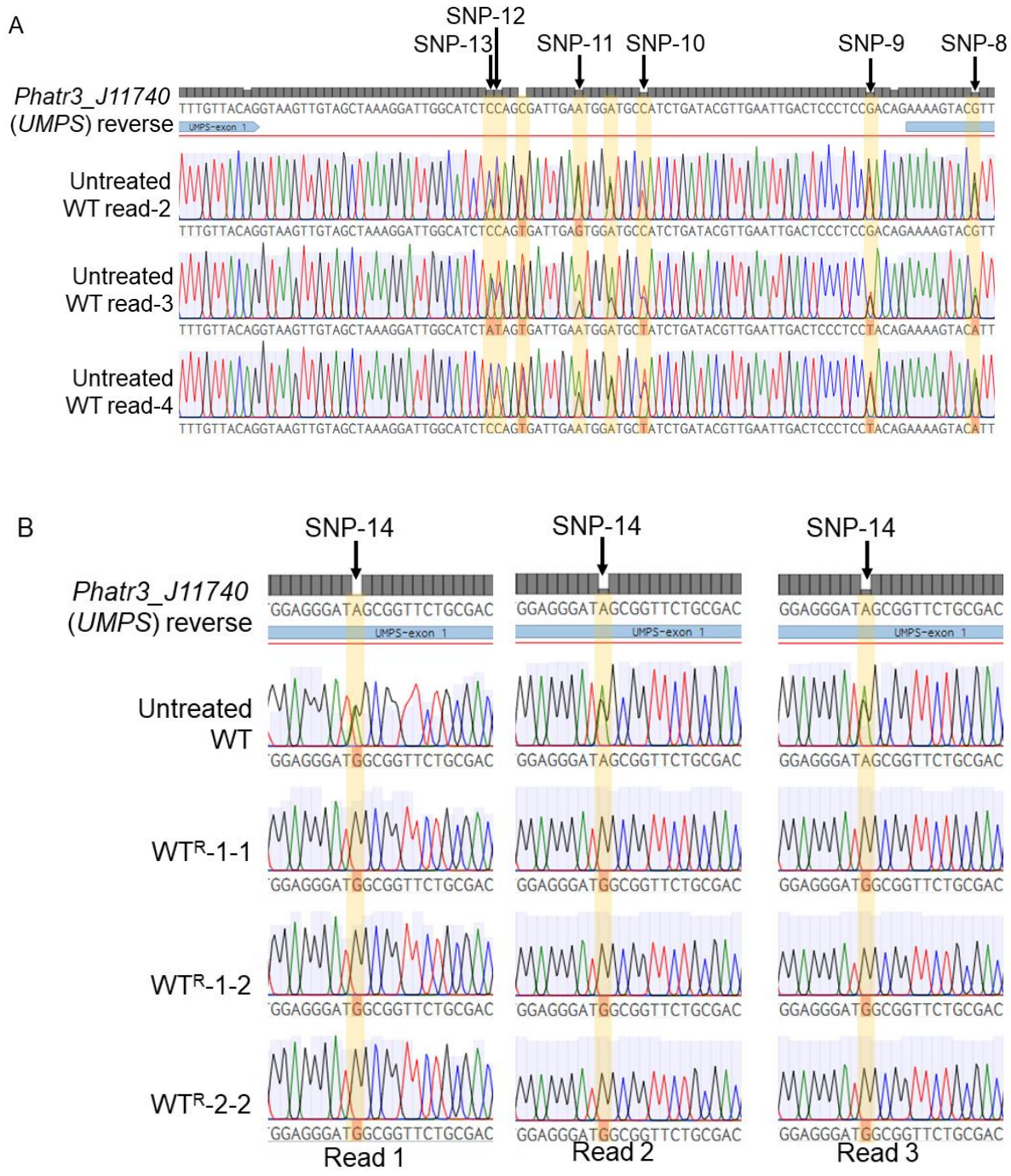


Figure 7. Analysis of *UMPS* genotype in untreated wild type *P. tricornutum* CCAP 1055/1 strain (A) The presence of doublet peaks at precise locations indicate single nucleotide polymorphisms characteristic of two non-identical gene copies (alleles) resulting in a heterozygous genotype in diploid organisms. These doublets occur in exogenic (blue) and intronic regions. The SNPs identified in this region are identical to those identified by Sakaguchi et al. (2011). (B) SNP-14 in wild type strain has been lost in all three WT^R strains as evident by the loss of the doublet peak confirmed across three individually sequenced reads per cell line, indicating a homozygous genotype.

Therefore, we rejected our hypothesis and concluded that the RURF phenotype in *P. tricornutum* CCAP1055/1 WT^R strains was not caused by the loss of a functional allele as reported by Sakaguchi et al. (2011), but instead was due to a larger chromosomal mutation which caused severe truncation of UMPS enzyme and the loss of the PPRT active domain. This is supported by the fact that we were not able to amplify the 3' end of *UMPS*, nor the neighouring 3' gene *Phatr3_45193* in any of the WT^R strains; but could amplify it successfully in the untreated WT samples (Figure 6A and B). Such large chromosomal mutations have indeed been reported in other species, such as the loss of entire chromosomes following 5-FOA exposure in *Candida albicans* (Wang et al., 2004; Wellington et al., 2006; Wellington & Rustchenko, 2005).

CONCLUSION

In this study, we set out to perform TGI in *P. tricornutum* at a putatively identified safe harbour locus, *ch1:2,477,260*, using CRISPR-Cas9 RNP technology. To date, there has been no validation of any putative safe harbour loci, nor any CRISPR-Cas9 mediated TGI in the model diatom, *P. tricornutum*. While we showed that the donor DNA designed for targeted integration and expression of geraniol synthase fused to mVenus could be expressed by RICE, we report no CRISPR-Cas9 driven targeted integration mutants due to the low efficiency of biolistic bombardment delivery of CRISPR-Cas9 RNP.

Importantly, we have demonstrated that selection using 300 µg/mL 5-fluoroorotic acid (5-FOA) can chemically induce mutations in *P. tricornutum* CCAP 1055/1, resulting in the requiring uracil resistance to 5-FOA (RURF) phenotype. The RURF phenotype has previously been be achieved in *P. tricornutum* strain UTEX LB 642 following chemical mutagenesis using N-ethyl-N-nitrosourea and selection with 5-FOA (Sakaguchi et al., 2011), and in *P. tricornutum* strain CCAP1055/1 following targeted gene knock-out of *UMPS* gene via CRISPR-Cas9 gene editing (Serif et al., 2018; Slattery et al., 2020). Serif et al. (2018) reported successful CRISPR-Cas9 gene editing following delivery of RNP targeting *UMPS* and selection using 100 µg/mL 5-FOA. Slattery et al. (2020) demonstrated CRISPR-Cas9 edited *UMPS* following extrachromosomal expression of CRISPR components and selection using zeocin or nourseothricin. They then demonstrated the RURF phenotype by spotting exconjugants onto plates containing 100 µg/mL 5-FOA. Taken altogether, it is possible that 5-FOA is only mutagenic at concentrations above 100 µg/mL 5-FOA; however, we found that such low concentrations had no impact on *P. tricornutum* cell growth and therefore would not be sufficient for selection. These results also suggest that it is possible that 5-FOA is

inappropriate for selection, as chemical mutagenesis cannot be excluded. However, as demonstrated by Slattery et al (2020) and in this study, 5-FOA can be used to discriminate a 5-FOA resistance phenotype from wild type phenotype following spot tests.

Although this is the first demonstration of 5-FOA mutagenesis in *P. tricornutum*, 5-FOA has shown mixed mutagenic effects in different species. 5-FOA is a commonly used selective agent in yeast synthetic biology including *S. cerevisiae* (Backhaus et al., 2017; Klein et al., 2014; Scott et al., 2018) and others, such as *Y. lipolytica* (Lv et al., 2019); as well as *E. coli* synthetic biology (Brandsen et al., 2018; Standage-Beier et al., 2015). However, 5-FOA is highly mutagenic in *C. albicans* (Wang et al., 2004; Wellington et al., 2006; Wellington & Rustchenko, 2005). There is also a report of 5-FOA inducing mutation in *Sulfolobus acidocaldarius* (Kondo et al., 1991) and in *S. cerevisiae* (Hao et al., 2016). Furthermore, 5-FOA has been used for generating spontaneous RURF phenotypes in the dinoflagellate *Symbiodinium SSB01* following exposure to 200 µg/mL 5-FOA (Ishii et al., 2018), and the red alga *Cyanidioschyzon merolae 10D* following treatment with 800 µg/mL 5-FOA (Minoda et al., 2004).

Given the early state of TGI in *P. tricornutum* and the small range of endogenous selectable markers validated in *P. tricornutum*, it is important that they are well characterised before they become widely used. This work demonstrates the risk of false positives or RURF strains, which can easily be generated by exposure to 5-FOA as demonstrated by our WT^R lines. Finally, these RURF strains are not appropriate for *UMPS* knock-in selection as described by Ishii et al. (2018) and Sakaguchi et al. (2011), as the chemically induced mutation is not contained within the *UMPS* locus. Further work to characterise the 5-FOA chemically induced mutagenised genome, such as whole-genome sequencing using MinION or PacBio, will help to provide

further guidance as to the appropriateness of 5-FOA for *P. tricornutum's* genetic engineering toolkit.

METHODS

Microbial strains and growth conditions

Phaeodactylum tricornutum CCAP1055/1 was grown in liquid ESAW (Berges et al., 2001) supplemented with 50 µg/mL zeocin (Invivogen, San Diego, CA, USA) where appropriate, under 100 µE m^{-2} s^{-1} light in 21 °C shaking at 95 rpm. *P. tricornutum* engineered strains containing *pPtPBR11_49p-CrGES-mVenus-FCBPt* episomes are described in Fabris et al. (2020). Where appropriate, *P. tricornutum* was cultured in ESAW containing 300 µg/mL 5-FOA and 50 µg/mL uracil, or ESAW supplemented with 50 µg/mL zeocin. *Escherichia coli* was grown in Luria broth supplemented with 100 µg/mL ampicillin.*Phaeodactylum tricornutum* CCAP1055/1 was grown in liquid ESAW (Berges et al., 2001) supplemented with 50 µg/mL zeocin (Invivogen, San Diego, CA, USA) where appropriate, under 100 µE m^{-2} s^{-1} light in 21 °C shaking at 95 rpm. *P. tricornutum* engineered strains containing *pPtPBR11_49p-CrGES-mVenus-FCBPt* episomes are described in Fabris et al. (2020). Where appropriate, *P. tricornutum* was cultured in ESAW containing 300 µg/mL 5-FOA and 50 µg/mL uracil, or ESAW supplemented with 50 µg/mL zeocin. *Escherichia coli* was grown in Luria broth supplemented with 100 µg/mL ampicillin.

Donor DNA preparation

The *pPtPBR11_49p-CrGES-mVenus-FCBPt* episomes were propagated in *E. coli* strain Top10 and purified by Monarch Plasmid Miniprep Kit (New England Biolabs, Hitchin, UK). To obtain the *CrGES-mVENUS_ShBle* donor DNA, we digested *pBR11_49p-CrGES-mVenus-FCBPt* episomal DNA using XbaI and PsiI to purge the *CEN/ARS/HIS* yeast centromeric DNA region. The digested DNA was separated by

gel electrophoresis and extracted and purified using Monarch PCR & DNA Cleanup Kit (New England Biolabs, Hitchin, UK).

CRISPR RNA design

The UMPS-1 and UMPS-3 crispr RNA sequences were generated by Serif et al. (2018). The ch1:247-A and ch1:247-A crispr RNA sequenced were designed using Cas-Designer www.rgenome.net/cas-designer (Park et al., 2015) and CRISPOR according to the following criteria: (i) the target site is located in the first exon; (ii) greater than 60% out-of-frame score (indicating likelihood of cut resulting in a frameshift); (iii) no mismatches and (iv) GC content 25-75% (Figure 6B). All four crispr RNAs were custom synthesized by IDT.

Cas9 RNP *in vitro* digestion of target DNA

All of the CRISPR components required were obtained from the Alt-R CRISPR Cas9 suite by IDT. This included the *Streptococcus pyogenes* Cas9 protein, containing three nuclear localization sequences (NLS) and a his tag, the tracrRNA and custom synthesized target specific crRNAs, both containing chemical modifications for increased resistance to cellular RNases. The tracrRNA and crRNA were mixed in equimolar concentrations, heated to 94°C for 2 minutes and allowed to cool to room temperature in order to form the duplex RNA. The duplex RNA was mixed with the Cas9 protein in a 1:1.3 molar ratio and incubated at 37°C for 30 min to allow the formation of the RNP. In order to obtain the target DNA, a region of genomic DNA flanking either end of the target site was amplified by PCR. One hundred nanograms of target DNA was incubated with 600 ng of Cas9 RNP or Cas9 protein at 37°C for 2 hours in a thermocycler with Cutsmart buffer in a 25 µL reaction. Half a microliter of RNase1 was added to the reaction and incubated at 37 °C for 20 minutes. DNA loading

buffer was added to stop the reaction and incubated at 37 °C for 20 minutes. The reaction was analysed by electrophoresis.

RNP proteolistic bombardment

The RNP proteolistic bombardment protocol described by Serif et al. (2008) was followed with the following changes: 1 µg donor DNA was added to RNP mix before mixed with microparticles.

Flow cytometry

Flow cytometry was performed using a BD CytoFLEX S flow cytometer (BD Biosciences). The cells were analysed at medium speed until 20,000 events were counted. FITC fluorescence was excited with a 488 nm laser and emission was acquired using a 525/40 nm optical filter.

PCR based analyses

PCR amplification was performed using Q5 high fidelity polymerase (New England Biolabs, Hitchin, UK) and PCR screening was performed using GoTaq Flexi DNA polymerase (Promega, Wisconsin, United States) according to the manufacturer's instructions. For high throughput PCR screening, single colonies were grown in 200 µl ESAW supplemented with relevant compounds for 10 days to increase biomass. A list of primers used can be found in Supplementary Table 3.

SUPPLEMENTARY TABLES

Suppl. Table 1. Details of the guide RNAs designed to target *Phaeodactylum tricornutum uridine-5'-monophosphate synthase (UMPS; Phatr3_J11740)*

Guide RNA name	5' – 3' Sequence including PAM	Specificity score*	GC %	Mismatches (0-1-2-3-4)	Exon target	Guide RNA functionality validation
UMPS-1	GATGTCAAGCGCGGCGACAT**TGG**	100	60%	0-0-0-0-0	1 of 3	Serif et al. 2018 and *in vitro* digest
UMPS-3	CAAGCTTGTTGGCTCGGAAT**GGG**	100	50%	0-0-0-0-0	2 of 3	Serif et al. 2018 and *in vitro* digest

Suppl. Table 2. Details of the guide RNAs designed to target *Phaeodactylum tricornutum* putative safe harbour locus *(ch1:2,477,260)*

Locus	Guide RNA name	5' – 3' Sequence including PAM	Specificity score*	GC %	Mismatches (0-1-2-3-4)	Exon target	Guide RNA functionality validation
Ch1:2,477,239	*1:247*-A	CAAAGATAACAGTCCTGCAG**AGG**	100	45%	0-0-0-0-0	NA	*In vitro* digest
Ch1:2,477,246	*1:247*-B	CTCAATCTCTCATCCTCTGC**AGG**	100	50%	0-0-0-0-1	NA	*In vitro* digest

Suppl. Table 3. Oligonucleotide primers utilised in this study.

Primer ID	Sequence 5'-3'	Description
40	CAACAAAGTGCTCCTGCAAA	Forward primer to amplify *UMPS* from *Phatr3_45195*
47	AAGATTCCCGGATCAAACAA	Forward primer to amplify *UMPS* from intergenic region between *Phatr3_45195* and *UMPS*
38	ATGGCCACCCCCTCTTTTCGATCA	Forward primer to amplify *UMPS* from *UMPS* start codon
75	GAAGAAAATCGCTGTGACGC	Forward primer to amplify *UMPS* from within exon 1 of *UMPS;* to amplify *UMPS* for RE analysis
74	GTCCGTAGCTTTGCTGATACC	Reverse primer to amplify *UMPS* until *Phatr3_45193;* to amplify *UMPS* for RE analysis
37	TTACACTCCGTATTCGTTTCGAT	Reverse primer to amplify *UMPS* until intergenic region between *UMPS* and *Phatr3_45193*
39	CCTCTGCTTTCCGCATGTAT	Reverse primer to amplify *UMPS* until *UMPS* stop codon
MF748	TCCGAAACGTTTTTCTGACA	Forward primer to amplify *ch1:247* for RE analysis
MF749	CCATGGTAGTCGGTGCTTCT	Reverse primer to amplify *ch1:247* for RE analysis
MF941	CCTCTGCTTTCCGCATGTAT	Sequencing *UMPS* amplicon
MF942	AAGTGTGCGACTCACGAATG	Sequencing *UMPS* amplicon
MF943	GAGCCGAATTTGAGAACACC	Sequencing *UMPS* amplicon
MF944	ATCCAACAAAATCGGCACAT	Sequencing *UMPS* amplicon
MF945	CAACACCAATTCGCTGATTC	Sequencing *UMPS* amplicon
MF946	GCTCAGCAGACCGAGAGTTC	Sequencing *UMPS* amplicon
MF947	AAGATTCCCGGATCAAACAA	Sequencing *UMPS* amplicon

SUPPLEMENTARY FIGURES

```
Seq_1   1     ATGGCCACCCCCTCTTTTCGATCAAAGCTTGAAGCTCGAGTCGCCGCAGTCAACTCTCTC  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1     ATGGCCACCCCCTCTTTTCGATCAAAGCTTGAAGCTCGAGTCGCCGCAGTCAACTCTCTC  60

Seq_1   61    TTGTGCGTTGGTCTAGACCCGCACGAGAAAGAGCTGTTTGCAGACGGATGGGAAGGCGTG  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   61    TTGTGCGTTGGTCTAGACCCGCACGAGAAAGAGCTGTTTGCAGACGGATGGGAAGGCGTG  120

Seq_1   121   CCGGAAGAAAATCGCTGTGACGCGGCCTTTACCTTTTGCAAAACGTTGGTCGACGCAACA  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   121   CCGGAAGAAAATCGCTGTGACGCGGCCTTTACCTTTTGCAAAACGTTGGTCGACGCAACA  180

Seq_1   181   TTGCCTTACACGGCCTGCTACAAACCCAATGCTGCCTTTTTCGAGGCGTTAGGCGATGGA  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   181   TTGCCTTACACGGCCTGCTACAAACCCAATGCTGCCTTTTTCGAGGCGTTAGGCGATGGA  240

Seq_1   241   GGGATAGCGGTTCTGCGACGAGTTTGTCAAAACATAATACCGGATGATGTGCCGATTTTG  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   241   GGGATAGCGGTTCTGCGACGAGTTTGTCAAAACATAATACCGGATGATGTGCCGATTTTG  300

Seq_1   301   TTGGATGTCAAGCGCGGCGACATTGGCTCGACCGCTGCGGCCTACGCCGAAGCGTGCTAT  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   301   TTGGATGTCAAGCGCGGCGACATTGGCTCGACCGCTGCGGCCTACGCCGAAGCGTGCTAT  360

Seq_1   361   GGTTTGGGTGCAGACTGTGTCACGCTTTCACCACTGATGGGATGGGACTCAGTCAGTCCC  420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   361   GGTTTGGGTGCAGACTGTGTCACGCTTTCACCACTGATGGGATGGGACTCAGTCAGTCCC  420

Seq_1   421   TTTGTTACAGGTAAGTTGTAGCTAAAGGATTGGCATCTCCAGCGATTGAATGGATGCCAT  480
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   421   TTTGTTACAGGTAAGTTGTAGCTAAAGGATTGGCATCTCCAGCGATTGAATGGATGCCAT  480

Seq_1   481   CTGATACGTTGAATTGACTCCCTCCGACAGAAAAGTACGTTCACAAAGGAGCATTTTTGC  540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   481   CTGATACGTTGAATTGACTCCCTCCGACAGAAAAGTACGTTCACAAAGGAGCATTTTTGC  540

Seq_1   541   TGTGCAAAACGTCAAATCCTGGATCCAACGATTTTTTAGCTCTGGGATTACGTTCAAATG  600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   541   TGTGCAAAACGTCAAATCCTGGATCCAACGATTTTTTAGCTCTGGGATTACGTTCAAATG  600

Seq_1   601   AATGTTTATACGAAAGAATTGCCAAGCTTGTTGGCTCGGAATGGGCTCAGCAGACCGAGA  660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   601   AATGTTTATACGAAAGAATTGCCAAGCTTGTTGGCTCGGAATGGGCTCAGCAGACCGAGA  660

Seq_1   661   GTTCATTGGGACTCGTTGTCGGGGCCACAGATCCAGTGGCCTTGTCCAAAGCGAGAAAGG  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   661   GTTCATTGGGACTCGTTGTCGGGGCCACAGATCCAGTGGCCTTGTCCAAAGCGAGAAAGG  720

Seq_1   721   CTGCAGGCGACGACACCTGGATTCTAGCACCCGGCGTTGGTGCTCAAGGTGGAGATCTTC  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   721   CTGCAGGCGACGACACCTGGATTCTAGCACCCGGCGTTGGTGCTCAAGGTGGAGATCTTC  780
```

```
Seq_1  781   TAGAAGCAGCGCAGGCTGGATTGAATACAAAGGGGACTTGCATGCTAATTCCCGTGTCTA  840
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  781   TAGAAGCAGCGCAGGCTGGATTGAATACAAAGGGGACTTGCATGCTAATTCCCGTGTCTA  840

Seq_1  841   GGGGTATCAGCAAAGCTACGGACCCAGCGCAGGCTGCAAAAGAATTGCAGGAGAGGATTC  900
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  841   GGGGTATCAGCAAAGCTACGGACCCAGCGCAGGCTGCAAAAGAATTGCAGGAGAGGATTC  900

Seq_1  901   AGAAAGCTCGGGACCAAGTCGTGGCCGCACACATGATAAAAAAGAGTTCAGACGAAGATA  960
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  901   AGAAAGCTCGGGACCAAGTCGTGGCCGCACACATGATAAAAAAGAGTTCAGACGAAGATA  960

Seq_1  961   TTAAACTCTATCAACGCGAGTTTCTTGAATTTAGTCTGTCTCTAGGTGTTCTCAAATTCG  1020
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  961   TTAAACTCTATCAACGCGAGTTTCTTGAATTTAGTCTGTCTCTAGGTGTTCTCAAATTCG  1020

Seq_1  1021  GCTCTTTTGTGCTGAAAAGCGGCCGCATCTCTCCATATTTTTTCAACGCCGGTCTTTTTG  1080
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1021  GCTCTTTTGTGCTGAAAAGCGGCCGCATCTCTCCATATTTTTTCAACGCCGGTCTTTTTG  1080

Seq_1  1081  CTTCTGGCGCTGCGTTAAGCAAGCTTGGGAAAGCCTATGCTTCGACTATCATGTCCTCGG  1140
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1081  CTTCTGGCGCTGCGTTAAGCAAGCTTGGGAAAGCCTATGCTTCGACTATCATGTCCTCGG  1140

Seq_1  1141  AATTATTGTAAGTGTGCTTTGTGTGTTTTTCTCTGCTGAACGGCAAAAATTCAAGAGAAG  1200
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1141  AATTATTGTAAGTGTGCTTTGTGTGTTTTTCTCTGCTGAACGGCAAAAATTCAAGAGAAG  1200

Seq_1  1201  GATGAGTATCCACTTGGTCCGTGTTACCGATCTGCCCCCACGTGAGTGGCAATGAGCAAA  1260
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1201  GATGAGTATCCACTTGGTCCGTGTTACCGATCTGCCCCCACGTGAGTGGCAATGAGCAAA  1260

Seq_1  1261  TTTTTTTCCAGTGGCCTGACTCTTGAAACAACATAGTCGATGATGACTCCTTTGGTCTTC  1320
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1261  TTTTTTTCCAGTGGCCTGACTCTTGAAACAACATAGTCGATGATGACTCCTTTGGTCTTC  1320

Seq_1  1321  TTTCACCTAATTTCTCCGGAAAGATGCCGGTCAACACCAATTCGCTGATTCGAAATTTTC  1380
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1321  TTTCACCTAATTTCTCCGGAAAGATGCCGGTCAACACCAATTCGCTGATTCGAAATTTTC  1380

Seq_1  1381  TGAGACTGTGTTTTGATTTAGTTCTATGGGACTATCATTGTTGTGAGCAGGCTTACCCAA  1440
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1381  TGAGACTGTGTTTTGATTTAGTTCTATGGGACTATCATTGTTGTGAGCAGGCTTACCCAA  1440

Seq_1  1441  CAAAATTCGTTTTCTTTTTCCTTCCAGAGCTGCTGGGCCCAACCAAGTCAATTTTGATGT  1500
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1441  CAAAATTCGTTTTCTTTTTCCTTCCAGAGCTGCTGGGCCCAACCAAGTCAATTTTGATGT  1500

Seq_1  1501  GATTTTTGGTCCTGCATACAAGGGTATTTCTCTAGGTGCTGTCGTTGGAAGCGCTCTGTA  1560
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  1501  GATTTTTGGTCCTGCATACAAGGGTATTTCTCTAGGTGCTGTCGTTGGAAGCGCTCTGTA  1560
```

```
Seq_1   1561  TAACGATTTTGAAGTAGATGTCGGTTTTGCGTATGACCGAAAAGAGGCAAAGGATCATGG  1620
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1561  TAACGATTTTGAAGTAGATGTCGGTTTTGCGTATGACCGAAAAGAGGCAAAGGATCATGG  1620

Seq_1   1621  GGAAGGTGGTAAATTGGTCGGGACTTCGTTGGAAGGAAAACGAGTTCTGATTGTAGATGA  1680
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1621  GGAAGGTGGTAAATTGGTCGGGACTTCGTTGGAAGGAAAACGAGTTCTGATTGTAGATGA  1680

Seq_1   1681  CGTAATCACAGCGGGAACCGCCATTCGTGAGTCGCACACTTTGCTCAACGATGTGGGTGC  1740
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1681  CGTAATCACAGCGGGAACCGCCATTCGTGAGTCGCACACTTTGCTCAACGATGTGGGTGC  1740

Seq_1   1741  TTTGCCAGTTGGAGTAGTTATTGCCCTCGATCGAGCCGAAATTCGCTCTATGGAGGACAA  1800
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1741  TTTGCCAGTTGGAGTAGTTATTGCCCTCGATCGAGCCGAAATTCGCTCTATGGAGGACAA  1800

Seq_1   1801  GATTTCCGCTGTTCAAGCAGTCGCACGAGATCTATCTCTTTTGGTCGTGTCAATTGTCAG  1860
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1801  GATTTCCGCTGTTCAAGCAGTCGCACGAGATCTATCTCTTTTGGTCGTGTCAATTGTCAG  1860

Seq_1   1861  TCTTCCTCAACTACAGACATTTCTCGAACGAAGTCCGGACTACGGCGATGAAACGCTGGA  1920
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2   1861  TCTTCCTCAACTACAGACATTTCTCGAACGAAGTCCGGACTACGGCGATGAAACGCTGGA  1920

Seq_1   1921  AAAAGTAATTAAGTATCGAAACGAATACGGAGTGTAA
              |||||||||||||||||||||||||||||||||||
Seq_2   1921  AAAAGTAATTAAGTATCGAAACGAATACGGAGTGTAA
```

Suppl. Figure 1. Sequence alignment highlighting single nucleotide polymorphisms (SNPs) identified within *P. tricornutum UMPS (Phatr3_J11740)* gene of UTEX LB 642 wild type strain (Seq_1) identified by Sakaguchi et al. (2011) and CCAP 1055/1 wild type strain (Seq_2) identified in this study. SNP-1 to SNP-16 are consistent between both strains (red text), whereas SNP-A, -B and -C were only present in *P. tricornutum* CCAP 1055/1 wild type strain (red text with yellow highlight). Exons (blue text), introns (black text) and active domains (orange text), OCT and PPRT indicate that the SNPs are found in exogenic and intronic regions, but not within the active domains of the enzyme.

```
C6L824_PHATR      MATPSFRSKLEARVAAVNSLLCVGLDPHEKELFADGWEGVPEENRCDAAFTFCKTLVDAT  60
Untreated         MATPSFRSKLEARVAAVNSLLCVGLDPHEKELFADGWEGVPEENRCDAAFTFCKTLVDAT  60
WTr-1-1           MATPSFRSKLEARVAAVNSLLCVGLDPHEKELFADGWEGVPEENRCDAAFTFCKTLVDAT  60
WTr-2-1           MATPSFRSKLEARVAAVNSLLCVGLDPHEKELFADGWEGVPEENRCDAAFTFCKTLVDAT  60
WTr-2-2           MATPSFRSKLEARVAAVNSLLCVGLDPHEKELFADGWEGVPEENRCDAAFTFCKTLVDAT  60
                  ************************************************************

                                                        ODC active domain
C6L824_PHATR      LPYTACYKPNAAFFEALGDGGIAVLRRVCQNIIPDDVPILLDVKRGDIGSTAAAYAEACY  120
Untreated         LPYTACYKPNAAFFEALGDGGIAVLRRVCQNIIPDDVPILLDVKRGDIGSTAAAYAEACY  120
WTr-1-1           LPYTACYKPNAAFFEALGDGGMAVLRRVCQNIIPDDVPILLDVKRGDIGSTAAAYAEACY  120
WTr-2-1           LPYTACYKPNAAFFEALGDGGMAVLRRVCQNIIPDDVPILLDVKRGDIGSTAAAYAEACY  120
WTr-2-2           LPYTACYKPNAAFFEALGDGGMAVLRRVCQNIIPDDVPILLDVKRGDIGSTAAAYAEACY  120
                  ********************:***************************************

C6L824_PHATR      GLGADCVTLSPLMGWDSVSPFVTEKYVHKGAFLLCKTSNPGSNDFLALGLRSNECLYERI  180
Untreated         GLGADCVTLSPLMGWDSVSPFVTEKYVHKGAFLLCKTSNPGSNDFLALGLRSNECLYERI  180
WTr-1-1           GLGADCVTLSPLMGWDSVSPFVTEKYVHKGAFLLCKTSNPGSNDFLALGLRSNECLYERI  180
WTr-2-1           GLGADCVTLSPLMGWDSVSPFVTEKYVHKGAFLLCKTSNPGSNDFLALGLRSNECLYERI  180
WTr-2-2           GLGADCVTLSPLMGWDSVSPFVTEKYVHKGAFLLCKTSNPGSNDFLALGLRSNECLYERI  180
                  ************************************************************

C6L824_PHATR      AKLVGSEWAQQTESSLGLVVGATDPVALSKARKAAGDDTWILAPGVGAQGGDLLEAAQAG  240
Untreated         AKLVGSEWAQQTESSLGLVVGATDPVALSKARKAAGDDTWILAPGVGAQGGDLLEAAQAG  240
WTr-1-1           AKLVGSEWAQQTESSLGLVVGATDPVALSKARKAAGDDTWILAPGVGAQGGDLLEAAQAG  240
WTr-2-1           AKLVGSEWAQQTESSLGLVVGATDPVALSKARKAAGDDTWILAPGVGAQGGDLLEAAQAG  240
WTr-2-2           AKLVGSEWAQQTESSLGLVVGATDPVALSKARKAAGDDTWILAPGVGAQGGDLLEAAQAG  240
                  ************************************************************

C6L824_PHATR      LNTKGTCMLIPVSRGISKATDPAQAAKELQERIQKARDQVVAAHMIKKSSDEDIKLYQRE  300
Untreated         LNTKGTCMLIPVSRGISKATDPAQAAKELQERIQKARDQVVAAHMIKKSSDEDIKLYQRE  300
WTr-1-1           LNTKGTCMLIPVSR--------------------------------------------------  254
WTr-2-1           LNTKGTCMLIPVSR--------------------------------------------------  254
WTr-2-2           LNTKGTCMLIPVSR--------------------------------------------------  254
                  **************

C6L824_PHATR      FLEFSLSLGVLKFGSFVLKSGRISPYFFNAGLFASGAALSKLGKAYASTIMSSELLAAGP  360
Untreated         FLEFSLSLGVLKFGSFVLKSGRISPYFFNAGLFASGAALSKLGKAYASTIMSSELLAAGP  360
WTr-1-1           ------------------------------------------------------------  254
WTr-2-1           ------------------------------------------------------------  254
WTr-2-2           ------------------------------------------------------------  254

C6L824_PHATR      NQVNFDVIFGPAYKGISLGAVVGSALYNDFEVDVGFAYDRKEAKDHGEGGKLVGTSLEGK  420
Untreated         NQVNFDVIFGPAYKGISLGAVVGSALYNDFEVDVGFAYDRKEAKDHGEGGKLVGTSLEGK  420
WTr-1-1           ------------------------------------------------------------  254
WTr-2-1           ------------------------------------------------------------  254
WTr-2-2           ------------------------------------------------------------  254

                  PPRT active domain
C6L824_PHATR      RVLIVDDVITAGTAIRESHTLLNDVGALPVGVVIALDRAEIRSMEDKISAVQAVARDLSL  480
Untreated         RVLIVDDVITAGTAIRESHTLLNDVGALPVGVVIALDRAEIRSMEDKISAVQAVARDLSL  480
WTr-1-1           ------------------------------------------------------------  254
WTr-2-1           ------------------------------------------------------------  254
WTr-2-2           ------------------------------------------------------------  254

C6L824_PHATR      LVVSIVSLPQLQTFLERSPDYGDETLEKVIKYRNEYGV  518
Untreated         LVVSIVSLPQLQTFLERSPDYGDETLEKVIKYRNEYGV  518
WTr-1-1           --------------------------------------  254
WTr-2-1           --------------------------------------  254
WTr-2-2           --------------------------------------  254
```

Supp. Figure 2. Multiple sequence alignment of UMPS (Uniprot Accession C6L824) protein to translated protein sequences obtained from the *UMPS*-frag amplicons of WT, WT^R-1-1, WT^R-2-1, WT^R-2-2 obtained in this study. All three WT^R strains demonstrate a single point mutation (green) causing an amino acid change from isoleucine to methionine and stunted coding region, as well as a truncated protein (254 amino acids instead of the full 518 amino acids) and the loss of the PPRT active domain (pink).

REFERENCES

Ainley, W. M., Sastry-Dent, L., Welter, M. E., Murray, M. G., Zeitler, B., Amora, R., … Petolino, J. F. (2013). Trait stacking via targeted genome editing. *Plant Biotechnology Journal*, *11*(9), 1126–1134. https://doi.org/10.1111/pbi.12107

Apt, K. E., Kroth-Pancic, P. G., & Grossman, A. R. (1996). Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Molecular and General Genetics*, *252*(5), 572–579. https://doi.org/10.1007/s004380050264

Backhaus, K., Ludwig-Radtke, L., Xie, X., & Li, S. M. (2017). Manipulation of the Precursor Supply in Yeast Significantly Enhances the Accumulation of Prenylated β-Carbolines. *ACS Synthetic Biology*, *6*(6), 1056–1064. https://doi.org/10.1021/acssynbio.6b00387

Bekker, A., Holland, H. D., Wang, P. L., Rumble, D., Stein, H. J., Hannah, J. L., … Beukes, N. J. (2004). Dating the rise of atmospheric oxygen. *Nature*, *427*(6970), 117–120. https://doi.org/10.1038/nature02260

Berges, J. A., Franklin, D. J., & Harrison, P. J. (2001). Evolution of an artificial seawater medium: Improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology*, *37*(6), 1138–1145. https://doi.org/10.1046/j.1529-8817.2001.01052.x

Bohlmann, J., & Keeling, C. I. (2008). Terpenoid biomaterials. *Plant Journal*, *54*(4), 656–669. https://doi.org/10.1111/j.1365-313X.2008.03449.x

Borowitzka, M. a. (2013). High-value products from microalgae-their development and commercialisation. *Journal of Applied Phycology*, *25*, 743–756. https://doi.org/10.1007/s10811-013-9983-9

Bourgeois, L., Pyne, M. E., & Martin, V. J. J. (2018). A Highly Characterized Synthetic Landing Pad System for Precise Multicopy Gene Integration in Yeast. *ACS Synthetic Biology*, *7*(11), 2675–2685. https://doi.org/10.1021/acssynbio.8b00339

Brandsen, B. M., Mattheisen, J. M., Noel, T., & Fields, S. (2018). A Biosensor Strategy for E. coli Based on Ligand-Dependent Stabilization. *ACS Synthetic Biology*, *7*(9), 1990–1999. rapid-communication. https://doi.org/10.1021/acssynbio.8b00052

Broddrick, J. T., Du, N., Smith, S. R., Tsuji, Y., Jallet, D., Ware, M. A., … Allen, A. E. (2019). Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom Phaeodactylum tricornutum. *New Phytologist*, *222*(3), 1364–1379. https://doi.org/10.1111/nph.15685

Cantos, C., Francisco, P., Trijatmiko, K. R., Slamet-Loedin, I., and ChadhaMohanty, P. K. (2014). Identification of “safe harbor” loci in indica rice genome by harnessing the property of zinc-finger nucleases to induce DNA damage and repair. Front. Plant Sci. 5:302. doi: 10.3389/fpls.2014.00302

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., … Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, *155*(7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001

Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., … Sharp, P. A. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, *160*(6), 1246–1260. https://doi.org/10.1016/j.cell.2015.02.038

Cheng, R. Bin, Lin, X. Z., Wang, Z. K., Yang, S. J., Rong, H., & Ma, Y. (2011). Establishment of a transgene expression system for the marine microalga Schizochytrium by 18S rDNA-targeted homologous recombination. *World Journal of Microbiology and*

Biotechnology, *27*(3), 737–741. https://doi.org/10.1007/s11274-010-0510-8

D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., … Matthijs, M. (2018). Engineering The Unicellular Alga Phaeodactylum tricornutum For High-Value Plant Triterpenoid Production. *Plant Biotechnology Journal*, 0–2. https://doi.org/10.1111/pbi.12948

D'Halluin, K., Vanderstraeten, C., Stals, E., Cornelissen, M., & Ruiter, R. (2008). Homologous recombination: A basis for targeted genome optimization in crop species such as maize. *Plant Biotechnology Journal*, *6*(1), 93–102. https://doi.org/10.1111/j.1467-7652.2007.00305.x

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., … Duchateau, P. (2014). Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nature Communications*, *5*(May), 1–7. https://doi.org/10.1038/ncomms4831

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6640–6645. https://doi.org/10.1073/pnas.120163297

Dicarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J., & Church, G. M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Research*, *41*(7), 4336–4343. https://doi.org/10.1093/nar/gkt135

Diner, R. E., Noddings, C. M., Lian, N. C., Kang, A. K., McQuaid, J. B., Jablanovic, J., … Weyman, P. D. (2017). Diatom centromeres suggest a mechanism for nuclear DNA acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(29), E6015–E6024. https://doi.org/10.1073/pnas.1700764114

Donoho, G., Jasin, M., & Berg, P. (1998). Analysis of Gene Targeting and Intrachromosomal Homologous Recombination Stimulated by Genomic Double-Strand Breaks in Mouse Embryonic Stem Cells. *Molecular and Cellular Biology*, *18*(7), 4070–4078. https://doi.org/10.1128/mcb.18.7.4070

Dudareva, N., Negre, F., Nagegowda, D. A., & Orlova, I. (2006). Plant volatiles: Recent advances and future perspectives. *Critical Reviews in Plant Sciences*, *25*(5), 417–440. https://doi.org/10.1080/07352680600899973

Eaton-Rye, J. J. (2011). Construction of Gene Interruptions and Gene Deletions in the Cyanobacterium Synechocystis sp. Strain PCC 6803. *Photosynthesis Research Protocols*, *684*, 363–374. https://doi.org/10.1007/978-1-60761-925-3

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., … Ralph, P. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.9b00455

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., … Goossens, A. (2014). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *The New Phytologist*, 521–535. https://doi.org/10.1111/nph.12917

Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., & Bowler, C. (1999). Transformation of nonselectable reporter genes in marine diatoms. *Marine Biotechnology*, *1*(3), 239–251. https://doi.org/10.1007/PL00011773

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, *281*(5374), 200–206.

https://doi.org/10.1126/science.281.5374.200

Ferenczi, A., Pyott, D. E., Xipnitou, A., & Molnar, A. (2017). Efficient targeted DNA editing and replacement in *Chlamydomonas reinhardtii* using Cpf1 ribonucleoproteins and single-stranded DNA. *Proceedings of the National Academy of Sciences*, *114*(51), 201710597. https://doi.org/10.1073/pnas.1710597114

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, *281*(5374), 237–240. https://doi.org/10.1126/science.281.5374.237

Gaidukov, L., Wroblewska, L., Teague, B., Nelson, T., Zhang, X., Liu, Y., … Weiss, R. (2018). A multi-landing pad DNA integration platform for mammalian cell engineering. *Nucleic Acids Research*, *46*(8), 4072–4086. https://doi.org/10.1093/nar/gky216

Gaj, T., Gersbach, C. A., & Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, *31*(7), 397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

George, J., Kahlke, T., Abbriano, R. M., Kuzhiumparambil, U., Ralp, P. J., & Fabris, M. (2020). Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology. *Frontiers in Bioengineering and Biotechnology*, *8*(June), 1–19. https://doi.org/10.3389/fbioe.2020.00513

Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., & Hegemann, P. (2017). Targeting of Photoreceptor Genes in Chlamydomonas reinhardtii via Zinc- finger Nucleases and CRISPR/Cas9. *Plant Cell Advance Publication. Published on October*, *4*. https://doi.org/10.1105/tpc.17.00659

Guri Giaever1, Angela M. Chu2, LiNi3, Carla Connelly4, Linda Riles5, Steeve Ve´ronneau6, Sally Dow7, Ankuta Lucau-Danila8, Keith Anderson1, Bruno Andre´9, Adam P. Arkin10, Anna Astromoff2, Mohamed El Bakkoury11, Rhonda Bangham3, Rocio Benito12, Sophie Bra, 2 & Mark Johnston. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, (418), 387–391.

Hamilton, M. L., Haslam, R. P., Napier, J. A., & Sayanova, O. (2014). Metabolic engineering of Phaeodactylum tricornutum for the enhanced accumulation of omega-3 long chain polyunsaturated fatty acids. *Metabolic Engineering*, *22*, 3–9. https://doi.org/10.1016/j.ymben.2013.12.003

Hamilton, M. L., Powers, S., Napier, J. A., & Sayanova, O. (2016). Heterotrophic production of omega-3 long-chain polyunsaturated fatty acids by trophically converted marine diatom phaeodactylum tricornutum. *Marine Drugs*, *14*(3). https://doi.org/10.3390/md14030053

Hao, H., Wang, X., Jia, H., Yu, M., Zhang, X., Tang, H., & Zhang, L. (2016). Large fragment deletion using a CRISPR/Cas9 system in Saccharomyces cerevisiae. *Analytical Biochemistry*, *509*, 118–123. https://doi.org/10.1016/j.ab.2016.07.008

Hempel, F., Bozarth, A. S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., … Maier, U. G. (2011). Microalgae as bioreactors for bioplastic production. *Microbial Cell Factories*, *10*(1), 81. https://doi.org/10.1186/1475-2859-10-81

Hong, S. G., Yada, R. C., Choi, K., Carpentier, A., Liang, T. J., Merling, R. K., … Dunbar, C. E. (2017). Rhesus iPSC Safe Harbor Gene-Editing Platform for Stable Expression of Transgenes in Differentiated Cells of All Germ Layers. *Molecular Therapy*, *25*(1), 44–53. https://doi.org/10.1016/j.ymthe.2016.10.007

Hou, L., Yau, Y. Y., Wei, J., Han, Z., Dong, Z., & Ow, D. W. (2014). An Open-Source System for in Planta Gene Stacking by Bxb1 and Cre Recombinases. *Molecular Plant*, *7*(12), 1756–1765. https://doi.org/10.1093/mp/ssu107

Huang, W., & Daboussi, F. (2017). Genetic and metabolic engineering in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160411. https://doi.org/10.1098/rstb.2016.0411

Ishii, Y., Maruyama, S., Fujimura-Kamada, K., Kutsuna, N., Takahashi, S., Kawata, M., & Minagawa, J. (2018). Isolation of uracil auxotroph mutants of coral symbiont alga for symbiosis studies. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-21499-3

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345. doi: 10.1038/nbt.4060

Jakočiunas, T., Rajkumar, A. S., Zhang, J., Arsovska, D., Rodriguez, A., Jendresen, C. B., … Keasling, J. D. (2015). CasEMBLR: Cas9-Facilitated Multiloci Genomic Integration of in Vivo Assembled DNA Parts in Saccharomyces cerevisiae. *ACS Synthetic Biology*, *4*(11), 1126–1134. https://doi.org/10.1021/acssynbio.5b00007

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., … Yuan, Y. J. (2017). Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metabolic Engineering*, *41*(March), 57–66. https://doi.org/10.1016/j.ymben.2017.03.005

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., … Ecker, J. R. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics*, *15*(1), 1–25. https://doi.org/10.1371/journal.pgen.1007819

Karas, B. J., Suzuki, Y., & Weyman, P. D. (2015). Strategies for cloning and manipulating natural and synthetic chromosomes. *Chromosome Research*, *23*(1), 57–68. https://doi.org/10.1007/s10577-014-9455-3

Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., … Weiss, R. (2014). CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nature Methods*, *11*(7), 723–726. https://doi.org/10.1038/nmeth.2969

Kilian, O., Benemann, C. S. E., Niyogi, K. K., & Vick, B. (2011). High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. *Proceedings of the National Academy of Sciences*, *108*(52), 21265–21269. https://doi.org/10.1073/pnas.1105861108

Klein, J., Heal, J. R., Hamilton, W. D. O., Boussemghoune, T., Tange, T. Ø., Delegrange, F., … Heim, J. (2014). Yeast synthetic biology platform generates novel chemical structures as scaffolds for drug discovery. *ACS Synthetic Biology*, *3*(5), 314–323. https://doi.org/10.1021/sb400177x

Kondo, S., Yamagishi, A., & Oshima, T. (1991). Thermophilic Archaebacterium Sulfolobus acidocaldarius by Use of 5-Fluoroorotic Acid. *Journal of Bacteriology*, *173*(23), 7698–7700.

Kuhlman, T. E., & Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research*, *38*(6). https://doi.org/10.1093/nar/gkp1193

Lan, E. I., Chuang, D. S., Shen, C. R., Lee, A. M., Ro, S. Y., & Liao, J. C. (2015). Metabolic engineering of cyanobacteria for photosynthetic 3-hydroxypropionic acid production from CO2 using Synechococcus elongatus PCC 7942. *Metabolic Engineering*, *31*, 163–

170. https://doi.org/10.1016/j.ymben.2015.08.002

Lauersen, K. J., Baier, T., Wichmann, J., Wördenweber, R., Mussgnug, J. H., Hübner, W., … Kruse, O. (2016). Efficient phototrophic production of a high-value sesquiterpenoid from the eukaryotic microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *38*, 331–343. https://doi.org/10.1016/j.ymben.2016.07.013

Lauersen, K. J., Wichmann, J., Baier, T., Kampranis, S. C., Pateraki, I., Møller, B. L., & Kruse, O. (2018). Phototrophic production of heterologous diterpenoids and a hydroxy-functionalized derivative from Chlamydomonas reinhardtii. *Metabolic Engineering*, *49*. https://doi.org/10.1016/j.ymben.2018.07.005

Levering, J., Dupont, C. L., Allen, A. E., Palsson, B. O., & Zengler, K. (2017). Integrated Regulatory and Metabolic Networks of the Marine Diatom Phaeodactylum tricornutum Predict the Response to Rising CO 2 Levels . *MSystems*, *2*(1), 1–12. https://doi.org/10.1128/msystems.00142-16

Li, H., Shen, C. R., Huang, C.-H., Sung, L.-Y., Wu, M.-Y., & Hu, Y.-C. (2016). CRISPR-Cas9 for the Genome Engineering of Cyanobacteria and Succinate Production. *Metabolic Engineering*, *38*(August), 293–302. https://doi.org/10.1016/j.ymben.2016.09.006

Li, M., & Borodina, I. (2015). Application of synthetic biology for production of chemicals in yeast Saccharomyces cerevisiae. *FEMS Yeast Research*, *15*(1). https://doi.org/10.1111/1567-1364.12213

Lieber, M. R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annual Review of Biochemistry*, *79*(1), 181–211. https://doi.org/10.1146/annurev.biochem.052308.093131

Linshiz, G., Jensen, E., Stawski, N., Bi, C., Elsbree, N., Jiao, H., … Hillson, N. J. (2016). End-to-end automated microfluidic platform for synthetic biology: From design to functional analysis. *Journal of Biological Engineering*, *10*(1), 1–15. https://doi.org/10.1186/s13036-016-0024-5

Lv, Y., Edwards, H., Zhou, J., & Xu, P. (2019). Combining 26s rDNA and the Cre-loxP System for Iterative Gene Integration and Efficient Marker Curation in Yarrowia lipolytica. *ACS Synthetic Biology*, *8*(3), 568–576. research-article. https://doi.org/10.1021/acssynbio.8b00535

Martin-Ortigosa, S., & Wang, K. (2014). Proteolistics: a biolistic method for intracellular delivery of proteins. *Transgenic Research*, *23*(5), 743–756. https://doi.org/10.1007/s11248-014-9807-y

Mijakovic, I., Petranovic, D., & Jensen, P. R. (2005). Tunable promoters in systems biology. *Current Opinion in Biotechnology*, *16*(3 SPEC. ISS.), 329–335. https://doi.org/10.1016/j.copbio.2005.04.003

Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T., & Tanaka, K. (2004). Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, *Cyanidioschyzon merolae* 10D. *Plant and Cell Physiology*, *45*(6), 667–671. https://doi.org/10.1093/pcp/pch087

Moosburner, M. A., Gholami, P., McCarthy, J. K., Tan, M., Bielinski, V. A., & Allen, A. E. (2020). Multiplexed Knockouts in the Model Diatom Phaeodactylum by Episomal Delivery of a Selectable Cas9. *Frontiers in Microbiology*, *11*(January), 1–13. https://doi.org/10.3389/fmicb.2020.00005

Pan, Q., Mustafa, N. R., Tang, K., Choi, Y. H., & Verpoorte, R. (2016). Monoterpenoid indole alkaloids biosynthesis and its regulation in Catharanthus roseus: a literature review

from genes to metabolites. *Phytochemistry Reviews*, *15*(2), 221–250. https://doi.org/10.1007/s11101-015-9406-4

Papaefthimiou, D., Diretto, G., Demurtas, O. C., Mini, P., Ferrante, P., Giuliano, G., & Kanellis, A. K. (2019). Heterologous production of labdane-type diterpenes in the green alga Chlamydomonas reinhardtii. *Phytochemistry*, *167*(July), 112082. https://doi.org/10.1016/j.phytochem.2019.112082

Papapetrou, E. P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L. M. S., … Sadelain, M. (2011). Genomic safe harbors permit high ??-globin transgene expression in thalassemia induced pluripotent stem cells. *Nature Biotechnology*, *29*(1), 73–81. https://doi.org/10.1038/nbt.1717

Pennisi, E. (2013). The CRISPR craze. *Science*, *341*(6148), 833–836. https://doi.org/10.1126/science.341.6148.833

Pflueger, C., Tan, D., Swain, T., Nguyen, T., Pflueger, J., Nefzger, C., … Lister, R. (2018). A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Research*, *28*(8), 1193–1206. https://doi.org/10.1101/gr.233049.117

Pichersky, E., & Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytologist*, *220*(3), 692–702. https://doi.org/10.1111/nph.14178

Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., … Tamagnini, P. (2015). Improving a Synechocystis-based photoautotrophic chassis through systematic genome mapping and validation of neutral sites. *DNA Research*, *22*(6), 425–437. https://doi.org/10.1093/dnares/dsv024

Pollak, B., Matute, T., Nunez, I., Cerda, A., Lopez, C., Kan, A., … Roscoff, S. B. De. (2019). Universal Loop assembly (uLoop): open, efficient, and species- agnostic DNA fabrication.

Pollier, J., Vancaester, E., Kuzhiumparambil, U., Vickers, C. E., Vandepoele, K., Goossens, A., & Fabris, M. (2019). A widespread alternative squalene epoxidase participates in eukaryote steroid biosynthesis. *Nature Microbiology*, *4*(2), 226–233. https://doi.org/10.1038/s41564-018-0305-5

Qian, S., Clomburg, J. M., & Gonzalez, R. (2019). Engineering Escherichia coli as a platform for the in vivo synthesis of prenylated aromatics. *Biotechnology and Bioengineering*, *116*(5), 1116–1127. https://doi.org/10.1002/bit.26932

Radakovits, R., Eduafo, P. M., & Posewitz, M. C. (2011). Genetic engineering of fatty acid chain length in Phaeodactylum tricornutum. *Metabolic Engineering*, *13*(1), 89–95. https://doi.org/10.1016/j.ymben.2010.10.003

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M. A., … Gunawardane, R. N. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular Biology of the Cell*, *28*(21), 2854–2874. https://doi.org/10.1091/mbc.E17-03-0209

Sadelain, M., Papapetrou, E. P., & Bushman, F. D. (2012). Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer*, *12*(1), 51–58. https://doi.org/10.1038/nrc3179

Sakaguchi, T., Nakajima, K., & Matsuda, Y. (2011). Identification of the UMP synthase gene by establishment of Uracil auxotrophic mutants and the phenotypic complementation system in the marine diatom Phaeodactylum tricornutum. *Plant Physiology*, *156*(1), 78–89. https://doi.org/10.1104/pp.110.169631

San Filippo, J., Sung, P., & Klein, H. (2008). Mechanism of eukaryotic homologous recombination. *Annual Review of Biochemistry*, *77*, 229–257. https://doi.org/10.1146/annurev.biochem.77.061306.125255

Scott, L. H., Mathews, J. C., Flematti, G. R., Filipovska, A., & Rackham, O. (2018). An Artificial Yeast Genetic Circuit Enables Deep Mutational Scanning of an Antimicrobial Resistance Protein. *ACS Synthetic Biology*, *7*(8), 1907–1917. research-article. https://doi.org/10.1021/acssynbio.8b00121

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., & Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-06378-9

Shalem. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells, *343*(January), 84–88.

Sharma, A. K., Nymark, M., Sparstad, T., Bones, A. M., & Winge, P. (2018). Transgene-free genome editing in marine algae by bacterial conjugation – comparison with biolistic CRISPR/Cas9 transformation. *Scientific Reports*, *8*(1), 14401. https://doi.org/10.1038/s41598-018-32342-0

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Chang, Y. K. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. *Nature Publishing Group*. https://doi.org/10.1038/srep27810

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Jeong, B. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. *Scientific Reports*, *6*(April), 27810. https://doi.org/10.1038/srep27810

Si, T., Chao, R., Min, Y., Wu, Y., Ren, W., & Zhao, H. (2017). Automated multiplex genome-scale engineering in yeast. *Nature Communications*, *8*(May), 1–12. https://doi.org/10.1038/ncomms15187

Slattery, S. S., Diamond, A., Wang, H., Therrien, J. A., Lant, J. T., Jazey, T., … Edgell, D. R. (2018). An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synthetic Biology*, acssynbio.7b00191. https://doi.org/10.1021/acssynbio.7b00191

Slattery, S. S., Wang, H., Kocsis, C., Urquhart, B. L., Bogumil, J., & Edgell, D. R. (2020). Cas9-generated auxotrophs of Phaeodactylum tricornutum are characterized by small and large deletions that can be complemented by plasmid-based genes.

Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H., … Moritz, T. (2016). Transcriptional Orchestration of the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron Limitation. *PLOS Genetics*, *12*(12), e1006490. https://doi.org/10.1371/journal.pgen.1006490

Spolaore, P., Joannis-Cassan, C., Duran, E., & Isambert, A. (2006). Commercial applications of microalgae. *Journal of Bioscience and Bioengineering*, *101*(2), 87–96. https://doi.org/10.1263/jbb.101.87

Standage-Beier, K., Zhang, Q., & Wang, X. (2015). Targeted Large-Scale Deletion of Bacterial Genomes Using CRISPR-Nickases. *ACS Synthetic Biology*, *4*(11), 1217–1225. https://doi.org/10.1021/acssynbio.5b00132

Sternberg, S. H., & Doudna, J. A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, *58*(4), 568–574. https://doi.org/10.1016/j.molcel.2015.02.032

Symington, L. S. (2002). Role of RAD52 Epistasis Group Genes in Homologous

Recombination and Double-Strand Break Repair. *Microbiology and Molecular Biology Reviews*, *66*(4), 630–670. https://doi.org/10.1128/mmbr.66.4.630-670.2002

Szita, N., Polizzi, K., Jaccard, N., & Baganz, F. (2010). Microfluidic approaches for systems and synthetic biology. *Current Opinion in Biotechnology*, *21*(4), 517–523. https://doi.org/10.1016/j.copbio.2010.08.002

Tetali, S. D. (2018). Terpenes and isoprenoids: a wealth of compounds for global use. *Planta*, *249*(1), 1–8. https://doi.org/10.1007/s00425-018-3056-x

Thakore, P. I., Song, L., Safi, A., Shivakumar, K., Kabadi, A. M., Reddy, T. E., … Gersbach, C. A. (2015). Highly Specific Epigenome Editing by CRISPR/Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nature Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630.Highly

Turnšek, J., Brunson, J. K., Deerinck, T. J., Horák, A., Bielinski, V. A., & Allen, A. E. (2019). Phytotransferrin endocytosis mediates a direct cell surface-to-chloroplast iron trafficking axis in marine diatoms, 1–93.

Wang, Y. K., Das, B., Huber, D. H., Wellington, M., Kabir, M. A., Sherman, F., & Rustchenko, E. (2004). Role of the 14-3-3 protein in carbon metabolism of the pathogenic yeast Candida albicans. *Yeast*, *21*(8), 685–702. https://doi.org/10.1002/yea.1079

Wellington, M., Kabir, M. A., & Rustchenko, E. (2006). 5-Fluoro-orotic acid induces chromosome alterations in genetically manipulated strains of Candida albicans. *Mycologia*, *98*(3), 393–398. https://doi.org/10.3852/mycologia.98.3.393

Wellington, M., & Rustchenko, E. (2005). 5-Fluoro-orotic acid induces chromosome alterations in Candida albicans. *Yeast*, *22*(1), 57–70. https://doi.org/10.1002/yea.1191

Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., Mccarthy, J. K., Heuberger, A. L., … Dupont, C. L. (2015). Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnology Journal*, *13*(4), 460–470. https://doi.org/10.1111/pbi.12254

Wichmann, J., Baier, T., Wentnagel, E., Lauersen, K. J., & Kruse, O. (2018). Tailored carbon partitioning for phototrophic production of (E)-α-bisabolene from the green microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *45*(October 2017), 211–222. https://doi.org/10.1016/j.ymben.2017.12.010

Xue, J., Niu, Y. F., Huang, T., Yang, W. D., Liu, J. S., & Li, H. Y. (2015). Genetic improvement of the microalga Phaeodactylum tricornutum for boosting neutral lipid accumulation. *Metabolic Engineering*, *27*, 1–9. https://doi.org/10.1016/j.ymben.2014.10.002

Zaslavskaia, L. A., Lippmeier, J. C., Shih, C., Ehrhardt, D., Grossman, A. R., & Apt, K. E. (2001). Trophic conversion of an obligate photoautotrophic organism through metabolic engineering. *Science*, *292*(5524), 2073–2075. https://doi.org/10.1126/science.160015

Zou, L. G., Chen, J. W., Zheng, D. L., Balamurugan, S., Li, D. W., Yang, W. D., … Li, H. Y. (2018). High-efficiency promoter-driven coordinated regulation of multiple metabolic nodes elevates lipid accumulation in the model microalga Phaeodactylum tricornutum. *Microbial Cell Factories*, *17*(1), 1–8. https://doi.org/10.1186/s12934-018-0906-y

Implications of geraniol accumulation on native terpenoids in *Phaeodactylum tricornutum*

Adapted from the published article:

Fabris, M., **George, J.,** Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., Vickers, E. C., and Ralph, P. J. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biology.* https://doi.org/10.1021/acssynbio.9b00455

Certificate of Authorship and Originality

This thesis chapter includes excerpts that have been published in *ACS Synthetic Biology*.

As a co-author of this publication who did not supervise Jestin George, I, Ana Cristina Jaramillo-Madrid, certify that Jestin George carried most of the work presented in this thesis chapter; where 70% of experiments presented in this thesis chapter were carried out, analysed and written up by Jestin George. The remaining 30% was carried out in collaboration with the authors of this manuscript.

Production Note:
Signature removed prior to publication.

Ana Cristina Jaramillo-Madrid

Certificate of Authorship and Originality

This thesis chapter includes excerpts that have been published in *ACS Synthetic Biology*.

As a co-author of this publication who did not supervise Jestin George, I, Dr Caitlin A. Lawson, certify that Jestin George carried most of the work presented in this thesis chapter; where 70% of experiments presented in this thesis chapter were carried out, analysed and written up by Jestin George. The remaining 30% was carried out in collaboration with the authors of this manuscript.

Production Note:
Signature removed prior to publication.

Dr Caitlin A. Lawson

Certificate of Authorship and Originality

This thesis chapter includes excerpts that have been published in *ACS Synthetic Biology*.

As a co-author of this publication who did not supervise Jestin George, I, Assoc. Prof. Claudia E. Vickers, certify that Jestin George carried most of the work presented in this thesis chapter; where 70% of experiments presented in this thesis chapter were carried out, analysed and written up by Jestin George. The remaining 30% was carried out in collaboration with the authors of this manuscript.

Production Note:
Signature removed prior to publication.

Assoc. Prof. Claudia E. Vickers

Certificate of Authorship and Originality

This thesis chapter includes excerpts that have been published in *ACS Synthetic Biology*.

As a co-author of this publication who did not supervise Jestin George, I, Dr Raffaela M. Abbriano, certify that Jestin George carried most of the work presented in this thesis chapter; where 70% of experiments presented in this thesis chapter were carried out, analysed and written up by Jestin George. The remaining 30% was carried out in collaboration with the authors of this manuscript.

Production Note:
Signature removed prior to publication.

Dr Raffaela M. Abbriano

Certificate of Authorship and Originality

This thesis chapter includes excerpts that have been published in *ACS Synthetic Biology*.

As a co-author of this publication who did not supervise Jestin George, I, Dr Unnikrishnan Kuzhiumparambil, certify that Jestin George carried most of the work presented in this thesis chapter; where 70% of experiments presented in this thesis chapter were carried out, analysed and written up by Jestin George. The remaining 30% was carried out in collaboration with the authors of this manuscript.

Production Note:
Signature removed prior to publication.

Dr Unnikrishnan Kuzhiumparambil

ABSTRACT

Terpenoids are ubiquitous in nature and constitute a variety of essential metabolites such as sterols and pigments, which are required for eukaryotic cell membrane regulation and photoautotrophy, respectively. Some medicinal plants, such as *Catharanthus roseus*, produce an exceptionally large array of bioactive secondary metabolites called monoterpenoid indole alkaloids (MIAs). Many MIAs hold therapeutic properties for a range of illnesses including cancer and hypertension. As plants naturally produce extremely small amounts of MIAs and their chemical synthesis is complex, industrially sourcing these compounds is costly and inefficient. Microbial hosts capable of producing MIA precursors through metabolic engineering are currently being sought; however, driving high flux through synthetic monoterpenoid pathways whilst maintaining sufficient flux through native essential terpenoid pathways is a delicate tuning process. Consequently, extensive metabolic engineering and bioprocessing approaches have been explored. To date, there is no information available regarding these approaches in any microalga, including the diatom *Phaeodactylum tricornutum*, as bacteria and yeast are the most widely engineered microorganisms for heterologous terpenoid production. Furthermore, there is currently no knowledge revealing the impact of heterologous monoterpenoid production on *P. tricornutum* native isoprenoid biosynthesis. Therefore, we designed strategies for enhanced production of the monoterpenoid geraniol in *P. tricornutum*. We demonstrated that extrachromosomal expression of *C. roseus geraniol synthase (CrGES)* fused to *Abies grandis geranyl diphosphate synthase* resulted in decreased geraniol accumulation, most likely due to improper folding of the fused enzymes. We also demonstrated that constitutive expression of CrGES did not perturb pigment and

sterol production and that changes to the photoperiod and cultivation time of a constitutively CrGES expressing cell line did not impact geraniol production. Altogether, our results demonstrate how extrachromosomal expression can be useful for faster synthetic biology design-build-test-learn cycle compared to the widely used approach of random integration and that terpenoid metabolism in diatoms could be particularly flexible and able to adapt to the installation of artificial metabolic sinks to drive flux through synthetic monoterpenoid pathways.

INTRODUCTION

Biochemical limitations of heterologous geraniol production in bacteria and yeast

Prior to the work conducted in our laboratory, geraniol had been heterologously produced in *E. coli* and *S. cerevisiae*. Consequently, strategies exploring different genetic designs to optimise geraniol production have only been explored in those species and never in any microalga. Simply installing the *GES* gene for heterologous geraniol production is not sufficient for generating high enough yields. This is because metabolism is regulated at many check points and consequently, it is often resistant to engineered changes in flux. In bacterial and yeast microorganisms, there are two main aspects in their native metabolic networks that hinder heterologous geraniol production.

First, *E. coli* only contains the MEP pathway and similarly, *S. cerevisiae* only contains the MVA pathway (Figure 1A and B). The presence of only a single IPP and DMAPP production precursor pathway makes competition between native pathways utilising GPP–like sterol biosynthesis–and the heterologous geraniol pathway difficult to overcome. Consequently, flux through the synthetic monoterpenoid pathway is often low, although this can be greatly improved with metabolic engineering (Fischer et al, 2011; Jiang et al., 2017; Liu et al., 2016; Shah et al., 2013; Zhou et al, 2015). Second, in *C. roseus* GPP biosynthesis (EC 2.5.1.1) and FPP biosynthesis (EC 2.5.1.10) occur via two catalytic domains which are present on two separate enzymes. In *S. cerevisiae* and *E. coli*, GPP and FPP active sites are present on a single, dual functioning enzyme called farnesyl diphosphate synthase (denoted as ERG20 in *S. cerevisiae* and FPPS in *E. coli*) (Figure 1). Having these enzymatic reactions occurring in such close

proximity means that very little opportunity for GPP to escape before conversion into FPP for sterol biosynthesis, leaving only a small pool available for flux through a heterologous monoterpenoid pathway (Zhao et al., 2016). For these reasons, extensive research in both of these species has explored many strategies for increasing GPP biosynthesis and carbon flux away from its conversion into FPP, without eradicating sterol biosynthesis completely, which would be lethal to the cell.

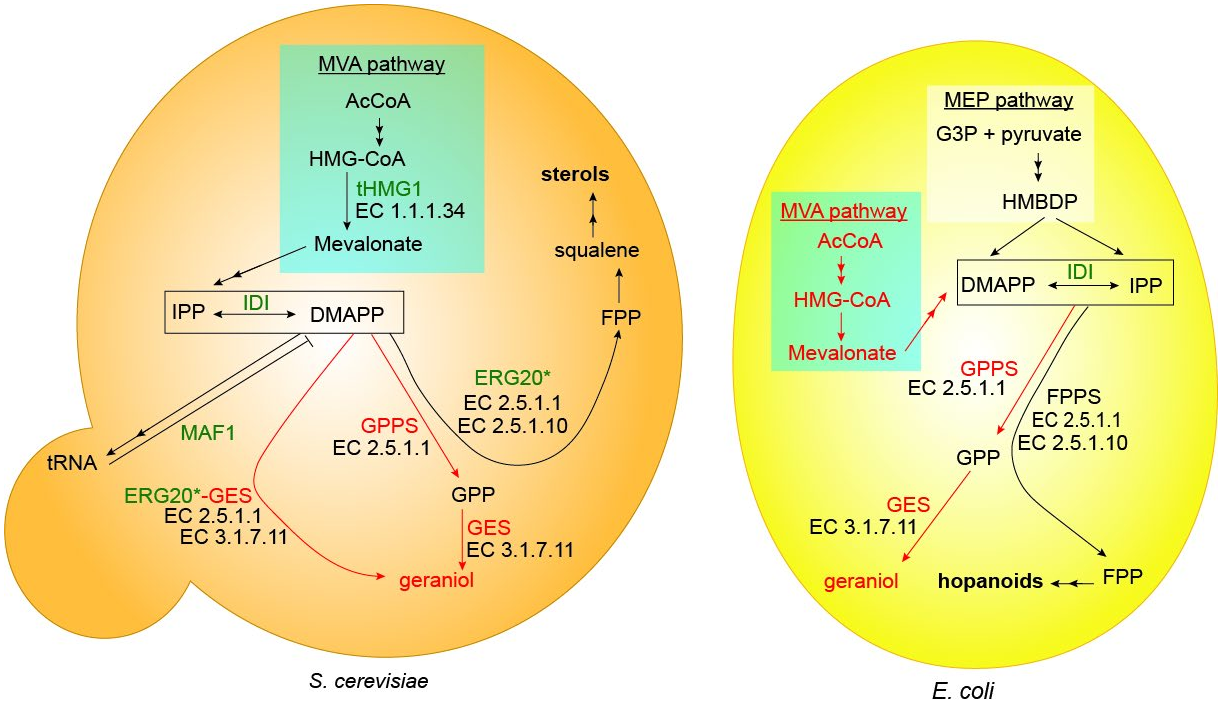


Figure 1. Schematic representation of metabolic engineering endeavours collated from various studies to drive heterologous geraniol production in yeast and bacteria. (A) Engineering approaches including exogenous expression (red) and exogenous overexpression (green) in *Saccharomyces cerevisiae,* which natively contains the MVA pathway (blue box). Overexpression of polymerase III transcription (MAF1) reduced DMAPP flux to tRNA synthesis and redirected it towards GPP synthesis, resulting in increased geraniol (Liu et al., 2013). Overexpression of rate-limiting enzymes IDI and tHMG1 increased flux to GPP synthesis resulting in increased geraniol (Zhao et al., 2016). Numerous studies have demonstrated that endogenous farnesyl diphosphate synthase (ERG20) can be modified (ERG20*) to favour GPP synthesis (EC 2.5.1.1) over FPP synthesis (EC 2.5.1.10) for increased monoterpenoid production (Ignea et al. 2014; Jiang et al., 2017; Zhao et al., 2016). Overexpression of exogenous plant GPP synthase (GPPS) caused decreased geraniol production (Zhao et al., 2016). Fusion of genes adjacent in the pathway, endogenous ERG20 mutant (ERG20*) and exogenous GES, resulted in increased geraniol production (Zhao et al., 2016; Jiang et al., 2017). (B) Engineering approaches including exogenous expression

(red) and exogenous overexpression (green) in *Escherichia coli*, which natively contains the MEP pathway (white box). Overexpression of genes involved in the MVA pathway (red) has resulted in an entirely synthetic MVA pathway (blue box) in *E. coli* for enhanced DMAPP and IPP synthesis and large yields of geraniol (Liu et al., 2016; Qian et al., 2019). Overexpression of expogenouus plant GPP synthase (GPPS) resulted in elevated geraniol production (Liu et al., 2016). *E. coli* do not naturally produce sterols, however, IPP and DMAPP are converted by FPP synthase (FPPS) into GPP (EC 2.5.1.1) and then FPP (EC 2.5.1.10) for production of sterol-like compounds called hopanoids.

Increasing GPP synthesis to improve heterologous geraniol production

In *S. cerevisiae,* various studies have investigated strategies to overcome rate-limiting steps in GPP synthesis (Liu et al., 2013; Zhao et al., 2016). Liu et al. (2013) demonstrated a 102% increase in geraniol following overexpression of repressor of RNA polymerase III transcription (MAF1) in *S. cerevisiae* expressing *Ocimum basilicum* GES (AY362553), compared to a parental strain. This is because both tRNA and GPP are synthesised from DMAPP; thus negative regulation of tRNA drove DMAPP flux towards GPP synthesis. Zhao et al. (2016) overexpressed native rate-limiting enzymes involved in the IPP and DMAPP biosynthesis pathway; 3-hydroxy-3-methylglutaryl-coenzyme A reductase 1 (HMG1) (854900; EC 1.1.1.34) and isopentenyl diphosphate delta isomerase (IDI1) (855986; EC 5.3.3.2). Following truncated HMG1 overexpression in a mutant strain expressing *Verbena officinalis* GES (KF951406), they reported an 83% increase in accumulation of geraniol and 34.6-fold increase in squalene compared to the parental strain. This suggested that while tHMG1-overexpression could increase carbon flux into the MVA pathway, it did not cause an increase GPP synthesis (Figure 1A). In the IDI1-overexpressing mutant, they reported 51% increased geraniol, but no significant difference in squalene, suggesting that IDI1 overexpression did indeed cause an increase GPP synthesis (Figure 1A).

Another approach to increase GPP synthesis is to install an additional IPP and DMAPP biosynthesis pathway. This has been demonstrated in *E. coli*, whereby a synthetic MVA pathway was successfully assembled and expressed to supplement the natively present MEP pathway resulting in impressive yields reaching 2 g/L (Liu et al., 2016) and 1.3 g/L (Qian et al., 2019) (Figure 1B).

Redistributing GPP flux towards heterologous geraniol production

There are also numerous approaches to redistribute flux away from GPP conversion into FPP (EC 2.5.1.10) in order to increase the free GPP available for heterologous monoterpenoid pathways. In 2014, Ignea et al. created a range of ERG20 mutant strains of *S. cerevisiae* which had reduced FPP biosynthesis capacity (EC 2.5.1.10) and instead strongly favoured GPP synthesis (EC 3.1.7.11) (Figure 1A). Of these, the most successful ERG20 mutant was able to produce a 340-fold increase in the monoterpenoid sabinene compared to the parental strain and subsequent studies have overexpressed this ERG20 mutant for enhanced geraniol production (Jiang et al., 2017; Zhao et al., 2016).

It is also possible to reduce the immediate conversion of GPP into FPP by installing a secondary, alternative exogenous GPP biosynthesis reaction (EC 3.1.7.11). In *E. coli,* Liu et al. (2016) reported a geraniol yield approximately double that of a control strain when engineered to express geranyl diphosphate synthase from the plant species *Abies grandis* (AgGPPS2; AF513112) (Figure 1B). A similar report in the cyanobacterium *Synechocystis sp. PCC 6803* demonstrated a 2.3-fold increase in the monoterpenoid limonene following AgGPPS2 expression (Lin et al., 2017). However, in *S. cerevisiae*, AgGPPS2 expression actually reduced the accumulation of geraniol by approximately 65% compared to the parental strain (Zhao et al., 2016) (Figure 1A).

This was confirmed with GPPS enzymes from three different plant species, *Abies grandis, Picea abies* (ACA21458) and *Catharanthus roseus* (JX417185).

Interestingly, while all GPPS expressing *S. cerevisiae* mutants showed decreased geraniol production, they also showed no change in squalene production. This suggested that heterologous production of GPPS cannot always alter the flux through a heterologous geraniol biosynthesis pathway, most likely due to (i) poor substrate utilisation by the heterologous GES, (ii) rapid utilisation by competing sterol biosynthesis pathways, or (iii) possible negative effects, such as toxicity, resulting from excessive heterologous GPP accumulation.

Increasing accessibility of GPP substrate to GES enzyme

Given that exogenous GPP synthesis can be ineffective at redistributing GPP flux away from FPP biosynthesis, studies have explored a strategy to increase accessibility of GPP to GES by fusing the GPP synthase enzyme (EC 3.1.7.11) directly to the GES enzyme, which can utilise the GPP for geraniol production (EC: 3.1.7.11). In *S. cerevisiae*, the GPP synthase-favouring ERG20 mutant described by Ignea et al. (2014) was fused to *Verbena officinalis* GES, improving geraniol accumulation by 1.7-fold compared to a strain which expressed these enzymes separately (Zhao et al., 2016). Likewise, Jiang et al. (2017) demonstrated a 15% increase in geraniol production in *S. cerevisiae* following expression of another ERG20 mutant fused to truncated *Verbena officinalis* GES.

Heterologous production of terpenoids in *P. tricornutum*

To date, there is significantly less knowledge available regarding terpenoid metabolism in *P. tricornutum* compared to medicinal plants, such as *C. roseus*, as well other microorganisms, namely *S. cerevisiae* and *E. coli*. Indeed, terpenoid engineering in *P. tricornutum* has only recently been reported (D'Adamo et al., 2018;

Fabris et al., 2020; George et al., 2020). In 2018, D'Adamo et al. demonstrated the heterologous production of the triterpenoid betulinic acid, and its precursor lupeol, reaching titres of approximately 0.1 mg/L. In 2020, we reported the first demonstration of heterologous production of the monoterpenoid geraniol following extrachromosomal expression of *Catharanthus roseus* GES (CrGES) reaching yields of 0.3 mg/L (Fabris et al., 2020). Subsequently, we demonstrated that randomly integrated chromosomal expression of CrGES resulted in titres up to 0.89 mg/L with no toxic effects, such as reduced growth (George et al., 2020).

Fabris et al. (2020) was also the first demonstration that *P. tricornutum* contains a free pool of cytosolic GPP. In *C. roseus,* the native GES enzyme contains a signal peptide localising it to the chloroplast to convert plastidial GPP into geraniol (EC: 3.1.7.11) (Simkin et al., 2013). In *P. tricornutum*, this 43 amino acid signal peptide is not functional and instead, exogenous CrGES localises to the cytosol where it converts GPP into geraniol (Fabris et al., 2020). However, it is still not known whether this GPP originates from cytosolic biosynthesis, or plastidial production and crosstalk into the cytosol, or a combination of both (Fabris et al., 2020). This is an example of where uncovering details in diatom terpenoid metabolism would be useful, as the enzymes involved with GPP synthesis in *P. tricornutum* have not yet been fully characterised.

Terpenoid biosynthesis in *P. tricornutum*

Diatoms are able to generate a diverse range of terpenoids. Over the last decade, numerous studies using 'omics analyses, modelling, inhibitors and genetic engineering have been useful for uncovering specific aspects of diatom terpenoid metabolism (D'Adamo et al., 2018; Fabris et al., 2014, 2012; Keeling et al., 2014; Pollier et al., 2019; Cvejic & Rohmer, 2000). Such studies have uncovered highly

unique aspects of *P. tricornutum's* terpenoid metabolism and a recent review has collated these findings (Athanasakoglou & Kampranis, 2019).

One of the major differences between diatoms and yeast or bacteria is that diatoms–including *P. tricornutum*–naturally bear both the MVA and MEP biosynthesis pathways (Figure 2). This is highly unusual in microalgae, of which most contain a plastidial MEP pathway, but lost a cytosolic MVA pathway through evolution (Lohr et al., 2012).

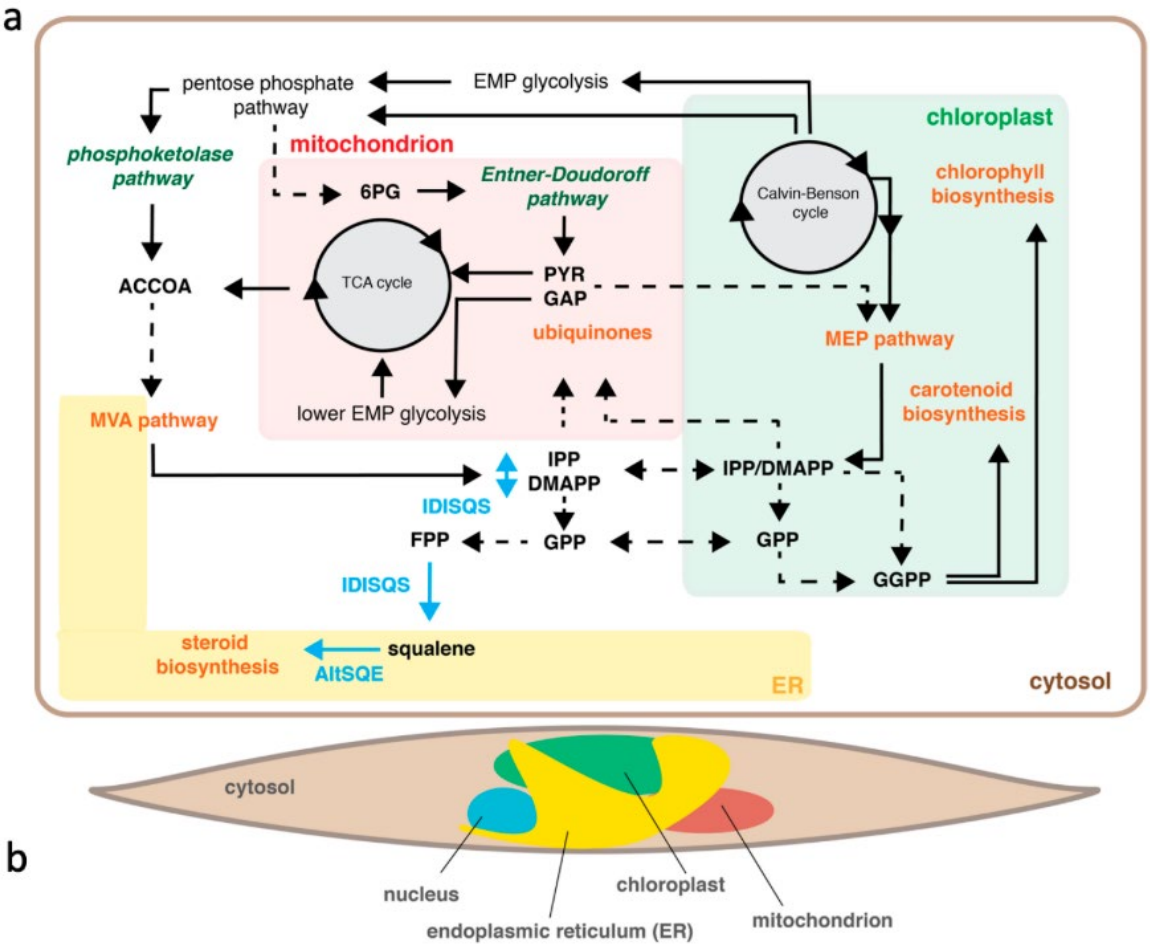


Figure 2. Graphic representation of a hypothetical terpenoid metabolic network in *P. tricornutum* strain CCMP/1055. The main terpenoid pathways are highlighted in orange; in light blue, distinctive diatom enzymes involved in diatom sterol biosynthesis; in green, distinctive diatom pathways hypothetically providing substrates to isoprenoid biosynthesis; dashed lines indicate hypothetical conversions and transport reactions. Abbreviations: ACCOA, acetyl-CoA; PYR, pyruvate; GAP, glyceraldehyde 3-phosphate; 6PG, 6-phosphogluconate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate; IPP isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; IDISQS, isopentenyl diphosphate

isomerase/squalene synthase; AltSQE, alternative squalene epoxidase. From Fabris et al. (2020).

In *P. tricornutum*, carbon and hydrogen isotope labelling experiments have shown that sterols are preferentially produced from acetate directed through the MVA pathway in the cytosol, whereas carotenoids are preferentially produced from fixed carbon flux through the MEP pathway in the chloroplast (Cvejic & Rohmer, 2000). Carotenoids are C40 tetraterpenoids that are synthesised from the condensation of two GGPP molecules (Figure 2). Carotenoids are classified as either carotenes, such as β-carotene, or xanthophylls, such as fucoxanthin, diadinoxanthin and diatoxanthin; all of which play essential roles in the photosystem for photoprotection via non-photochemical quenching (Kuczynska et al., 2015). Fucoxanthin, which is produced from a key intermediate β-carotene, is the main photosynthetic pigment in diatoms and is in high demand as a nutraceutical, particularly as an antioxidant (Athanasakoglou & Kampranis, 2019). GGPP is also the precursor of chlorophylls, which are highly conserved throughout photosynthetic organisms due to their central role in converting light energy into chemical energy in photosystem I and distributing excitation energy between photosystem I and II (Figure 2) (Owens, 1986).

In the cytosol, IPP is converted to DMAPP in a reversible reaction by isopentenyl diphosphate isomerase/squalene synthase (IDISQS). Unusually, *P. tricornutum* isopentenyl diphosphate isomerase is fused to squalene synthase and therefore, also converts FPP to squalene in the endoplasmic reticulum. Isopentenyl diphosphate isomerisation and squalene synthesis are not consecutive reactions in squalene biosynthesis and it is hypothesised that these enzymes are fused in *P. tricornutum* to commit flux through sterol biosynthesis (Athanasakoglou & Kampranis, 2019). Squalene is the precursor of all eukaryotic sterols and its conversion to 2,3-

oxidosqualene occurs via a unique squalene epoxidase called alternative squalene epoxidase (AltSQE), only recently discovered in *P. tricornutum* but now known to be present in various other organisms which do not bear a conventional squalene epoxidase (Figure 2) (Pollier et al., 2019).

Diatoms also bear three glycolytic pathways: a Embden–Meyerhof–Parnas (EMP) pathway, which is also present in chlorophytes; a functional Entner-Douroroff pathway, more conventionally associated with prokaryotic metabolism; and a putative phosphoketolase pathway, which is common in fungal metabolism (Figure 2) (Fabris et al., 2012; Hildebrand et al., 2013). Pyruvate and G3P are produced via the Entner-Douroroff pathway and are precursors for MEP pathway; likewise, acetyl-CoA is produced via the EMP pathway–and could also be produced via the putative phosphoketolase pathway–and is a precursor for MVA pathway (Fabris et al., 2012; Meadows et al., 2016). Both Entner-Douroroff and the putative phosphoketolase pathways also release an ATP and NADPH molecule, which are required by both MVA and MEP pathways (Vavitsas et al., 2018).

Although many unusual aspects of diatom terpenoid metabolism have been uncovered, there are still many unknown aspects of these pathways; such as (i) cross talk between cytosolic and plastidial pools of IPP, DMAPP, GPP and (ii) GPP biosynthesis (dashed lines in Figure 2).

P. tricornutum as a heterologous monoterpenoid biofactory

Alongside an incomplete understanding of native diatom terpenoid metabolism, there is also no knowledge yet of engineering strategies able to elevate monoterpenoid production in *P. tricornutum*. However, our results have already uncovered important aspects of *P. tricornutum* which can inform rational design approaches to such

investigations. First, we have shown that extrachromosomal expression is a more appropriate genetic engineering tool for metabolic engineering in *P. tricornutum* because it results in highly consistent mutants and no inadvertent integration of plasmid DNA (George et al., 2020). This is of particular importance when elucidating the influence of heterologous geraniol biosynthesis on diatom native terpenoid metabolism and consequently assessing this microorganism as a potential MIA biofactory. Second, *P. tricornutum* natively contains a free pool of GPP (Fabris et al., 2020), which–unlike *S. cerevisiae* and *E. coli*–is not the main limiting factor of geraniol production. Instead, our work showed that increasing the level of CrGES expression resulted in significantly increased geraniol production (George et al., 2020). This suggested that CrGES availability and/or access to GPP is a more limiting factor in *P. tricornutum's* capacity for heterologous geraniol production than GPP availability.

With all of this in mind, we used extrachromosomal expression of exogenous *Abies grandis* GPP synthase (AgGPPS2; AF513112) fused to *Catharanthus roseus* geraniol synthase (CrGES) to assess the impact of geraniol production on both sterol and pigment diatom biosynthesis. Given the lack of knowledge regarding GPP synthesis and flux in *P. tricornutum*, it is not yet known how geraniol expression might alter flux to both sterol and pigment biosynthesis. Therefore, we aimed to increase flux through heterologous geraniol production in *P. tricornutum* by both increasing geraniol synthase expression and availability to GPP, as well as by prolonging the cultivation period. We assessed the impact of increased geraniol on key diatom sterols squalene, cycloartenol, cholesterol campesterol and brassicasterol; and pigments *ß*-carotene, fucoxanthin, diatoxanthin, diadinoxanthin, and chlorophylls *a* and *c*. Like geraniol, both sterols and pigments depend on endogenously available GPP. Therefore, we hypothesised that increased geraniol would either decrease production of a selection

of essential sterol triterpenoids, or pigment sesquiterpenoids; or a combination of these. Such knowledge is important not only for informing future rational designs to increase flux to more complex monoterpenoid pathways, but also for elucidating *P. tricornutum's* metabolic flexibility.

RESULTS AND DISCUSSION

Constitutive, extrachromosomal expression of CrGES resulted in increased geraniol accumulation compared to induced expression

The synthetic biology design-build-test-learn pipeline depends heavily on the need to compare and validate many genetic parts in a reliable, high-throughput manner. Given that most genetic engineering strategies in *Phaeodactylum tricornutum* depend on randomly integrated chromosomal expression (RICE), this has not been possible. In Chapter 3, we robustly demonstrated this, showing that RICE is associated with highly dissimilar transgene expression outputs and massive transgene insertion islands that can disrupt protein coding regions. Attempting to build a synthetic biology pipeline that validated genetic parts by RICE would be unreliable and extremely inefficient, as clones are subject to position effect, resulting in extremely diverse clones using the same parts (George et al., 2020).

Unlike RICE, we also demonstrated that extrachromosomal expression (EE) is associated with highly consistent, reproducible mean expression of transgenes (Chapter 3) and that bacterial conjugation does not result in inadvertent plasmid DNA integration into the genome. These characteristics of EE are extremely useful for comparing the performance of different genetic parts, such as promoter elements, modified GES enzymes, and the effect of multigene combinatorial design associated with metabolic engineering.

We previously demonstrated that *geraniol synthase* fused to reporter gene, *mVenus (GES-mVenus)* was extrachromosomally expressed in *P. tricornutum* using an inducible promoter, *AP1* (Chapter 3). We chose an inducible configuration because heterologous geraniol production can be toxic in some species (Jiang et al., 2017), and has never before been reported in *P. tricornutum*. The work presented in Chapter

3 demonstrated that geraniol produced at titres as high as 0.89 mg/L in strain RICE_GmV-41 had no toxic effects, such as reduced growth. This suggested that geraniol production in *P. tricornutum* is not limited by availability of free GPP, but rather by expression levels of *GES-mVenus*. However, strain RICE_GmV-41 depended on random integration of GES-mVenus as well as inducible expression. Consequently, it is unknown how these large, highly concatenated integration events impact the global metabolic network. Furthermore, inducible expression driven by phosphate starvation would also cause metabolic stress and also influence the metabolic network, making it impossible to decipher changes in flux due to geraniol production alone. Therefore, we compared various constitutive promoter sequences using extrachromosomal expression on order to increase GPP flux through geraniol synthesis.

We demonstrated that constitutive, extrachromosomal expression of Cr*GES-mVenus* resulted in mVenus fluorescence within the cytosol throughout the cultivation period (Fabris et al., 2020). Two of the four promoter sequences tested, (*Phatr3_J21659*; Fabris et al., 2020) and (*Phatr3_J49202;* Pollak et al., 2019) showed the highest mVenus expression and geraniol accumulation (Fabris et al., 2020).

In order to test if constitutive, high extrachromosomal expression of Cr*GES-mVenus* in *P. tricornutum* is associated with elevated geraniol production, we set up a full-scale batch cultivation experiment with wild type and engineered cell lines and sampled each culture for cell growth rates and mVenus fluorescence over 7 days, and cumulative geraniol production over the full course of the experiment. We compared three independent exconjugants transformed with *Phatr3_J49202p_CrGES-mVenus* (referred to as EE_49p-GmV) and three transformed with *Phatr3_J21659_CrGES-mVenus* (referred to as EE_21p-GmV). We included one control cell line for each promoter sequence, *Phatr3_J49202p_mVenus* (referred to as EE_49p-mV) and

Phatr3_J21659_mVenus (referred to as EE_21p-mV) and one wild type strain (Figure 3).

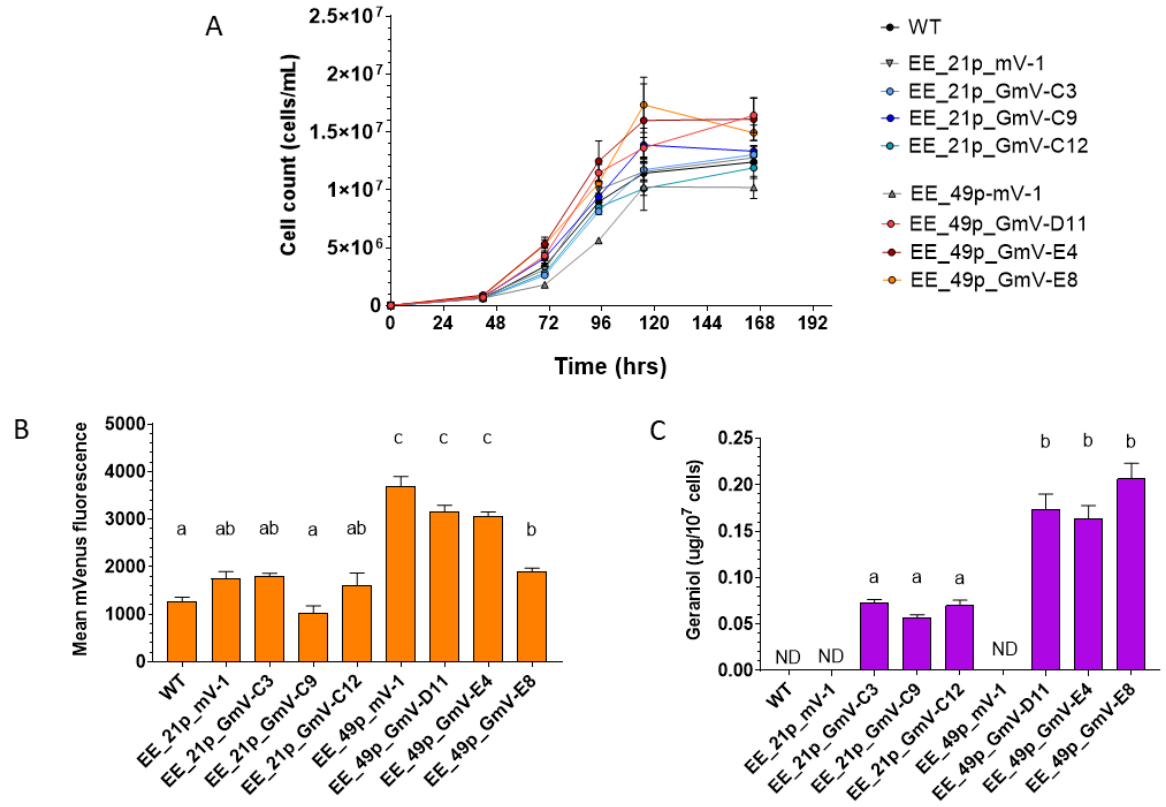


Figure 3. Constitutive production of geraniol in *P. tricornutum*. (A) Growth profile of wild type diatoms and cell lines harbouring either *49202p_GES-mVenus*, *49202p_mVenus*, *21659p_GES-mVenus*, or *21659p_mVenus* (n = 3, error bar indicates SEM) (B) representative mean mVenus fluorescence of control and transgenic strains after 96 hours of cultivation in ESAW medium (20,000 cells analysed for each cell line), mean and SEM are depicted (n = 3); (C) final geraniol yield in CrGES-mVenus expressing diatoms after 168 hours, compared to empty vector and wild type controls. ND: not detected. Mean and SEM are depicted (n = 3). Identical letters denote no statistically significant differences among groups using the Tukey method. From Fabris et al. (2020).

As expected, EE_49p-GmV-E8, EE_49p-GmV-E4 and EE_49p-GmV-D11 cell lines all showed mVenus fluorescence signals significantly higher than wild type auto fluorescence (Figure 3B). This correlated with increased geraniol accumulation, the highest titre reaching 0.21 µg/10^7 cells, which corresponds to 309 µg/L at the density measured at harvest of 1.49 x 10^7 cells/mL (EE_49p-GmV-E8). This is three times

higher than that accumulated in EE_21p-GmV-C3, EE_21p-GmV-C9 and EE_21p-GmV-C12 strains, in which the highest titre only reached 0.07 µg/10^7 cells (95 µg/L) in strain EE_21p-GmV-C3 (Figure 3C). This supported our previous finding that the production of heterologous geraniol in *P. tricornutum* may be proportional to *GES-mVenus* expression levels, instead of being limited by GPP precursors, as it is in yeast (Oswald, Fischer, Dirninger, & Karst, 2007). Therefore, we proceeded to apply rational design using the *Phatr3_J49202p* promoter sequence to further increase geraniol accumulation using the consistent, non-disruptive approach of extrachromosomal expression.

Expression of AgGPPS2-CrGES fusion enzyme does not increase geraniol production

As previously outlined, there are many rational design strategies to apply to the diatom synthetic geraniol producing pathway, such as improving substrate utilisation and overexpressing rate limiting enzymes. We have previously shown that increased expression of Cr*GES-mVenus* results in increased geraniol both via randomly integrated inducible chromosomal expression (George et al., 2020) and via extrachromosomal, constitutive expression (Fabris et al., 2020). This evidence strongly suggests that geraniol productivity in *P. tricornutum* is more likely to depend on expression of *CrGES* and not on low GPP availability, and that *P. tricornutum* might be able to tolerate higher levels of GPP production than yeast (Brennan et al., 2012; Zhao et al., 2016) and bacteria (Shah et al., 2013; Wang et al., 2010). Therefore, we aimed to increase the accessibility of CrGES to the expectedly large free pool of GPP substrate by physically linking a geranyl diphosphate synthase from *Abies grandis* (AgGPPS2) to geraniol synthase from *Catharanthus roseus* (CrGES), to avoid

diffusion of substrate as well as competition with other GPP catabolising enzymes, such as FPP synthase involved in sterol biosynthesis.

In order to test if the exogenous Cr*GES-AgGPPS2* fusion enzyme would be able to increase flux towards production of geraniol, we introduced into *P. tricornutum* the codon-optimised, full-length coding sequence of *CrGES* (Fabris et al., 2020; George et al., 2020) without a stop codon, fused at its carboxy-terminal to a flexible linker λ (GSTSSGSG) and the codon-optimised, full-length coding sequence of *AgGPPS2* without any mVenus reporter enzyme. Given the higher GES-mVenus fluorescence reported above following expression driven by the *Phatr3_J49202* promoter sequence, we modified the genetic design to include both the 5' and 3' *Phatr3_J49202* regulatory regions, instead of just the *Phatr3_J49202* promoter sequence with the *Phatr3_J18049* (FCBPt) terminator sequence, as was used in all previous genetic designs used in our laboratory for studying *CrGES* (Fabris et al., 2020; George et al., 2020). The resultant pPTBR11_Phatr3_*J49202p-CrGES-AgGPPS2* construct (Figure 4A) was conjugated into wild type *P. tricornutum* with high efficiency of zeocin-resistant exconjugants *EE_49p_CrGES-AgGPPS2* associated with this transformation strategy resulting. A control strain was transformed in the same way using pPTBR11_*Phatr3_J49202p-CrGES-Phatr3_J49202t* construct (Figure 4A) without the AgGPPS2 fusion enzyme also resulted in a high efficiency of *EE_49p_CrGES* exconjugants.

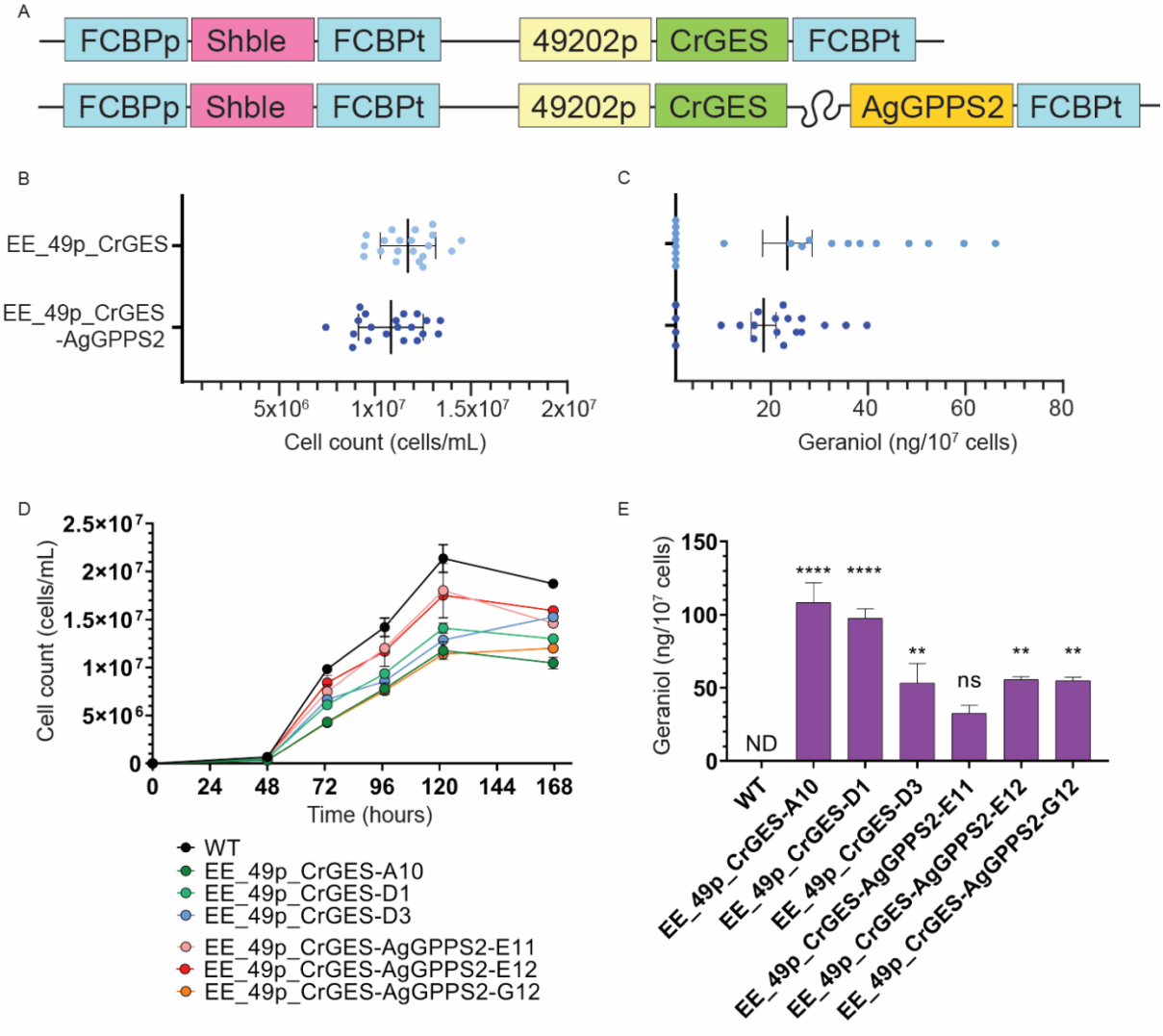


Figure 4. Extrachromosomal expression of *Catharanthus roseus geraniol synthase* (*CrGES)* fused to *Abies grandis geranyl diphosphate synthase (AgGPPS2)* in *P. tricornutum.* (A) Schematic representation of genetic constructs designed and built in this study, pPTBR11_Phatr3_*J49202p-CrGES* (11,530 bp in size) and pPTBR11_Phatr3_*J49202p-CrGES-AgGPPS2* (12,703 bp in size). (B) Cell counts and (C) geraniol quantification at time of harvesting from 20 EE_49p_CrGES-AgGPPS2 cell lines and 20 EE_49p_CrGES cell lines analysed in the initial screening. (D) Growth and (E) geraniol quantification at time of harvesting the three best EE_49p_CrGES-AgGPPS2 and EE_49p_CrGES engineered strains and wild type control following full-scale batch cultivation experiment. (n = 3, error bars indicate SEM).

We screened 20 randomly selected *EE_49p_CrGES-AgGPPS2* and 20 randomly selected *EE_49p_CrGES* independent exconjugants using a batch cultivation experiment sampled after 7 days in order to identify any superior geraniol producing

strains (Figure 4B and C) before subsequent analysis. We then analysed the three highest geraniol accumulating *EE_49p_CrGES-AgGPPS2* cell lines and three highest *EE_49p_CrGES* cell lines in a full-scale batch cultivation. Interestingly, the *EE_49p_CrGES-AgGPPS2* cell lines showed a 55.3% lower geraniol accumulation (average of 47.70 ng/10^7 cells) compared to *EE_49p_CrGES* cell lines (average of 86.29 ng/10^7 cells). This was consistent in both the screening analysis (Figure 4C) and full-scale batch cultivation (Figure 4E). This suggested that the exogenous AgGPPS2 might not be functional in *P. tricornutum*, or that the fusion CrGES-AgGPPS2 protein was folded incorrectly, despite the use of a flexible linker. In this case, it is not unlikely that geraniol accumulation would be lower, as the misfolding would impact either the CrGES activity, or the AgGPPS2 activity, or both. Any impact on the CrGES activity would reduce its capacity to convert the native GPP into geraniol resulting in decreased geraniol production compared to *EE_49p_CrGES* cell lines. While GPPS2 expression has been shown to increase geraniol accumulation (Jiang et al., 2017; Zhao et al., 2016), other reports have shown similar reductions of approximately 65% in geraniol accumulation, similar to what we report here (Zhao et al., 2016). While it is possible that the linker may not have been long enough to prevent any interferences with the catalytic centres of AgGPPS2 or CrGES, this design was previously validated (Jiang et al., 2017).

Two additional aspects of this experiment were unexpected. First, the *EE_49p_CrGES* cell lines obtained produced an average geraniol yield (86.29 ng/10^7 cells) approximately 40% lower than those with the same genetic configuration, EE_49p_GmV (~200 ng/10^7 cells), with the exception of the mVenus fusion and *FCB*P terminator. This suggested that perhaps *Phatr3_J49202* promoter sequence might perform better with *FCBP* terminator sequence; as the *Phatr3_J49202* regulatory

regions, as well as the protein coding sequence, are currently uncharacterised. Alternatively, there may be some benefit to expressing exogenous CrGES fused to a smaller, stabilising protein such as mVenus, which we removed here. Second, it was surprising that all the transgenic cell lines grew to lower cell densities than wild type (Figure 4D). This has never been observed before, as *GES-mVenus* expressing *P. tricornutum* lines usually show growth rates that are similar to or faster than wild type controls, when engineered both random integration and extrachromosomal expression. This could indicate that the genetic modifications associated with EE_49p_CrGES strains impacted cell fitness. Given these findings, we proceeded to explore geraniol optimisation strategies using the highest geraniol-producing EE strain we had obtained to date, EE_49p_GmV-E8, which reached a geraniol productivity of 0.21 µg/10^7 cells, corresponding to 309 µg/L.

Photoperiod and prolonged cultivation do not affect heterologous monoterpenoid production

While genetic design strategies have been explored in non-photosynthetic yeast and bacteria, there are no studies exploring the effect of photoperiod on geraniol accumulation. A recent investigation of heterologous terpenoid production in the green chlorophyte *Chlamydomonas reinhardtii* demonstrated that cultivation in a photoperiodic light regime as opposed to continuous light resulted in increased terpenoid accumulation (Lauersen et al., 2018). However, it has also been shown that most isoprenoid enzymes seem to be active in the light phase (Smith et al., 2016). Therefore, the increased terpenoid accumulation reported by Lauersen et al. (2018) might be explained by less photosynthetic metabolic demand associated with continuous light regimes, and therefore more metabolic resources might be shifted towards terpenoid production. We hypothesised that prolonged cultivation time could

increase terpenoid accumulation due to the slowing of growth associated with photoperiodic light regime (Lauersen et al., 2018). However, these results pertain to the synthesis of heterologous compounds from farnesyl diphosphate (FPP) and geranylgeranyl diphosphate (GGPP) in an organism phylogenetically and biochemically distant from diatoms. Nothing is currently known about the regulation and fluctuation of prenylphosphate pools in diatoms.

To assess whether these two primary cultivation parameters have effects on the production of monoterpenoids in *P. tricornutum*, we profiled the geraniol production in the highest geraniol-yielding exconjugant cell line, EE_49p_GmV-E8, following cultivation in either continuous light or following a 12:12 hours photoperiod at the same light intensity, either for seven or for ten days, thus prolonging the stationary phase by three additional days.

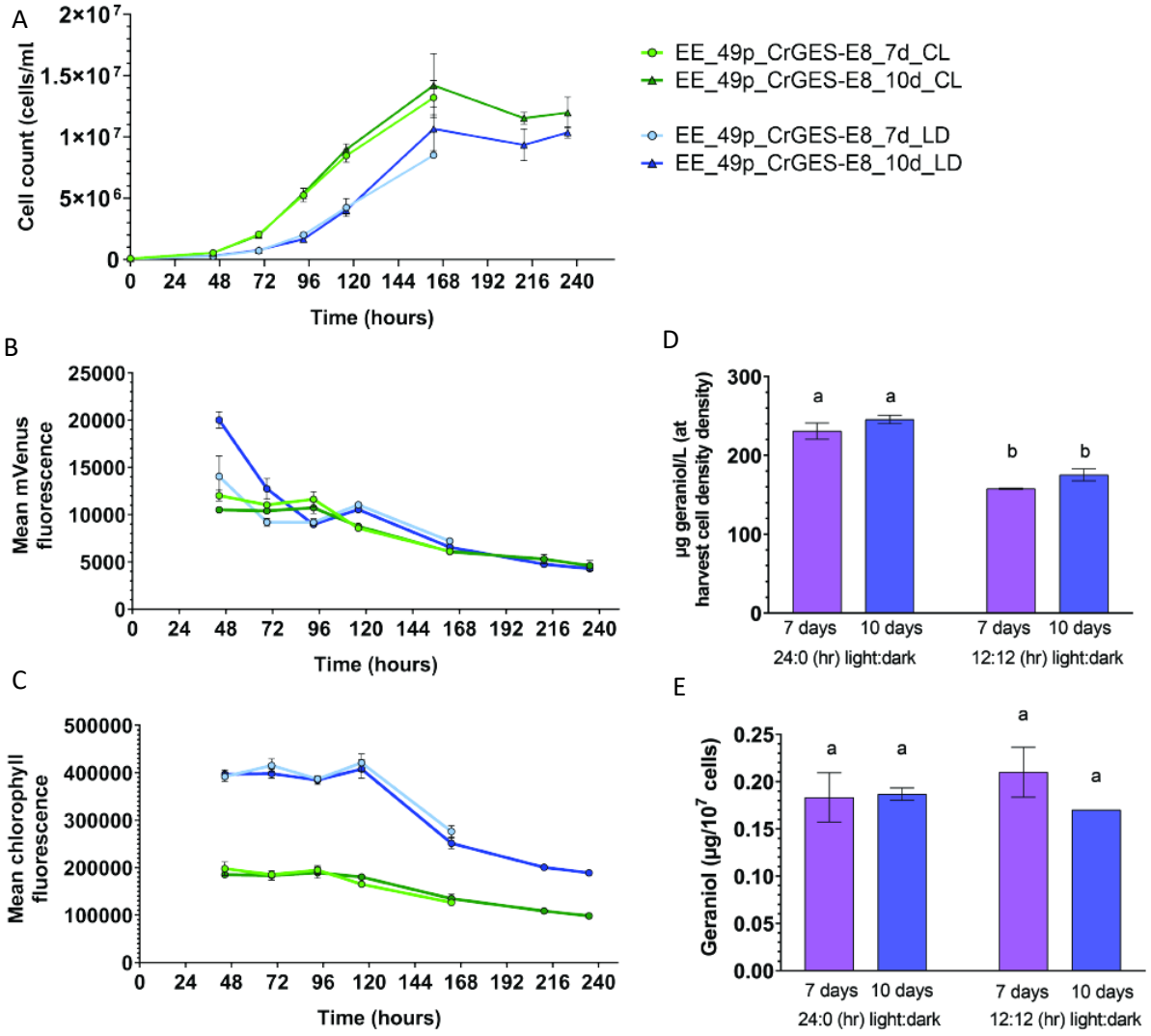


Figure 5. Effects of light regime and cultivation time on the geraniol yield in the *P. tricornutum* strain 4-GESmV-E8. (A) growth profile (B) mVenus mean fluorescence (C) chlorophyll fluorescence, of EE_49p_GmV-E8 diatoms either grown in continuous light (CL green) or light/dark (LD, blue) regime, for 7 (lighter colours) or 10 (darker colours) days (D) and (E) overall geraniol yield expressed in µg//L and in µg product/10^7 cells, respectively (n = 3, error bars indicate standard error of the mean). Identical letters denote no statistically significant differences among groups using the Tukey method**.** From Fabris et al. (2020).

Cultures grown in continuous light (CL) regime exhibited slightly higher final cell densities than those grown with a 12-hour photoperiod (LD), whereas densities did not significantly change in cell lines grown in the same light regime (Figure 5A). As expected cell lines grown in CL exhibited a higher chlorophyll fluorescence profile

throughout the experiment (Figure 5C). The expression of the GES-mVenus fusion protein, monitored by the analysis of mVenus mean fluorescence, was consistent in all conditions, with the highest expression in early exponential phase and a gradual decline in stationary phase (Figure 5B). Cultures grown either in CL or with a LD cycle showed significant variation in geraniol production consistent with the different final cell densities of these cultures (Figure 5D), with higher yields reported after 10 days for cultures grown in CL (245.56 µg/L) compared to those grown in LD (175.22 µg/L). However, when considering the overall µg of geraniol produced per 10^7 cells, cultures exhibited no significant differences (Figure 5E). These results show that the production of geraniol in *P. tricornutum* is minimal in the late stationary phase. This could be explained by the nature of the promoter of the gene *Phatr3_J49202*, for which expression peaks in the exponential growth phase in the tested conditions (Fabris et al., 2020). Moreover, primary cellular metabolism–of which MVA and MEP are a part– is more transcriptionally active during the exponential phase of *P. tricornutum* (Smith et al., 2016). These results correlate with the productivity profile reported in the heterologous production of betulinic acid in *P. tricornutum* (D'Adamo et al., 2018), and are in contrast with heterologous production of terpenoids in *C. reinhardtii* (Lauersen *et al.*, 2018a). With the tested genetic constructs, in *P. tricornutum*, factors such as light regime and prolonged stationary phase, seem to not have significant effects on the yield of geraniol based on cell density (Figure 5D), while continuous light regimes allow the diatoms to grow faster and to a slightly higher density.

Cytosolic expression of GES-mVenus does not affect pigment and sterol content in *P. tricornutum*

In higher plants, the plastidial MEP pathway provides precursors IPP and DMAPP to the biosynthesis of photosynthetic pigments and monoterpenoids. Recent studies in

the diatom *Haslea ostrearia* proposed the synthesis of GPP as an intermediate to GGPP in the chloroplast (Athanasakoglou et al., 2018) and to FPP in the cytosol (Figure 6A). While *P. tricornutum* accumulates a pool of GPP in the cytosol, it is not clear whether this is fully synthetised in the cytosol alone, or if plastidial IPP and DMAPP are transported into the cytosol to contribute to GPP biosynthesis (Figure 6A). Alternatively, GPP itself may be partially or completely exported from the chloroplast to the cytosol. Hence, we investigated whether the presence of a functional CrGES in the diatom cytosol would draw precursor moieties from the biosynthesis of pigments and sterols and perturb their homeostasis. To address this, we profiled the composition of representative carotenoids (ß-carotene, fucoxanthin, diatoxanthin, diadinoxanthin) and chlorophylls (*a* and *c*), in the representative EE_49p_GmV-E8 diatom cell line, compared to a transgenic and a wild type control as part of the experiment described in Figure 3. Apart from ß-carotene, the EE_49p_GmV-E8 line showed a slightly lower pigment content compared to the controls, however, no significant differences were observed (Figure 6B).

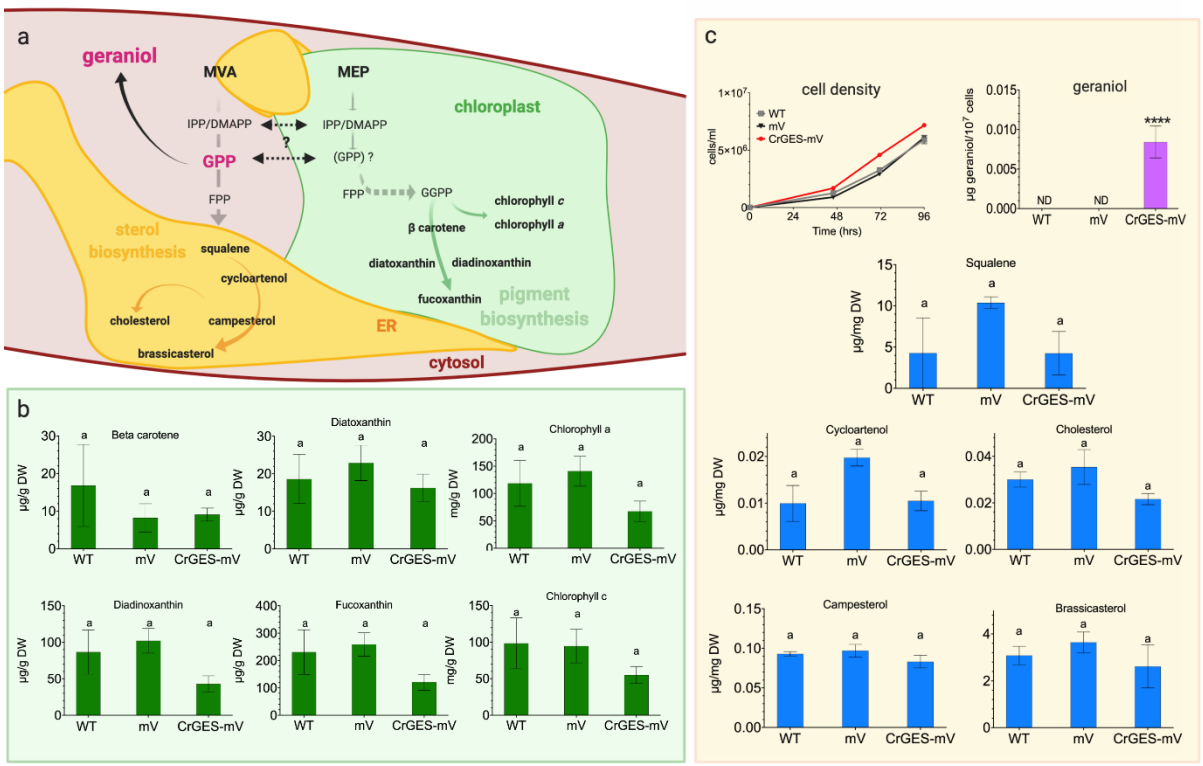


Figure 6. Interaction of cytosolic geraniol biosynthesis with the endogenous terpenoid pathways in *P. tricornutum*. (A) schematic representation of hypothetical substrate subcellular allocation between geraniol, sterol and pigment biosynthesis; (B) accumulation of photosynthetic pigments in wild type (WT) and transgenic diatom lines EE-*49p-GmV-E8* (*GES-mV*) and *49202p_mVenus-1* (*mV*), after 192 hours of cultivation (C) growth profile and accumulation of geraniol in WT and representative transgenic diatom lines GES-mV and mV cultures, sampled after 96 hours of cultivation, and accumulation of main triterpenoids (n = 3, mean and SEM are depicted, asterisks indicate significant differences with control lines, identical letters denote no statistically significant differences among groups using the Tukey method). From Fabris et al. (2020).

It has been determined that diatom sterol metabolism is mostly active during the exponential growth phase (D'Adamo et al., 2018). Therefore, we designed an experiment to specifically evaluate the effect of geraniol production on the triterpenoid composition (squalene, cycloartenol, cholesterol campesterol and brassicasterol) in EE_49p_GmV-E8, EE_49p_mV-1, and wild type *P. tricornutum* in exponential phase, which in our settings correspond to 96 hours of cultivation. As we reported with pigment comparisons, the production of geraniol did not significantly affect the growth or the content of squalene and sterols of the analysed diatom cell lines (Figure 6C).

We detected 0.07 µg geraniol per 10^7 cells, correlating to 0.25 µg/µg biomass or $2.47x10^{-7}$ µg/g biomass. The pigments analysed ranged from 10 – 250 µg/g biomass, whereas the sterols detected ranged from 0.00001 – 0.01 µg/g biomass (Figure 6B and C). Considering these results, we suggest that either the cytosolic GPP pool converted to geraniol by GES-mVenus is not, or is only marginally, involved in the biosynthesis of pigments and sterols. This would occur by the sequential phosphorylation of GPP into FPP and geranylgeranyl diphosphate (GGPP), respectively. GGPP is the isoprenoid precursor for pigment biosynthesis, including those we tested here (Figure 6A). Alternatively–and more likely–the expression levels of *GES-mVenus* that we report here are below the threshold where pigment and triterpenoid content are impacted.

CONCLUSION

Terpenoid metabolic engineering has only recently been investigated the model pennate diatom *P. tricornutum* (D'Adamo et al., 2018; Fabris et al., 2020; George et al., 2020). Consequently, there is significantly less understanding about native diatom terpenoid metabolism and systems biology compared to that of other more extensively explored in microorganisms, namely *E. coli* (Alper et al., 2005; King et al., 2017; Maury et al., 2008; Ward et al., 2018) and *S. cerevisiae* (Asadollahi et al., 2009; Chemler et al., 2006; Forster et al., 2003; Meadows et al., 2016).

Herein, we used the recently developed approach of extrachromosomal expression to compare metabolic engineering and bioprocessing approaches to increase heterologous production of the monoterpenoid geraniol in *P. tricornutum*. First, we attempted to increase GPP substrate availability and utilisation by expressing two adjacent enzymes fused to one another: geranyl diphosphate synthase (AgGPPS2) (AF513112; EC 2.5.1.1) which produces GPP from IPP and/or DMAPP in the plant *Abies grandis,* and geraniol synthase (CrGES) (Caros003727.1; EC 3.1.7.11) which converts GPP into geraniol in the plant *Catharanthus roseus*. This resulted in 55.3% decreased geraniol accumulation compared to a control cell line that did not contain AgGPPS2, most likely due to improper folding of the fused exogenous enzymes. It is not surprising that CrGES expressed in various configurations shows a large variation in efficiencies: reaching geraniol titres of up to 210 ng/10^7 cells when expressed as CrGES-mVenus fusion enzyme (Fabris et al., 2020), 54.9 ng/10^7 cells when expressed as AgGPPS2-CrGES fusion enzyme, 108.2 ng/10^7 cells when expressed as a free enzyme and 4.08 ng/10^7 cells when expressed as a truncated protein (Fabris et al., 2020). Therefore, further work should also explore protein engineering, particularly to improve AgGPPS2-CrGES fusion enzyme activity in *P. tricornutum*.

Next, we showed that a dark:light (12h:12h) photoperiod had no impact on geraniol production compared to continuous light cultivation. Similarly, we reported no impact geraniol production following an elongated cultivation period extending beyond stationary phase. This was not surprising, as primary cellular metabolism–of which MVA and MEP pathways are a part–is more transcriptionally active during the exponential phase of *P. tricornutum* (Smith et al., 2016). These results correlate with the productivity profile reported in the heterologous production of betullinic acid in *P. tricornutum* (D'Adamo et al., 2018), and are in contrast with heterologous production of terpenoids in *C. reinhardtii* (Lauersen *et al.*, 2018a).

Finally, we investigated the impact of exogenous CrGES expression and activity on other native *P. tricornutum* terpenoids, namely sterols and pigments. Such information is important for future metabolic engineering approaches as well as expanding the understanding of native terpenoid metabolism in this widely studied model diatom. Because there is very little information available regarding the biosynthesis of GPP in *P. tricornutum*, it is not known if this cytosolic pool is created entirely in the cytosol via the MVA pathway, or partially or completely by the MEP pathway in the chloroplast and transported into the cytosol. We reported that neither sterol nor pigment biosynthesis was significantly impacted by heterologous geraniol production, most likely due to the expression levels of *GES-mVenus* being below the threshold where pigment and sterol content are impacted. This suggested that terpenoid metabolism in diatoms could be particularly flexible and able to adapt to the installation of artificial metabolic sinks as hypothesised in the green alga *C. reinhardtii* (Lauersen et al., 2016). If confirmed, this might represent another promising trait for terpenoid-based engineering in *P. tricornutum*.

Further work is needed to better elucidate diatom isoprenoid metabolism, which appears to be quite different from other organisms, as well as to better characterise GPP flux through native and heterologous metabolic pathways in *P. tricornutum*. There are numerous alternative rational designs to consider for this approach–none of which have been tested in any diatom but have been validated in other species. These include, but are not limited to, adding a heterologous MVA pathway (Liu et al., 2016; Qian et al., 2019); testing the yeast farnesyl diphosphate synthase (ERG20) mutant, which favours GPP synthesis over FPP synthesis (Ignea et al., 2014); overexpressing key enzymes such as HMG-CoA reductase in the MVA pathway (under investigation in our laboratory) and determining the specificity of *P. tricornutum's* native prenyltransferases for informing future engineering approaches (under investigation in our laboratory). Such approaches could overcome the diatom's natural cellular regulation, even if redundant. Given that we demonstrated high expression following RICE and highly concatenated *CrGES* gene arrangement, it would also be appropriate to explore multigene copy expression on geraniol production in *P. tricornutum*. This approach has been validated in *S. cerevisiae*, whereby expression of an *ERG20* mutant fused to *CrGES* was supplemented with an additional free copy of the *ERG20* mutant gene, resulting in approximately 20% increased geraniol accumulation (Jiang et al., 2017).

Such knowledge is important for better elucidating the ecological impact of these important marine primary producers in a changing climate, as well as for biotechnological production of high-value terpenoids such as monoterpenoid indole alkaloids, which are difficult to produce at high titres in *E. coli* and *S. cerevisiae (REFS).* Finally, throughout all the experiments in this study, we demonstrated how extrachromosomal expression can be useful for faster synthetic biology design-build-

test-learn cycle compared to RICE, owing to the fact that this non-integrative approach results in highly consistent transgenic phenotypes across clones (George et al., 2020). In this way, the work described here offers an important proof-of-concept for future more complex synthetic biology ventures beyond terpenoid metabolic engineering.

METHODS

Microbial strains and growth conditions

Phaeodactylum tricornutum CCAP1055/1 was grown in liquid ESAW (Berges et al., 2001) supplemented with 50 µg/mL zeocin (Invivogen, San Diego, CA, USA) where appropriate, under 100 µE m^{-2} s^{-1} light in 21 °C shaking at 95 rpm. *Escherichia coli* was grown in Luria broth supplemented with 100 µg/mL ampicillin. For analyses, transgenic *P. tricornutum* strains were cultured in ESAW without antibiotic supplementation for 7 or 10 days under constant 100 µE m^{-2} s^{-1} light in 21 °C or 12:12 hr photoperiod. Cell counts were obtained by sampling 200 µL of culture over the growth period and analysing them by flow cytometry using CytoFLEX S (Beckman Coulter). Chlorophyll fluorescence was detected using 690/50 nm filter. Compensation of chlorophyll channel was set to 0.3.

Cloning and genetic construct assembly

The *Catharanthus roseus geraniol synthase (CrGES)* and *Abies grandis geranyl diphosphate (AgGPPS2)* genes were codon optimised (https://github.com/rafabbriano/Diatom_tools/blob/master/Pt_codon_optimize_v.2.ipynb) and synthesised with a flexible linker separating them by Twist Bioscience. Episomes were constructed using Gibson assembly cloning kit (New England Biolabs, Hitchin, UK) and primers described in Supplementary Table 1. All episomes were propagated in *E. coli* strain Top10 and purified by Monarch Plasmid Miniprep Kit (New England Biolabs, Hitchin, UK). PCR amplification was performed using Q5 high fidelity polymerase (New England Biolabs, Hitchin, UK) and PCR screening was performed using GoTaq Flexi DNA polymerase (Promega, Wisconsin, United States) according to

the manufacturer's instructions. Plasmid coding sequences were validated by Sanger sequencing (Macrogen Korea, Seoul, Korea).

Diatom conjugation

E. coli containing *pTA-Mob* (Karas et al., 2015) and the episome of interest were used for conjugation with *P. tricornutum* according to protocol described by Diner et al. (2016). The complete list of episomes used are described in Supplementary Table 2. The cell mixture was scraped and plated onto 3-5 fresh ½ ESAW zeocin agar plates and left for 10-15 days when single colonies appeared. Single colonies generated by conjugation were picked and inoculated into individual wells of 96-well round bottom plates containing 200 µL of ESAW supplemented with 50 µg/mL zeocin. The EE generated cell lines were incubated at 21 °C with 100 µE m^{-2} s^{-1} light for 1 week to adjust to liquid growth, after which they were subcultured every 4 days.

Cell count analysis

Induced cells were screened by flow cytometry using CytoFLEX S (Beckman Coulter).

Geraniol capture and analysis

Geraniol was captured by the addition of 1.6 mL of isopropyl myristate (IM; Sigma Aldrich, Australia) to 50 mL of diatom culture at the time of inoculation. This ratio was upscaled accordingly for sterol analysis experiments (3.4 mL IM in 100 mL culture). In all experiments, the IM layer was harvested by centrifugation for 10 minutes at 4,500 g at the end of the cultivation, diluted with ethylacetate (1:3) and directly analysed by GC-MS. To extract geraniol and any other monoterpenoids from *P. tricornutum* biomass, 50 mL of diatom cultures were harvested by centrifugation at 4,500 g for 5 minutes, washed with 2 mL PBS and pellets snap frozen in liquid nitrogen. Pellets were freeze-dried and extracted using 500 µL of ethyl acetate:hexane (1:1) and

homogenised for 1 min by bead beating (Next Advance, Troy, USA). The mixture was stored at 4°C for 1h, centrifuged at 4°C for 10 min (10000 rpm), supernatant collected and analysed directly for geraniol and other monoterpenoids. Samples were run on a GC 2010 (Shimadzu Corporation, Kyoto Japan) equipped with an AOC-20is autosampler (Shimadzu Corporation). The column used was an SH-Rxi-5Sil MS fused silica capillary column (30.0 m x 0.25 mm x 0.25 µm). The carrier gas was Helium, used at a constant flow of 1.0 mL/min and an injection volume of 1 µL. The temperature gradient of the oven was 70ºC for 1 min, then sequentially increased at the rate of 30ºC per minute to 200ºC and then further to 320ºC at 40ºC per minute. The GC-MS-QP2020 (Shimadzu Corporation, Kyoto Japan) operating in electron impact mode at 70 eV was used. The injector temperature was maintained at 280ºC and ion source temperature at 230°C. Quantitative analysis of geraniol was run in Selective Ion Monitoring (SIM) mode, where selected ions were monitored for 20 ms each. The ions monitored were m/z 121, 136 and 154. The peak areas were converted into concentrations in comparison with calibration curves plotted against a range of known concentrations of geraniol standard (Sigma Aldrich, Australia). MS was run in full scan mode for monoterpenoid analysis, with a mass range of m/z 50 to 600. Peaks were monitored by matching their mass spectra with those of the NIST17 mass spectral database.

Pigment analysis

Diatom cultures (50 mL) were harvested by centrifugation at 4,500 g for 5 minutes, washed with 2 mL PBS and pellets snap frozen in liquid nitrogen. Pellets were freeze-dried and subsequently extracted in the dark with 1.5 mL acetone (cold) in an amber coloured glass vial. Samples were vortexed 3 times for 30 sec, with 1 minute interval in between during which the temperature was maintained at 4°C and stored at -20ºC

overnight. Pigment extracts were then filtered through 0.2 µm PTFE 13 mm syringe filters and stored in -80°C until analysis. An Agilent 1290 HPLC system equipped with a binary pump with integrated vacuum degasser, thermostated column compartment modules, Infinity 1290 autosampler and PDA detector was used for the analysis. Column separation was performed using an Agilent's Zorbax Eclipse XDB C8 HPLC 4.6 mm × 150 mm and guard column using a gradient of TBAA: Methanol mix (30:70) (solvent A) and Methanol (Solvent B) as follows: 0–22 min, from 5 to 95% B; 22–29 min, 95% B; 29-31 min, 5% B; 31-40 min, column equilibration with 5% B. Column temperature was maintained at 55°C. A complete pigment spectrum from 270 to 700 nm was recorded using PDA detector with 3.4 nm bandwidth. Calibration was performed using individual pigment standards which were purchased from DHI (Denmark).

Sterols analysis

After 48 hours of growth, biomass was harvested by centrifuging at 4000 g for 10 minutes. Diatom pellets were washed with DMSO to eliminate salt excess, freeze-dried to determine dry matter weight, and kept at −80°C until sterol extraction. For sterol extraction, dry cell matter was heated in 1 mL of 10% KOH ethanolic solution at 90°C for one hour. Sterols were extracted from cooled material in three volumes of 400 µL of hexane. 0.4 µg 5a-cholestane was added to each sample as internal standard. Hexane fractions were dried under a gentle N_2 stream, and derivatized with 50 µL of 99% BSTFA + 1% TMCS at 70°C for one hour. The resulting extractions were resuspended in 50 µL of fresh hexane prior to GC-MS injection. Gas chromatography/mass spectrometry (GC-MS) analysis was performed using a GC-MS-QP2020 (Shimadzu Corporation, Kyoto Japan) equipped with an AOC-20is autosampler (Shimadzu Corporation). The column used was an SH-Rxi-5Sil MS fused

silica capillary column (30.0 m x 0.25 mm x 0.25 μm) operating in electron impact mode at 70 eV. The following settings were used: oven temperature initially set to 50°C, with a gradient from 50°C to 250°C (15.0°C min^{-1}), and then from 250°C to 310°C (8°C min^{-1}, hold 10 min); injector temperature = 250°C; carrier gas helium flow = 0.9 mL min^{-1}. A split-less mode of injection was used, with a purge time of 1 min and an injection volume of 2 μL. Mass spectrometer operating conditions were as follows: injector temperature of 280°C and an ion source temperature of 230°C. The scan range was m/z 50-650. Sterol peaks were identified based on retention time, mass spectrum, and representative fragment ions compared to the retention times and mass spectrum of authentic standards. The NIST (National Institute of Standards and Technology) library was also used as reference. The area of the peaks and deconvolution analysis was carried out using the default settings of the Automated Mass Spectral Deconvolution and Identification System AMDIS software (v2.6, NIST). Peak area measurements were normalized by both the weight of dry matter prior to extraction, and the within-sample peak area of the internal standard 5a-cholestane. Sterol standards used to calibrate and identify GC-MS results in this study included: cholest-5-en-3-β-ol (cholesterol); (22E)-stigmasta-5,22-dien-3β-ol (stigmasterol); stigmast-5-en-3-β-ol (sitosterol); campest-5-en-3-β-ol (campesterol); (22E)-ergosta-5,22-dien-3-β-ol (brassicasterol); (24E)-stigmasta-5,24-dien-3β-ol (fucosterol); 9,19-Cyclo-24-lanosten-3β-ol (cycloartenol);, 5-α-cholestane; and the derivatization reagent bis(trimethyl-silyl) trifluoroacetamide and trimethylchlorosilane (99% BSTFA + 1% TMCS) and were obtained from Sigma-Aldrich, Australia.

SUPPLEMENTARY TABLES

Suppl. Table 1. Oligonucleotide primers utilised in this study. GA, Gibson Assembly primer

Primer ID	Sequence 5'-3'	Description
MF864	gagcagactctagagtcgacctgcaAACGGTACGTCAGATCCC	GA - 49p (Phatr3_J49202) promoter-Fwd [SbfI] (pPTBR11 3')
MF865	cgtagcggccatATTGGTGGTGCGGAAAGAG	GA - 49p (Phatr3_J49202) promoter-Rv (CrGES 5')
MF866	ccgcaccaccaatATGGCCGCTACGATCAGTAAC	GA - CrGES-Fwd (49p 3')
MF909	gatcgaatcaGAAGCAGGGCGTGAAAAAC	GA - CrGES-Rv (49t 5')
MF910	gccctgcttctgaTTCGATCGACGAGCTTAC	GA - 49t (Phatr3_J49202) terminator-Fwd (CrGES 3')
MF869	cgaggaagcggaagagatgcctgcaATCGCATCGACGGTCTTC	GA - 49p (Phatr3_J49202) terminator-Rv [SbfI] (pPTBR11 5')
MF867	tcgatcgaatcaTGAGCCGTTTTGACGAAAAG	GA - CrGES-AgGPPS2-Rv (49t 5')
MF868	tcaaaacggctcatgaTTCGATCGACGAGCTTAC	GA - 49t (Phatr3_J49202) terminator-Fwd (CrGES-AgGPPS2 3')

Suppl. Table 2. List of episomes utilised in this study.

Episome ID	Assembly
pPtPBR11_49202p-mVenus	Primers used described in Fabris et al. 2020
pPtPBR11_49202p_CrGES-mVenus	Primers used described in Fabris et al. 2020
pPtPBR11_21659p_mVenus	Primers used described in Fabris et al. 2020
pPtPBR11_21659p_CrGES-mVenus	Primers used described in Fabris et al. 2020
pPtPBR11_49202p_CrGES	Primers used described in Table S1
pPtPBR11_49202p_CrGES-AgGPPS2	Primers used described in Table S1

REFERENCES

Alper, H., Jin, Y. S., Moxley, J. F., & Stephanopoulos, G. (2005). Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metabolic Engineering*, *7*(3), 155–164. https://doi.org/10.1016/j.ymben.2004.12.003

Asadollahi, M. A., Maury, J., Patil, K. R., Schalk, M., Clark, A., & Nielsen, J. (2009). Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. *Metabolic Engineering*, *11*(6), 328–334. https://doi.org/10.1016/j.ymben.2009.07.001

Athanasakoglou, A., Grypioti, E., Michailidou, S., Ignea, C., Makris, A. M., Kalantidis, K., … Kampranis, S. C. (2018). Isoprenoid biosynthesis in the diatom *Haslea ostrearia*. *New Phytologist*. https://doi.org/10.1111/nph.15586

Athanasakoglou, A., & Kampranis, S. C. (2019). Diatom isoprenoids: Advances and biotechnological potential. *Biotechnology Advances*, *37*(8). https://doi.org/10.1016/j.biotechadv.2019.107417

Berges, J. A., Franklin, D. J., & Harrison, P. J. (2001). Evolution of an artificial seawater medium: Improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology*, *37*(6), 1138–1145. https://doi.org/10.1046/j.1529-8817.2001.01052.x

Brennan, T. C. R., Turner, C. D., Krömer, J. O., & Nielsen, L. K. (2012). Alleviating monoterpene toxicity using a two-phase extractive fermentation for the bioproduction of jet fuel mixtures in Saccharomyces cerevisiae. *Biotechnology and Bioengineering*, *109*(10), 2513–2522. https://doi.org/10.1002/bit.24536

Carqueijeiro, I., Brown, S., Chung, K., Dang, T. T., Walia, M., Besseau, S., … Courdavault, V. (2018). Two tabersonine 6,7-epoxidases initiate lochnericine-derived alkaloid biosynthesis in catharanthus roseus. *Plant Physiology*, *177*(4), 1473–1486. https://doi.org/10.1104/pp.18.00549

Chemler, J., Yan, Y., & Koffas, M. (2006). Biosynthesis of isoprenoids, polyunsaturated fatty acids and flavonoids in Saccharomyces cerevisiae. *Microbial Cell Factories*, *5*(1), 20. https://doi.org/10.1186/1475-2859-5-20

Chen, W., & Viljoen, A. M. (2010). Geraniol - A review of a commercially important fragrance material. *South African Journal of Botany*, *76*(4), 643–651. https://doi.org/10.1016/j.sajb.2010.05.008

Cvejic, J. H., & Rohmer, M. (2000). CO 2 as main carbon source for isoprenoid biosynthesis via the mevalonate-independent methylerythritol 4-phosphate route in the marine diatoms Phaeodactylum tricornutum and Nitzschia ovalis. *Phytochemistry*, *53*, 21–28.

D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., … Matthijs, M. (2018). Engineering The Unicellular Alga Phaeodactylum tricornutum For High-Value Plant Triterpenoid Production. *Plant Biotechnology Journal*, 0–2. https://doi.org/10.1111/pbi.12948

De Luca, V., Salim, V., Thamm, A., Masada, S. A., & Yu, F. (2014). Making iridoids/secoiridoids and monoterpenoid indole alkaloids: Progress on pathway elucidation. *Current Opinion in Plant Biology*, *19*, 35–42. https://doi.org/10.1016/j.pbi.2014.03.006

Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E., & Weyman, P. D. (2016). Refinement of the Diatom Episome Maintenance Sequence and Improvement of

Conjugation-Based DNA Delivery Methods. *Frontiers in Bioengineering and Biotechnology*, *4*(August). https://doi.org/10.3389/fbioe.2016.00065

Dudareva, N., Negre, F., Nagegowda, D. A., & Orlova, I. (2006). Plant volatiles: Recent advances and future perspectives. *Critical Reviews in Plant Sciences*, *25*(5), 417–440. https://doi.org/10.1080/07352680600899973

Dufourc, E. J. (2008). Sterols and membrane dynamics. *Journal of Chemical Biology*, *1*(1–4), 63–77. https://doi.org/10.1007/s12154-008-0010-6

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., … Ralph, P. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.9b00455

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., … Goossens, A. (2014). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *The New Phytologist*, 521–535. https://doi.org/10.1111/nph.12917

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., & Baart, G. J. E. (2012). The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant Journal*, *70*(6), 1004–1014. https://doi.org/10.1111/j.1365-313X.2012.04941.x

Fischer, M. J. C., Meyer, S., Claudel, P., Bergdoll, M., & Karst, F. (2011). Metabolic engineering of monoterpene synthesis in yeast. *Biotechnology and Bioengineering*, *108*(8), 1883–1892. https://doi.org/10.1002/bit.23129

Forster, J., Famili, I., Fu, P., Palsson, B., & Jens, N. (2003). Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Research*, 244–253. https://doi.org/10.1101/gr.234503.complex

George, J., Kahlke, T., Abbriano, R. M., Kuzhiumparambil, U., Ralp, P. J., & Fabris, M. (2020). Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology. *Frontiers in Bioengineering and Biotechnology*, *8*(June), 1–19. https://doi.org/10.3389/fbioe.2020.00513

Hebert Jair Barrales-Cureño, César Reyes Reyes, Irma Vásquez García, Luis Germán López Valdez, Adrián Gómez De Jesús, Juan Antonio Cortés Ruíz, Leticia Mónica Sánchez Herrera, María Carmina Calderón Caballero, Jesús Antonio Salazar Magallón, J. E. P. and J. M. M. (2012). Alkaloids of Pharmacological Importance in Catharanthus roseus. *Intech*, 13. https://doi.org/10.1016/j.colsurfa.2011.12.014

Hildebrand, M., Abbriano, R. M., Polle, J. E. W., Traller, J. C., Trentacoste, E. M., Smith, S. R., & Davis, A. K. (2013). Metabolic and cellular organization in evolutionarily diverse microalgae as related to biofuels production. *Current Opinion in Chemical Biology*, *17*(3), 506–514. https://doi.org/10.1016/j.cbpa.2013.02.027

Ignea, C., Pontini, M., Maffei, M. E., Makris, A. M., & Kampranis, S. C. (2014). Engineering monoterpene production in yeast using a synthetic dominant negative geranyl diphosphate synthase. *ACS Synthetic Biology*, *3*(5), 298–306. https://doi.org/10.1021/sb400115e

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., … Yuan, Y. J. (2017). Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metabolic Engineering*, *41*(March), 57–66. https://doi.org/10.1016/j.ymben.2017.03.005

Karas, B. J., Diner, R. E., Lefebvre, S. C., McQuaid, J., Phillips, A. P. R., Noddings, C. M., … Weyman, P. D. (2015). Designer diatom episomes delivered by bacterial conjugation. *Nature Communications*, *6*, 6925. https://doi.org/10.1038/ncomms7925

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, *12*(6). https://doi.org/10.1371/journal.pbio.1001889

King, J. R., Woolston, B. M., & Stephanopoulos, G. (2017). Designing a New Entry Point into Isoprenoid Metabolism by Exploiting Fructose-6-Phosphate Aldolase Side Reactivity of Escherichia coli. *ACS Synthetic Biology*, *6*(7), 1416–1426. https://doi.org/10.1021/acssynbio.7b00072

Kitano, H. (2002). Systems biology: A brief overview. *Science*, *295*(5560), 1662–1664. https://doi.org/10.1126/science.1069492

Kuczynska, P., Jemiola-Rzeminska, M., & Strzalka, K. (2015). Photosynthetic pigments in diatoms. *Marine Drugs*, *13*(9), 5847–5881. https://doi.org/10.3390/md13095847

Lauersen, K. J., Baier, T., Wichmann, J., Wördenweber, R., Mussgnug, J. H., Hübner, W., … Kruse, O. (2016). Efficient phototrophic production of a high-value sesquiterpenoid from the eukaryotic microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *38*, 331–343. https://doi.org/10.1016/j.ymben.2016.07.013

Lauersen, K. J., Wichmann, J., Baier, T., Kampranis, S. C., Pateraki, I., Møller, B. L., & Kruse, O. (2018). Phototrophic production of heterologous diterpenoids and a hydroxy-functionalized derivative from Chlamydomonas reinhardtii. *Metabolic Engineering*, *49*. https://doi.org/10.1016/j.ymben.2018.07.005

Lin, P. C., Saha, R., Zhang, F., & Pakrasi, H. B. (2017). Metabolic engineering of the pentose phosphate pathway for enhanced limonene production in the cyanobacterium Synechocysti s sp. PCC. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-17831-y

Liu, J., Zhang, W., Du, G., Chen, J., & Zhou, J. (2013). Overproduction of geraniol by enhanced precursor supply in Saccharomyces cerevisiae. *Journal of Biotechnology*, *168*(4), 446–451. https://doi.org/10.1016/j.jbiotec.2013.10.017

Liu, W., Xu, X., Zhang, R., Cheng, T., Cao, Y., Li, X., … Xian, M. (2016). Engineering Escherichia coli for high-yield geraniol production with biotransformation of geranyl acetate to geraniol under fed-batch culture. *Biotechnology for Biofuels*, *9*(1), 1–8. https://doi.org/10.1186/s13068-016-0466-5

Lohr, M., Schwender, J., & Polle, J. E. W. (2012). Isoprenoid biosynthesis in eukaryotic phototrophs: A spotlight on algae. *Plant Science*, *185–186*, 9–22. https://doi.org/10.1016/j.plantsci.2011.07.018

Maury, J., Asadollahi, M. A., Møller, K., Schalk, M., Clark, A., Formenti, L. R., & Nielsen, J. (2008). Reconstruction of a bacterial isoprenoid biosynthetic pathway in Saccharomyces cerevisiae. *FEBS Letters*, *582*(29), 4032–4038. https://doi.org/10.1016/j.febslet.2008.10.045

Meadows, A. L., Hawkins, K. M., Tsegaye, Y., Antipov, E., Kim, Y., Raetz, L., … Tsong, A. E. (2016). Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature*, *537*(7622), 694–697. https://doi.org/10.1038/nature19769

Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., … Werck-Reichhart,

D. (2014). The seco-iridoid pathway from Catharanthus roseus. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4606

Nagegowda, D. A., & Gupta, P. (2020). Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant Science*, *294*(February), 110457. https://doi.org/10.1016/j.plantsci.2020.110457

Oswald, M., Fischer, M., Dirninger, N., & Karst, F. (2007). Monoterpenoid biosynthesis in Saccharomyces cerevisiae. *FEMS Yeast Research*, *7*(3), 413–421. https://doi.org/10.1111/j.1567-1364.2006.00172.x

Oudin, A., Courtois, M., Rideau, M., & Clastre, M. (2007). The iridoid pathway in Catharanthus roseus alkaloid biosynthesis. *Phytochemistry Reviews*, *6*(2–3), 259–276. https://doi.org/10.1007/s11101-006-9054-9

Oudin, A., Mahroug, S., Courdavault, V., Hervouet, N., Zelwer, C., Rodríguez-Concepción, M., … Burlat, V. (2007). Spatial distribution and hormonal regulation of gene products from methyl erythritol phosphate and monoterpene-secoiridoid pathways in Catharanthus roseus. *Plant Molecular Biology*, *65*(1–2), 13–30. https://doi.org/10.1007/s11103-007-9190-7

Owens, T. G. (1986). Light-Harvesting Function in the Diatom Phaeodactylum tricornutum . *Plant Physiology*, *80*(3), 739–746. https://doi.org/10.1104/pp.80.3.739

Pan, Q., Mustafa, N. R., Tang, K., Choi, Y. H., & Verpoorte, R. (2016). Monoterpenoid indole alkaloids biosynthesis and its regulation in Catharanthus roseus: a literature review from genes to metabolites. *Phytochemistry Reviews*, *15*(2), 221–250. https://doi.org/10.1007/s11101-015-9406-4

Pichersky, E., & Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytologist*, *220*(3), 692–702. https://doi.org/10.1111/nph.14178

Pollak, B., Matute, T., Nuñez, I., Cerda, A., Lopez, C., Vargas, V., … Federici, F. (2019). Universal Loop assembly (uLoop): open, efficient, and species-agnostic DNA fabrication. *BioRxiv*. https://doi.org/10.1101/744854

Pollier, J., Vancaester, E., Kuzhiumparambil, U., Vickers, C. E., Vandepoele, K., Goossens, A., & Fabris, M. (2019). A widespread alternative squalene epoxidase participates in eukaryote steroid biosynthesis. *Nature Microbiology*, *4*(2), 226–233. https://doi.org/10.1038/s41564-018-0305-5

Qian, S., Clomburg, J. M., & Gonzalez, R. (2019). Engineering Escherichia coli as a platform for the in vivo synthesis of prenylated aromatics. *Biotechnology and Bioengineering*, *116*(5), 1116–1127. https://doi.org/10.1002/bit.26932

Rai, A., Smita, S. S., Singh, A. K., Shanker, K., & Nagegowda, D. A. (2013). Heteromeric and homomeric geranyl diphosphate synthases from catharanthus roseus and their role in monoterpene indole alkaloid biosynthesis. *Molecular Plant*, *6*(5), 1531–1549. https://doi.org/10.1093/mp/sst058

Shah, A. A., Wang, C., Chung, Y. R., Kim, J. Y., Choi, E. S., & Kim, S. W. (2013). Enhancement of geraniol resistance of Escherichia coli by MarA overexpression. *Journal of Bioscience and Bioengineering*, *115*(3), 253–258. https://doi.org/10.1016/j.jbiosc.2012.10.009

Shah, A. A., Wang, C., Yoon, S. H., Kim, J. Y., Choi, E. S., & Kim, S. W. (2013). RecA-mediated SOS response provides a geraniol tolerance in Escherichia coli. *Journal of Biotechnology*, *167*(4), 357–364. https://doi.org/10.1016/j.jbiotec.2013.07.023

Simkin, A. J., Miettinen, K., Claudel, P., Burlat, V., Guirimand, G., Courdavault, V., …

Clastre, M. (2013). Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. *Phytochemistry*, *85*, 36–43. https://doi.org/10.1016/j.phytochem.2012.09.014

Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H., … Moritz, T. (2016). Transcriptional Orchestration of the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron Limitation. *PLOS Genetics*, *12*(12), e1006490. https://doi.org/10.1371/journal.pgen.1006490

Thabet, I., Grégory Guirimand, G., Guihur, A., Lanoue, A., Courdavault, V., Papon, N., … Clastre, M. (2012). Characterization and subcellular localization of geranylgeranyl diphosphate synthase from Catharanthus roseus. *Molecular Biology Reports*, *39*(3), 3235–3243. https://doi.org/10.10.1007/s11033-011-1091-9

Vavitsas, K., Fabris, M., & Vickers, C. E. (2018). Terpenoid Metabolic Engineering in, (Figure 1). https://doi.org/10.3390/genes9110520

Vickers, C. E., Bongers, M., Liu, Q., Delatte, T., & Bouwmeester, H. (2014). Metabolic engineering of volatile isoprenoids in plants and microbes. *Plant, Cell and Environment*, *37*(8), 1753–1775. https://doi.org/10.1111/pce.12316

Wang, C., Kim, J.-H., & Kim, S.-W. (2014). Synthetic Biology and Metabolic Engineering for Marine Carotenoids: New Opportunities and Future Prospects. *Marine Drugs*, *12*(9), 4810–4832. https://doi.org/10.3390/md12094810

Wang, C. T., Liu, H., Gao, X. S., & Zhang, H. X. (2010). Overexpression of G10H and ORCA3 in the hairy roots of Catharanthus roseus improves catharanthine production. *Plant Cell Reports*, *29*(8), 887–894. https://doi.org/10.1007/s00299-010-0874-0

Wang, C., Yoon, S. H., Shah, A. A., Chung, Y. R., Kim, J. Y., Choi, E. S., … Kim, S. W. (2010). Farnesol production from Escherichia coli by harnessing the exogenous mevalonate pathway. *Biotechnology and Bioengineering*, *107*(3), 421–429. https://doi.org/10.1002/bit.22831

Ward, V. C. A., Chatzivasileiou, A. O., & Stephanopoulos, G. (2018). Metabolic engineering of Escherichia coli for the production of isoprenoids. *FEMS Microbiology Letters*, *365*(10), 1–9. https://doi.org/10.1093/femsle/fny079

Yamamoto, K., Takahashi, K., Mizuno, H., Anegawa, A., Ishizaki, K., Fukaki, H., … Mimura, T. (2016). Cell-specific localization of alkaloids in Catharanthus roseus stem tissue measured with Imaging MS and Single-cell MS. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(14), 3891–3896. https://doi.org/10.1073/pnas.1521959113

Zebec, Z., Wilkes, J., Jervis, A. J., Scrutton, N. S., Takano, E., & Breitling, R. (2016). Towards synthesis of monoterpenes and derivatives using synthetic biology. *Current Opinion in Chemical Biology*, *34*, 37–43. https://doi.org/10.1016/j.cbpa.2016.06.002

Zhao, J., Bao, X., Li, C., Shen, Y., & Hou, J. (2016). Improving monoterpene geraniol production through geranyl diphosphate synthesis regulation in Saccharomyces cerevisiae. *Applied Microbiology and Biotechnology*, *100*(10), 4561–4571. https://doi.org/10.1007/s00253-016-7375-1

Zhou, J., Wang, C., Yang, L., Choi, E. S., & Kim, S. W. (2015). Geranyl diphosphate synthase: An important regulation point in balancing a recombinant monoterpene pathway in Escherichia coli. *Enzyme and Microbial Technology*, *68*, 50–55. https://doi.org/10.1016/j.enzmictec.2014.10.005

General conclusion

Exploring genetic engineering strategies to enable heterologous monoterpenoid production in model microalgae, *Chlamydomonas reinhardtii* and *Phaeodactylum tricornutum*

GENERAL CONCLUSION

The overarching aim of this thesis was to investigate and characterise genetic engineering strategies for heterologous production of the monoterpenoid geraniol in the pennate diatom *P. tricornutum.* This work lays foundations for more complex synthetic biology ventures in diatoms not only for monoterpenoid production. The impact of this thesis is therefore broad, including characterisation of the genetic engineering technologies required for the production of monoterpenoids in diatoms and ventures beyond terpenoid engineering. First, this work offers new approaches for high-throughput phenotyping large libraries of transformants at the protein expression level and transgenome characterisation via third generation long-read whole-genome sequencing with implications for diatom genetic engineering. Second, this thesis provides new information about the widely used method of randomly integration chromosomal expression (RICE) and the chemical mutagenic effect of a recently demonstrated selectable agent 5-fluoorotic acid (5-FOA) which has relevance for diatom genetics and functional genetics research. Third, it highlights challenges and new directions for genome editing, namely by providing the first putative safe harbour loci for targeted integration and critical reflection of CRISPR-Cas9 editing technology in the microalgal context. Finally, this thesis provides evidence for the possibility of using diatoms for heterologous terpenoid production, highlighting the need to better understand native diatom terpenoid biosynthesis.

Phenotypic characterisation of first and next generation genetic engineering in *P. tricornutum*

In Chapter 3, we engineered *P. tricornutum* for the heterologous production of the monoterpenoid geraniol, a key precursor of monoterpenoid indole alkaloids (MIAs) that

include high-demand pharmaceuticals, such as vinblastine and vincristine, by using both the newly developed episome-based extrachromosomal expression (EE) and well-established randomly integrated chromosomal expression (RICE) approaches. Following expression of a *Catharanthus roseus* geraniol synthase enzyme fused to a fluorescent reporter mVenus (CrGES-mVenus), we developed an efficient high-throughput phenotyping approach to screen hundreds of colonies by flow cytometry for fast and accurate fluorescence detection, which was used as a proxy for *CrGES-mVenus* expression.

This revealed that randomly integrated chromosomal expression of *CrGES-mVenus* resulted in highly dissimilar phenotypes between cell lines, whereby approximately one quarter showed no expression and only 1% showed maximum expression of approximately 650-fold that of WT auto fluorescence. Similar dissimilarities between transformant cell lines were also seen in an mVenus only control, strongly suggesting that this disparity between mutants was the result of RICE and not related to the heterologous production of geraniol. Our high throughout phenotyping screen could also be used to compare cells within a clonal population, and showed that RICE can result in a clonal population exhibiting normally distributed mVenus fluorescence, as well as bimodal and even trimodal distributions. Finally, this analysis revealed that some RICE cell lines were able to maintain stable expression of transgenes even without selective pressure over a three week cultivation period. Interestingly, the mean mVenus fluorescence of a clonal population did not correlate with the distribution within the population nor with expression stability. Altogether, these findings offer the first quantitative demonstration of ‘position effect’ associated with RICE, the first generation of genetic engineering in diatoms which has been widely relied upon for decades and is still widely used today.

In almost complete opposition to the RICE phenotype, our analysis showed that 100% of EE exconjugants expressing *mVenus* and *mVenus* fused to *CrGES (CrGES-mVenus)* demonstrated mVenus fluorescence significantly higher than auto fluorescence of WT cells. This is a substantial improvement over RICE, which we showed can result in up to a quarter of transformants exhibiting extremely low to no transgene expression. Furthermore, EE resulted in highly consistent mVenus fluorescence of approximately 250-fold that of WT auto fluorescence, even when expressed as a fusion protein with CrGES. These results are the first quantitative demonstration that EE is not subject to positional effects, which is important for the synthetic biology design-build-test-learn cycle and validating genetic parts reliably.

Interestingly, we also detected that the mVenus fluorescence within a single EE clone was not distributed as discreetly as in RICE, but instead showed broad variegation. This highlights the need for more work to be done to uncover the particularly limited knowledge regarding episome stability, copy number and segregation patterns, and re-arrangement in diatoms, which we did not observe but has been reported (Slattery et al., 2018). Such characterisations will help to fully exploit EE as a synthetic biology platform in diatoms, particularly as invaluable resource for genetic parts validation and modular assembly, and even automation of the design-build-test-learn cycle. These aspects of more complex synthetic biology strategies are crucial for heterologous production of high-value products such as monoterpenoids.

Heterologous production of geraniol in *P. tricornutum*

In Chapter 3, we showed that randomly integrated chromosomal expression of CrGES-mVenus resulted in 9.4-fold higher mVenus fluorescence and 4.2-fold increased geraniol production than extrachromosomal expression of CrGES-mVenus

(George et al., 2020). The RICE cell line reaching a maximum titre of 304.4 ng/10^7 cells, equivalent to 0.89 mg/L with no toxic effects, such as reduced growth; whereas the top performing EE cell line produced only 72.5 ng/10^7 cells, equivalent to 0.15 mg/L (George et al., 2020). Herein, we also showed that increasing mVenus fluoresce demonstrated by three superior mVenus fluorescing RICE cell lines correlated with increased geraniol accumulation, suggesting that even at the highest geraniol titre (304.4 ng/10^7 cells), we still had not reached the threshold of GPP substrate consumption, or CrGES or geraniol accumulation. In Chapter 5, we showed that extrachromosomal expression of CrGES-mVenus be improved up to 210 ng/10^7 cells, which corresponds to 0.31 mg/L using an uncharacterised constitutive promoter driving *Phatr3_J49202* (Pollak et al., 2019; Fabris et al., 2020). Taken together, these results strongly show that CrGES availability and/or access to GPP is a more limiting factor in *P. tricornutum's* capacity for heterologous geraniol production than GPP availability.

This is particularly useful when comparing *P. tricornutum* to the more widely used workhorses, *E. coli* and *S. cerevisiae*, as both of these organisms use an FPP synthase that converts DPP and IPP into GPP and then immediately into FPP, leaving a very small free pool of GPP to escape for heterologous production of monoterpenoids. Consequently, highly complex metabolic engineering is required to increase flux through GPP synthesis in these non-algal species (Brown et al., 2015; Dziggel et al., 2017; Willrodt et al., 2014). It also highlights the unusual biochemistry and metabolism of this species, and more work should be done to better elucidate GPP synthesis and its fate in the cell. For example, various diatom species have been recorded to produce and emit monoterpenoids naturally, however, the biochemistry

involved with diatom monoterpenoid metabolism is poorly understood (Shaw et al., 2010; Yassaa et al., 2008).

Therefore, in order to increase GPP substrate availability and utilisation by CrGES, we expressed two adjacent enzymes fused to one another: geranyl diphosphate synthase (AgGPPS2) (AF513112; EC 2.5.1.1) which produces GPP from IPP and/or DMAPP in the plant *Abies grandis,* and geraniol synthase (CrGES) (Caros003727.1; EC 3.1.7.11) which converts GPP into geraniol in the plant *Catharanthus roseus*. This resulted in 55.3% decreased geraniol accumulation compared to a control cell line that did not contain AgGPPS2, most likely due to improper folding of the fused exogenous enzymes. Then, we showed that a dark:light (12h:12h) photoperiod had no impact on geraniol production compared to continuous light cultivation. Similarly, we reported no impact geraniol production following an elongated cultivation period extending beyond stationary phase. This was not surprising, as primary cellular metabolism–of which MVA and MEP pathways are a part–is more transcriptionally active during the exponential phase of *P. tricornutum* (Smith et al., 2016). These results correlate with the productivity profile reported in the heterologous production of betulinic acid in *P. tricornutum* (D'Adamo et al., 2018), and are in contrast with heterologous production of terpenoids in *C. reinhardtii* (Lauersen *et al.*, 2018a).

The significantly higher geraniol yields achieved by RICE over EE also indicate that there must be something unique to chromosomal integration that facilitates high transgene expression. This is also particularly important at a time when next generation tools are being developed; namely extrachromosomal expression and targeted genomic integration (TGI) by programmable endonucleases.

A move towards precision editing and CRISPR-Cas9

CRISPR-Cas9 is a programmable endonuclease originating from bacterial immunity which has been demonstrated to be a highly effective genome editing tool for TGI. Beyond genomic DNA editing, CRISPR has been adapted for epigenetic modification (Pflueger et al., 2018; Thakore et al., 2015), gene localisation (Chen et al., 2013; Roberts et al., 2017), genome-wide screening (Chen et al., 2015; Shalem, 2014), regulating gene circuits in synthetic biology (Kiani et al., 2014), and potential for *in vivo* gene therapy (Xue et al., 2016), making it a revolutionary discovery. Consequently, CRISPR technology has had significant impact on molecular research of most model species, accelerating discoveries in industrial biotechnology, plant breeding, and disease investigation and treatment (Sternberg & Doudna, 2015; Stovicek et al., 2017).

Following the phenotyping analyses conducted in Chapter 3, we hypothesised that genetic engineering in *P. tricornutum* would ideally combine the reproducibility of EE with the high expression achievable through RICE, which could be achieved by targeted genomic integration (TGI). Early examples of TGI have been demonstrated by delivering donor DNA which contains flanks that align to the site of integration and is driven by the natural DNA repair process, homology driven repair (HDR). HDR occurs at very low frequency and is often considered too low to be feasible in many organisms. However, HDR mediated integration is much more likely to occur when a double stranded break is present and numerous reports have now shown that endonuclease-driven TGI is possible. Here, an endonuclease programmed to target the integration location of interest is co-delivered with the donor DNA to drive integration. Recent developments in endonuclease technology have made TGI more feasible and have mostly been validated using TALENs, Zinc Fingers and CRISPR

Cas9. Successful TGI examples are highly applicable for synthetic biology and have repeatedly been demonstrated in yeast. For example, endonuclease-driven TGI has allowed the introduction of the carotenoid pathway using 15 DNA parts integrated at three targeted chromosomal locations and a strain producing tyrosine using 10 parts integrated at two loci using CRISPR Cas9 (Jakočiunas et al., 2015).

TGI has only recently been demonstrated in the model microalgae *C. reinhardtii* and *P. tricornutum* following the co-delivery of a programmable endonuclease able to generate a double stranded break at the target site. TGI can be used for insertional KO, which aims to integrate selectable marker DNA into a gene of interest to disrupt it, or targeted KI, which aims to integrate transgenes of interest into reliable safe harbour loci or 'neutral sites'. Neutral sites in cyanobacteria have been used for targeted integration in metabolic engineering for multigene pathway assembly (Bentley et al., 2014) and dual knock-in knock-out modifications (Li et al., 2016). Synthetic "landing pads" are useful for gene stacking via "domino cloning", but also depend on the knowledge of robust, reliable safe harbour loci prior to being feasibly applied to diatoms (Karas et al., 2015b). However, there are no putative safe harbours known for *C. reinhardtii*, and until our work in Chapter 3, there were none for *P. tricornutum* either.

Given the lack of knowledge regarding RICE mechanisms and genomic outputs of RICE and the lack of any diatom safe harbour loci, we used Oxford Nanopore third generation whole genome sequencing to interrogate the superior, high mVenus fluorescing and geraniol yielding biolistic-bombarded transgenic *P. tricornutum* cell lines generated in Chapter 3. Herein, we aimed to explore integrated transgene arrangements, integration locations, and associated genetic architecture, as has been

recently done in *Arabidopsis thaliana* and mouse models (Jupe et al., 2019; Nicholls et al., 2019) but never before for any microalga.

We reported that these superior diatom cell lines bared highly concatenated arrangements of exogenous DNA–hundreds of megabases in length–present as vast islands within or nearby predicted protein-coding genes (George et al., 2020). These findings raised a concern about this widespread method of generating transgenic diatom cell lines, as disrupting numerous protein coding regions can introduce unknown changes to *P. tricornutum* physiology that may not be easily detected. This is a particularly relevant issue in functional genetics studies involving overexpression, knock-in, knock-down or knock-out constructs, which are traditionally delivered by biolistics, and randomly integrated in the genome of diatoms. On the contrary, it is not yet known if such large, highly concatenated integration events might be a factor in transgene stability and expression. In such scenario, RICE via biolistic bombardment, might be preferable over EE for obtaining high expressing cell lines. Furthermore, although it has been shown that high copy number and transgene tandem repeats can cause transcriptional silencing of transgene cassettes in other organisms (Kaufman et al., 2008; Moritz et al., 2016), our findings highlight the need to explore copy number and transgene arrangement optimisation in more detail, as this may well not be the case in *P. tricornutum*. This study provided for the first time the direct genetic evidence that random chromosomal integration–routinely used for functional genetics and biotechnology for more than two decades in diatom research–can result in massive insertion islands and extremely complex, concatenated genetic re-arrangements in the genome, and that these can result in particularly productive phenotypes (George et al., 2020).

Whilst it is generally accepted that exogenous DNA delivered by biolistic bombardment randomly integrates in diatom chromosomes, this work also suggested that the implications of this may have previously been overlooked, particularly at a time when CRISPR-Cas9 technology is being developed. There is a general concern in CRISPR research to monitor and prevent off-target cutting by CRISPR-Cas9 itself; however, our results demonstrate that off-target effects from random integration of exogenous constructs such as vector backbone and DNA-encoded CRISPR-Cas9 components, could be just as much cause for concern. In this way, generating a precise knock-in or knock-out genotype by randomly integrating CRISPR-Cas9 components is suboptimal. As suggested by other works (Sharma et al., 2018; 702 Stukenberg et al., 2018), our findings clearly demonstrate the need to move towards non-integrative alternatives, such as episomal expression (Slattery et al., 2018, 2020) and ribonucleoprotein delivery (Serif et al., 2018). Therefore, we also interrogated an EE exconjugant. Our results demonstrated that self-replicating episomes do not integrate in diatom genome and produce highly consistent phenotypes, offering an ideal platform for synthetic biology, as they do not have random integration events which could cause unpredictable disruptions to native diatom biology.

In Chapter 3, we identified four putative safe-harbour or neutral loci that could be tested for targeted integration in *P. tricornutum*. In RICE_GmV-41, fragments of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid were inserted at two unique genomic loci, *ch1: 2,477,260* and *ch11: 316,959 – 317,016* (George et al., 2020). Both integration islands occurred at intergenic regions in the genome; however, they are both flanked by predicted protein coding genes. Neither of these islands disrupted the protein coding regions of these neighbouring genes and we did not detect any growth defective phenotypes for these cell lines. However, the close proximity of the islands

to these neighbouring genes means that the integration events may have affected their associated endogenous regulatory regions.

In transformant RICE_GmV-47, two integration events were localised to *ch9: 865,083 - 865,119* and *ch10: 609,260 - 609,276* (George et al., 2020). Both of these loci harbour predicted single-exon protein coding regions *Phatr3_J46300* and *Phatr3_J46528*, respectively, with no predicted functional annotations, nor similarity to known protein domains (Finn et al., 2011). Disruption of protein coding genes is not unusual for safe harbours, as seen in human cell lines (e.g. *CCR5* and *ROSA26* loci) and mouse cell lines (e.g. *Rosa26* locus), which all occur within protein coding regions. Furthermore, regions that may currently appear to be intergenic or non-functional may be re-categorised in the future, as more information about “junk DNA”, transcripts without function (TUFs) and unannotated regulatory regions are discovered (Gingeras, 2007). Interestingly, all four integration events were contained within unique sites across the entire genome of both cell lines, instead of occurring in a more scattered arrangement at a high number of locations, as has been demonstrated following biolistic bombardment in the plants *Oryza sativa* and *Zea mays* (Liu et al., 2018).

Third generation long-read whole genome sequencing provided a wealth of information for interrogating the superior geraniol producing cell lines and opened up new questions with relevance for developing TGI strategies in the future. This is because it is not yet apparent whether the increased geraniol productivity in RICE cell lines was due to (1) stable integration at favourable regions in the genome, known as safe harbours, or benefits related to other genomic characteristics such as chromosomal packing or epigenetic markers; (2) dramatic rearrangement of the transgenes resulting in enhanced chimeric regulatory regions; (3) a combination of these; or (4) yet to be determined factors. Should the integration location or

architecture be a significant contributor –and considering the concerns of stability– it stands that targeting these sites for TGI would be highly desirable. Therefore, in Chapter 5 we attempted to target *CrGES-mVenus* into the putative safe harbour *ch1: 2,477,660* using CRISPR-Cas9 ribonucleoprotein targeting these regions. However, we were unable to identify any knock-in mutants due to low efficiency of this technology. Future work should investigate the four putative loci described in Chapter 3, as they could offer invaluable starting point to develop engineering strategies based on targeted chromosomal integration, as demonstrated in many other organisms, but not yet in microalgae. Finally, the utility of long-read technology in mapping exogenous DNA insertion sites, as showcased in our manuscript, is of significant interest for a broad audience as it is applicable beyond microalgal research.

CRISPR-Cas9 revolution in a microalgal context: hype and reproducibility

The success of CRISPR-Cas9 editing technology across organisms and fields of molecular biology research has resulted in widespread hype and acceptance that CRISPR technology is ‘quick, easy and cheap’. While this somewhat true, it is only quick, easy and cheap when it can be robustly replicated in numerous laboratories. Given that this is not being demonstrated yet, we set out to categorise the CRISPR-Cas9 ribonucleoprotein (RNP) work flow for editing *C. reinhardtii* to better understand the complexities that have not been discussed in the literature in Chapter 2. This work flow informed the development of two new RNP-based CRISPR-Cas9 editing approaches for *C. reinhardtii* in Chapter 2: RNP-proteolistic bombardment and lipofection, neither of which had been demonstrated in any microalga at the time of this study, as well as two previously published strategies using electroporation (Baek et al., 2016; Shin et al., 2016). Over the course of this research, a new publication

demonstrated a RNP-proteolistic bombardment for *P. tricornutum* (Serif et al., 2018). Given that we had identified putative safe harbour loci in this species, we investigated CRISPR-Cas9 ribonucleoprotein (RNP) genetic engineering in both the model microalgae, *Chlamydomonas reinhardtii* in Chapter 2 and *Phaeodactylum tricornutum* in Chapter 4. We developed three optimised strategies never before reported in these organisms; namely RNP-proteolistic bombardment and lipofection-mediated delivery for gene knock-out in *C. reinhardtii*; and RNP-proteolistic bombardment for targeted gene knock-in in *P. tricornutum*. For all three of these strategies, we were unable to detect a single CRISPR-Cas9 driven mutation. Furthermore, we investigated three published strategies for CRISPR genome editing in both species, also with no success (Baek et al., 2016; Serif et al., 2018; Shin et al., 2016). It is evident that these publications have indeed generated CRISPR-Cas9 mediated mutations in these microalgal genomes, given the robust, peer reviewed evidence and sequencing data provided in each. However, it cannot be overlooked that there is a reproducibility issue, not only in our laboratory, but across the community. This is supported by two observations.

First, the lack of publications within the community following any of these key studies is extremely surprising. This is because the time taken to conduct a CRISPR microalga experiment –should these protocols be reproducible– is in the order of 3 - 5 weeks based on Shin et al. (2016), Baek et al. (2016) and Greiner et al. (2017). Given the power of CRISPR technology to drive precision genome editing to answer functional genetics questions, improve strains for biotechnology by gene-knockout or for targeted knock-in by NHEJ (which is simpler to achieve than HR driven knock-in), it is highly likely that at least a couple of studies would now be available, four years after these formative proof-of-concept publications.

Second, it is clear from our work that the enrichment and screening are crucial nodes in the CRISPR-Cas9 work flow and current studies may not have addressed these issues. Direct enrichment and screening depend on endogenous markers, whereby colonies obtained are the result of efficient CRISPR RNP delivery and editing activity, such as a discernible phenotype or resistance trait.

For example, in Chapter 2 we used RNPs targeting *C. reinhardtii Mg-protoporphyrin IX S-adenosyl methionine O-methyl transferase (ChlM*; Cre12.g498550). *ChlM* is a useful endogenous marker gene, as the ChlM enzyme converts magnesium-protoporphyrin IX into magnesium-protoporphyrin IX 13-monomethyl esterchlorophyll (EC 2.1.1.11) as one of the first steps in the chlorophyll-a biosynthesis pathway (Meinecke et al., 2010). Knocking out *ChlM* generates a low chlorophyll-a mutant, giving rise to a light green phenotype under low light conditions, easily identifiable for screening (Meinecke et al., 2010; Shin et al., 2016). We confirmed the correct assembly and activity of the RNPs we designed to target *ChlM* using *in vitro* digest and showed that cells treated with RNPs showed light green phenotype following both RNP-proteolistic bombardment and lipofection, unlike cells that were treated without the RNP. This light green, low chlorophyll fluorescence phenotype was detected both visually and by flow cytometry analysis, respectively but T7E1 analysis, RFLP analysis and Sanger sequencing confirmed that these colonies were all false positives.

Likewise, in Chapter 4, we used RNP-proteolistic bombardment to drive targeted integration of CrGES donor DNA into the putative safe harbour locus in *P. tricornutum*, *ch1: 2,477,660.* Herein, we co-delivered donor DNA and RNPs targeting both the putative safe harbour locus as well as the endogenous marker gene *uridine-5'-monophosphate synthase (UMPS; Phatr3_J11740*). UMPS has recently been shown to be a useful endogenous marker for *P. tricornutum*, as *UMPS* knock-out mutants are

tolerant to 5-fluoroorotic acid (5-FOA) and are uracil auxotrophic (Sakaguchi et al., 2011; Serif et al., 2018). However, we identified 5-FOA and uracil dependent mutants in our negative control treatments, in which single colonies of wild type *P. tricornutum* cells which were not bombarded were able to grow and survive 5-FOA selection. Similarly, cells bombarded with RNPs targeting the putative safe harbour alone and selected for using exposure to 5-FOA also resulted in single resistant colonies. Herein, we report the first demonstration of the mutagenic effects of 5-FOA on *P. tricornutum*, which has also been demonstrated in other organisms (Wang et al., 2004; Wellington et al., 2006; Wellington & Rustchenko, 2005*;* Ishii et al., 2018; Minoda et al., 2004). Again, this work is timely, as more researchers begin using this promiscuous endogenous marker (Slattery et al., 2020) since its demonstration with CRISPR-Cas9 (Serif et al., 2018).

Even though we faced issues with false positives in both *C. reinhardtii*, we were very surprised to not detect a single *C. reinhardtii* mutant, as this species is theoretically an ideal CRISPR-Cas9 candidate cell line. *C. reinhardtii* is haploid, which overcomes issues regarding heterogeneity in edits, has cell-wall deficient mutants amenable for chemical delivery strategies such as lipofection, and is unicellular, which means it does not experience tissue mosaicism associated with more complex eukaryotes such as zebrafish, mice and plants. Instead, only one cell has to be edited for whole organism to be edited. Unfortunately, our results do not help to elucidate at which point the problem is occurring: be it at the intracellular delivery level, the CRISPR-Cas9 double stranded break induction level, or the DNA repair and editing level. This is because of the lack of a reliable endogenous target gene to knock-out. However, this work underscored that there is an urgent need to develop CRISPR-Cas9 protocols for genome editing in microalgae that are reproducible in numerous labs. Particular focus

should be made to reliable enrichment and/or screening processes, as the work-load to optimise CRISPR-Cas9 RNP protocols solely by edit detection (using T7E1 analysis or RFLP analysis) becomes unmanageable, both regarding cost and labour intensity. Finally, this work confirms that developing such protocols in microalgae is uniquely more complex than other species and future research should investigate the biologicals reasons for this. Herein, we investigated strategies to optimise delivery protocols because research is unable to control CRISPR-Cas9 molecular editing at this stage. However, future research should investigate optimisation at the protein activity level, including but not limited to protein engineering Cas9, which is a bacterial proteins that might not function optimally in a nonbacterial, microalgal cellular environment. Indeed, low delivery efficiency is less problematic when a high efficiency editing system is available.

P. tricornutum as a promising chassis organism for monoterpenoid production

In Chapter 5, we used the recently developed approach of extrachromosomal expression to compare metabolic engineering and bioprocessing approaches to increase heterologous production of the monoterpenoid geraniol in *P. tricornutum*. First, we attempted to increase GPP substrate availability and utilisation by expressing two adjacent enzymes fused to one another: geranyl diphosphate synthase (AgGPPS2; AF513112; EC 2.5.1.1) which produces GPP from IPP and/or DMAPP in the plant *Abies grandis,* and geraniol synthase (CrGES; Caros003727.1; EC 3.1.7.11) which converts GPP into geraniol in the plant *Catharanthus roseus*. This resulted in 55.3% decreased geraniol accumulation compared to a control cell line that did not contain AgGPPS2, most likely due to improper folding of the fused exogenous enzymes. Then, we showed that a dark:light (12h:12h) photoperiod had no impact on

geraniol production compared to continuous light cultivation. Similarly, we reported no impact geraniol production following an elongated cultivation period extending beyond stationary phase. This was not surprising, as primary cellular metabolism–of which MVA and MEP pathways are a part–is more transcriptionally active during the exponential phase of *P. tricornutum* (Smith et al., 2016). These results correlate with the productivity profile reported in the heterologous production of betullinic acid in *P. tricornutum* (D'Adamo et al., 2018), and are in contrast with heterologous production of terpenoids in *C. reinhardtii* (Lauersen *et al.*, 2018a).

Finally, we investigated the impact of exogenous CrGES expression and activity on other native *P. tricornutum* terpenoids, namely sterols and pigments. Such information is important for future metabolic engineering approaches as well as expanding the understanding of native terpenoid metabolism in this widely studied model diatom. Because there is very little information available regarding the biosynthesis of GPP in *P. tricornutum*, it is not known if this cytosolic pool is created entirely in the cytosol via the MVA pathway, or partially or completely by the MEP pathway in the chloroplast and transported into the cytosol. We reported that neither sterol nor pigment biosynthesis was significantly impacted by heterologous geraniol production, most likely due to the expression levels of *GES-mVenus* being below the threshold where pigment and sterol content are impacted. This suggested that terpenoid metabolism in diatoms could be particularly flexible and able to adapt to the installation of artificial metabolic sinks as hypothesised in the green alga *C. reinhardtii* (Lauersen et al., 2016). If confirmed, this might represent another promising trait for terpenoid-based engineering in *P. tricornutum*.

Future work

To the best of our knowledge, this work reports for the first time the suitability and utility of third generation long-read whole-genome sequencing to reveal the previously unknown nature of chromosomal integration sites, that would not have been feasible with conventional short-read sequencing in microalgae. Future work investigating trans-genomes, such as low expression RICE cell lines or epigenetic modifications including DNA methylation patterns (Jain et al., 2018; Jupe et al., 2019), would build upon this knowledge to help uncover mechanisms driving transgene integration in diatoms. Such knowledge is important for developing better functional genomics tools including targeted genome editing. Our research primarily aimed at tracking specific, known transgenic constructs in EE and RICE transgenic diatoms cell lines. Long-read whole-genome sequencing technology can also be used to identify changes to the genome independent of an integration event, such as large translocations (Jupe et al., 2019) and deletions (Nicholls et al., 2019), purely due to the disruptive nature of the DNA delivery method. Our work lays the basis for future research efforts specifically focused on these relevant aspects, currently unknown in diatom, to investigate the impact of biolistic bombardment itself on genome integrity.

Further work is needed to drive higher geraniol productivity to clearly elucidate GPP flux through native and heterologous metabolic pathways in *P. tricornutum*. There are numerous alternative rational designs to consider for this approach –none of which have been tested in any diatom– such as adding a heterologous MVA pathway (Liu et al., 2016; Qian et al., 2019) or testing the yeast farnesyl diphosphate synthase (ERG20) mutant, which favours GPP synthesis over FPP synthesis (Ignea, Pontini, Maffei, Makris, & Kampranis, 2014). Such approaches could overcome the diatom's natural cellular regulation, even if redundant. Given that we demonstrated high

expression following RICE and highly concatenated *CrGES* gene arrangement, it would also be appropriate to explore multigene copy expression on geraniol production in *P. tricornutum*. This approach has been validated in *S. cerevisiae*, whereby expression of an *ERG20* mutant fused to *CrGES* was supplemented with an additional free copy of the *ERG20* mutant gene, resulting in approximately 20% increased geraniol accumulation (Jiang et al., 2017).

Such knowledge is important for better elucidating the ecological impact of these important marine primary producers in a changing climate, as well as for biotechnological production of high-value terpenoids such as monoterpenoid indole alkaloids, which are difficult to produce at high titres in *E. coli* and *S. cerevisiae. P. tricornutum* is a widely used model species which can be used to apply knowledge to many other diatoms. This makes it's a strong candidate to explore terpenoid production, and possibly a useful industrial model in the future, depending on how diatom-specific synthetic biology tools and knowledge advance. Finally, throughout all the experiments in this study, we demonstrated how extrachromosomal expression can be useful for faster synthetic biology design-build-test-learn cycle compared to RICE, owing to the fact that this non-integrative approach results in highly consistent transgenic phenotypes across clones (George et al., 2020). In this way, the work described here offers an important proof-of-concept for future more complex synthetic biology ventures beyond terpenoid metabolic engineering.

REFERENCES

Baek, K., Kim, D. H., Jeong, J., Sim, S. J., Melis, A., Kim, J.-S., … Bae, S. (2016). DNA-free two-gene knockout in Chlamydomonas reinhardtii via CRISPR-Cas9 ribonucleoproteins. *Scientific Reports*, *6*, 30620. https://doi.org/10.1038/srep30620

Brown, S., Clastre, M., Courdavault, V., & O'Connor, S. E. (2015). De novo production of the plant-derived alkaloid strictosidine in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3205–3210. https://doi.org/10.1073/pnas.1423555112

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., … Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, *155*(7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001

Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., … Sharp, P. A. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, *160*(6), 1246–1260. https://doi.org/10.1016/j.cell.2015.02.038

D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., … Matthijs, M. (2018). Engineering The Unicellular Alga Phaeodactylum tricornutum For High-Value Plant Triterpenoid Production. *Plant Biotechnology Journal*, 0–2. https://doi.org/10.1111/pbi.12948

Dziggel, C., Schafer, H., & Wink, M. (2017). Tools of pathway reconstruction and production of economically relevant plant secondary metabolites in recombinant microorganisms. *Biotechnology Journal*, *12*(1). https://doi.org/10.1002/biot.201600145

Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R. M., … Ralph, P. (2020). Extrachromosomal genetic engineering of the marine diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.9b00455

George, J., Kahlke, T., Abbriano, R. M., Kuzhiumparambil, U., Ralp, P. J., & Fabris, M. (2020). Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology. *Frontiers in Bioengineering and Biotechnology*, *8*(June), 1–19. https://doi.org/10.3389/fbioe.2020.00513

Ignea, C., Pontini, M., Maffei, M. E., Makris, A. M., & Kampranis, S. C. (2014). Engineering monoterpene production in yeast using a synthetic dominant negative geranyl diphosphate synthase. *ACS Synthetic Biology*, *3*(5), 298–306. https://doi.org/10.1021/sb400115e

Ishii, Y., Maruyama, S., Fujimura-Kamada, K., Kutsuna, N., Takahashi, S., Kawata, M., & Minagawa, J. (2018). Isolation of uracil auxotroph mutants of coral symbiont alga for symbiosis studies. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-21499-3

Jakočiunas, T., Rajkumar, A. S., Zhang, J., Arsovska, D., Rodriguez, A., Jendresen, C. B., … Keasling, J. D. (2015). CasEMBLR: Cas9-Facilitated Multiloci Genomic Integration of in Vivo Assembled DNA Parts in Saccharomyces cerevisiae. *ACS Synthetic Biology*, *4*(11), 1126–1134. https://doi.org/10.1021/acssynbio.5b00007

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., … Yuan, Y. J. (2017). Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metabolic Engineering*, *41*(March), 57–66.

https://doi.org/10.1016/j.ymben.2017.03.005

Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., … Weiss, R. (2014). CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nature Methods*, *11*(7), 723–726. https://doi.org/10.1038/nmeth.2969

Lauersen, K. J., Baier, T., Wichmann, J., Wördenweber, R., Mussgnug, J. H., Hübner, W., … Kruse, O. (2016). Efficient phototrophic production of a high-value sesquiterpenoid from the eukaryotic microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *38*, 331–343. https://doi.org/10.1016/j.ymben.2016.07.013

Liu, W., Xu, X., Zhang, R., Cheng, T., Cao, Y., Li, X., … Xian, M. (2016). Engineering Escherichia coli for high-yield geraniol production with biotransformation of geranyl acetate to geraniol under fed-batch culture. *Biotechnology for Biofuels*, *9*(1), 1–8. https://doi.org/10.1186/s13068-016-0466-5

Meinecke, L., Alawady, A., Schroda, M., Willows, R., Kobayashi, M. C., Niyogi, K. K., … Beck, C. F. (2010). Chlorophyll-deficient mutants of Chlamydomonas reinhardtii that accumulate magnesium protoporphyrin IX. *Plant Molecular Biology*, *72*(6), 643–658. https://doi.org/10.1007/s11103-010-9604-9

Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T., & Tanaka, K. (2004). Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, *Cyanidioschyzon merolae* 10D. *Plant and Cell Physiology*, *45*(6), 667–671. https://doi.org/10.1093/pcp/pch087

Pflueger, C., Tan, D., Swain, T., Nguyen, T., Pflueger, J., Nefzger, C., … Lister, R. (2018). A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Research*, *28*(8), 1193–1206. https://doi.org/10.1101/gr.233049.117

Pollak, B., Matute, T., Nuñez, I., Cerda, A., Lopez, C., Vargas, V., … Federici, F. (2019). Universal Loop assembly (uLoop): open, efficient, and species-agnostic DNA fabrication. *BioRxiv*. https://doi.org/10.1101/744854

Qian, S., Clomburg, J. M., & Gonzalez, R. (2019). Engineering Escherichia coli as a platform for the in vivo synthesis of prenylated aromatics. *Biotechnology and Bioengineering*, *116*(5), 1116–1127. https://doi.org/10.1002/bit.26932

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M. A., … Gunawardane, R. N. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular Biology of the Cell*, *28*(21), 2854–2874. https://doi.org/10.1091/mbc.E17-03-0209

Sakaguchi, T., Nakajima, K., & Matsuda, Y. (2011). Identification of the UMP synthase gene by establishment of Uracil auxotrophic mutants and the phenotypic complementation system in the marine diatom Phaeodactylum tricornutum. *Plant Physiology*, *156*(1), 78–89. https://doi.org/10.1104/pp.110.169631

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., & Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-06378-9

Shalem. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells, *343*(January), 84–88.

Shaw, S. L., Gantt, B., & Meskhidze, N. (2010). Production and Emissions of Marine Isoprene and Monoterpenes: A Review. *Advances in Meteorology*, *2010*(1), 1–24.

https://doi.org/10.1155/2010/408696

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Chang, Y. K. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. *Nature Publishing Group*. https://doi.org/10.1038/srep27810

Slattery, S. S., Diamond, A., Wang, H., Therrien, J. A., Lant, J. T., Jazey, T., … Edgell, D. R. (2018). An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synthetic Biology*, acssynbio.7b00191. https://doi.org/10.1021/acssynbio.7b00191

Slattery, S. S., Wang, H., Kocsis, C., Urquhart, B. L., Bogumil, J., & Edgell, D. R. (2020). Cas9-generated auxotrophs of Phaeodactylum tricornutum are characterized by small and large deletions that can be complemented by plasmid-based genes.

Smith, S. R., Gillard, J. T. F., Kustka, A. B., McCrow, J. P., Badger, J. H., Zheng, H., … Moritz, T. (2016). Transcriptional Orchestration of the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron Limitation. *PLOS Genetics*, *12*(12), e1006490. https://doi.org/10.1371/journal.pgen.1006490

Sternberg, S. H., & Doudna, J. A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, *58*(4), 568–574. https://doi.org/10.1016/j.molcel.2015.02.032

Stovicek, V., Holkenbrink, C., & Borodina, I. (2017). CRISPR/Cas system for yeast genome engineering: advances and applications. *FEMS Yeast Research*, *17*(5), 1–16. https://doi.org/10.1093/femsyr/fox030

Thakore, P. I., Song, L., Safi, A., Shivakumar, K., Kabadi, A. M., Reddy, T. E., … Gersbach, C. A. (2015). Highly Specific Epigenome Editing by CRISPR/Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nature Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630.Highly

Wang, Y. K., Das, B., Huber, D. H., Wellington, M., Kabir, M. A., Sherman, F., & Rustchenko, E. (2004). Role of the 14-3-3 protein in carbon metabolism of the pathogenic yeast Candida albicans. *Yeast*, *21*(8), 685–702. https://doi.org/10.1002/yea.1079

Wellington, M., Kabir, M. A., & Rustchenko, E. (2006). 5-Fluoro-orotic acid induces chromosome alterations in genetically manipulated strains of Candida albicans. *Mycologia*, *98*(3), 393–398. https://doi.org/10.3852/mycologia.98.3.393

Wellington, M., & Rustchenko, E. (2005). 5-Fluoro-orotic acid induces chromosome alterations in Candida albicans. *Yeast*, *22*(1), 57–70. https://doi.org/10.1002/yea.1191

Willrodt, C., David, C., Cornelissen, S., Bühler, B., Julsing, M. K., & Schmid, A. (2014). Engineering the productivity of recombinant Escherichia coli for limonene formation from glycerol in minimal media. *Biotechnology Journal*, *9*(8), 1000–1012. https://doi.org/10.1002/biot.201400023

Xue, H. Y., Zhang, X., Wang, Y., Xiaojie, L., Dai, W. J., & Xu, Y. (2016). In vivo gene therapy potentials of CRISPR-Cas9. *Gene Therapy*, *23*(7), 557–559. https://doi.org/10.1038/gt.2016.25

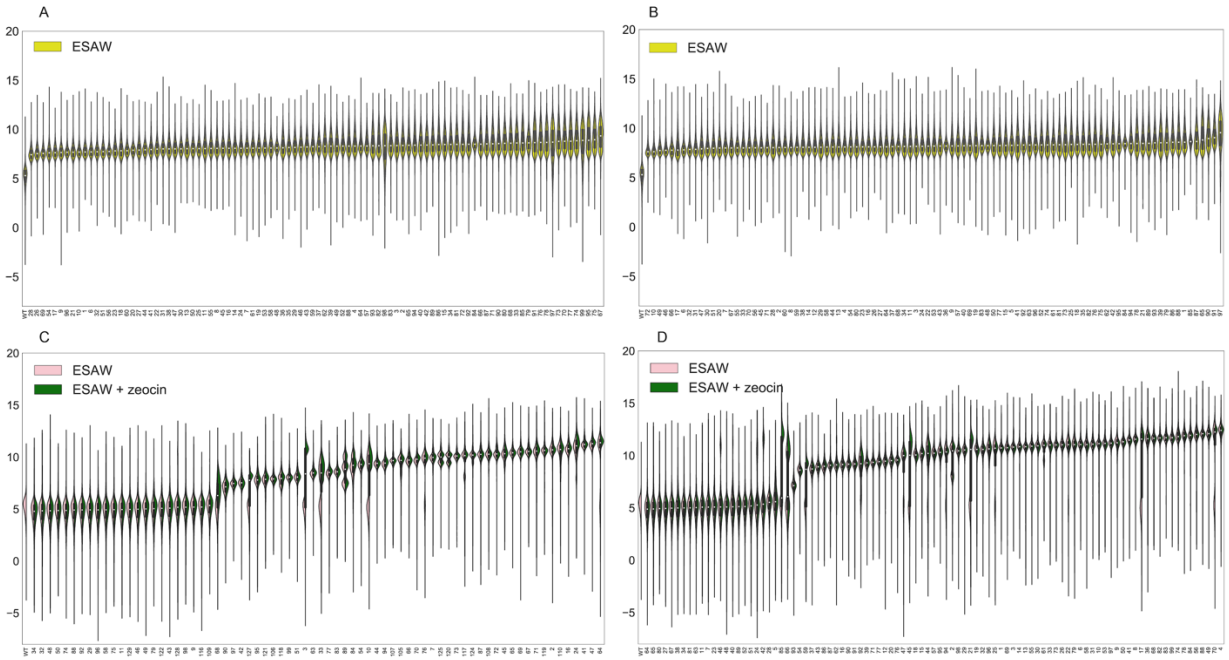
Yassaa, N., Peeken, I., Zöllner, E., Bluhm, K., Arnold, S., Spracklen, D., & Williams, J. (2008). Evidence for marine production of monoterpenes. *Environmental Chemistry*, *5*(6), 391. https://doi.org/10.1071/EN08047

Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology

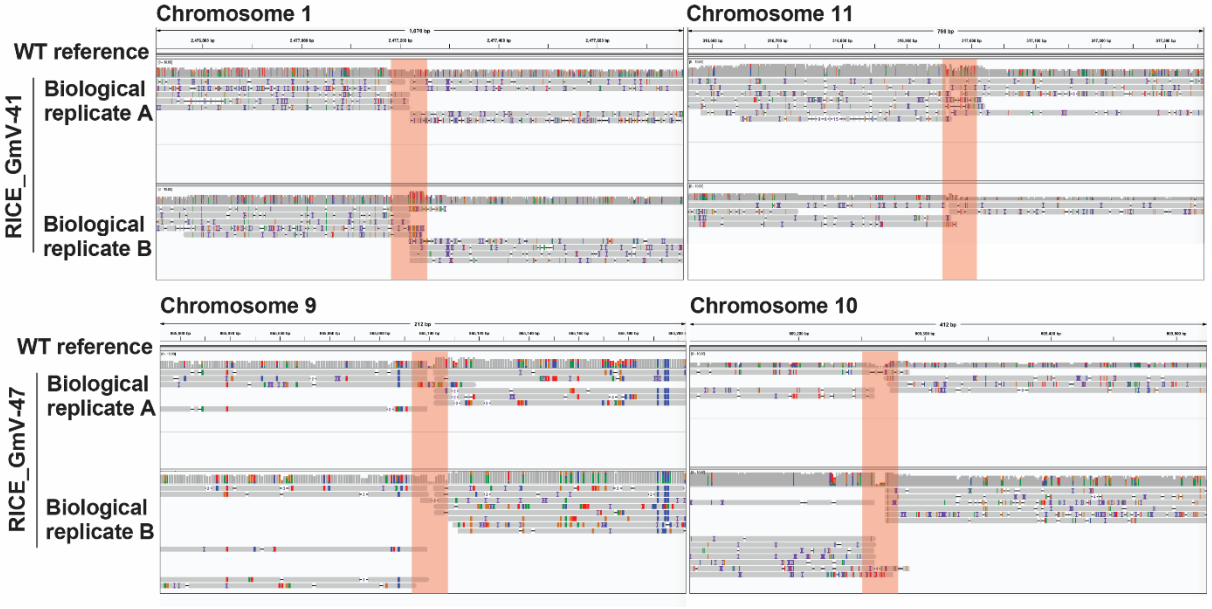
1

2 Supplementary File 1

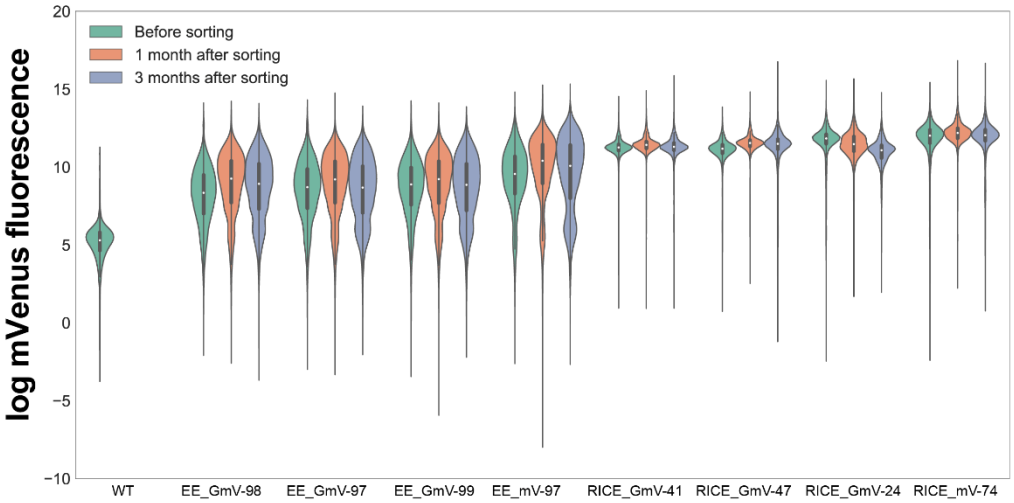
3



Suppl. Figure 1 mVenus fluorescence intensities of transgenic *P. tricornutum* extrachromosomal expression (EE) and randomly integrated chromosomal expression (RICE) libraries. Violin plots indicate mVenus fluorescence intensity per cell, of all cell lines for each library. (a) EE_GmV; (b) EE_mV; (c) RICE_GmV; (d) RICE_mV. Pink indicates selection free growth conditions, green and yellow indicates zeocin selection growth conditions, cell lines are ranked by mean mVenus intensity (n = 20,000 cells for each cell line).



Suppl. Figure 2 Integration sites in RICE_GmV-41 and RICE_GmV-47 cell lines. Reads obtained from transgenic diatom cell lines that show homology with the *pUC19_AP1pCrGES-mVenus* construct are aligned to *P. tricornutum* wild type reference genome, ASM15095v2. Replicate sequencing experiments per cell line are indicated by either A or B notation. Nucleotides corresponding to the wild type reference genome appears at the top of each frame, with the aligned MinION reads below. Orange bars broadly indicate the regions where integration islands occur.



Suppl. Figure 3 mVenus fluorescence population distribution of exconjugants and random integration transformant cell lines used in geraniol analysis. Population distributions for before, 1 month after, and 3 months after sorting by FACS are shown; n = 20,000 cells.

ADDENDUM II: LIST OF THESIS TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

ADDENDUM III: LIST OF THESIS FIGURES

CHAPTER 1

monoterpenoid-indole alkaloid (MIA) synthesis occurring in various cellular compartments and cell types of the medicinal plant, *Catharanthus roseus*. The universal terpenoid precursor molecules, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are synthesised via two independent metabolic pathways: the mevalonate (MVA) pathway (blue box) in the cytosol (grey compartment) and the methylerythritol phosphate (MEP) pathway (light green box) in the chloroplast (dark green compartment). The MVA pathway uses acetyl coenzyme A (AcCoA) to produce mevalonate, which is converted into IPP and DMAPP. In the chloroplast, the MEP pathway uses glyceraldehyde-3-phosphate (G-3-P) and pyruvate to produce (E)-4-Hydroxy-3-methyl-but-2-enyl pyrophosphate (HMBDP), which can be converted into IPP and DMAPP. DMAPP is converted to geranyl diphosphate (GPP) by GPP synthase (EC 2.5.1.1) in the plastid. Plastidial GPP is converted to the monoterpenoid geraniol by geraniol synthase (EC 3.1.7.11). Geraniol is the first monoterpenoid (white box) and produced in the chloroplasts of *C. roseus* internal phloem-associated parenchyma (IPAP) cells and vascular cells. It is then used for the production of another key monoterpenoid, secologanin, which is produced in the vacuoles (orange compartment) of epidermal cells. The secologanin monoterpenoid is then condensed with tryptamine, an indole produced in the shikimate pathway, to form the base molecule, strictosidine, in the Pictet–Spengler reaction. Strictosidine is the universal precursor of all MIAs, including key therapeutic molecules: vindolicine, anhydrovinblastine, vincristine, ajmalicine, tabersonine, catharanthine, vindoline and vinblastine. GPP is also used to make farnesyl diphosphate (FPP), a precursor for squalene and all sterols, which are produced in the endoplasmic recticulum (yellow compartment). Likewise, GPP is also used to make FPP in the plastid, where it acts as a precursor for production of photosynthetic and accessory pigments, such as chlorophylls and carotenoids, respectively.

Figure 6. Biosynthesis of terpenoids. The pathways have been conceptually separated into four modules, which is representative of the modularised method many metabolic engineers use to approach isoprenoid pathway engineering. Mevalonate (MVA) and methyl-D-erythritol phosphate (MEP) pathways lead to IPP (isopentenyl pyrophosphate) and DMAPP (dimethylallyl pyrophosphate) (module I). Additions of IPP produce higher-order prenyl phosphates (module II), dephosphorylation (often coincident with or followed by bond rearrangement and/or cyclisation) to form specialized terpenoid backbones (module III), chemical decorations, and other modifications to yield end products. Note that not all end products undergo decorations of the carbon skeleton. From Vavitsas et al. (2018).

Figure 7. Enzymatic reactions in the MVA and MEP pathways in plants and synthesis of the short-chain prenyl diphosphates. The MVA pathway is shown in yellow; the MEP pathway is shown in green. AACT; Acetyl-CoA c-acetyltransferase, HMGS; hydroxy-methylglutaryl-CoA synthase, HMGR; hydroxyl-methylglutaryl-CoA reductase, MVK; mevalonate kinase, PMK; phospho-mevalonate kinase, MVD; mevalonate disphosphate decarboxylase, DXS; 1-deoxy-D- xylulose 5-phosphate synthase, DXR; 1-deoxy-D-xylulose 5-phosphate reductoisomerase, MCT; 2-C-methyl-D-erythritol-4-phosphate-cytidylyltransferase, CMK; 4- diphosphocytidyl-2c-methyl-d-erythritol kinase, MDS; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, HDS; (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase, HDR; 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase, IPPI; isopentenyl diphosphate isomerase, GPPS; geranyl diphosphate synthase, FPP; farnesyl diphosphate synthase, GGPP; geranylgeranyl diphosphate synthase. From Vranová et al. (2013).

CHAPTER 2

Figure 1. Schematic representation of CRISPR-Cas9 editing technology. The orange guide RNA component of the CRISPR-Cas9 complex is made up of a 20 nucleotide long guide region and a longer regulatory region that helps the guide to assemble with the Cas9 enzyme. The blue Cas9 enzyme contains a HNH and RuvC endonuclease domain for double stranded DNA digestion. After delivery into the target species, the CRISPR-Cas9 ribonucleoprotein complex scans the host genome to find the region that aligns to the guide RNA, which is always upstream of a protospacer adjacent motif (yellow). The double stranded break can be repaired by non-homologous end joining (NHEJ) resulting in various possibilities of INDELs (insertions, deletions, or a combination). Alternatively, if donor DNA or single stranded donor oligonucleotide (ssODN) containing homologous regions to the cut site is supplied to the cell, the repair can be mended by homology driven (or directed) repair (HDR), also commonly referred to as homologous recombination. Numerous reports have also demonstrated that donor DNA without homologous flanks can be integrated by NHEJ. Adapted from https://www.sinobiological.com/crispr-cas9-system.html

Figure 2. Schematic representation of CRISPR-Cas9 technology based on either (A) conventional plasmid delivery system or (B) ribonucleoprotein (RNP) delivery system. From https://viromer-transfection.com/crispr.

Figure 3. Schematic representation of the various workflow strategies involved with generating a CRISPR-Cas9 mutant. Step 1. Selection of the type of endonuclease such as Cas9 or CfpI, and design of the guide RNA targets, which requires the use of software such as CRISPOR. It is also recommended to test 2-3 guide RNAs for those which have not previously been validated. Step 2. In vitro assembly of the guide RNA and endonuclease to form an active ribonucleoprotein (RNP). To confirm correct assembly and activity of the RNP, an in vitro digestion should be performed, whereby a PCR amplicon containing the target site is incubated with the RNP and the product is screened by electrophoretic mobility shift assay. Step 3. Intracellular delivery of the tested RNP into microalgal cells results in a mixture of transformed mutants and untransformed (background) cells. Step 4a. Enrichment involves removing background cells and can occur via direct method, in which the phenotype used for enrichment is the result of the CRISPR-Cas9 edit itself. Indirect methods rely on a phenotype that is not the direct result of the CRISPR-Cas9 edit, but instead relies on a proxy such as use of fluorescently tagged CRISPR-Cas9 RNP for FACS, or expression of a selectable marker which has been integrated during co-transformation with RNP. In cases where enrichment is not appropriate or not possible, screening can be used. Step 4b. Here, background cells are not removed but instead can be distinguished from the RNP edited cell lines. Screening should be directly related to CRISPR-Cas9 edit, such as colour change following knock-out of ChlM gene. Screening can also be used in conjunction with indirect enrichment, as indirect enrichment can result in a mixture of both edited and non-edited cell lines. Step 4 can be performed either using step 4a, or step 4b, or both. Step 5. Because the intracellular presence of CRISPR-Cas9 can cause some daughter cells to be repaired by NHEJ causing INDELs and others to be repaired faithfully with no edits, mutants may present as colony mosaics and consequently must be subcloned in order to generate single cell lineages (Huang & Daboussi, 2017). Step 6. Subclones can be screened for CRISPR-Cas9 induced mutation via electrophoretic mobility shift assay following various analyses of the target DNA. For example, if the experiment was designed for targeted integration or if two endonucleases were used for

cutting a fragment out of a target region, conventional PCR can be used to screen for the size change mutants. Alternatively, where targeted knock-out via NHEJ-mediated repair was expected, the T7E1 assay or restriction digestion analysis can be used. Step 7. Mutants identified as positive for CRISPR-Cas9 mutation in the screen must be sequenced in order to determine the sequence and confirm the nature of the CRISPR edit.

Figure 4. CRISPR-Cas9 RNP design and validation targeting *C. reinhardtii ChlM* gene. (A) Scheme of the tetrapyrrole biosynthetic pathway in *C. reinhardtii*, in which ChlM catalyses the conversion of Mg-protoporphyrin IX into Mg-protoporphyrin IX 13-monomethyl esterchlorophyll for the eventual production of chlorophyll a. Adapted from Meinecke et al., 2010. (B) Graphic representation of the genetic sequences for guide RNA *ChlM*-1 and guide RNA *ChlM*-2 with their respective 20 nucleotide recognition sites (pink and red, respectively) alongside protospacer adjacent motifs (PAMs) (black). Both guide RNAs target the first exon of *C. reinhardtii ChlM* gene. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break location, three base pairs upstream of the PAM site. (C) Electrophoretic mobility shift assay following *in vitro* digest of RNP-*ChlM*-1 and RNP-*ChlM*-2 activity to validate that RNP components were correctly assembled and are able to recognise and cut their target regions. The 350 bp amplicon showed no shift when incubated with Cas9 protein alone, but the presence of smaller bands when incubated with fully assembled CRISPR-Cas9 RNPs.

Figure 5. Method development for proteolistic bombardment of *C. reinhardtii* with CRISPR-Cas9 targeting *ChlM* gene. (A) Schematic representation of experimental design FITC-labelled antibody (FITC-Ab) was coated onto tungsten Microparticles and bombarded into cells. The cell biomass was scraped off bombarded agar plate a 24 hours post treatment and analysed by flow cytometry to determine the delivery efficiency based on FITC fluorescence. (B) In order to determine that proteolistic treatment did not prevent the cells from being able to recover on selectable agent, the cells were co-bombarded with the FITC-Ab and resistance DNA cassette. Treated cells with FITC-Ab and donor DNA survived after being plated on TAP supplemented with zeocin, demonstrating that this delivery method allowed for cells to recover even after selective treatment. (C) Flow cytometry analysis demonstrated the percentage of FITC positive cells following various treatments.

Figure 6. (A) Schematic flow diagram illustrating the proteolistic delivery of RNP-*ChlM*-2 and editing analysis. After co-delivery with RNP-*ChlM*-2 and FITC-Ab, cell biomass was scraped off bombarded agar plate and sorted for FITC fluorescence. Once sorted cells have grown for increased biomass, each sorted population was split and either plated in low cell density dilution for isolating single colonies, or used for genomic DNA extraction on pooled population for T7E1 analysis. (B) T7E1 analysis of pooled proteolistic control (bombarded with FITC-Ab only) and pooled RNP-*ChlM*-2 sample. The T7E1 commercially supplied positive and negative controls infdicate that the analysis was successful and that no editing occurred. (C) Photographs of diluted cell plating following proteolistic delivery. The images suggest that co-delivery sample has the expected light green phenotype associated with *ChlM* gene knockout, compared to FITC-Ab only control. (D) Sanger sequencing alignment results of the *ChlM* amplicon amplified from wild type *C. reinhardti 137c*, a single colony from the FITC-Ab only control proteolistic bombardment and ten colonies from RNP-*ChlM*-2 co-delivery bombardment showing no mismatches and therefore unsuccessful editing. Red indicates target sequence, underlined text indicates PAM motif.

Figure 7. Method development for lipofection of cell wall-deficient *C. reinhardtii* cc-503 with CRISPR-Cas9 targeting *ChlM* gene. (A) Schematic representation of experimental design where RNP-*ChlM*-2 is complexed with lipofectamine to make a RNP-*ChlM*-2:lipid complex. When the complex comes into contact with the cell membrane of the cc-503 cell (green), it fused and enters the cell, releasing the CRISPR-Cas9 cargo intracellularly. Cells which successfully receive this cargo are edited resulting in a *ChlM* knock out (pink). Daughter cells of the *ChlM* knocked out mutants have a characteristic light green phenotype due to reduced chlorophyll content which can be detected by flow cytometry. The mixture of edited and non-edited cells is analysed by flow cytometry to detect the percentage of cells with reduced chlorophyll 24 and 48 hours after lipofection. Flow cytometry gate P5 indicated cells which have lost chlorophyll fluorescence. (B) Flow cytometry analysis of lipofection method optimisation. Cells receiving RNP-*ChlM*-2:lipid complex show significantly increased percentage of cells in low chlorophyll gate P5 compared to cells exposed to the inactive Cas9 enzyme:lipid complex. N = 3, error bars indicate SEM, significance analysis by two way ANOVA where * indicates $p<0.05$ and ** indicates $p<0.01$.

Figure 8. CRISPR-Cas9 RNP design and validation targeting *C. reinhardtii ARG7* gene. (A) Graphic representation of the genetic sequences for RNP-*ARG7*-1 and RNP-*ARG7*-2 guide RNAs showing two 20 nt recognition sites (teal and green, respectively) alongside protospacer adjacent motifs (PAMs) (black, underlined). RNP-*ARG7*-1 targeted the first exon of *C. reinhardtii ChlM* gene and RNP-*ARG7*-2 targeted the second of the 14 exon *ARG7* gene. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break location, which is always three base pairs upstream of the PAM site. (B) *In vitro* digest analysis of RNP-*ARG7*-1 and RNP-*ARG7*-2 activity to validate that RNP components were correctly assembled in vitro and are able to recognise and cut their target regions. The 600 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the presence of smaller bands when incubated with fully assembled CRISPR-Cas9 RNPs indicating correct assembly. (C) The electrophoretic mobility shift assay following both T7E1 and RFLP analysis showed no CRISPR-Cas9 edited mutants. The electrophoretic assay was performed at three dilutions and intentionally overexposed in order to detect any faint smaller bands.

Suppl. Figure 1. Indirect enrichment and screening of *C. reinhardtii* following biolistic delivery of RNP-*ChlM*-2. (A) Cells bombarded with RNP-*ChlM*-2 only plated on TAP agar plates show normal growth indicating cells were able to survive biolistic treatment. (B) Cells bombarded with pChlamy4 DNA only plated on TAP agar plates supplemented with 50 mg/L zeocin (TAP-Z) show single colony growth. (C) Cells bombarded with RNP-*ChlM*-2 and pChalmy4 DNA plated on TAP-Z agar plates showing no single colonies but instead a thick mat of cells.

Suppl. Figure 2. Fluorometry analysis of *C. reinhardtii* single colonies following co-delivery of RNP-*ChlM*-1 or RNP-*ChlM*-2 with pChlamy4 DNA, where colonies were selected on TAP supplemented with 50 mg/L zeocin (TAP-Z). (A) Photographs of selection plates excited by 450nm light during pulse amplitude modulation (PAM) fluorometry. (B) Chlorophyll fluorescence of three randomly selected colonies per selection plate measured by flow cytometry (n=2 selection plates for wild type (WT) controls; n=10 selection plates for bombarded samples).

CHAPTER 3

Figure 1. mVenus fluorescence intensities of transgenic P. tricornutum extrachromosomal expression (EE) and randomly integrated chromosomal expression (RICE) libraries. (A) Fold change of mean mVenus fluorescence of RICE_GmV and RICE_mV transformant libraries and EE_GmV and EE_mV libraries compared to wild type auto-fluorecence. Peach indicates CrGES-mVenus transgenic cell lines and teal indicates mVenus transgenic cell lines. Statistical comparisons were made using Kruskal-Wallis non-parametric ANOVA and Dunn's post-hoc test. For EE_GmV library n = 96 cell lines total, EE_mV library n = 96 cell lines total, RICE_GmV library n = 74 cell lines total and RICE_mV library n = 95 cell lines total. (B) Percentage of pooled RICE libraries (green) compared to percentage of pooled EE libraries (yellow) binned according to mean mVenus fluorescence fold change. (C–F) Violin plots indicate the per cell mVenus fluorescence intensity of ten representative cell lines for each library (C) EE_GmV (D) EE_mV (E) RICE_GmV (F) RICE_mV, ranked from lowest to mean mVenus expression (n = 20,000 cells for each cell line). (G,H) Representative cell lines from transgene stability analysis for (G) RICE_GmV and (H) RICE_mV libraries. Pink indicates selection free growth conditions, green indicates zeocin selection growth conditions, cell lines are ranked by mean mVenus intensity (n = 20,000 cells for each cell line).

Figure 2. Graphic representation of exogenous DNA constructs in extrachromosomal and chromosomal DNA of the transgenic cell lines, based on long-read sequencing. Only reads aligning to both exogenous DNA and wild type *P. tricornutum* genome, and not those aligning to the genome alone are depicted. (A) EE_GmV-97 transformant showed no reads which aligned to both exogenous episomal *DNA pPtPBR11_AP1p_CrGES-mVenus* (yellow), and the reference *P. tricornutum* genome, indicating that no exogenous DNA was integrated into the genome. Instead, some reads showed alignment (red) only to episome DNA, suggesting these reads came from episomal DNA which was extracted and analysed with genomic DNA. (B) Transformant RICE_GmV-41 generated by biolistic bombardment showed reads which aligned to both exogenous RICE plasmid, *pUC19_AP1p_CrGES-mVenus*, and the reference *P. tricornutum* genome, indicating a frequency of only two integration islands, 41-1 and 41-11, occurring throughout the whole genome. Island 41-1 occurred on chromosome 1 where the longest left border read (LB-R) and right border read (RB-R) collectively indicated that this island was a minimum of 43 Kbp in size. Island 41-11 occurred on chromosome 11 and was spanned by a single read, left-right border read (LRB-R), which aligned to the reference genome at both left and right borders (light blue), as well as the exogenous RICE plasmid. Red indicates alignment in sense orientation and dark blue indicates alignment in antisense orientation, representing the highly concatenated integration events observed. (C) RICE_GmV-47 transformant showed reads which aligned to both exogenous RICE plasmid, *pUC19_AP1p_CrGES-mVenus*, and the reference *P. tricornutum* genome, indicating a frequency of only two integration islands, 47-9 and 47-10, occurring throughout the whole genome. Island 47-9 occurred on chromosome 9 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 124 kB in size. Island 47-10 occurred on chromosome 10 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 87 Kbp in size.

Figure 3. Graphic representation of rearrangements of exogenous DNA in *P. tricornutum* chromosomes, based on long-read sequencing. Red channels show alignment in sense orientation and blue channels show alignment in antisense orientation. Regions that are not

highlighted did align to the plasmid, but with below-threshold for hit length of percent identity used for the visualisation, which was performed manually. (A) Alignment of a left-right border read (LRB-Read) (top) from integration event 41-11 to RICE plasmid *pUC19_AP1p_CrGES-mVenus* (bottom) and to the wild type *P. tricornutum* genome (green). (B) A single 97.5 Kbp read (bottom) with no regions of similarity to the P. tricornutum wild type reference genome aligned to the RICE plasmid *pUC19_AP1p_CrGES-mVenus* (top). (C) Integration island 47-9 made up by two reads; the left border read (LB-Read) (middle) contains approximately 42 Kbp aligned to the RICE plasmid (bottom) and 3 Kbp aligned to the *P. tricornutum* wild type reference genome (top). The right border read (RB-Read) (middle) contains approximately 82 Kbp of aligned to the RICE plasmid and 2 Kbp aligned to the *P. tricornutum* wild type reference genome. (D) Alignments of left and right border reads to each other for integration island 41-1, 47-9, and 47-10. These reads do not align to each other to "close" the integration island, suggesting that some "filler" reads are missing.

Figure 4. Geraniol production in three selected RICE_GmV and EE_GmV cell lines. N = 3, error bars represent SEM, statistical comparisons were made using one-way ANOVA and Tukey's multiple comparisons post-hoc test. Significance is demonstrated with asterisks, where ****p ≤ 0.0001; ***p ≤ 0.001; and *p ≤ 0.05. (A) Growth curve for all cell lines. (B) mVenus fluorescence intensity 24 h after induction. (C) geraniol produced after 72 h induction.

CHAPTER 4

Figure 1. Graphic representation of uridine-5'-monophosphate synthase (*UMPS*; Phatr3_J11740) involved in de novo pyrimidine biosynthesis. (A) The heterozygous *UMPS* genotype from *P. tricornutum* UTEX LB 642 wild type strain sequenced by Sakaguchi et al. (2011) confirmed the presence of a functional and non-functional *UMPS* allele. Both the sequenced heterozygous and theoretical homozygous phenotypes will contribute to a functional UMPS enzyme able of converting oroate (dark pink hexagon) into uracil (light pink triangle) and survival of these strains in ESAW media with and without uracil supplementation. However, when supplemented with fluoroorotic acid (5-FOA) (orange hexagon), an analogue of oroate, *UMPS* will produce toxic 5-fluorouracil (5-FU) (yellow pentagon), which kills the cells. (B) On the contrary, a homozygous strain bearing both non-functional *UMPS* alleles–or a *UMPS* knock-out genotype–is not able to survive in ESAW unless uracil is supplemented. Furthermore, this mutant is able to survive in the presence of 5-FOA, as it is unable to use this substrate. (C) The biosynthesis of uridine-5'-phosphate from oroate in *P. tricornutum* is catabolised by the UMPS enzyme (Phatr3_J11740) in two subsequent reactions; the conversion of oroate to orotidine-5'-phosphate (EC 2.4.2.10) and then to uridine-5'-phosphate (EC 4.1.1.23).

Figure 2. CRISPR-Cas9 RNP design and validation targeting *P. tricornutum UMPS* gene and putative safe harbour locus, the intergenic region *ch1:2,477,260*. (A) Graphic representation of the genetic sequences for RNP-UMPS-1 (red) and RNP-UMPS-3 (pink) guide RNAs targeting *UMPS*, and RNP-ch1:247-A (blue) and RNP-ch1:247-B (purple) guide RNAs targeting near to the putative safe harbour locus, *ch1:2,477,260*. All guide RNA recognition sites are 20 nt in length occurring alongside protospacer adjacent motifs (PAMs) (black, underlined). The ch1:247 locus occurs between predicted protein coding gene *Phatr3_J8770*

and *Phatr3_J54066*. The scissor icon and dashed line indicate the predicted CRISPR-Cas9 induced double stranded break location, which is always three base pairs upstream of the PAM site. (B) *In vitro* digest analysis of RNP-*UMPS*-1 and RNP-*UMPS*-3 activity to validate that RNP components were correctly assembled *in vitro* and are able to recognise and cut their target regions. The 740 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the expected 543 bp and 197 bp bands following incubation with fully assembled CRISPR-Cas9 RNP-*UMPS*-1. Likewise, the expected 516 bp and 224 bp bands occurred after incubation with RNP-*UMPS*-3. In the double digest with both RNP-*UMPS*-1 and RNP-*UMPS*-3, the expected 319 bp, 224 bp, and 197 bp bands are present, as well as the 516 bp band from RNP-*UMPS*-3 digest, indicating that some fragments were not fully digested. (C) *In vitro* digest analysis of RNP-*ch1:247*-B and RNP-*ch1:247*-B activity to validate that RNP components were correctly assembled *in vitro* and are able to recognise and cut their target regions. The 669 bp amplicon showed no gel shift when incubated with Cas9 protein alone, but the expected 565 bp and 104 bp bands following incubation with fully assembled RNP-*ch1:247*-A. The same expected pattern is seen following incubation with fully assembled RNP-*ch1:247*-B, however it appears RNP-*ch1:247*-B is more efficient than RNP-*ch1:247*-A, due to the faint undigested 669 bp band present in RNP-*ch1:247*-A. The double digest reaction would not allow for both RNP-*ch1:247*-A and RNP-*ch1:247*-B to digest the amplicon, as the activity of one would destroy the recognition site required for the other. However, we see that one of the two is able to digest the amplicon, which is useful to ensure *in vivo* digest, should one RNP be inactive.

Figure 3. CRISPR-Cas9 RNP proteolistic bombardment does not mediate targeted genomic integration (TGI). (A) Number colonies appearing on selection plates (n = 2 for control transformations and n = 6 for CRISPR-Cas9 proteolistic transformations, error indicates SEM). Small specs which could be tiny colonies or could be non-living background were the only ‘colony-like’ spots on the TGI integration plates. (B, C) Growth of picked colonies in liquid culture confirming that TGI tiny specs were not colonies but just background from two independent experiments, where experiment 1 is shown in B and experiment 2 is shown in C). (D) PCR verifying that the donor DNA retained in mutant populations following both RICE and EE. (E) mVenus fluorescence of mutants generated by RICE and EE of donor DNA was able retained in mutant populations following both RICE and EE expressed by RICE and EE.

Figure 4. Electrophoretic mobility shift assay to screen sub-colonies for CRISPR-Cas9 driven knock-out of *UMPS* gene following proteolistic bombardment with two RNPs targeting *UMPS* (2% agarose). Primers annealing to regions approximately 200 bp up and downstream of the recognition sites of RNP-*UMPS*-1 and RNP-*UMPS*-3 were used to detect an edit in these regions. Edited cell lines are expected to have variable sized amplicons between these regions due to INDELs caused by NHEJ mediated repair of CRISPR-Cas9 driven double stranded breaks. The unedited wild type amplicon is expected to be 740 bp.

Figure 5. RURF phenotype in WT^R and Pt_RNP-Dual strains. (A) Growth analysis of untreated wild type, WT^R and Pt_RNP-Dual cultured in ESAW media supplemented with 5-FOA and uracil (dotted lines) compared to ESAW media without supplementation (solid lines). (B) Uracil auxotrophy analysis of 81 sub-colonies starved of uracil prior to dilution plating on ESAW agar plates with and without uracil.

Figure 6. Analysis of *UMPS* genotype in untreated wild type *P. tricornutum* CCAP 1055/1 strain, WT^R strains and a Pt_RNP-Dual strain. (A) Graphic representation of the *UMPS* locus

and neighbouring predicted protein coding regions *Phatr3_J45195* and *Phatr3_J45193*. Red markers indicate left primers and blue markers indicate right primers. Amplicons obtained from WT^R and Pt_RNP-Dual strains are depicted above the *UMPS* locus in purple fragments, where the light purple fragment indicates the region which was not amplifiable. Amplicons obtained from untreated wild type are depicted below the *UMPS* locus. Amplicons are named based on the forward and reverse primer ID codes used to amplify those products. (B) Amplicons from all six PCR amplifications.

Figure 7. Analysis of *UMPS* genotype in untreated wild type *P. tricornutum* CCAP 1055/1 strain (A) The presence of doublet peaks at precise locations indicate single nucleotide polymorphisms characteristic of two non-identical gene copies (alleles) resulting in a heterozygous genotype in diploid organisms. These doublets occur in exogenic (blue) and intronic regions. The SNPs identified in this region are identical to those identified by Sakaguchi et al. (2011). (B) SNP-14 in wild type strain has been lost in all three WT^R strains as evident by the loss of the doublet peak confirmed across three individually sequenced reads per cell line, indicating a homozygous genotype.

Suppl. Figure 1. Sequence alignment highlighting single nucleotide polymorphisms (SNPs) identified within *P. tricornutum UMPS (Phatr3_J11740)* gene of UTEX LB 642 wild type strain (Seq_1) identified by Sakaguchi et al. (2011) and CCAP 1055/1 wild type strain (Seq_2) identified in this study. SNP-1 to SNP-16 are consistent between both strains (red text), whereas SNP-A, -B and -C were only present in *P. tricornutum* CCAP 1055/1 wild type strain (red text with yellow highlight). Exons (blue text), introns (black text) and active domains (orange text), OCT and PPRT indicate that the SNPs are found in exogenic and intronic regions, but not within the active domains of the enzyme.

Supp. Figure 2. Multiple sequence alignment of UMPS (Uniprot Accession C6L824) protein to translated protein sequences obtained from the *UMPS*-frag amplicons of WT, WT^R-1-1, WT^R-2-1, WT^R-2-2 obtained in this study. All three WT^R strains demonstrate a single point mutation (green) causing an amino acid change from isoleucine to methionine and stunted coding region, as well as a truncated protein (254 amino acids instead of the full 518 amino acids) and the loss of the PPRT active domain (pink).

CHAPTER 5

Figure 1. Schematic representation of metabolic engineering endeavours collated from various studies to drive heterologous geraniol production in yeast and bacteria. (A) Engineering approaches including exogenous expression (red) and exogenous overexpression (green) in *Saccharomyces cerevisiae,* which natively contains the MVA pathway (blue box). Overexpression of polymerase III transcription (MAF1) reduced DMAPP flux to tRNA synthesis and redirected it towards GPP synthesis, resulting in increased geraniol (Liu et al., 2013). Overexpression of rate-limiting enzymes IDI and tHMG1 increased flux to GPP synthesis resulting in increased geraniol (Zhao et al., 2016). Numerous studies have demonstrated that endogenous farnesyl diphosphate synthase (ERG20) can be modified (ERG20*) to favour GPP synthesis (EC 2.5.1.1) over FPP synthesis (EC 2.5.1.10) for increased monoterpenoid production (Ignea et al. 2014; Jiang et al., 2017; Zhao et al., 2016).

Overexpression of exogenous plant GPP synthase (GPPS) caused decreased geraniol production (Zhao et al., 2016). Fusion of genes adjacent in the pathway, endogenous ERG20 mutant (ERG20*) and exogenous GES, resulted in increased geraniol production (Zhao et al., 2016; Jiang et al., 2017). (B) Engineering approaches including exogenous expression (red) and exogenous overexpression (green) in *Escherichia coli*, which natively contains the MEP pathway (white box). Overexpression of genes involved in the MVA pathway (red) has resulted in an entirely synthetic MVA pathway (blue box) in *E. coli* for enhanced DMAPP and IPP synthesis and large yields of geraniol (Liu et al., 2016; Qian et al., 2019). Overexpression of expogenouus plant GPP synthase (GPPS) resulted in elevated geraniol production (Liu et al., 2016). *E. coli* do not naturally produce sterols, however, IPP and DMAPP are converted by FPP synthase (FPPS) into GPP (EC 2.5.1.1) and then FPP (EC 2.5.1.10) for production of sterol-like compounds called hopanoids.

Figure 2. Graphic representation of a hypothetical terpenoid metabolic network in *P. tricornutum* strain CCMP/1055. The main terpenoid pathways are highlighted in orange; in light blue, distinctive diatom enzymes involved in diatom sterol biosynthesis; in green, distinctive diatom pathways hypothetically providing substrates to isoprenoid biosynthesis; dashed lines indicate hypothetical conversions and transport reactions. Abbreviations: ACCOA, acetyl-CoA; PYR, pyruvate; GAP, glyceraldehyde 3-phosphate; 6PG, 6-phosphogluconate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate; IPP isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; IDISQS, isopentenyl diphosphate isomerase/squalene synthase; AltSQE, alternative squalene epoxidase. From Fabris et al. (2020).

Figure 3. Constitutive production of geraniol in *P. tricornutum*. (A) Growth profile of wild type diatoms and cell lines harbouring either *49202p_GES-mVenus*, *49202p_mVenus*, *21659p_GES-mVenus*, or *21659p_mVenus* (n = 3, error bar indicates SEM) (B) representative mean mVenus fluorescence of control and transgenic strains after 96 hours of cultivation in ESAW medium (20,000 cells analysed for each cell line), mean and SEM are depicted (n = 3); (C) final geraniol yield in CrGES-mVenus expressing diatoms after 168 hours, compared to empty vector and wild type controls. ND: not detected. Mean and SEM are depicted (n = 3). Identical letters denote no statistically significant differences among groups using the Tukey method. From Fabris et al. (2020).

Figure 4. Extrachromosomal expression of *Catharanthus roseus geraniol synthase* (*CrGES)* fused to *Abies grandis geranyl diphosphate synthase (AgGPPS2)* in *P. tricornutum.* (A) Schematic representation of genetic constructs designed and built in this study, pPTBR11_Phatr3_*J49202p-CrGES* (11,530 bp in size) and pPTBR11_Phatr3_*J49202p-CrGES-AgGPPS2* (12,703 bp in size). (B) Cell counts and (C) geraniol quantification at time of harvesting from 20 EE_49p_CrGES-AgGPPS2 cell lines and 20 EE_49p_CrGES cell lines analysed in the initial screening. (D) Growth and (E) geraniol quantification at time of harvesting the three best EE_49p_CrGES-AgGPPS2 and EE_49p_CrGES engineered strains and wild type control following full-scale batch cultivation experiment. (n = 3, error bars indicate SEM).

Figure 5. Effects of light regime and cultivation time on the geraniol yield in the *P. tricornutum* strain 4-GESmV-E8. (A) growth profile (B) mVenus mean fluorescence (C) chlorophyll fluorescence, of EE_49p_GmV-E8 diatoms either grown in continuous light (CL green) or light/dark (LD, blue) regime, for 7 (lighter colours) or 10 (darker colours) days (D) and (E) overall geraniol yield expressed in µg//L and in µg product/10^7 cells, respectively (n = 3, error bars indicate standard error of the mean). Identical letters denote no statistically significant differences among groups using the Tukey method**.** From Fabris et al. (2020).

Figure 6. Interaction of cytosolic geraniol biosynthesis with the endogenous terpenoid pathways in *P. tricornutum*. (A) schematic representation of hypothetical substrate subcellular allocation between geraniol, sterol and pigment biosynthesis; (B) accumulation of photosynthetic pigments in wild type (WT) and transgenic diatom lines EE-*49p-GmV-E8* (*GES-mV*) and *49202p_mVenus-1* (*mV*), after 192 hours of cultivation (C) growth profile and accumulation of geraniol in WT and representative transgenic diatom lines GES-mV and mV cultures, sampled after 96 hours of cultivation, and accumulation of main triterpenoids (n = 3, mean and SEM are depicted, asterisks indicate significant differences with control lines, identical letters denote no statistically significant differences among groups using the Tukey method). From Fabris et al. (2020).