
Electronic Thesis and Dissertation Repository

7-25-2023 10:00 AM

Advances in Phaeodactylum tricornutum nuclear engineering

Mark Pampuch, *Western University*


Supervisor: Karas, Bogumil J., *The University of Western Ontario*

Co-Supervisor: Gloor, Gregory B., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biochemistry

© Mark Pampuch 2023

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>

 Part of the [Biochemistry Commons](#), [Biotechnology Commons](#), [Genetics Commons](#), [Genomics Commons](#), [Molecular Biology Commons](#), and the [Molecular Genetics Commons](#)

Recommended Citation

Pampuch, Mark, "Advances in Phaeodactylum tricornutum nuclear engineering" (2023). *Electronic Thesis and Dissertation Repository*. 9831.

<https://ir.lib.uwo.ca/etd/9831>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

The marine diatom *Phaeodactylum tricornutum* has the potential to become an excellent platform for the sustainable production of valuable compounds and pharmaceuticals, but currently large-scale engineering of this organism remains a challenge due to factors like inefficient genetic transformation protocols and a lack of accurate genomic data. This thesis addresses these two bottlenecks by (i) optimizing an electroporation protocol to *P. tricornutum* and (ii) remapping genomic data from a scaffolded genome assembly to a telomere-to-telomere genome assembly. An optimized transformation protocol was developed that could consistently transform blunt-ended DNA with overhangs and yielded up to 1000+ colony forming units per transformation. The method of transgene integration has also been determined to be random integration via non-homologous end joining. Furthermore, the genome coordinates have been updated for 56,624 out of 69,070 annotated genome features to determine their location on the most accurate genome assembly currently available for this organism. In conclusion, the advances made here will streamline genetic engineering for this organism and enables large scale nuclear genome engineering efforts.

Keywords

Phaeodactylum tricornutum, algae, electroporation, selectable markers, nuclear genome engineering, genome annotation, synthetic biology

Summary for Lay Audience

Our planet is being destroyed in humanity's quest for natural resources, leading to things like deforestation, biodiversity loss, and climate change. To ensure that we can fulfill a demand for natural resources in the future and maintain our civilization, we need to find an alternative and sustainable ways to produce natural resources without causing any more harm to our planet. To address this, researchers have proposed using engineered microorganisms, specifically photosynthetic organisms like marine microalgae, as cell factories to produce compounds of interest. One of the best candidate microbes for this purpose is called *Phaeodactylum tricornutum*. By engineering this microbe, researchers have been able to use it to produce chemicals involved in creating plastics, pharmaceuticals, and even COVID-19 diagnostic tests. Many genetic tools have been developed for this microbe to make engineering easier, but there is still a lot that needs to be done to improve this microorganism's potential for industrial use.

My thesis focuses on improving methods to deliver and integrate custom DNA into this organism and also to accurately identify where in the organisms DNA all of its genes and other genetic information lie, as new technology has shown that where researchers previously believed all the genes were located is not actually accurate. I have been able to establish a simple, efficient, and reliable protocol to introduce custom DNA into this organism's genetic code. I've also demonstrated that this method can integrate custom DNA randomly into the organism's genetic code and can be used to inactivate and investigate the function of the organism's genes. I've also identified the correct location of 56,624 out of 69,070 genes and gene-like features in this organism's genetic code and have provided potential reasons for each of the remaining 12,446 features as to why finding out where they are actually located within the organism's DNA is a challenge. Overall, the progress I've made in this thesis brings us closer to being able to introduce larger fragments of DNA with more complicated instructions into this organism and boosts its potential for use in creating valuable resources in an environmentally friendly way.

Co-Authorship Statement

Regarding Chapter 1:

M. Pampuch and B. J. Karas conceived of experiments and M. Pampuch performed experiments and analyzed results. Under the supervision of M. Pampuch, an undergraduate student G. Tran assisted in performing many electroporations. G. Tran also designed primers to amplify *loxP* constructs and designed and amplified fragments for plasmids containing *Cre* recombinases. E. J. Walker performed yeast assembly and screening for *Cre* containing plasmids. A work-study student A. Kaneshan aided with culture maintenance and DNA isolation under the supervision of M. Pampuch. Sequencing experiments were performed with aid from J. Biltcliffe at the London Regional Genomics Center.

Regarding Chapter 2:

M. Pampuch conceptualized the remapping strategy and wrote the majority of the scripts used. An undergraduate student D. Zahid aided in writing many scripts with logic outlined by M. Pampuch, specifically with respect to data-preprocessing, performing global alignments on BLAST results, and assessing accuracy of results.

Acknowledgments

There are many people I'd like to thank for making my grad school journey a memorable one. First and foremost, my supervisor Dr. Bogumil Karas. Thank you for giving me a chance and taking me on to be a part of your ambitious research. Your optimism and love for science is contagious and I'm very thankful to have had a supervisor who was pretty much available on-demand whenever I had results to discuss.

To the many fantastic and brilliant people got to call my lab mates. Emma J. Walker, I consider you my lab mentor. Thank you for all your work helping me out when I first was getting started and for your continued support. Daniel Nucifora, your mind for puns is wizard-like, an honest talent that I hope you keep well preserved. To Jordyn and Samir, I wish I could have got to know you both more. I enjoyed the chats we had in the lab when it was just us around. You both helped me take my mind off the science when it was needed. And the lab alum, Stephanie Brumwell and Ryan Cochrane. I wish I didn't join the team as you were both heading out. The lab feels a little emptier without your charming presences. Grad school is a busy time in our lives, and I only wish we got to spend more quality time together. The past two years have flown by. I wish you all nothing but success in the future, and I hope that our paths will cross again someday.

To my team of undergrads: Danish Zahid thank you for all your help with bioinformatics and Garvin Tran thank you for all your help with the wet work. I wish you all the best in your professional schools. And last but not least thank you Agi for all your help with all my experiments. I think I speak for everyone when I say you brought a unique vibe to the lab but it was very welcome. I'll have fond memories of our late night and weekend sessions in the lab.

To the many other friends I've made in the department, specifically from the Edgell lab, the O.D. lab, and the Schild-Poulter lab: Thank you for all the levity and the escape from science

you provided. I echo what I said before. I only wish we had more time to spend. I wish you all the best and hope that we'll all stay connected into the far future.

To Farah, I feel very fortunate that I got to go through this experience with you by my side. You have been my best friend for many years, and I appreciate all the moments we have gotten to spend together. You've brought joy to my life in times when I was feeling low and brought me back down to earth when I was overextending myself. I know that you're a beautiful person inside and out and there's no one else I can see myself sharing my life with. I love you.

And last but not least to my parents: I can't tell you enough how much I appreciate you both. From leaving your country having no money and only your first language to building the life that you have now is remarkable. You've both sacrificed so much so that me and Adaś can have a chance at picking our futures and I can't wait for the day I can repay you both. I love you both more than you know.

Table of Contents

Abstract	ii
Keywords	iii
Summary for Lay Audience	iv
Co-Authorship Statement.....	v
Acknowledgments	vi
Table of Contents	viii
List of Tables	xi
List of Figures	xii
List of Abbreviations	xiv
List of Appendices	xviii
Chapter 1	1
1 Introduction	1
1.1 Synthetic Biology.....	1
1.2 Biomanufacturing	2
1.3 <i>Phaeodactylum Tricornutum</i> and the Synthetic Diatoms Project.....	4
1.4 Scope of this Thesis	7
1.4.1 Developing and Optimizing a Transformation Protocol to the <i>Phaeodactylum tricornutum</i> Nuclear Genome	7
1.4.2 Remapping Genome Annotation Data from an Old <i>Phaeodactylum</i> <i>tricornutum</i> Genome Assembly to the Telomere-to-Telomere Genome Assembly	8
1.5 References.....	8

Chapter 2.....	12
2 Developing and Optimizing a Transformation Protocol to the <i>Phaeodactylum</i> <i>tricornutum</i> Nuclear Genome	12
2.1 Introduction.....	12
2.2 Materials and Methods.....	13
2.2.1 Microbial Strains and Growth Conditions	13
2.2.2 DNA Preparation for Transformation	14
2.2.3 DNA Isolation.....	17
2.2.4 PCR Analysis of <i>P. tricornutum</i> Transformants.....	20
2.2.5 Spot Plating of <i>P. tricornutum</i> Transformants.....	21
2.2.6 DNA Sequencing	22
2.2.7 Bioinformatics Analysis.....	22
2.3 Results.....	24
2.3.1 Optimizing an electroporation protocol to <i>P. tricornutum</i>	24
2.3.2 Assaying a double marker transformation	31
2.4 Discussion.....	48
2.5 References.....	48
Chapter 3.....	55
3 Remapping Genome Annotation Data from an Old <i>P. tricornutum</i> Genome Assembly to the Telomere-to-Telomere Genome Assembly.....	58
3.1 Introduction.....	58
3.2 Materials and Methods.....	60
3.2.1 Data Extraction and Clean Up	60

3.2.2	Sequence Extraction and Prefiltering.....	61
3.2.3	Local and Global Alignments	61
3.2.4	Assembly to Assembly Mapping	62
3.2.5	Filtering and Calling Reciprocal Best Hits	63
3.2.6	Updating Coordinates	64
3.2.7	Calculating Success Rate Chromosome by Chromosome	64
3.2.8	Data availability	64
3.3	Results.....	65
3.4	Discussion.....	73
3.5	References.....	76
Chapter 4	79
4	General Discussion	79
4.1	Transformation and Remapping Data enable Large Scale Genome Engineering Efforts	79
4.2	Conclusions.....	80
4.3	References.....	81
Appendices	82
Curriculum Vitae	107

List of Tables

Table 2-1: Colony or streak counts from various transformation and re-streaking plates.....	37
---	----

Appendix B

Table B-1: Oligonucleotides used in this study.	87
--	----

Table B-2: Genetic parts used in this study.	89
---	----

Appendix C

Table C-1: Data for all electroporations performed in this study.	91
--	----

List of Figures

Figure 1-1: Planetary hazards and boundaries.	3
Figure 2-1: Cross section of plates for <i>P. tricornutum</i> electroporations.....	25
Figure 2-2: 1% agarose gel of genotyping experiment performed on cell lines transformed with the <i>nat</i> single selectable marker cassette, passed two times following initial selection.	26
Figure 2-3: 1% agarose gel of genotyping experiment performed on cell lines transformed with the <i>nat</i> single selectable marker cassette, passed seven times following initial selection.	27
Figure 2-4: 1% agarose gel of inverse PCR experiment performed on cell lines transformed with the <i>nat</i> single selectable marker cassette.	29
Figure 2-5: 1% agarose gel of genotyping experiment performed on cell lines where genomic DNA was found in inverse PCR sequencing reads.....	30
Figure 2-6: Cross section of selection plates for <i>P. tricornutum</i> electroporations.	32
Figure 2-7: Bar charts of transformation frequencies calculated without using total colony forming unit estimates.....	34
Figure 2-8: Bar charts of transformation frequencies calculated using total colony forming unit estimates.....	35
Figure 2-9: Genotyping experiment performed on cell lines transformed with the <i>shBle</i> -T2A- <i>nat</i> double selectable marker cassette, passed two times following initial selection.....	38
Figure 2-10: Spot plate of transformation lines.	39
Figure 2-11: Sequencing results from cell lines where integration of transformed cassettes was observed.	41
Figure 2-12: 1% agarose gel of genotyping experiment performed on <i>fcp</i> gene cluster.	43
Figure 2-13: 1% agarose gel of genotyping experiment performed on <i>fcpD</i> gene.	44

Figure 2-14: Genotyping experiment performed on Z-Z ₁ cell line.	46
Figure 2-15: Genotyping experiment performed on Z-Z ₂ cell line.	48
Figure 3-1: Genome-wide summary of remapped features to the telomere-to-telomere assembly of <i>P. tricornutum</i>	68
Figure 3-2: Chromosome-wide summary of remapped features to the telomere-to-telomere assembly of <i>P. tricornutum</i>	73
 Appendix A	
Figure A-1: Schematic of important transformation cassettes used in this study.	82
Figure A-2: Potential coverage plot of sequencing reads obtained from transformation line Z-Z ₁	82
Figure A-3: Potential coverage plot of sequencing reads obtained from transformation line Z-Z ₂	83
Figure A-4: Potential coverage plot of sequencing reads obtained from transformation line N-N ₁	83
Figure A-5: Potential coverage plot of sequencing reads obtained from transformation line N-N ₂	84
Figure A-6: Flowchart of remapping process.	85
Figure A-7: Diagram describing genome coordinate modifications that need to be made to mappings to remove overlaps from old assembly to new assembly mapping.	86

List of Abbreviations

Ω	Ohm(s)
$^{\circ}\text{C}$	Degree(s) Celsius
μF	Microfarad(s)
μg	Microgram(s)
μm	Micrometer(s)
μmol	Micromole(s)
μL	Microliter(s)
3'-UTR	Three prime untranslated region
5'-UTR	Five prime untranslated region
A	Adenine
$A_{260/230}$	Absorbance ratio at wavelengths of 260 and 230 nm
$A_{260/280}$	Absorbance ratio at wavelengths of 260 and 280 nm
BED	Browser Extensible Data (filetype)
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
C	Cytosine
<i>cat</i>	Chloramphenicol acetyltransferase gene
CCAP	Culture Collection of Algae and Protozoa
CDS	Coding sequence of a gene
CFU	Colony-forming unit
CM	Chloramphenicol
cm	Centimeter(s)
<i>Cre</i>	P1 bacteriophage enzyme that causes recombination
CTAB	Cetrimonium bromide
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EMBOSS	European Molecular Biology Open Software Suite
EtOH	Ethanol
<i>f/2</i>	Enriched seawater medium (" <i>f</i> medium") reduced by half

FASTA	File format for DNA, RNA, and/or amino acid sequences
FASTQ	File format for biological sequences with quality scores
<i>fcpA</i>	Fucoxanthin chlorophyll a/c binding protein A
<i>fcpB</i>	Fucoxanthin chlorophyll a/c binding protein B
<i>fcpC</i>	Fucoxanthin chlorophyll a/c binding protein C
<i>fcpD</i>	Fucoxanthin chlorophyll a/c binding protein D
G	Guanine
g	Gram(s)
gDNA	Genomic DNA
GFF	General feature format (filetype)
GXL	GC -rich templates, ex cess template, long amplicons
h	Hour(s)
kbp	Kilobase pairs(s)
<i>k</i> -mer	A subsequence of DNA of length <i>k</i>
kV	Kilovolts(s)
L	Liter(s)
L1	Enriched seawater (<i>f</i> /2 medium with more trace metals)
LB	Lysogeny broth
<i>loxP</i>	Locus of X(cross)-over in P1 bacteriophage
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
M	Molar
m	Meter(s)
MAPQ	Mapping quality score
Mbp	Megabase pair(s)
mg	Milligram(s)
min	Minute(s)
mL	Milliliter(s)
mM	Millimolar
mm	Millimeter(s)
mRNA	Messenger RNA
MPX	Multiplex
ms	Millisecond(s)

N	Notation for adenine, cytosine, guanine, thymine, or uracil
N/A	Not applicable
<i>nat</i>	<i>N</i> -acetyltransferase gene
ncRNA	Non-coding RNA
NEB	New England Biolabs
NHEJ	Non-homologous end joining
nM	Nanomolar
nm	Nanometer(s)
NP	Nitrate phosphate
NTC	Nourseothricin
OD	Optical density
OD ₇₃₀	Optical density at a wavelength of 730 nm
ONT	Oxford Nanopore Technologies
p	Plasmid
PAF	Pairwise mapping format (filetype)
PCR	Polymerase chain reaction
pH	Potential of hydrogen
Phatr3	<i>Phaeodactylum tricornutum</i> annotation 3
PPFD	Photosynthetic photon flux density
RCF	Relative centrifugal force
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RPM	Revolutions per minute
rRNA	Ribosomal RNA
s	Second(s)
sddH ₂ O	Sterile double-distilled water
<i>shBle</i>	Bleomycin-resistance gene
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
ssssDNA	Single-stranded salmon sperm DNA
T	Thymine
T2A	<i>Thosea asigna</i> virus 2A
T2T	Telomere-to-telomere

TAE	Tris-acetic acid-EDTA
TC	Time constant
TE	Tris-EDTA
tRNA	Transfer RNA
V	Volt(s)
v/v	Volume per volume
w/v	Weight per volume
WT	Wild type
ZEO	Zeocin

List of Appendices

Appendix A: Supplementary Figures.....	82
Appendix B: Primers and Genetic Parts	87
Appendix C: Electroporation Data.....	91

Chapter 1

1 Introduction

1.1 Synthetic Biology

Synthetic biology is a field of science that lies at the intersection of engineering and biology. Its core tenet is the design-build-test-learn cycle, whereby researchers are able to develop organisms with completely novel traits by first designing large-scale modifications to an organism's genetic code, "build" these organisms by integrating these edits into the organism's genome, test the effect of the newly introduced genetic material on the organism, and make novel discoveries about how these modifications affect the organism. Synthetic biology is therefore highly dependent on a researcher's ability to modify an organism's genetic code, either by altering it directly within the organism (*in vivo*) or by synthesizing novel DNA molecules and utilizing some method to introduce them directly to the organism. As technologies improve, the scale at which researchers are able to modify an organism's DNA has greatly increased. In past years, researchers would usually perform genetic modifications on the scale of one to a couple genes at a time, whereas recently various studies have been performed whereby modifications are performed on a "genome-scale", introducing tens to hundreds of modifications to an organism's genetic code at once^{1,2}. This has brought about the term "whole genome engineering". Various organisms have had their whole genome re-engineered, resulting in novel traits such as (i) organism's possessing a "minimized" genome, whereby all genetic content not essential for the organism's survival having been removed³, (ii) organism's possessing a "refactored" genetic code, wherein the organism's internal biological language has been modified making it resistant to many natural pathogens such as virus⁴, and (iii) organism's possessing a "humanized" genome, whereby an organisms genes are replaced with their

closest human counterpart to establish a more human-like intracellular environment⁵. Being able to introduce large scale modifications to an organism's genome has enabled researchers to ask grand questions and gain invaluable insights into the biological workings of many organisms, but it also instills tremendous industrial potential to design specific organisms to address real-world issues.

1.2 Biomanufacturing

For 12,000 years prior to the Industrial Revolution, the global mean temperature has only varied only plus or minus 1°C. This stable period in our earth's history gave way to stable oceans, reliable weather patterns, and has made human civilization possible⁶. However, the stress humans have placed on the planet, specifically in the quest for resources has already disrupted this balance. A study by Rockstrom et al.,⁷ identified different planetary hazards and proposed a safe threshold for all these is for humanity to operate within. The researchers showed that many safe operating boundaries for the planet have been transgressed back in 2009 (Figure 1-1).

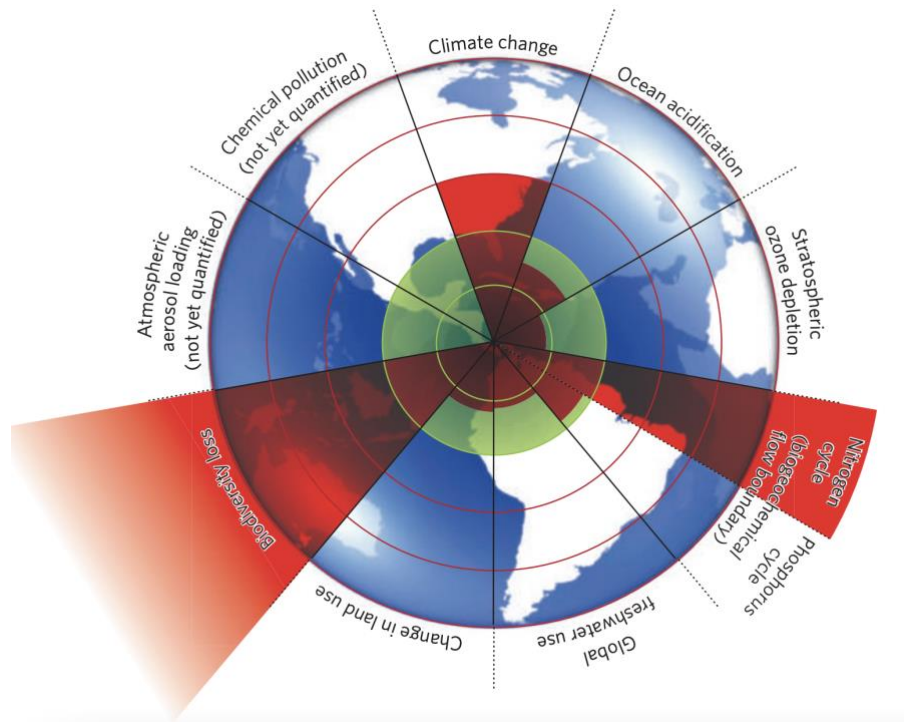


Figure 1-1: Planetary hazards and boundaries. Green shading indicates safe operating thresholds for each potential hazard. Red shading indicates where the researchers quantified the operating level was for each of these hazards back in 2009. Figure obtained from Rockstrom et al,⁷.

In the years since the release of the study, the situation has only worsened as green technology initiatives have not been rolled at a rate concordant with climate scientists demands⁸. This is because traditional manufacturing often depends on highly energy intensive and environmentally damaging mechanical and chemical processes. In order to help restore the health of the planet and meet the needs of an increasing global population, humanity needs to find alternative methods for producing natural resources. This is why a transition to a bio-based economy has been a proposed solution^{9,10}.

Biomanufacturing is a subdomain of synthetic biology that involves utilizing engineered organisms for the production of valuable compounds. It leverages the self-replicating nature of biology to scale up the manufacture process of desired products, whereby a production strain of an organism is developed through the introduction of a gene or

metabolic pathway encoding the desired product, and then cultures of the strain can generate this product provided the correct media is supplemented. This concept has been termed “microbial cell factories”. Some of the compounds microbial cell factories have been used to produce include pharmaceuticals such as insulin¹¹, plastic polymers¹², petrochemicals¹³, and even alternative meat products¹⁴.

This kind of biomanufacturing is currently primarily done in engineered *Escherichia coli* and yeasts, however this is more a result of legacy reasons rather than any innate biological properties that make these organisms better for biomanufacturing. Yeast and *E. coli* have been model organisms for decades therefore a lot of research has been performed on their underlying biology and many genetic tools have been developed for them which make them relatively easy to engineer. However, there are some significant disadvantages to using them as industrial production platforms. They are heterotrophic organisms, meaning they require a sugar or lipid-based carbon source to maintain their survival. Industrially scaling these organisms would create greater demands for sugars and would also contribute to a lot of the problems we face with modern agriculture. Researchers have proposed utilizing photosynthetic organisms as a production platform to circumvent this issue. These production platforms can utilize carbon dioxide as a carbon source and can act as carbon sinks when industrial scaled for manufacturing⁹.

1.3 *Phaeodactylum Tricornutum* and the Synthetic Diatoms Project

Microalgae have emerged as leading candidates in terms of photosynthetic organisms to be used as production platforms, and chief amongst them are those belonging to the class Bacillariophyceae, also known as diatoms. Diatoms are a very ecologically important group of algae, have comprise nearly half of all organic material in the oceans and

contribute to a large proportion of oxygen in our atmosphere¹⁵. They outcompete a lot of other microorganisms for space and resources, and their ability to live in a multitude of niches indicates a lot of metabolic capabilities that can be tapped into to optimize resource production⁹. One unique trait diatoms possess is the ability to metabolize silicon and incorporate it into their cell wall, creating an outer glass-like shell known as a silica frustule. This causes a problem for scientists and bioengineers working with these organisms as studying and genetically manipulating organisms with silica incorporated into their cell walls has proven to be a challenge¹⁶. However, a subset of diatoms exist that do not require silicon in their cell wall, one of these being the model diatom *Phaeodactylum tricornutum*⁹. On top of being a well-studied diatom, *P. tricornutum* holds a lot of potential to make a great chassis for biomanufacturing due to its tolerance to high pH, ability to grow under low light conditions, and ability to outcompete other microalgae in outdoor cultivars⁹. Some of the compounds that *P. tricornutum* has been used to produce are native compounds, such as omega-3 fatty acids, and non-native compounds, some of these being polyhydroxybutyrates, which are polyesters used in making biodegradable plastics, plant terpenoids used for anti-tumor and anti-inflammatory medications, and monoclonal antibodies⁹. Another remarkable thing about *P. tricornutum* is that unlike other production platforms like yeast and *E. coli*, it contains glycosylation pathways that are very similar to those found in humans, which gives it value as platform for pharmaceutical compounds¹⁷. One example of this is Slattery et al.,¹⁷ wherein the researchers were able transgenically express SARS-CoV-2 spike proteins in *P. tricornutum* to be used in coronavirus serological tests for humans.

Despite the potential that *P. tricornutum* holds for developing clean technologies, performing genetic manipulations in this organism remains a challenge. The whole host of genetic tools have been developed for *P. tricornutum*, including well characterized endogenous promoters and terminators^{18,19}, inducible expression systems²⁰, *P. tricornutum*-specific introns¹⁹, CRISPR/Cas systems^{19,21,22}, auxotrophic strains²³, and a selection/counter-selection system orthologous to the URA3/5-FoA system in yeast²³, however this suite of tools only enable relatively few genetic edits to be made at a time, and using them to perform larger-scale genome engineering would be very laborious and require many iterations. In order to accelerate *P. tricornutum*'s use as a production platform, a large-scale genome reengineering project titled the Synthetic Diatoms project was proposed to alleviate many of the current engineering constraints and accelerate strain development for biotechnological purposes²⁴. The rationale for this project is that currently reading DNA is inexpensive and relatively simple, rapid improvements are being made in the field of DNA synthesis, and an outstanding challenge remains in delivering and integrating large pieces of DNA. The primary mechanism for large scale genome engineering proposed in the Synthetic Diatoms project is refactor the nuclear genome of *P. tricornutum* into smaller chromosomes that can be captured on multi-host shuttle plasmids and stably maintained within organisms that contain more robust DNA engineering capabilities. Following genetic manipulations to the chromosomes of interest, the chromosome can be reintroduced into *P. tricornutum* through bacterial conjugation, whereby some method of removing the chromosome from the vector backbone and curing the native chromosome would need to be employed. Such a method enables many genetic

modifications and whole metabolic pathways to be developed with relative ease and introduced into *P. tricornutum* with a single step.

Such a large-scale genome re-engineering project is primarily made possible by three significant advances in *P. tricornutum* research. (i) A telomere-to-telomere assembly of the *P. tricornutum* genome was recently created, determining the true chromosome count of this organism's nuclear genome²⁵, (ii) multi-host shuttle plasmids that can stably maintain large portions of *P. tricornutum* DNA in *P. tricornutum* and other model organisms like *E. coli* and *Saccharomyces cerevisiae* have been developed^{26,27}, and (iii) an efficient method of delivering large pieces of DNA into *P. tricornutum* has been achieved using bacterial conjugation²⁶. Although these advancements have streamlined engineering efforts for this organism, various bottlenecks still exist in the engineering process that limit its biotechnological potential.

1.4 Scope of this Thesis

This thesis consists of methods and protocols developed to overcome various engineering constraints and acquire missing information within *P. tricornutum*'s genetic code in order to accelerate diatom research and provide a better foundation for a large-scale genome engineering project to be attempted.

1.4.1 Developing and Optimizing a Transformation Protocol to the *Phaeodactylum tricornutum* Nuclear Genome

This chapter describes the creation of an efficient and reliable transformation protocol for *P. tricornutum* and an investigation of how DNA is integrated and maintained within this organism.

1.4.2 Remapping Genome Annotation Data from an Old *Phaeodactylum tricornutum* Genome Assembly to the Telomere-to-Telomere Genome Assembly

This chapter describes the process of extracting all the genomic data from an outdated *P. tricornutum* reference genome and accurately remapping the genomic coordinates to the most accurate reference genome that currently exists for this organism.

1.5 References

1. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
2. Thompson, D. B. *et al.* The Future of Multiplexed Eukaryotic Genome Engineering. *ACS Chem. Biol.* **13**, 313–325 (2018).
3. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
4. Nyerges, A. *et al.* A swapped genetic code prevents viral infections and gene transfer. *Nature* **615**, 720–727 (2023).
5. Kachroo, A. H., Vandelloo, M., Greco, B. M. & Abdullah, M. Humanized yeast to model human biology, disease and evolution. *Dis. Model. Mech.* **15**, dmm049309 (2022).
6. Steffen, W. *et al.* Sustainability. Planetary boundaries: guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
7. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
8. Costandi, M. Climate change extinction threat to double the number at risk. *Nat. Afr.* (2023) doi:10.1038/d44148-023-00136-w.

9. Butler, T., Kapoore, R. V. & Vaidyanathan, S. *Phaeodactylum tricornutum*: A Diatom Cell Factory. *Trends Biotechnol.* **38**, 606–622 (2020).
10. Naseri, G. A roadmap to establish a comprehensive platform for sustainable manufacturing of natural products in yeast. *Nat. Commun.* **14**, 1916 (2023).
11. Ladisch, M. R. & Kohlmann, K. L. Recombinant human insulin. *Biotechnol. Prog.* **8**, 469–478 (1992).
12. Zhang, X., Lin, Y., Wu, Q., Wang, Y. & Chen, G.-Q. Synthetic Biology and Genome-Editing Tools for Improving PHA Metabolic Engineering. *Trends Biotechnol.* **38**, 689–700 (2020).
13. Jarboe, L. R. *et al.* Metabolic engineering for production of biorenewable fuels and chemicals: contributions of synthetic biology. *J. Biomed. Biotechnol.* **2010**, 761042 (2010).
14. Wang, G., Wu, X. & Yin, Y. Synthetic biology-driven customization of functional feed resources. *Trends Biotechnol.* **40**, 777–780 (2022).
15. Roberts, K., Granum, E., Leegood, R. C. & Raven, J. A. Carbon acquisition by diatoms. *Photosynth. Res.* **93**, 79–88 (2007).
16. Rogato, A. *et al.* *Phaeodactylum tricornutum* as a model organism for testing the membrane penetrability of sulphonamide carbonic anhydrase inhibitors. *J. Enzyme Inhib. Med. Chem.* **34**, 510–518 (2019).
17. Slattery, S. S. *et al.* Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom *Phaeodactylum tricornutum* for pandemic diagnostics. *Sci. Rep.* **12**, 7010 (2022).

18. Windhagauer, M. *et al.* Characterisation of novel regulatory sequences compatible with modular assembly in the diatom *Phaeodactylum tricornutum*. *Algal Res.* **53**, 102159 (2021).
19. Slattery, S. S. *et al.* An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **7**, 328–338 (2018).
20. Kassaw, T. K., Paton, A. J. & Peers, G. Episome-Based Gene Expression Modulation Platform in the Model Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **11**, 191–204 (2022).
21. Stukenberg, D., Zauner, S., Dell'Aquila, G. & Maier, U. G. Optimizing CRISPR/Cas9 for the Diatom *Phaeodactylum tricornutum*. *Front. Plant Sci.* **9**, (2018).
22. Serif, M. *et al.* One-step generation of multiple gene knock-outs in the diatom *Phaeodactylum tricornutum* by DNA-free genome editing. *Nat. Commun.* **9**, 3924 (2018).
23. Slattery, S. S. *et al.* Plasmid-based complementation of large deletions in *Phaeodactylum tricornutum* biosynthetic genes generated by Cas9 editing. *Sci. Rep.* **10**, 13879 (2020).
24. Pampuch, M., Walker, E. J. L. & Karas, B. J. Towards synthetic diatoms: The *Phaeodactylum tricornutum* Pt-syn 1.0 project. *Curr. Opin. Green Sustain. Chem.* **35**, 100611 (2022).
25. Giguere, D. J. *et al.* Telomere-to-telomere genome assembly of *Phaeodactylum tricornutum*. *PeerJ* **10**, e13607 (2022).

26. Karas, B. J. *et al.* Designer diatom episomes delivered by bacterial conjugation. *Nat. Commun.* **6**, 6925 (2015).
27. Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E. & Weyman, P. D.
Refinement of the Diatom Episome Maintenance Sequence and Improvement of
Conjugation-Based DNA Delivery Methods. *Front. Bioeng. Biotechnol.* **4**, 65 (2016).

Chapter 2

2 Developing and Optimizing a Transformation Protocol to the *Phaeodactylum tricornutum* Nuclear Genome

2.1 Introduction

The ability to genetically modify an organism is often predicated by the ease of which it is possible to deliver and express foreign DNA. However, in *P. tricornutum*, the efficiency in established transformation mechanism pale in comparison to other model organisms, constraining its potential as a synthetic biology chassis¹. Exogenous DNA has historically been delivered to *P. tricornutum* via two primary methods: bacterial conjugation and biolistic transformation². DNA delivery through bacterial conjugation is mediated by developing a conjugative vector harboring the transgenic material of interest and having expression occur on an extrachromosomal fragment termed an episome. This has been shown to be an effective method for reliably introducing large pieces of DNA and has been shown to be immune from positional effects disrupting gene expression². However, this approach has certain drawbacks. Certain studies have shown higher levels of leaky expression on genes introduced on episomes compared to those integrated into the nuclear genome³. Additionally, the type of genetic cargo is limited by the fact that it must not be toxic or overly metabolically taxing on the donor bacteria in order for this delivery mechanism to occur⁴. On the other hand, biolistic transformation is the direct integration of genetic material randomly into the organism's genome. It is performed by lacing genetic material of interest onto gold or tungsten particles and penetrating the cell host cell via a high-pressure blast. This type of transformation has been shown to be an effective method to deliver DNA to organelles or organisms with rigid cell walls, however it suffers from a range of side-effects such as fragmentation of the DNA cargo, off-target effects, and

varying expression levels². Electroporation has often been viewed as a gentler alternative for random DNA delivery, especially into the nuclear genome. Various studies have outlined electroporation protocols to *P. tricornutum*^{5,6}, but the reported transformation frequencies are orders of magnitude lower than those reported for in other organisms⁷, indicating that there may be potential for vast improvements to be made. Therefore, the objective outlined here is to optimize an electroporation protocol for DNA delivery to *P. tricornutum*.

2.2 Materials and Methods

2.2.1 Microbial Strains and Growth Conditions

Phaeodactylum tricornutum (Culture Collection of Algae and Protozoa CCAP 1055/1) was grown in enriched seawater (L1) media without silica. Cultures were grown at 18°C under cool white, fluorescent lights (75 $\mu\text{mol m}^{-2} \text{s}^{-1}$ of PPFD) and a photoperiod of 16 h light: 8 h dark. Cultures that were grown in falcon tubes had their caps loosened to ensure they received sufficient airflow. To make 1 L of L1 media without silica, 1 L of aquil salts, 2 mL of a 500X nitrate phosphate (NP) stock solution, 1 mL of a 1000X trace metal stock solution and 500 μL of a 2000X *f/2* vitamins stock solution was mixed. 1 L of aquil salts comprised of 500 mL of anhydrous and 500 mL of hydrous salts. Anhydrous salts solution was comprised of 838mM NaCl, 58mM Na₂SO₄, 19mM KCl, 5mM NaHCO₃, 1.7mM KBr, 970 μM H₃BO₃, and 143 μM NaF. Hydrous salts solution was comprised of 109mM MgCl₂ · 6H₂O and 21mM CaCl₂ · 2H₂O. NP stock solution was comprised of 4.4M NaNO₃ and 180mM NaH₂PO₄ · H₂O. 1000X trace metal stock solution was comprised of 12mM FeCl₃ · 6H₂O, 12mM Na₂EDTA · 2H₂O, 9.8 μM CuSO₄ · 5H₂O, 78 μM Na₂MoO₄ · 2H₂O, 76.5 μM ZnSO₄ · 7H₂O, 42 μM CoCl₂ · 6H₂O, 910 μM MnCl₂ · 4H₂O, 10 μM H₂SeO₃, 10 μM NiSO₄ · 6H₂O, 10 μM Na₃VO₄ and 10 μM K₂CrO₄. 2000X *f/2*

vitamins stock solution was comprised of 593mM thiamine-HCl, 4nM biotin, and 0.7nM cyanocobalamin. Media was adjusted to a pH of 8.0 with 2M NaOH or 3M HCl. L1 media was sterilized through vacuum filtration through a 0.2- μ m filter.

2.2.2 DNA Preparation for Transformation

The *nat* single selectable marker cassette was amplified using primers BK2510_F and BK2510_R using pPtGE27⁸ as template DNA. The *shBle*-T2A-*nat* double selectable marker cassette was amplified using primers BK2821_F and BK2821_R using pPtGE32⁸ as template DNA. pAL fragment used as a transformation control was amplified using primers BK1980_F and BK1981_R. All primer sequences can be found in Table B-1

. PCR was performed using PrimeSTAR® GXL DNA Polymerase (Takara Bio) according to manufacturer's instructions. PCR purification was performed using a modified version of the EZ-10 Spin Column PCR Products Purification Kit (BioBasic). 400 μ L of PCR product was pooled into two 1.7 mL Eppendorf tubes (200 μ L of solution into each tube). 1.2 mL of B3 buffer with isopropanol was added into each tube. 750 μ L of the solution was transferred into two EZ-10 spin columns and incubated at room temperature for 2 minutes. The solution was centrifuged at 10,000 RPM for 2 minutes. The flow-through was transferred back into the EZ-10 spin column, left to incubate at room temperature for another 2 minutes, and centrifuged again at 10,000 RPM for 2 minutes. This process was repeated on the flow-through one more time. The remaining 650 μ L of PCR product / B3 solution was transferred into the same two EZ-10 spin columns, and the same three incubation and centrifugation steps were performed. Once complete, 750 μ L of Wash Solution was added into the EZ-10 spin column, centrifuged at 10,000 RPM for 2 minutes and flow-through was discarded. The previous wash step was performed one more time.

The sample was centrifuged at 10,000 RPM for 1 more minute to remove any residual Wash Buffer from the spin column. The spin column was then transferred to a fresh 1.7 mL Eppendorf tube. To elude the DNA, 30 μ L of ddH₂O pre-heated to 80°C was added into the EZ-10 spin column, left to incubate at room temperature for 10 minutes, and then the sample centrifuged at 10,000 RPM for 2 minutes. DNA was measured on the DeNovix DS-11 Series Spectrophotometer / Fluorometer using the dsDNA application on microvolume mode and the leftover ddH₂O used for elution to blank the samples.

DNA Transfer to *P. tricornutum* Transformants via Electroporation

600 mL of shaking *P. tricornutum* cultures were harvested at early- to mid-exponential phase (OD₇₃₀ of 0.2 – 0.38 or 2×10^6 – 5×10^6 cells/mL). All the following steps were performed aseptically. 600 mL of culture was split into twelve 50 mL falcon tubes and centrifuged at 2000 RCF for 15 min at 4°C. From each falcon tube, the supernatant was decanted gently into a waste beaker without disturbing the pellet. Leftover media was removed via aspiration. 1 mL of ice-cold 375mM D-Sorbitol was added to one of the 50 mL falcon tubes and the cell pellet was mixed by pipetting. To ensure a proper resuspension of the pellet, the mixing was performed with the following points in mind. Pipetting was performed not too vigorously to ensure that the cells did not shoot up into the pipette, however it was performed forcefully enough to dislodge the pellet. The cells were also physically scraped off the sides of the tube with the pipette tip to ensure they got mixed into the solution. Once the pellet was resuspended, the entire mixture was transferred into one of the remaining 50 mL falcon tubes and the pellet in that tube was resuspended as previously described. This process was repeated until all twelve pellets were resuspended in a small volume of ice-cold 375mM D-Sorbitol. The resuspended cells were split evenly

into two sterile 1.7 mL Eppendorf tubes (was split into more than two tubes if the final volume of the solution containing the resuspended cells was more than 3 mL). The cells were centrifuged for 5 min at 2000 RCF, the supernatant was removed via aspiration, and the cells were resuspended in 750 μ L – 1000 μ L of 375mM D-Sorbitol. This step was repeated four more times until the cells have been washed a total of five times in 375mM D-Sorbitol. After the last wash, the pelleted cells were resuspended in a small volume of ice-cold 375mM D-Sorbitol (approximately 300 μ L – 400 μ L). In another 1.7 μ L Eppendorf tube, a 1:250 dilution of washed cells was made using 375mM D-Sorbitol as the diluent and this dilution was measured on the DeNovix CellDrop™ Automated Cell Counter using the propidium iodide (PI) app to calculate the concentration of washed cells. The total cells/mL output was used as the concentration value. The concentration of the washed cells was adjusted to 2×10^9 cells/mL using ice-cold 375mM D-Sorbitol as the diluent. 100 μ L aliquots of washed cells were made into sterile 1.7 μ L Eppendorf tubes. Into each aliquot of washed cells, 3 μ g or up to 6 μ L of DNA to be transformed was added. A 10 μ g/ μ L solution of single-stranded salmon sperm DNA (ssssDNA) was pre-heated at 95°C for 10 min and cooled on ice for 5 min. Following cooling, 4 μ L of the sssDNA solution was added to each transformation mixture. Each transformation solution was gently flicked to mix the components and then incubated on ice for 10 min. The transformation solutions were transferred into 2mm electroporation cuvettes pre-cooled on ice, dispensing the solutions into the loading wells and against the side of the electroporation cuvettes to ensure a clean transfer. The electroporation cuvettes were wiped to with a kimwipe to remove any moisture on the electrodes and then placed inside a Gene Pulser Xcell™ Electroporator (Bio-Rad). Cells were pulsed with the following electrical

parameters: 500 V, 50 μ F, 400 Ω . Immediately after pulsing, 1 mL of L1 media was added to the electrocuvette and cells quickly resuspended by pipetting. Resuspended cells were then transferred to a 15 mL falcon tube containing 9 mL of L1 media. Cultures were then incubated for 24h at 18°C under cool white, fluorescent lights ($75 \mu\text{mol m}^{-2} \text{s}^{-1}$ of PPFD) and a photoperiod of 16 h light: 8 h dark with the falcon tube caps loosened. Following this recovery period, transformed cultures were centrifuged at 2000 RCF at 18°C for 15 min. The supernatants were decanted gently into a waste beaker without disturbing the pellets. Leftover media was removed via aspiration. Cell pellets were resuspended in 1 mL of L1 media and desired dilutions of cells (typically 10^{-1}) were plated on plates containing 1% agar, $\frac{1}{2}$ L1 media, and desired concentration of antibiotics. Plated cultures were placed into a clean plastic square tray to retain moisture inside the plates and then incubated at 18°C under cool white, fluorescent lights ($75 \mu\text{mol m}^{-2} \text{s}^{-1}$ of PPFD) and a photoperiod of 16 h light: 8 h. Colonies appeared after 10 – 14 days.

2.2.3 DNA Isolation

P. tricornutum DNA was isolated using a modified alkaline lysis protocol or a phenol chloroform isolation protocol. The steps performed for alkaline lysis were as follows: to isolate DNA from liquid cultures, 5 – 10 mL of liquid culture was harvested during exponential phase. To isolate DNA from solid plates, streaked cultures (1 – 2 cm in length) were scraped into 1 mL of L1 media and mixed well via pipetting. Cells were pelleted at 4000 RCF for 10 min and the supernatant was discarded. Cells were resuspended in 250 μ L of resuspension buffer, consisting of 235 μ L of Buffer P1 (QIAGEN) with RNase A, 5 μ L of a 100 mg/mL hemicellulase stock solution, 5 μ L of a 25 mg/mL lysozyme stock solution, and 5 μ L of a zymolase solution (consisting of 200mg of zymolase 20T (US Biological Life Sciences), 9 mL of ddH₂O, 1 mL of 1M Tris pH 7.5, and 10 mL of 50%

glycerol). Samples were mixed by gently pipetting and were then incubated at 37°C for 1h. Following incubation, 250 µL of Buffer P2 (QIAGEN) was added to lyse the cells and samples were vortexed for 30s each. 250 µL of Buffer P3 (QIAGEN) was added to neutralize the reaction and samples were mixed well via shaking and inversion. Samples were spun down at 14,000 RPM for 10 minutes and the supernatant was transferred to sterile 1.7 mL Eppendorf tubes. To precipitate the DNA, 750 µL of ice-cold isopropanol was added to the samples and the samples were mixed via inversion and spun down at 14,000 RPM for 10 minutes. The supernatant was removed and 750 µL of ice-cold 70% ethanol (EtOH) was added to the samples and the samples were mixed via inversion and spun down at 14,000 RPM for 10 minutes. The supernatant was removed and residual EtOH was removed from the samples by gently tapping them on a paper towel, and by aseptically aspirating any visible EtOH in the tube, being mindful to not aspirate near the DNA pellet. Samples were left uncapped for 15 minutes to allow for any residual EtOH to evaporate. 50 µL of ddH₂O pre-heated to 56°C was then added to the samples and left to dissolve overnight at 4°C. DNA was measured on the DeNovix DS-11 Series Spectrophotometer / Fluorometer using the dsDNA application on microvolume mode and the leftover ddH₂O used for elution to blank the samples.

The steps performed for phenol chloroform extraction were as follows: 50 – 200 mL of liquid culture was harvested during exponential or stationary phase. The culture was spun down at 4000 RCF for 10 min at 4°C and the supernatant was discarded. The remaining pellet was resuspended in TE buffer and added dropwise into a mortar pre-cooled at 80°C and surrounded by liquid nitrogen. The frozen droplets were crushed and grinded into a fine powder using a pestle. Liquid nitrogen was added periodically to the surrounding

environment of the mortar to keep cells cool. The ground up cells were transferred into a 15 mL falcon tube and 2 mL of lysis buffer containing 1.4 M NaCl, 200 mM Tris-HCl, 50 mM EDTA, 2% (w/v) CTAB, RNase A (250 µg/mL) and Proteinase K (100 µg/mL) was added. The solution was mixed via inversion and incubated for 1h at 37°C. The solution was then centrifuged at 3000 RCF for 10 min and the supernatant was transferred into a new 15 mL falcon tube. One volume of UltraPure™ Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v; Invitrogen™) was added to the solution and mixed via inversion. The solution was then centrifuged at 3000 RCF for 10 min and the aqueous phase was transferred into a new 15 mL falcon tube. One volume of chloroform was added and mixed via inversion. The solution was then centrifuged at 3000 RCF for 10 min. Up to 450 µL of the aqueous phase was transferred to a sterile 1.7 mL Eppendorf tube. Two volumes of ice-cold 100% EtOH and a 1/10th volume of 3M NaAc pH 5.2 were added and the solution was mixed via inversion. The supernatant was removed and 500 µL of ice-cold 70% EtOH was added to the solution. The solution was mixed via inversion and spun down at 14,000 RPM for 10 minutes. This step was repeated for a total of two 70% EtOH washes. The supernatant was removed and residual EtOH was removed from the samples by gently tapping them on a paper towel, and by aseptically aspirating any visible EtOH in the tube, being mindful to not aspirate near the DNA pellet. Samples were left uncapped for 15 minutes to allow for any residual EtOH to evaporate. 50 µL of sddH₂O pre-heated to 56°C was then added to the samples and left to dissolve overnight at 4°C. DNA was measured on the DeNovix DS-11 Series Spectrophotometer / Fluorometer using the dsDNA application on microvolume mode and the leftover sddH₂O used for elution to blank the samples.

2.2.4 PCR Analysis of *P. tricornutum* Transformants

Genotyping of the cell lines transformed with the *nat* single selectable marker cassette, the *fcpA* region, the *fcpD* region, and the integration loci was performed using PrimeSTAR® GXL DNA Polymerase (Takara Bio) according to manufacturer's instructions and primers BK2598_F, BK2598_R, BK2632_F, BK2632_R, BK2935_F, BK2935_R, BK2998_F, BK2998_R, BK3000_F, BK3000_R, BK3003_F, BK3003_R, BK3004_F, BK3004_R, BK3006_F, BK3006_R, BK3009_F, and BK3009_R (Table B-1). Genotyping of cell lines transformed with the *shBle*-T2A-*nat* double selectable marker cassette and the *fcp* gene clusters was performed using the QIAGEN Multiplex PCR Kit according to manufacturer's instructions. The primer mix for the *shBle*-T2A-*nat* cassette genotyping contained BK2923_F, BK2923_R, BK2928_F, BK2928_R, BK2937_F, and BK2937_R (Table B-1). The primer mix for *fcp* gene cluster genotyping contained BK2935_F, BK2935_R, BK2936_F, BK2936_R, BK2937_F, BK2937_R, BK2938_F, and BK2938_R (Table B-1). Inverse PCR was done by first performing *SalI*-HF and *NcoI*-HF digests on DNA isolated from transformation lines. Digestion was performed using 2 µL of rCutsmart buffer (NEB), 0.4 µL of *SalI*-HF or *NcoI*-HF enzyme solution (NEB), 7.6 µL of *sddH₂O*, and 10 µL of isolated DNA per reaction. *SalI*-HF digestion was performed for 1h at 37°C and inactivated at 65°C for 20min. *NcoI*-HF digestion was performed for 1h at 37°C and inactivated at 80°C for 20min. Ligation was performed using 1 µL of 10X T4 Ligase buffer (NEB), 0.5 µL of T4 Ligase (NEB), 16.5 µL of *sddH₂O*, and 2 µL of restriction digest product per reaction. PCR was performed using PrimeSTAR® GXL DNA Polymerase (Takara Bio) according to manufacturer's instructions. The elongation time was set 4min based on product size estimates through restriction digest analysis. The primers used for

amplification were BK2618_F and BK2618_R (Table B-1). All samples were prepped and sent out for sanger sequencing using BK2618_F.

2.2.5 Spot Plating of *P. tricornutum* Transformants

P. tricornutum cultures were transferred to individual falcon tubes and centrifuged at 2000 RCF for 15 min at 18°C. The supernatants were decanted gently into a waste beaker without disturbing the cell pellets. Leftover media was removed via sterile aspiration. Cell pellets were then resuspended in 1 mL of L1 media. The concentrations of the resuspended cells were measured on a DeNovix CellDrop™ Automated Cell Counter using the propidium iodide (PI) app to calculate the concentration of washed cells. The total cells/mL output was used as the concentration value. All the following steps were performed aseptically. The resuspended cells were all diluted with L1 media so that the concentration of all the cultures was equal to the lowest concentration measured on the cell counter. 100 µL of the cell suspensions were then transferred to wells on a 96-well PCR plate. For every desired dilution, 90 µL of L1 media was added to another well in the PCR plate. For each sample, 10 µL of cell suspension was mixed with 90 µL of L1 media to make the first dilution. 10 µL of the first set of diluted cells were mixed with 90 µL of L1 media to make the second set of dilutions. This process was repeated until all samples were diluted to their desired amounts. The final volume for all diluted and undiluted samples in the well plate was 90 µL. Plates containing 1% agar, ½ L1 media, and desired concentrations of antibiotics were prepared and air-dried in a biosafety cabinet for 1h before use. Once dried, 5 µL of the samples were plated onto the plates by a flame. The samples were airdried for 15 min before being placed into sterile plastic square trays and then incubated at 18°C under cool white, fluorescent lights (75 µmol m⁻² s⁻¹ of PPFD) and a photoperiod of 16 h light: 8 h. Plates were pictures 14 days after plating.

2.2.6 DNA Sequencing

DNA sequencing was performed using the Oxford Nanopore Technologies (ONT) MinION R9.4.1 Flow Cell with the Rapid Barcoding Kit (SQK-RBK004) on *P. tricornutum* genomic DNA (gDNA). Briefly, gDNA concentration was measured using a Qubit™ 2.0 Fluorometer (Thermo Fisher Scientific) using the Qubit™ dsDNA High Sensitivity buffer and gDNA purity assayed using A_{260/230} and A_{260/280} readings obtained from the DeNovix DS-11 Series Spectrophotometer / Fluorometer using the dsDNA application on microvolume mode. Sequencing library preparation was performed according to the ONT Rapid sequencing gDNA - barcoding (SQK-RBK004) protocol. Sequencing runs were performed for 24 – 72h. Basecalling was performed using the ONT Guppy basecaller v6.5.7 using the super accuracy model.

2.2.7 Bioinformatics Analysis

In order to locate the transformed constructs in the sequencing reads, raw sequencing reads were converted from FASTQ to FASTA file formats and FASTA files were passed into the makeblastdb application of the command line blast package (v2.6.0) to create a BLAST database of the sequencing reads. A FASTA file containing the sequence of the transformed constructs was searched against the BLAST database of the sequencing reads using the blastn application of the command line blast package. BLAST outputs were manually verified by importing the sequencing reads into Benchling and manually annotating the transformation construct elements. The sequences upstream and downstream of the integration events were inputted into BLAST and searched against the *P. tricornutum* telomere-to-telomere genome assembly (2021, GenBank accession ID: GCA_914521175.1) with the *P. tricornutum* mitochondria and chloroplast genome assemblies appended (NCBI Reference Sequences: NC_016739.1 and NC_008588.1,

respectively) and the Diatom Consortium genome assembly (2008, GenBank accession ID: GCA_000150955.2) with mitochondria and chloroplast genome assemblies appended to ensure integration sites identified by BLAST were accurate and consistent across genome assemblies. In order to design primers for genotyping, the upstream and downstream regions of the integration sites, as well as the transformation construct elements that were present in the sequencing reads, were reconstructed using the sequence information from the telomere-to-telomere and the sequence of the transformed constructs. Primers were then designed with respect to the sequence observed in the sequencing reads, and also with respect to the reconstructed sequences in order to mitigate any potential errors in the sequencing reads from affecting PCR success. To assess potential sequence coverage, all the sequencing reads for each sample were pooled and aligned to the *P. tricornutum* telomere-to-telomere genome assembly with the *P. tricornutum* mitochondria and chloroplast genome assemblies appended using minimap2 (v2.22-r1101, using the -x map-ont flag to optimize for Oxford Nanopore Sequencing reads and outputted in PAF format). Potential coverage was visualized using a modified coverage plot function obtained from the pafr library (v0.0.2). Genome-wide restriction analysis was performed using the restriction enzyme package from biopython (v1.80). Briefly, a restriction digest analysis was performed on the transformed constructs using all commercial enzymes. Enzymes that had zero cut sites were isolated and then a restriction digest analysis was performed on the *P. tricornutum* telomere-to-telomere and Diatom Consortium genome assemblies using this batch of enzymes. Enzymes that demonstrated the shortest average and median product sizes and that also had non-ambiguous restriction sites (only A's, C's, G's, or T's present)

and were capable of being heat-inactivated were selected as candidates for the inverse PCR screens to locate transgenes.

2.3 Results

2.3.1 Optimizing an electroporation protocol to *P. tricornutum*

The electroporation protocol detailed in Kassaw et al.³ was chosen as a starting point in order to optimize an electroporation protocol for *P. tricornutum*. Several attempts at replicating the protocol were made using plasmids harboring nourseothricin resistance (pPtGE27 and pPtGE31, from Slattery et al.⁸) and linearized with restriction enzymes prior to transformations (using *XhoI* and *BmtI*-HF, respectively), however all of these transformation attempts yielded 0 colony-forming units (CFUs) up to 21 days following transformation. A linear DNA construct containing just the nourseothricin resistance gene (*nat*), flanked by a promoter/terminator pair shown to exhibit strong gene expression in *P. tricornutum* (*fcpD* promoter and *fcpA* terminator, both endogenous regulatory elements to *P. tricornutum*⁸) was then amplified and used to reattempt this protocol (Figure A-1A). *XhoI* restriction sites were added to the end of the constructs via PCR primers in order to have the ability to generate 5' overhangs within the transformation construct as all transformations performed in Kassaw et al. were linearized with restriction enzymes that produced 5' overhangs (including *XhoI*). Attempts to replicate this protocol using both purified PCR products (blunt ended) and constructs with 5' overhangs introduced yielded 0 CFU's as well.

Various modifications to different electroporation parameters were then performed in an attempt to achieve transformation success in *P. tricornutum*. These parameters included varying the concentration of DNA added, varying the volume of DNA containing solution added, harvesting cell cultures at different growth stages, adding various carrier DNA

types, and modifying the electrical parameters used for the transformations. Altering the electrical parameters to generate an electrical pulse at 500V with time constant (TC) of 10 – 25 ms was the modification that seemingly had the most effect on transformation success. A final transformation protocol that reliably produced colonies on selection plates was established and consisted of using cells harvested from a culture at OD₇₃₀ at 0.25 – 0.38, pre-heating single-stranded salmon sperm DNA (ssssDNA) to 95°C prior to use as carrier DNA and pulsing the cells at the following electrical parameters: 500V, 50µF, and 400Ω. These conditions yielded transformations that produced over 1000 CFU's per selection plate, with both blunt ended DNA and DNA with 5' overhangs (Figure 2-1). Information on all electroporation and electroporation attempts can be found in Table C-1.

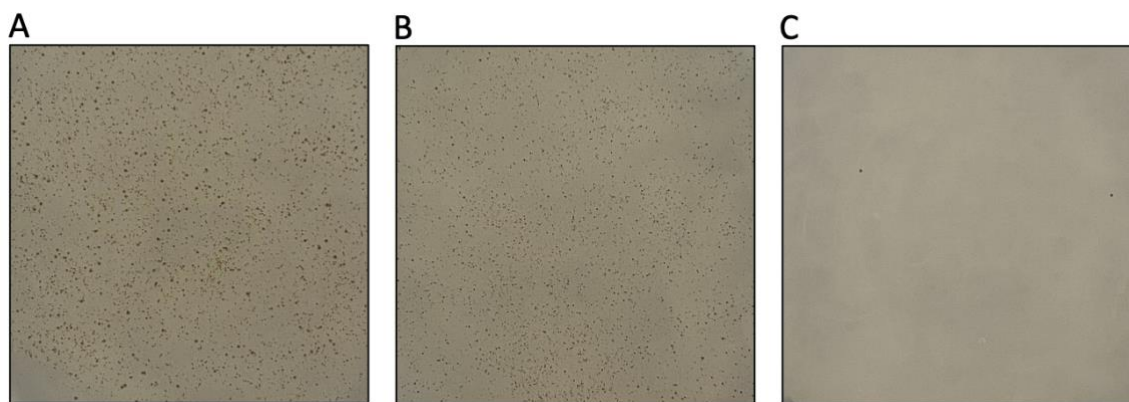


Figure 2-1: Cross section of plates for *P. tricornutum* electroporations. (A) Electroporation performed using the *XhoI.site-fcpD.Promoter-nat-fcpA.Terminator-XhoI.site* construct with 5' overhangs created via restriction digestion with *XhoI*. 1000+ CFUs are present on the entire plate. (B) Electroporation performed using the *XhoI.site-fcpD.Promoter-nat-fcpA.Terminator-XhoI.site* construct with blunt ends. 1000+ CFUs are present on the entire plate. (C) Negative control electroporation performed using only carrier DNA (ssssDNA). 3 CFU's are present on the entire plate. For all transformations, 1/10th volume of cell cultures was plated on ½ L1 plates containing 200 µg/mL of nourseothricin. Pictures were taken 12 days following plating.

Ten transformation lines were then generated from three independent transformations to assess whether evidence of the transformed selection marker could be observed in DNA samples isolated from each transformation line. To mitigate the possibility of untransformed transformations constructs present on the selection plates contaminating the DNA sample isolated from each transformant culture, cultures were passed twice (once by restreaking a single colony on solid media, and once by creating liquid cultures from the streaked colony) before any DNA isolation was performed. DNA was then isolated from the transformation lines and genotyping was performed by amplifying a 283 bp region of the *nat* CDS. Agarose gel electrophoresis showed presence of the *nat* gene in samples from all the transformation lines (Figure 2-2).

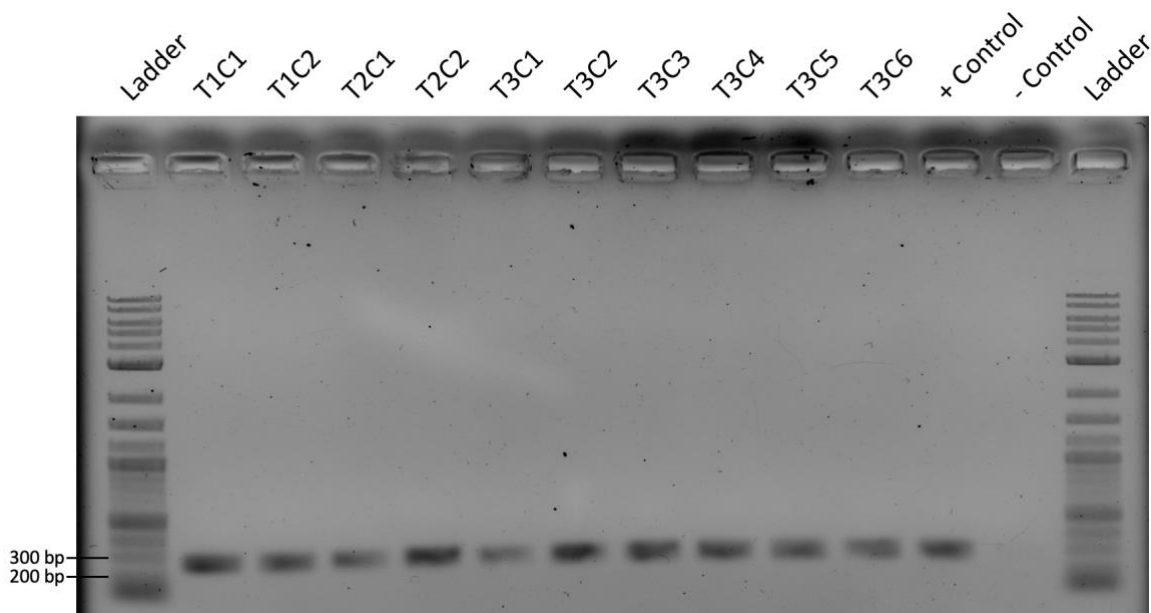


Figure 2-2: 1% agarose gel of genotyping experiment performed on cell lines transformed with the *nat* single selectable marker cassette, passed two times following initial selection. Genotyping was performed by amplifying a 283 bp region of the *nat* CDS. “T” denotes which independent transformation and “C” denotes which colony from that transformation the cell line was generated from. The positive control lane contains a genotyping PCR performed on DNA isolated from a cell line containing the *nat* gene on an episomal vector (pSS94). The negative control lane contains a genotyping PCR performed on high quality wild-type DNA.

However, after continuing to pass the same cultures five more times in liquid, presence of the *nat* gene was only able to be seen in 6/10 transformation lines, indicating potential instability of the construct within the genome (Figure 2-3).

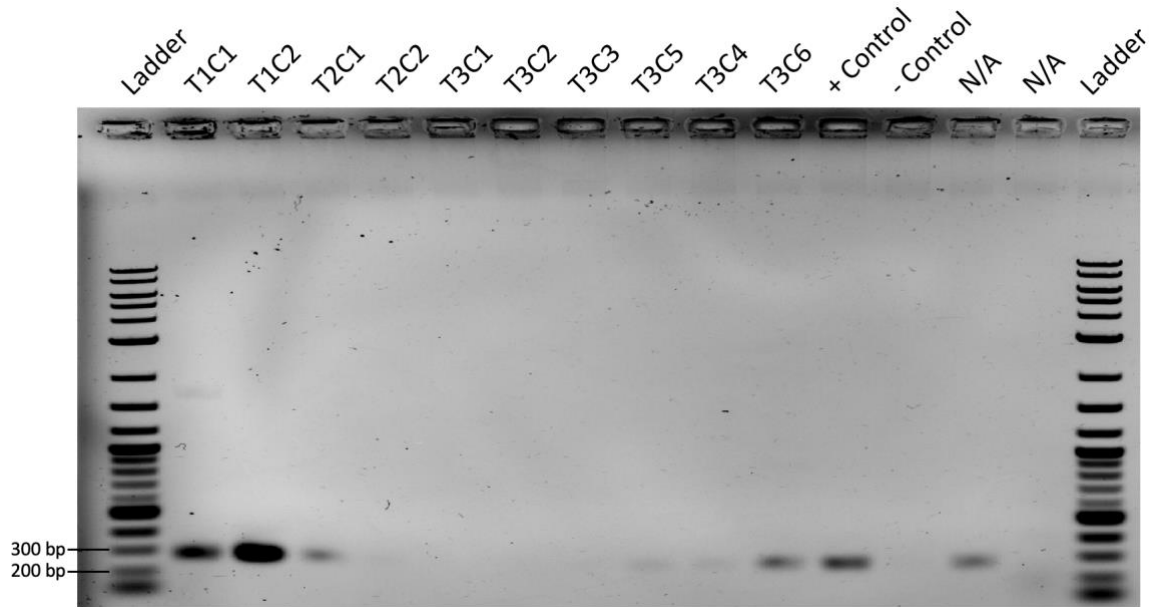


Figure 2-3: 1% agarose gel of genotyping experiment performed on cell lines transformed with the *nat* single selectable marker cassette, passed seven times following initial selection. Genotyping was performed by amplifying a 283 bp region of the *nat* CDS. “T” denotes which independent transformation and “C” denotes which colony from that transformation the cell line was generated from. The positive control lane contains a genotyping PCR performed on DNA isolated from a cell line containing the *nat* gene on an episomal vector (pSS94). The negative control lane contains a genotyping PCR performed on high quality wild-type DNA. N/A denotes lanes and samples that are not applicable to this portion of the study.

To determine concretely if the transformed construct was being integrated into the genome and if so how it was being integrated, an inverse PCR strategy was applied. This strategy consists of fragmenting genomic DNA (gDNA) of the transformed organism using a restriction enzyme that does not have a restriction site inside of the transgenic construct, circularizing the DNA fragments through ligation, then amplifying the gDNA adjacent to where the construct was integrated using primers that bind to the transgenic construct

facing outwards to enrich this region for Sanger sequencing. This strategy was attempted on the same ten transformation lines, once using *SalI*-HF and once using *NcoI*-HF to fragment the isolated gDNA. These restriction enzymes were selected because they are the commercial enzymes that contained the most restriction sites within the *P. tricornutum* genome without having any present in the transformed construct, making the fragments as small as possible to give the polymerase in the PCR stage the greatest chance of replicating the entire fragment. They were also chosen because contain a continuous and unique restriction site, meaning no alternative bases can be present in the restriction site for the enzyme to cleave that region. This provides the highest chance of ligation to work as all the bases on at the cut sites will be perfectly complimentary. Finally, these enzymes were capable of being inactivated by heating, preventing them from interfering with ligation. Genome-wide fragmentation of *P. tricornutum* DNA with *SalI*-HF produces an average fragment size of 2534 bp and a median fragment size of 1680 bp. Fragmentation with *NcoI*-HF produces an average fragment size of 2841 bp and a median fragment size of 1863 bp. The size of the transformation construct containing the *nat* gene was 1430 bp. The expected band sizes for a gDNA fragment containing the transformed construct digested by either enzyme would therefore be approximately 3000 bp or higher if the whole construct was inserted. For all ten transformation lines that inverse PCR was performed on, only one line showed potential evidence of a transformed construct captured within a gDNA fragment (Line T1C1). Fragments from this line that were derived from *SalI*-HF and *NcoI*-HF digestions were both able to be amplified, although the band sizes observed were below the expected based on the mean and median estimates (ranging from approximately 1000 to 1500 bp; Figure 2-4).

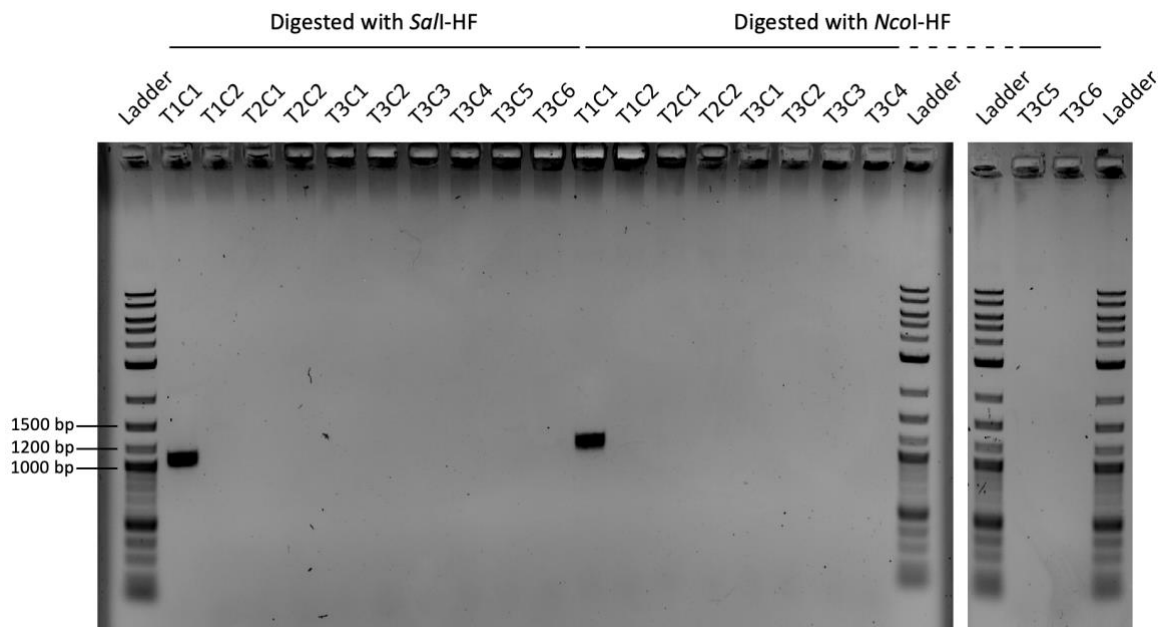


Figure 2-4: 1% agarose gel of inverse PCR experiment performed on cell lines transformed with the *nat* single selectable marker cassette. “T” denotes which independent transformation and “C” denotes which colony from that transformation the cell line was generated from.

Sanger sequencing was then performed on purified inverse PCR products. gDNA not present in the transformed construct was observed in one of the sequencing reads (NCBI Reference Sequence: NC_011670.1, region: 137119 – 137159). This indicated that the construct was likely integrated into the *P. tricornutum* nuclear genome at chromosome 2, directly upstream of the *fcpA* gene. Since the transformation construct contained an endogenous *fcpA* promoter to drive *nat* expression, this indicated that transformed constructs may have been getting integrated through homologous recombination at regions of native homology. Genotyping was performed at the *fcpA* locus in to validate the sequencing results. Primers were designed to amplify a 1264 bp region of gDNA encapsulating the *fcpA* gene. If the transformation construct had been recombined within that region, a noticeable band size difference would have been expected to be perceived on

an agarose gel between wild-type gDNA and gDNA from T1C1 as the *nat* CDS was 96 bp shorter than the *fcpA* CDS. However no perceivable difference was observed (Figure 2-5). Determining where and how the transformed constructs were being integrated into the genome remained inconclusive through this set of experiments.

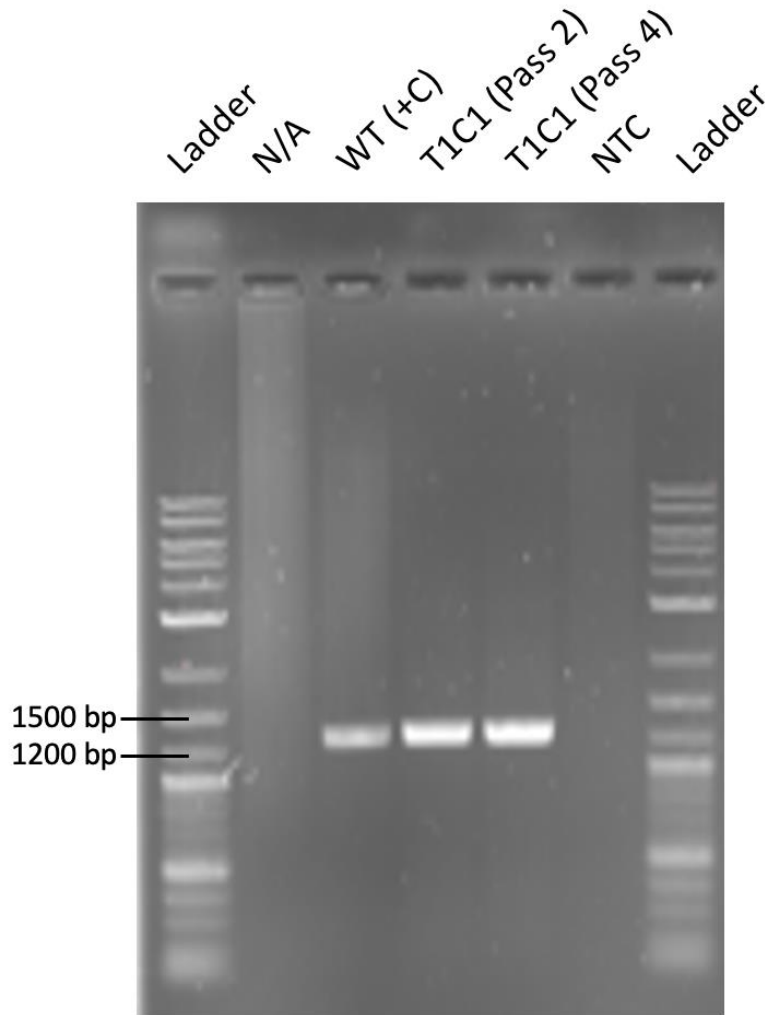
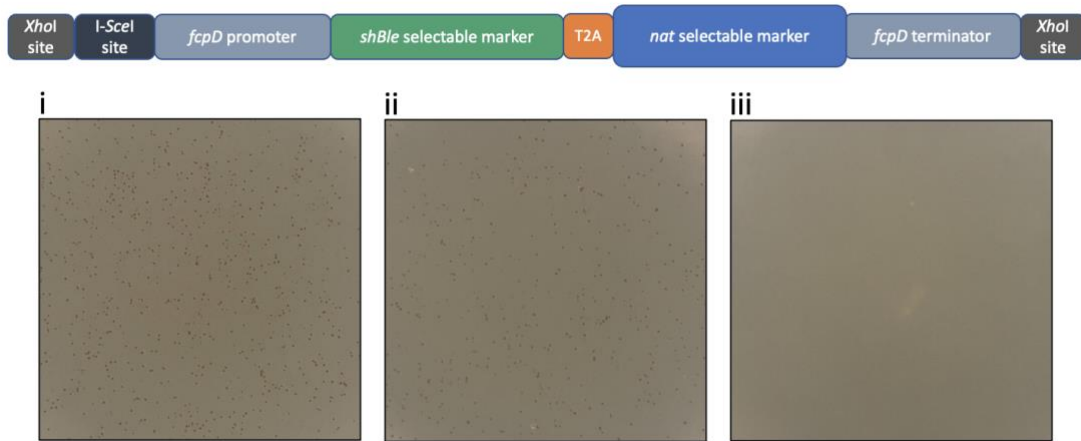


Figure 2-5: 1% agarose gel of genotyping experiment performed on cell lines where genomic DNA was found in inverse PCR sequencing reads. Genotyping was performed by PCR amplification of the whole *fcpA* gene. WT denotes genotyping performed on wild-type genomic DNA. “T1C1” denotes genotyping performed on the cell line where genomic DNA was found in inverse PCR sequencing reads. Genotyping was performed on DNA isolated from this line after the second and fourth pass. NTC denotes the no template control PCR. N/A denotes lanes and samples that are not applicable to this portion of the study.

2.3.2 Assaying a double marker transformation

Having established a reliable protocol for introducing DNA to the nuclear genome of *P. tricornutum* using electroporation, the next objectives were to conclusively determine where the DNA was being integrated and how transgenic lines performed over wild-type cultures. For these purposes, a double selection marker cassette encoding the zeocin resistance gene (*shBle*) linked to the *nat* gene via a T2A self-cleaving peptide linker sequence was used for the following transformations in order to attempt to mitigate natural resistance from arising by employing antibiotic switching during passing. The construct also contained native regions of homology from the *P. tricornutum*, containing the promoter and terminator regions for the gene encoding the fucoxanthin-chlorophyll a-c binding protein D (*fcpD*). The construct also contained restriction sites for *XhoI* at the ends, and also contained a single restriction site for *I-SceI* upstream of the *fcpD* promoter which was added for downstream experiments (Figure A-1B). Using the same electroporation conditions as previous established and blunt-ended DNA, transformations that yielded over 1000 CFU's per selection plate were able to be achieved when plated on nourseothricin selection (Figure 2-6A). However, there was an approximately 7-fold decrease in CFU's when the same transformed cultures were plated on zeocin selection (Figure 2-6B). To assess whether introduction of foreign DNA could induce an immune response that bolsters *P. tricornutum*'s resilience to antibiotics, a portion of non-coding DNA for *P. tricornutum* that was of similar size to the double marker construct was amplified from a plasmid containing replicative elements for the bacteria *Acholeplasma laidlawii*, pAL1, and used in place of the double selectable marker cassette in transformation negative controls. All of the negative control transformations performed using the pAL1 fragment yielded 0 CFU's (Figure 2-6, Table C-1).

A



B

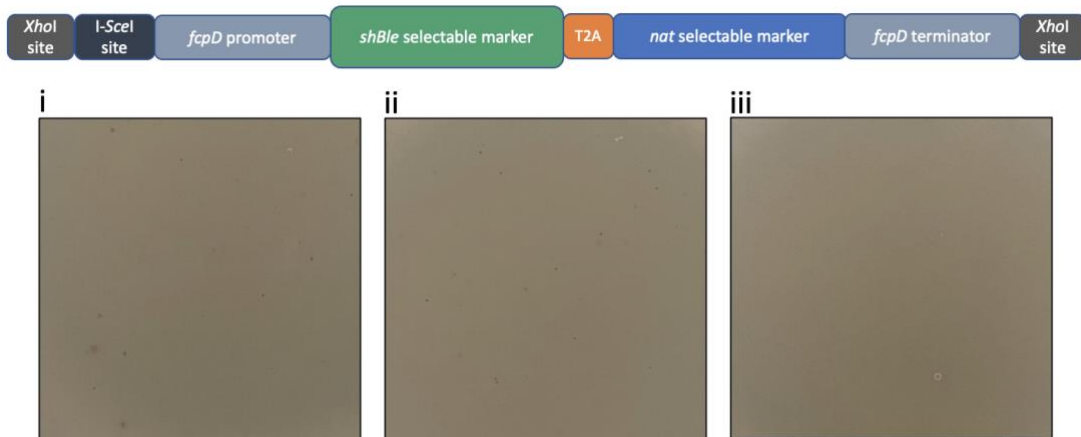


Figure 2-6: Cross section of selection plates for *P. tricornutum* electroporations. Electroporations performed using the *XhoI*.site-*fcpD*.Promoter-*shBle*-T2A-*nat*-*fcpA*.Terminator-*XhoI*.site construct with blunt ends, plated on two selectable backgrounds. (A) Electroporations plated on ½ L1 plates containing 200 µg/mL of nourseothricin. (i). First replicate. 1322 CFU's are present on the entire plate. (ii). Second replicate. 782 CFU's are present on the entire plate. (iii) Negative control transformation (transformed with non-coding DNA from pAL backbone). 0 CFU's are present on the entire plate. (B) Electroporations plated on ½ L1 plates containing 25 µg/mL of zeocin. Replicates belong to the same electroporations performed in section A. (i). First replicate. 195 CFU's are present on the entire plate. (ii). Second replicate. 107 CFU's are present on the entire plate. (iii) Negative control transformation. 0 CFU's are present on the entire plate. For all transformations, 1/10th volume of cell cultures were plated on selective plates. Pictures were taken 10 days following plating.

After demonstrating that electroporation of double selectable marker cassette can instill antibiotic resistance in *P. tricornutum*, the next objectives were to assess transformation frequency, the performance of the transgenic lines, and how the constructs were being integrated into the genome. Electroporations of the double selectable marker were performed across six biological replicates using 160 ng/ μ L of purified double selectable cassette DNA. Transformation frequencies were calculated relative to transformed DNA and to total number of cells transformed. Transformation frequencies relative to transformed DNA were attained by dividing the number of CFUs that were observed 14 days following plating over the micrograms of DNA added to selection plates divided by the initial dilution factor. The average transformation frequency calculated for transformants selected on 200 μ g/mL of nourseothricin was approximately 107 CFUs/ μ g of DNA. For transformants selected on 25 μ g/mL of zeocin, the calculated frequency was approximately 23 CFUs/ μ g (Figure 2-7). Transformation frequencies relative to transformed DNA were attained by dividing the number of CFUs that were observed 14 days following plating on selective plates multiplied by the dilution factor over the total number of CFUs in the recovery media as determined through serial dilutions on non-selective media over the μ g of DNA transformed. The average transformation frequency calculated for transformants selected on 200 μ g/mL of nourseothricin was approximately 1.91×10^{-5} CFUs/ μ g. For transformants selected on 25 μ g/mL of zeocin, the calculated frequency was approximately 3.53×10^{-6} CFUs/ μ g (Figure 2-8). It should be noted that fewer colonies were observed in this set of transformations than expected based on the previous double selectable marker transformations. This may have been due to growth issues of the starting cultures as they exhibited slower growth than most other starting

cultures used. Due to time constraints, these experiments could not be reperformed and downstream experiments were continued on these transformation lines.

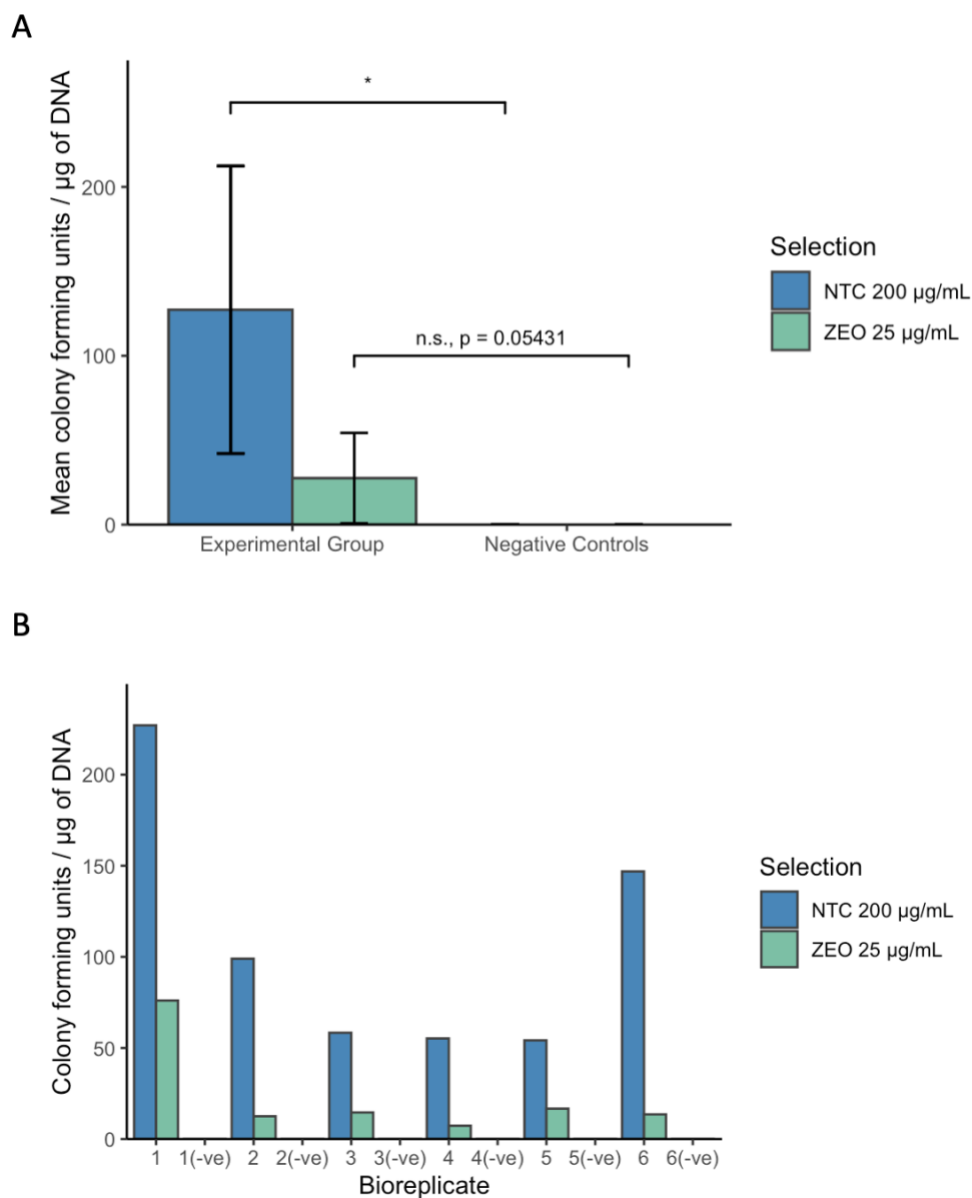


Figure 2-7: Bar charts of transformation frequencies calculated without using total colony forming unit estimates. Calculations were based of CFU's observed 14 days following plating on plates containing either nourseothricin (NTC) or zeocin (ZEO) antibiotics. Transformation frequencies as calculated by transformed colony forming units per μg of transformed DNA. **(A)** Bar chart depicting the average transformation frequencies for each treatment and experimental group. Independent sample t-tests were computed (n.s. – not significant, $* = p \leq 0.05$) based on $N = 6$ biological replicates. **(B)** Bar chart depicting calculated transformation frequencies for each biological replicate performed. “-ve” denotes frequencies calculated on negative control plates for each biological replicate.

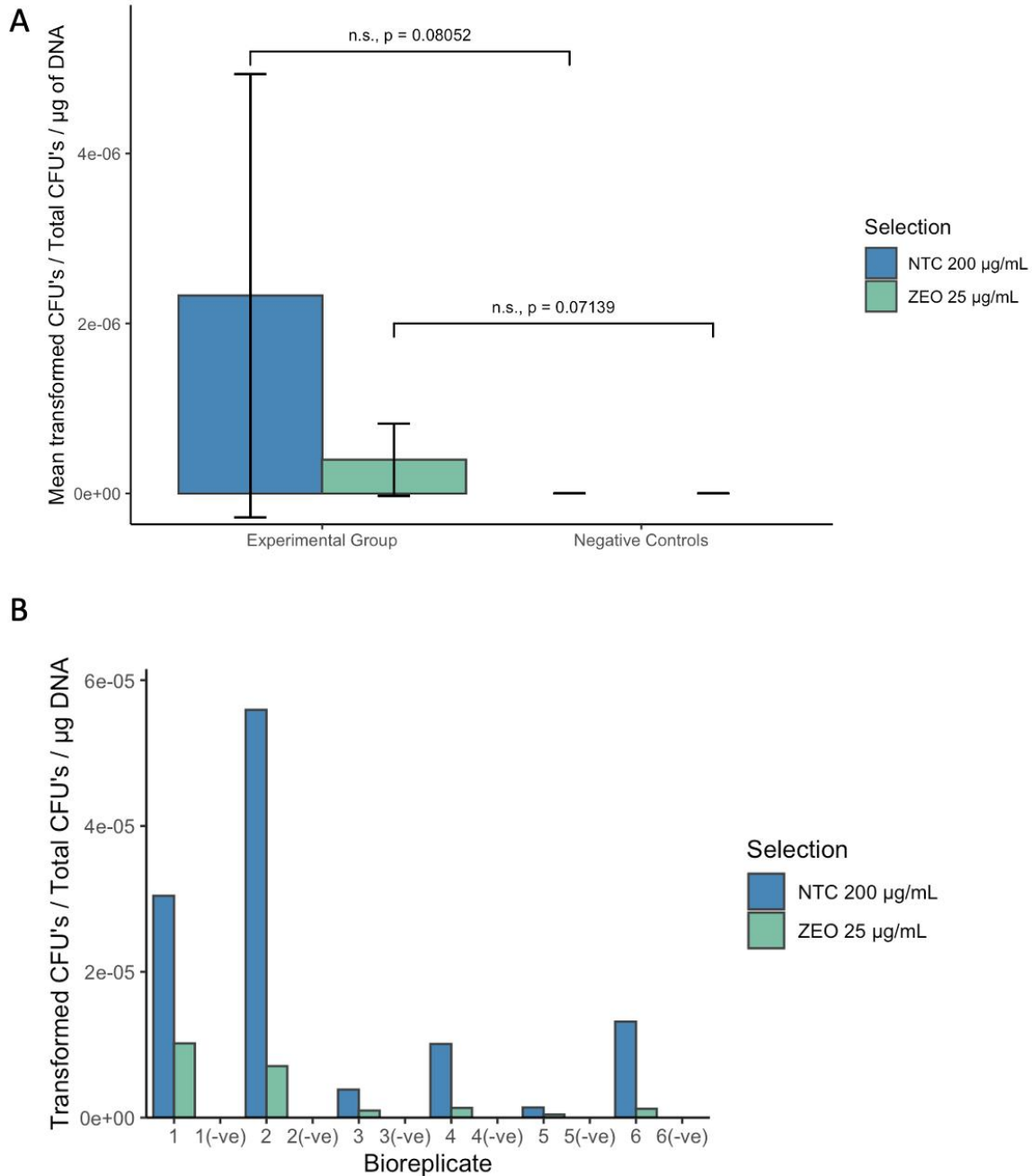
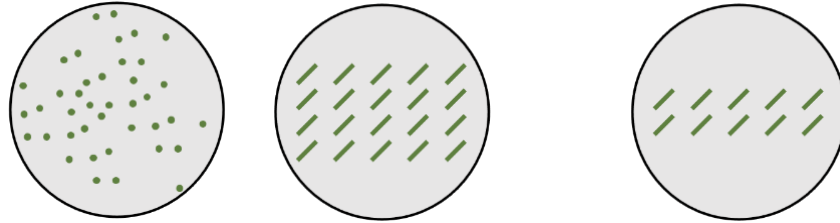


Figure 2-8: Bar charts of transformation frequencies calculated using total colony forming unit estimates. Calculations were based of CFU's observed 14 days following plating on plates containing either nourseothricin (NTC) or zeocin (ZEO) antibiotics. Transformation frequencies as calculated by transformed colony forming units per total colony forming units in transformed culture per μg of transformed DNA. **(A)** Bar chart depicting the average transformation frequencies for each treatment and experimental group. Independent sample t-tests were computed (n.s. – not significant) based on $N = 6$ biological replicates. **(B)** Bar chart depicting calculated transformation frequencies for each biological replicate performed. “-ve” denotes frequencies calculated on negative control plates for each biological replicate.

Transformant colonies were then repatched on differing orders of antibiotics to assess whether selection order had an effect on the proliferation of transformation lines. Up to 20 colonies were picked from the initial selection plate 21 days after plating. These colonies were repatched on a ½ L1 plate containing the same antibiotic at the same concentration as was used in the initial selection plate (either 200 µg/mL of nourseothricin or 25 µg/mL of zeocin). 14 days after repatching, up to 10 streaks were repatched onto a ½ L1 plate containing the alternate antibiotic. It was determined that although selection with 200 µg/mL of nourseothricin initially yielded more CFUs than selection with 25 µg/mL of zeocin, a majority of the colonies picked could not be proliferated on successive passes (Table 2-1). Colonies initially selected for with nourseothricin were also very compact and firmly bound to the plate media, making noticeably very difficult to pick. 3/120 colonies were able to be proliferated from nourseothricin-containing plates to another nourseothricin-containing plate, however all three of these lines were able to be proliferated after alternating the antibiotic to be zeocin. On the other hand, a majority of the colonies initially selected for with zeocin were able to be proliferated. 89/99 colonies were able to be proliferated onto plates containing the same selection, and after altering the antibiotic to nourseothricin, 51/55 streaks were able to be proliferated. Colonies initially selected for with zeocin were also phenotypically very noticeably different to those selected for by nourseothricin, being more globular in nature and much easier to pick (Table 2-1).

Table 2-1: Colony or streak counts from various transformation and re-streaking plates. Six independent transformations were performed and selected for on 1% agar and ½ L1 plates containing either 25 µg/mL of zeocin (ZEO) or 200 µg/mL of nourseothricin (NTC). Up to 20 colonies were picked from the transformation plates and streaked out on an initial plate containing the same antibiotics as the plates the colonies were derived from. After 2 weeks, the number of streaks that exhibited growth was recorded and up to 10 streaks were re-streaked onto another plate containing the opposite antibiotic. Cultures were left to grow for 2 weeks and then the number of streaks that exhibited growth was once again recorded.



Biological replicate	Initial Colonies		Restreak 1		Restreak 2	
	ZEO	NTC	ZEO → ZEO	NTC → NTC	ZEO → ZEO → NTC	NTC → NTC → ZEO
1	78	263	19/20	0/20	10/10	0/0
2	13	119	11/13	1/20	9/10	1/1
3	18	64	15/18	1/20	9/10	1/1
4	8	58	5/8	1/20	4/5	1/1
5	20	60	20/20	0/20	10/10	0/0
6	21	169	19/20	0/20	9/10	0/0

Four liquid cultures were then generated from initial repatches, two from streaked cultures on a first pass plate containing zeocin, and two from plate containing nourseothricin. After two successive passes in liquid, DNA was then isolated from the transformation lines and genotyping was performed by amplifying a 150 bp region of the *shBle* CDS, a 235 bp portion of the *nat* CDS, and an 865 bp portion of the *P. tricornutum* chromosome 2 (*fcpB* gene; used as positive control for the multiplex PCR). Agarose gel electrophoresis showed presence of the double selection marker cassette in samples from all the transformation lines (Figure 2-9).

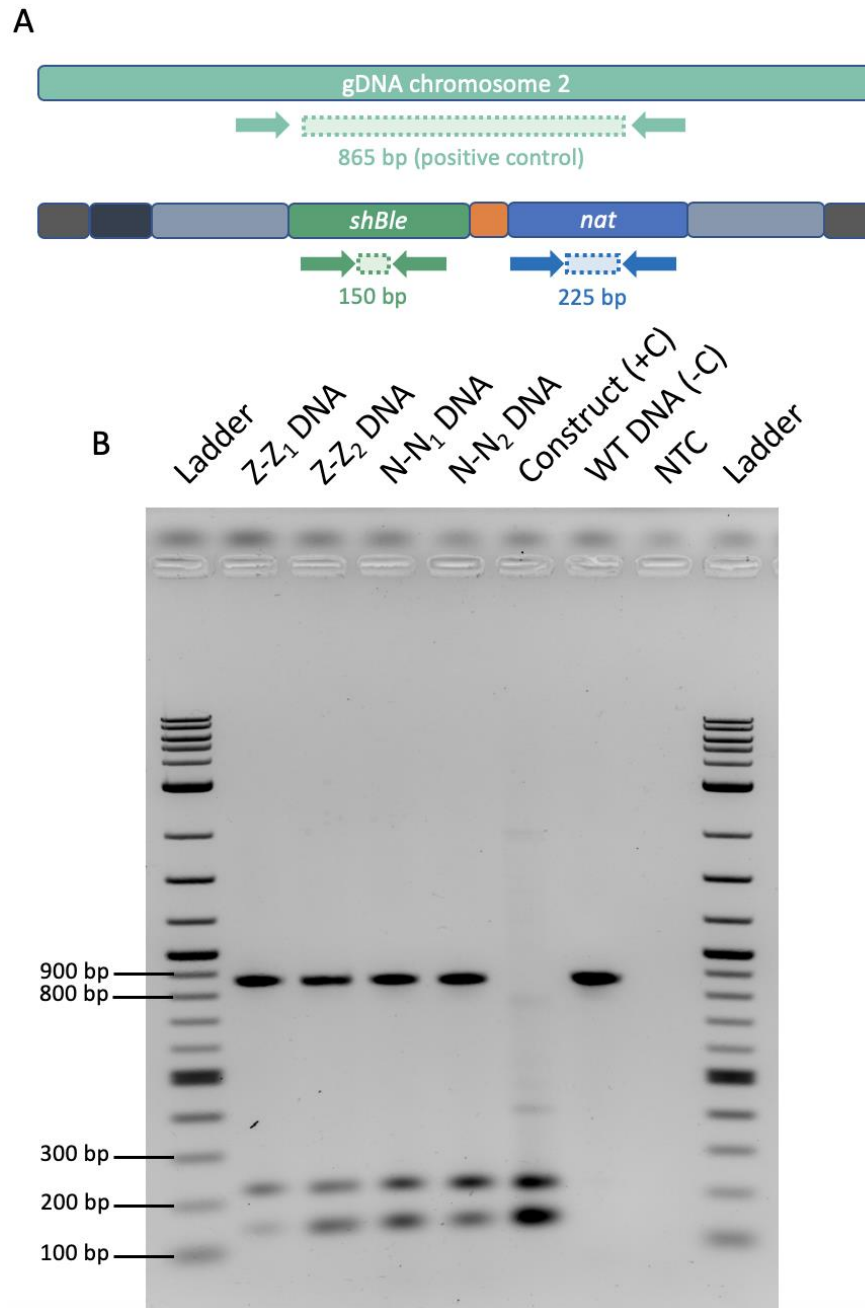


Figure 2-9: Genotyping experiment performed on cell lines transformed with the *shBle*-T2A-*nat* double selectable marker cassette, passed two times following initial selection. (A) Schematic detailing genotyping regions. (B) 1% agarose gel of genotyping experiment. Z-Z₁ and Z-Z₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. N-N₁ and N-N₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing nourseothricin. Construct refers to PCR performed using purified *shBle*-T2A-*nat* cassette DNA as a template. WT denotes two independent wild-type cultures. NTC denotes the no template control PCR.

The maintenance and efficacy of the double marker cassette was then assayed through spot plating. The four transformation lines, as well as two wild type cultures were passed four times in non-selective L1 media, with passing having been performed every 14 days. The cultures were then adjusted to a concentration of 2×10^6 cells/mL prior to diluting and plated on $\frac{1}{2}$ L1 with no antibiotics, with either zeocin or nourseothricin, or both antibiotics. Similar growth was observed across all cell lines, with the exception of the wild-type controls which exhibited no growth in the presence of antibiotics (Figure 2-10).

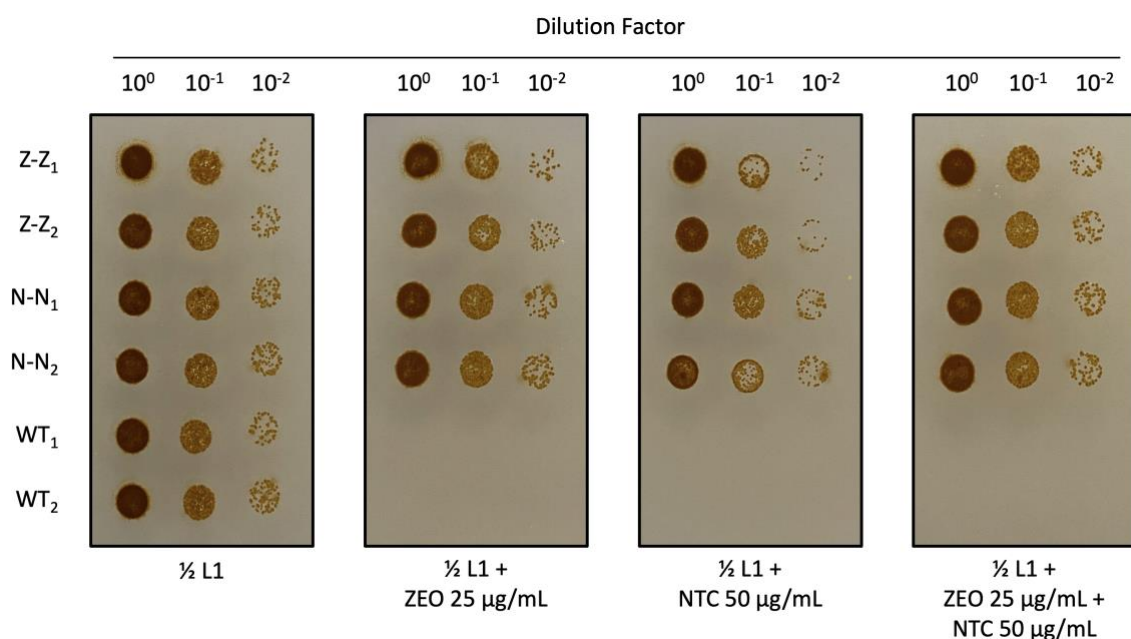


Figure 2-10: Spot plate of transformation lines. Four transformation lines were spotted, and two independently grown cultures of wild-type cells were spotted as well on plates containing 1% agar and $\frac{1}{2}$ L1 media and/or zeocin (ZEO) or nourseothricin (NTC) antibiotics. Z-Z₁ and Z-Z₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. N-N₁ and N-N₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing nourseothricin. WT₁ and WT₂ denote two independent wild-type cultures. Pictures were taken 14 days after spot plating.

Although the genotyping and growth experiments strongly indicated the double marker cassette had been integrated into the *P. tricornutum* genome, no concrete evidence had been procured to indicate where the constructs were being integrated. To this end,

sequencing was performed on by isolating high molecular weight DNA from the four transformation lines using phenol chloroform isolation and then performing nanopore sequencing. The sequencing reads were assembled into a BLAST database and then a BLAST alignment was performed using the transformation construct as the query sequence to locate the region of the inserts. Two instances of genomic integration were observed in the four transformation lines, one in line Z-Z₁ and one in line Z-Z₂. The lack of integration events observed in transformation lines N-N₁ and N-N₂ were likely a result of low sequencing coverage rather than an absence of transgene integration, as genome wide coverage estimates indicates large regions of the *P. tricornutum* genome was not present in the sequencing reads for all sequenced lines (Figures A-2 – A-5).

For the constructs that were found to have integrated, there were a few things to note. For the integration event found in line Z-Z₁, the construct was inserted as a partial concatemer. The entirety of one construct can be found followed by a partial construct that terminates within the *nat* gene directly downstream of it. The constructs are seemingly fused at the *XhoI* site at the end of the first construct and at the start of the next, as only one restriction site at this region was observed in the sequencing read (the restriction site was also seemingly altered and no longer functional). Additionally, the region directly upstream of the insert was found to map to chromosome 5 (NCBI Reference Sequence: OU594946.1, region: 904322– 918802), however the region directly downstream of the insert mapped to chromosome 3 (NCBI Reference Sequence: OU594944.1, region: 1386512 – 1389764). The region of insertion lies in an intergenic region in both of these chromosomes (Figure 2-11B). For the integration event found in line Z-Z₂, the insertion mapped to a region within chromosome 9 on the antisense strand (NCBI Reference Sequence: OU594950.1, upstream

region: 462762 – 475385, downstream region: 473782 – 483277), however the inserted region included only a partial construct, with the *fcpD* promoter being partially truncated. The integration region also consisted of a 16 bp sequence directly upstream and 260 bp sequence directly downstream of the integration site that showed no significant similarity to the *P. tricornutum* genome or the transformed cassette (Figure 2-11C). The integration site was also determined to fall directly inside of the CDS of a predicted gene, PHATRDRAFT_45242. A BLAST search of the predicted protein encoded by this gene reveals 11 potential orthologs, all belonging to pennate diatom species, however no known function of this gene or its orthologs is known at present.



Figure 2-11: Sequencing results from cell lines where integration of transformed cassettes was observed. (A) Schematic of original transformed construct. (B) Integration event captured on sequencing read for DNA isolated from Z-Z₁ cell line. (C) Integration event captured on sequencing read for DNA isolated from Z-Z₂ cell line.

Because full genome coverage was not obtained in the sequencing experiments, there may have been multiple integration events that were not captured by the sequencing reads, specifically at the regions within the genome that exhibited a high level of homology to the

double selection marker cassette. Gene expression on the double selection marker cassette was driven by the *fcpD* promoter and terminator, which are native to chromosome 2 of *P. tricornutum*. Furthermore, the *fcpD* gene falls within a gene cluster of four highly similar genes (*fcpA*, *fcpB*, *fcpC*, *fcpD*) where a high degree of homology is shared with the transformed double selection marker cassette. No sequencing reads were identified that spanned this region, therefore genotyping was performed at these regions for the four transformation lines to assess whether evidence of integration could be detected.

Genotyping was performed by amplifying a 1928 bp region flanking the *fcpD* gene, an 1183 bp region flanking the *fcpC* gene, an 865 bp region flanking the *fcpB* gene, and a 1453 bp region flanking the *fcpA* gene in a multiplex PCR reaction. It was predicted that integration of the double selection marker cassette at any of these loci would have resulted in a larger amplicon at that given locus as the dual coding regions in the double marker cassette are larger than all the coding regions of all the *fcp* genes in the gene cluster.

Agarose gel electrophoresis showed clear bands corresponding to the expected sizes of the wild type *fcpA*, *fcpB*, and *fcpC* genes in all four transformation lines (Figure 2-12). The amplified region corresponding to the *fcpD* gene remained ambiguous as the observed bands were very faint or non-existent therefore genotyping was performed again for this locus using a different polymerase that is more suitable for larger amplicons (GXL). For all four sequenced transformation lines, agarose gel electrophoresis showed bands corresponding to the expected size of the wild type *fcpD* gene as well (Figure 2-13).

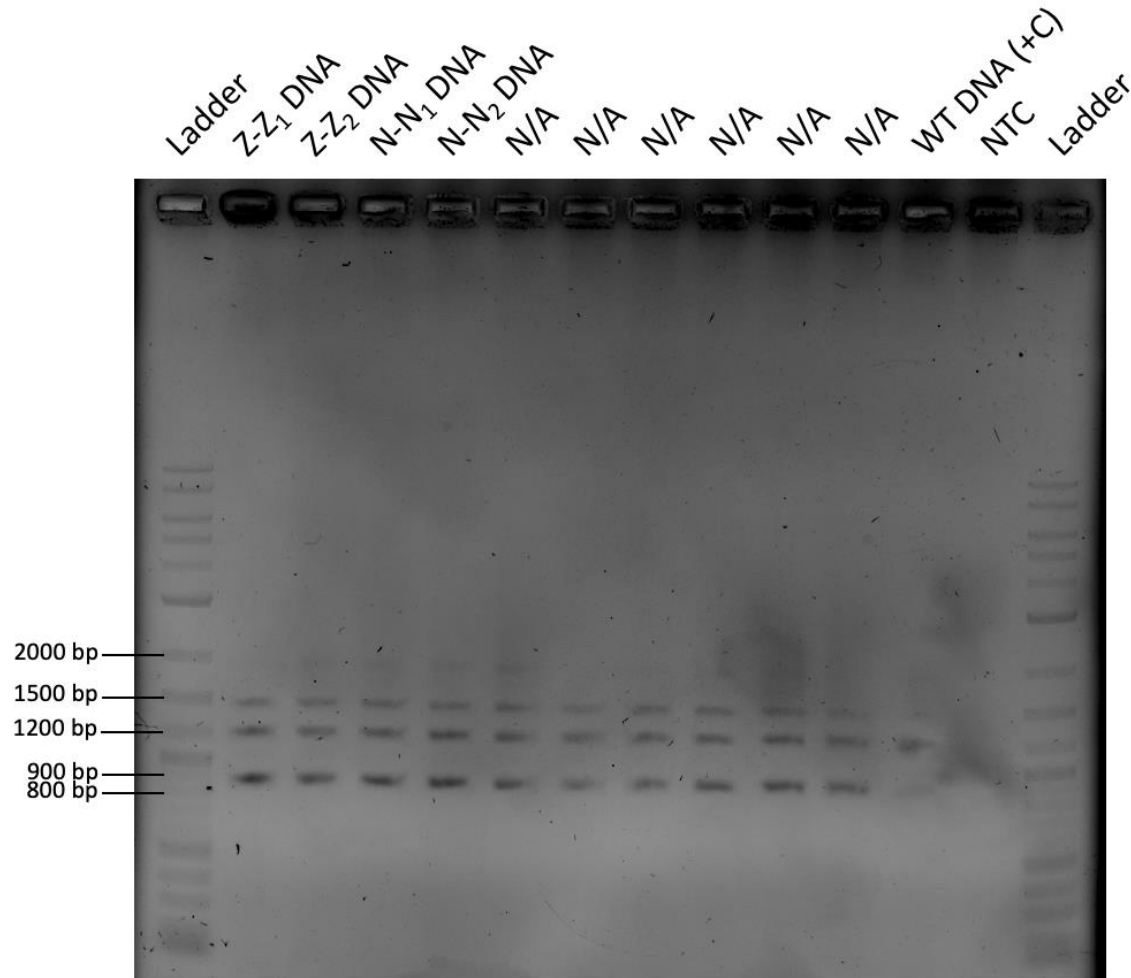


Figure 2-12: 1% agarose gel of genotyping experiment performed on *fcp* gene cluster. Genotyping was performed by PCR amplification of the regions flanking the *fcpA*, *fcpB*, *fcpC*, and *fcpD* genes (1453, 865, 1183, and 1928 bp, respectively). Z-Z₁ and Z-Z₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. N-N₁ and N-N₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing nourseothricin. WT denotes genotyping performed on wild-type genomic DNA. NTC denotes the no template control PCR. N/A denotes lanes and samples that are not applicable to this portion of the study.

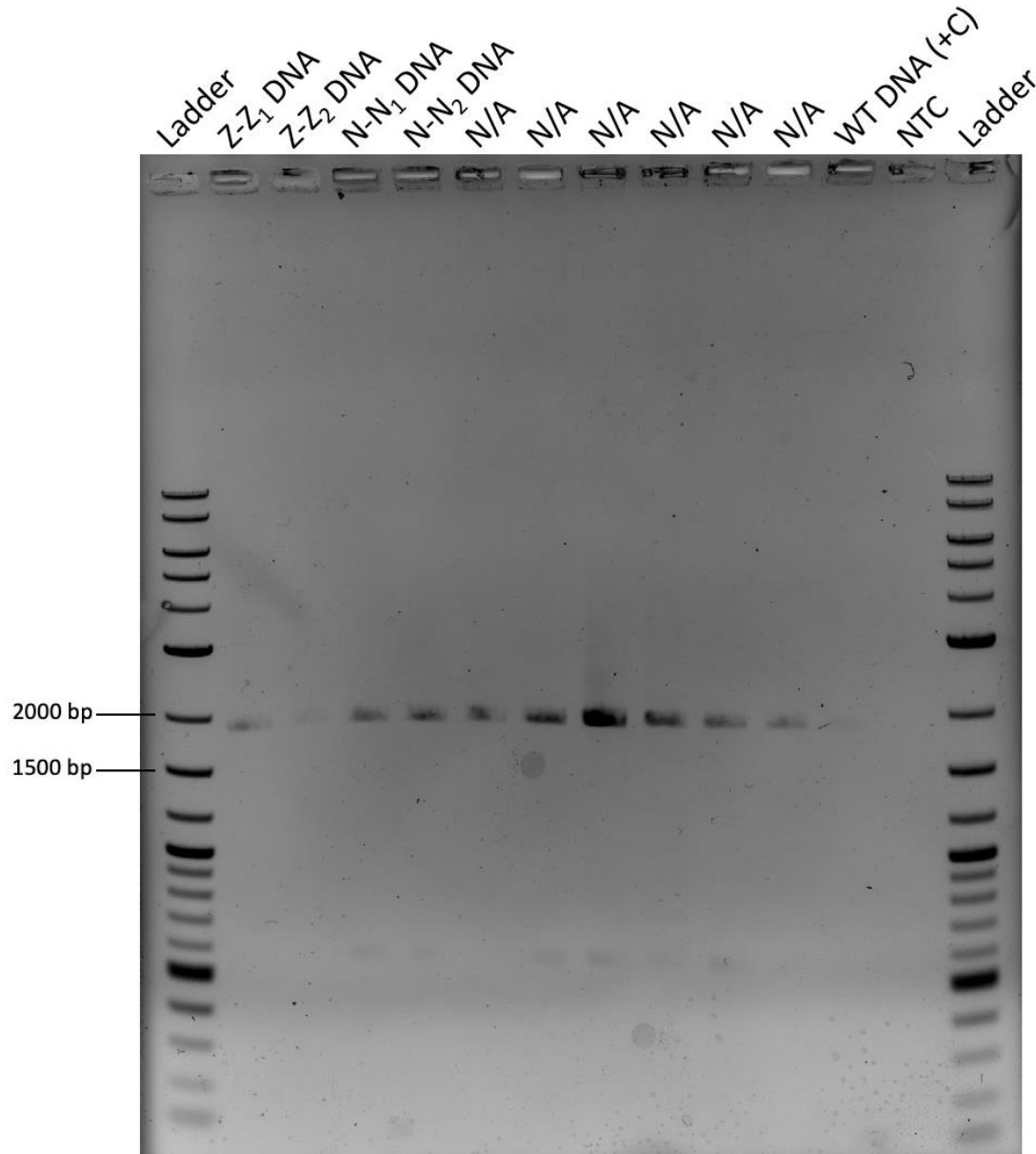


Figure 2-13: 1% agarose gel of genotyping experiment performed on *fcpD* gene. Z-Z₁ and Z-Z₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. N-N₁ and N-N₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing nourseothricin. WT denotes genotyping performed on wild-type genomic DNA. NTC denotes the no template control PCR. N/A denotes lanes and samples that are not applicable to this portion of the study.

A final set of genotyping experiments were performed to validate the sequencing results. For the transformation line Z-Z₁, primers were designed to amplify from the sequence found upstream of the construct on the sequencing read (chromosome 5) to the *shBle* CDS, to amplify from the *nat* CDS to the sequence directly downstream of the construct on the sequencing read (chromosome 3), and to amplify the entire insertion (chromosome 5 to chromosome 3) (Figure 2-14A). Amplicons were only expected to be produced in PCR reactions that used DNA isolated from the Z-Z₁ line, even in the reactions flanking the whole insertion event, as DNA isolated from wild type *P. tricornutum* was not expected to have a contiguous DNA sequence from chromosome 5 to chromosome 3. Agarose gel electrophoresis showed expected band sizes for all PCRs performed on DNA from Z-Z₁, and no bands for all PCRs performed on DNA from wildtype cultures and no template control reactions (Figure 2-14B).

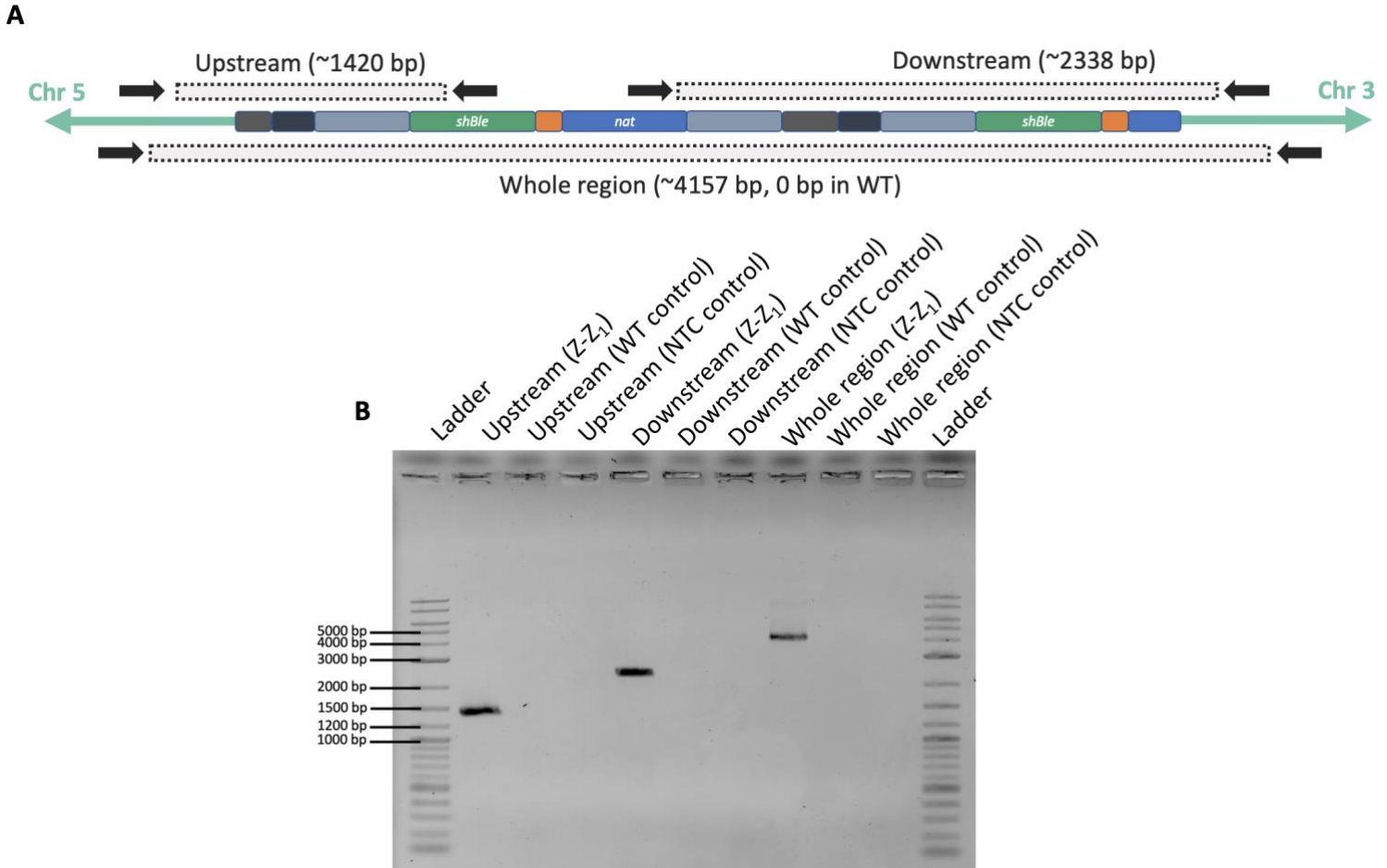


Figure 2-14: Genotyping experiment performed on Z-Z₁ cell line. (A) Schematic detailing genotyping regions. (B) 1% agarose gel of genotyping experiment. Z-Z₁ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. WT denotes two independent wild-type cultures. NTC denotes the no template control PCR. N/A denotes lanes and samples that are not applicable to this portion of the study.

For the transformation line Z-Z₂, primers were designed to amplify from the sequence found upstream of the construct on the sequencing read (chromosome 9) to the *shBle* CDS, to amplify from the *nat* CDS to the sequence directly downstream of the construct on the sequencing read (chromosome 9), and to amplify the entire insertion (chromosome 9 to chromosome 9) (Figure 2-15A). Amplicons containing portions of the *shBle* CDS and *nat* CDS were on only expected to be produced in PCR reactions that used DNA isolated from

the Z-Z₂ line and not in wild type DNA. Since *P. tricornutum* is a diploid organism, the double marker cassette could be present in one or both homologs at the integration locus. In the case that the marker is heterozygously inserted, two types of amplicons should be produced. One matching the size of the one produced by wildtype DNA (approximately 1000 bp), and one approximately 1880 bp larger, relating to the size of the of insertion observed in the sequencing library. In the case that the marker is inserted homozygously, only the latter type of amplicon should be produced. Agarose gel electrophoresis showed expected band sizes for all PCRs performed on DNA from Z-Z₂, and showed two bands in the lane containing the PCR products spanning the entire insertion locus, indicative of heterozygous integration (Figure 2-15B).

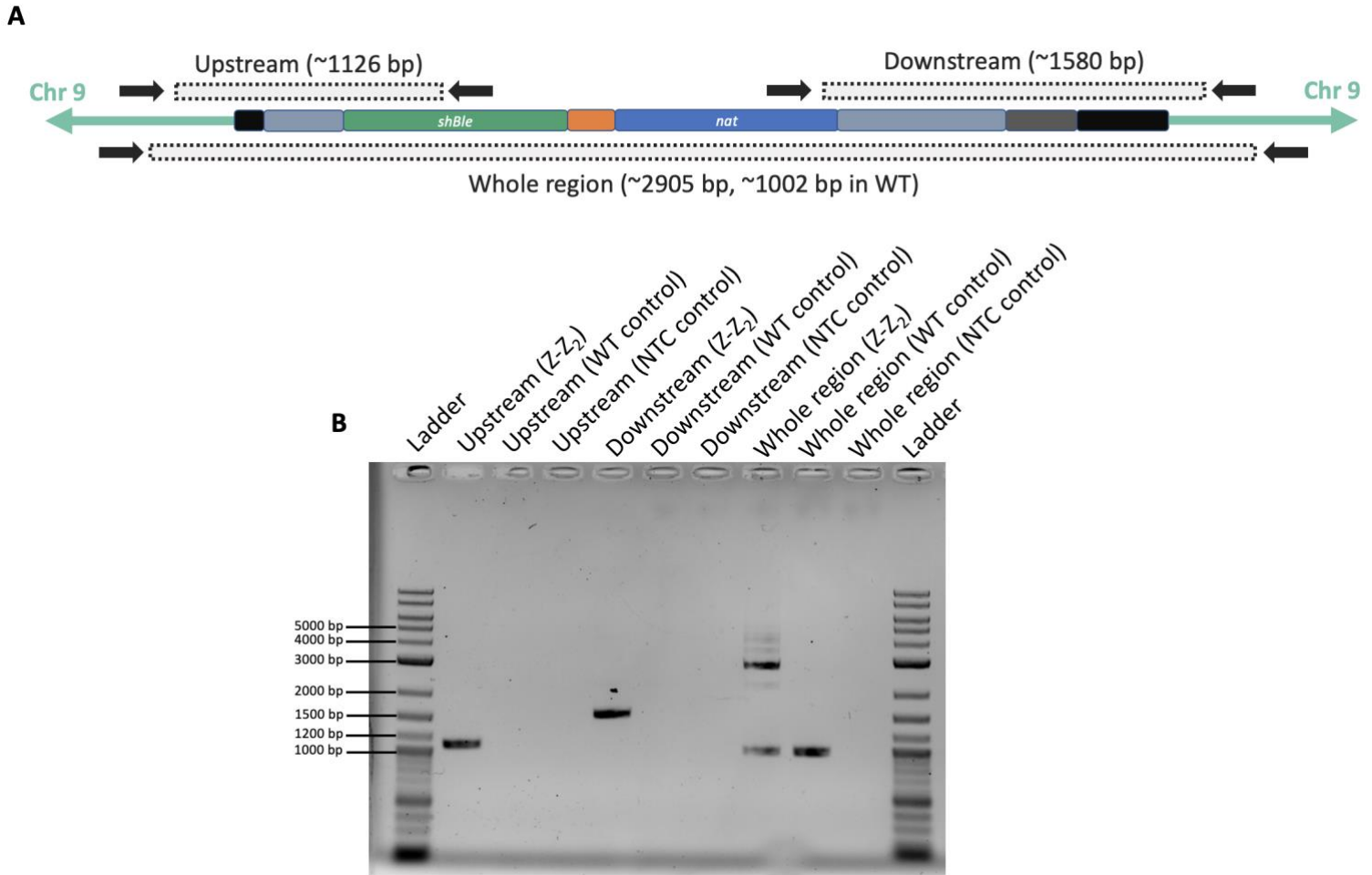


Figure 2-15: Genotyping experiment performed on Z-Z₂ cell line. (A) Schematic detailing genotyping regions. (B) 1% agarose gel of genotyping experiment. Z-Z₂ denote individual cultures that was generated from the streaked culture on a first pass plate containing zeocin. WT denotes two independent wild-type cultures. NTC denotes the no template control PCR. N/A denotes lanes and samples that are not applicable to this portion of the study.

2.4 Discussion

The protocol established here for *P. tricornutum* electroporation constitutes another effective method for transforming transgenic DNA into this organism. Previous studies have reported obtaining a maximum transformation efficiency (no averages were reported in any study) of 3×10^{-5} CFU's/ μ g of DNA, whereas the average transformation frequency calculated here was of 3.53×10^{-6} CFU's/ μ g per transformation when plated on the same

antibiotic selection^{6,9}. While the transformation efficiency calculated here is not higher than those reported in other established protocols, all attempts to replicate the other transformation protocols were unsuccessful. The protocol established here reliably produced a consistent number of transformants. It should also be noted that data collected to calculate transformation frequencies were only collected in one set of transformations, however this set of transformations generated far fewer transformants than other transformations performed of the same marker, potentially due to poor growth conditions of the starting cultures. All calculated transformation frequencies from each initial selective background, with the exception of the frequencies calculated from transformants on nourseothricin excluding total CFU's in transformed culture estimates, showed no significant difference between the experiment when analyzed with independent sample t-tests. This is believed to be consequence of the poor growth observed in the initial transformant cultures and not an indication of the lack of efficacy of the transformation protocol. Due to time constraints these frequencies were reported however it may be possible that more replicates of this experiment could yield a higher average transformation frequency for this protocol. The primary difference between the *P. tricornutum* electroporation protocol established here and the ones performed in other studies is the electrical parameters used to pulse the cells. Typical electrical parameters that have been used in other studies were using an applied voltage of 500V, a capacitance of 25 μ F, and 400 Ω of resistance, which yielded a time constant (TC) of approximately 10ms when using 2mm electrocuvettes. The parameters used here were the same except with the capacitance modified to be 50 μ F. This generally produced a TC ranging between 20 – 25 ms, depending on how effectively the salts from the sample media were washed away. The time constant

is a critical component to electroporation success, as it dictates how pore formation occurs on the cell membrane, which in turn dictates what how DNA enters the cell. Electroporation works by disrupting the transmembrane potential across the lipid bilayer using an applied voltage. This causes the lipids to reorient themselves to form hydrophilic pores in the membrane. The duration of the pulse dictates the size of the pores and the length of time they remain open for (and if the process is still reversible)¹⁰. The TC is a measure of the time it takes for the applied voltage to be reduced by one third of its initial value. For many established transformation protocols to other organisms, a time constant of 5 – 10 ms is typically all that is required to achieve efficient electroporation. Longer pulse durations may be detrimental as the aqueous pores formed may take too long to close or may never close, leading to cell death via over-leakage of ions and metabolites out of the cell¹⁰. The results shown here indicate that *P. tricornutum* is susceptible to longer pulse durations and that this may improve its DNA uptake efficiency. A recent study that also sought to optimize electroporation in *P. tricornutum* indicated that *P. tricornutum* transformation efficiency is most optimal when the TC is around 8 ms when applying 500V, however the researchers only tested the aforementioned electrical parameters that yielded a TC of 10 ms¹. Interestingly, the researchers did investigate higher TC effects on another diatom species (*Nannochloropsis oceanica*) and determine that using electrical parameters that generated a time constant of 21 – 25 ms led to optimal transformation efficiency¹, exactly what was observed in this thesis. The researchers also determined a few extra electroporation parameters that may have influenced diatom transformation efficiency. Electroporations performed to *N. oceanica* were performed using an initial voltage of 2.2 kV as opposed to 500V. While longer pulse durations generally correlate to larger aqueous

pore formation, higher voltage generally correlates to an increased number of pores formed across the cell membrane¹⁰. It could be possible that *P. tricornutum* may be tolerant to a higher applied voltage in conjunction with all the other established electrical parameters and that this may further improve transformation efficiency. Furthermore, the researchers determined that electroporations performed in *P. tricornutum* grown in the absence of silica (as they were grown in this study was well) were up to five times more efficient when the cultures were exposed to continuous light as opposed to being grown in light/dark cycles¹. Adding saponins (plant derived, sterol binding compounds that induce pore formation) also increased electroporation efficiency in *P. tricornutum* two-fold¹. By combining all these factors, it may be possible to drastically improve electroporation to *P. tricornutum* beyond what has been accomplished so far.

Another consistent observation throughout all transformations performed with the *shBle-T2A-nat* was that selecting transformants on a nourseothricin selectable background resulted in a much higher number of initial colonies compared to selecting with zeocin, however propagating cells from plates containing zeocin was much more reliable than vice versa. Initially it was believed that this may have been due to *P. tricornutum*'s ability to develop spontaneous resistance to nourseothricin, however because in most transformation experiments no colonies were observed on any negative control plates this assumption seemed unlikely. Due to the noticeable phenotypic differences in *P. tricornutum* colonies found on nourseothricin and zeocin containing plates, another possible explanation is that the two antibiotics induce different cell morphologies. *P. tricornutum* cells are known to exist in three main morphotypes: fusiform, oval, and triradiate¹¹. These different morphologies are believed to be adaptations to different life stages of *P. tricornutum*, either

as a free floating organism (planktonic) or sedimented to a surface (benthic)¹². The ovular morphotype of *P. tricornutum* is most typically associated with benthic life stages, and only this morphotype has been shown to secrete adhesive mucilage into its extracellular matrix¹³. The difficulty for propagating colonies on nourseothricin could be due to colonies being harder to pick and streak effectively because they are more “stuck” to the agar plates as opposed to those selected for with zeocin. Whether nourseothricin induces secretion of adhesives in *P. tricornutum* requires further investigation.

A large focus of this thesis was to locate where and how the transformed constructs were being integrated, as this was a persistent challenge over the course of this research. Although *P. tricornutum* is known for exhibiting high rates of mitotic recombination between homologous chromosomes¹⁴, very little evidence of targeted homologous recombination has ever been observed^{1,15}. When performing the inverse PCRs to attempting to locate the transgenes, one of the transformation lines showed potential evidence of homologous recombination, as a small region of gDNA that mapped at the locus of the *fcpA* terminator but was not present in the *fcpA* terminator on the transformed cassette was found in the sequencing read. However, this integration event could not be verified via genotyping. In the confirmed integration events, the transgenes were found to integrate randomly. The presence of DNA absent in the transgene constructs or in the genome that was found directly upstream and downstream of the integration implies that the transgenes were likely being integrated via non-homologous end joining (NHEJ). This is consistent with results from other studies that showed NHEJ is the primary mechanism of insertion when biolistic transformation is used on *P. tricornutum*². *P. tricornutum* has shown to exhibit a lot of genome instability¹⁴, with double stranded breaks likely occurring

frequently in the genome, giving transformed DNA in the vicinity an opportunity to integrate in the genome during the repair process. This is supported by the results as in one of the two cases of confirmed integration, the sequences directly upstream and downstream of the transgenic construct map to different chromosomes of the reference genome. This implies that the transgenic construct was integrated during a translocation event that was occurring. Another important detail in the integration events was that neither of the transgenic constructs integrated as they were transformed. In the first integration event, the construct was inserted as a partial concatemer, with 2 constructs seemingly have been fused together during the insertion. Concatemer formation has been shown to occur in other eukaryotes when exogenous DNA is introduced, and has also found to be a consequence of NHEJ¹⁶. This phenomenon has also been observed in *P. tricornutum* for transgenic DNA delivered via biolistic transformation². In the other integration event, the construct was truncated before insertion. Since this construct inserted within the PHATRDRAFT_45242 gene locus, its unknown whether gene expression was being driven by the *fcpD* promoter in the gene locus, or the promoter upstream of the native gene. However, the possibility that truncated inserts can still be expressed through regulatory elements of non-essential or partially essential genes, and the fact the transgenes can integrate as concatemers, poses a challenge for genetic engineering when the quantity of the inserted sequences is crucial. This problem manifested itself in this study, as a proposed method for locating transgene inserts could not be assessed due to this phenomenon. Transgenic constructs were designed to have one I-SceI restriction site because this restriction site is absent in the rest of the *P. tricornutum* genome. The rationale was that next generation sequencing adapters cannot be ligated onto a DNA molecule if it is missing a phosphate group on the 5' terminus¹⁷. If

isolated DNA library is first dephosphorylated and then subsequently digested with I-*SceI*, then the sequencing adapters can only be ligated to phosphate groups revealed by the I-*SceI* digestion. In this way sequencing reads would only be acquired upstream and downstream of the I-*SceI* introduced on the construct and the location of insertion could be easily uncovered. This method depends on only one I-*SceI* restriction site having been inserted per integration site, as the lack of a site would mean this region would not be sequenced and more than one site would mean that only the flanking genomic regions would be sequenced but it would be impossible to see how the fragment inserted as a concatemer. The final detail of note from the sequencing data is that for each cell line where integration was confirmed, integration was only found at one genomic locus. Whether or not multiple integration events typically occur via electroporation could not be resolved in this study as the coverage obtained from sequencing reads could did not span 100% of the genome. A study by George et al. conducted a similar experiment on two transgenic *P. tricornutum* lines transformed through biolistic transformation, and observed that two separate integration loci were present for each cell line². Further research is required to determine whether this phenomenon is consistent in transformation via electroporation.

In conclusion, the results obtained in this chapter improve our ability to perform genetic manipulations on *P. tricornutum* provide novel insights into how these constructs get integrated in the nuclear genome. Future studies can be performed to assess the long-term stability of transgenic inserts and whether or not integration is completely random or whether certain points in the genome show a bias for integrating exogenous DNA.

2.5 References

1. Poveda-Huertes, D., Patwari, P., Günther, J., Fabris, M. & Andersen-Ranberg, J. Novel transformation strategies improve efficiency up to 10-fold in stramenopile algae. *Algal Res.* 103165 (2023) doi:10.1016/j.algal.2023.103165.
2. George, J. *et al.* Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology. *Front. Bioeng. Biotechnol.* **8**, 513 (2020).
3. Kassaw, T. K., Paton, A. J. & Peers, G. Episome-Based Gene Expression Modulation Platform in the Model Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **11**, 191–204 (2022).
4. Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E. & Weyman, P. D. Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods. *Front. Bioeng. Biotechnol.* **4**, 65 (2016).
5. Niu, Y.-F. *et al.* Transformation of diatom *Phaeodactylum tricornutum* by electroporation and establishment of inducible selection marker. *BioTechniques* **52**, 1–3 (2012).
6. Zhang, C. & Hu, H. High-efficiency nuclear transformation of the diatom *Phaeodactylum tricornutum* by electroporation. *Mar. Genomics* **16**, 63–66 (2014).
7. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel. PEDS* **23**, 155–159 (2010).
8. Slattery, S. S. *et al.* An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **7**, 328–338 (2018).

9. Hu, H. & Pan, Y. Electroporation Transformation Protocol for *Phaeodactylum tricornutum*. in *Electroporation Protocols: Microorganism, Mammalian System, and Nanodevice* (eds. Li, S., Chang, L. & Teissie, J.) 163–167 (Springer US, 2020).
doi:10.1007/978-1-4939-9740-4_17.
10. Shi, J. *et al.* A Review on Electroporation-Based Intracellular Delivery. *Mol. Basel Switz.* **23**, 3044 (2018).
11. Chuberre, C. *et al.* Comparative Proteomic Analysis of the Diatom *Phaeodactylum tricornutum* Reveals New Insights Into Intra- and Extra-Cellular Protein Contents of Its Oval, Fusiform, and Triradiate Morphotypes. *Front. Plant Sci.* **13**, 673113 (2022).
12. De Martino, A. *et al.* Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom *Phaeodactylum tricornutum*. *Protist* **162**, 462–481 (2011).
13. Willis, A., Chiovitti, A., Dugdale, T. M. & Wetherbee, R. Characterization of the extracellular matrix of *Phaeodactylum tricornutum* (Bacillariophyceae): structure, composition, and adhesive characteristics. *J. Phycol.* **49**, 937–949 (2013).
14. Bulankova, P. *et al.* Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. *Curr. Biol. CB* **31**, 3221-3232.e9 (2021).
15. Angstenberger, M., Krischer, J., Aktaş, O. & Büchel, C. Knock-Down of a ligIV Homologue Enables DNA Integration via Homologous Recombination in the Marine Diatom *Phaeodactylum tricornutum*. *ACS Synth. Biol.* **8**, 57–69 (2019).
16. Dai, J., Cui, X., Zhu, Z. & Hu, W. Non-homologous end joining plays a key role in transgene concatemer formation in transgenic zebrafish embryos. *Int. J. Biol. Sci.* **6**, 756–768 (2010).

17. McDonald, T. L. *et al.* Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun.* **12**, 3586 (2021).

Chapter 3

3 Remapping Genome Annotation Data from an Old *P. tricornutum* Genome Assembly to the Telomere-to-Telomere Genome Assembly

3.1 Introduction

Phaeodactylum tricornutum was one of the first two diatoms to have their nuclear genomes sequenced, with the first draft of the genome having been assembled by the Diatom Consortium in 2008¹. This assembly was constructed using paired-end whole genome shotgun sequencing. This sequencing method relies on randomly fragmenting genomic DNA molecules and sequencing the fragments in small portions without alignment to a reference map. The short reads of DNA are then hierarchically assembled into larger sequences known as contigs based on similar overlapping regions of other sequencing reads, and many contigs can be further assembled into larger fragments called scaffolds². The problem with this technology is that it is reliant on very short sequencing reads, which cannot unambiguously resolve large repetitive regions or the genome. As a result, regions such as centromeres, telomeres, and transposable element rich regions are often the least accurate regions in these kinds of assemblies, and assemblies generated using this approach are often referred to as “scaffold-level” assemblies as whole chromosomes cannot confidently be assembled³.

The Diatom Consortium assembly was re-analyzed by another group using long-read sequencing technologies and the researchers concluded that this genome assembly was likely fragmented and over-estimated the number of chromosomes in this organism’s nuclear genome. This group of researchers were unable to quantify the exact number of chromosomes the organism contained, in large part because assembly algorithms often

produce contig- and/or scaffold-level assemblies and provide no information on the number of chromosomes assembled⁴. Using nanopore sequencing and a novel approach termed “long-read-karyocounting”, the number of chromosomes within the *P. tricornutum* nuclear genome was determined to be 25⁵. A full telomere-to-telomere (T2T) length assembly of the nuclear genome further confirmed this value and is now established as the most accurate assembly of the nuclear genome⁵. However, in the time spanning the creation of the first draft assembly and the creation of the T2T assembly, dozens of RNA-Seq experiments, as well as some proteomics and epigenetic studies were performed, and all of this data was used to construct gene models and genome annotations in relation to the Diatom Consortium assembly. The most up-to-date and high-quality annotation effort was published in 2018 and resulted in the *Phaeodactylum tricornutum* annotation 3 (Phatr3), a collection of data consisting of 12,236 gene models and 69,809 total genome features⁶. In order to perform large-scale genome engineering and/or develop chromosomal or partial chromosomal replacement methods, it is vital to have an accurate map of where the genes of the organisms exist, especially those that are critical for the survival of the organism, in order to not unintentionally perturb or disrupt existing gene networks. Aligning genome models from one version an organisms’ reference genome to another can also elucidate any potential differences in the DNA sequences between two assemblies at key loci, and further investigation into the discrepancies can elucidate whether any sequence alterations can exist within a given feature or whether a sequence error exists in either of the assemblies. Therefore, the objective outlined here is to accurately update the genome coordinates for each feature in the Phatr3 annotation data and determine where it belongs on the *P. tricornutum* T2T assembly.

3.2 Materials and Methods

3.2.1 Data Extraction and Clean Up

Sequence data for the Diatom Consortium genome assembly (2008, GenBank accession ID: GCA_000150955.2), as well as sequence data for the mitochondria and chloroplast genome assemblies (NCBI Reference Sequences: NC_016739.1 and NC_008588.1, respectively), was extracted from the resource page of DiatOmicBase (¹). The data for this assembly (entitled the *Phaeodactylum tricornutum* annotation 3, or Phatr3) was also extracted from this resource page. The sequence data for the *P. tricornutum* telomere-to-telomere genome assembly (2021, GenBank accession ID: GCA_914521175.1) was extracted from NCBI. Phatr3 annotation data for mapped to the GCA_000150955.2 assembly was stored in a general feature format (GFF) file entitled *Phaeodactylum_tricornutum.ASM15095v2.52.gff3*. To prepare this file for downstream analyses, all header data was removed ensuring all lines in the file contained 9 tab-separated columns. The data in the last column, hereby referred to as the metadata, was also transformed to remove any characters, such as symbols and spaces, that might have been interpreted as special characters by certain programming languages and have caused problems in the downstream data analysis pipeline. Additionally, there were a few annotated features that had identical metadata to other features. To ensure that each annotated feature could be uniquely identified, a unique id value was generated and appended to the end of each metadata. This unique id value corresponded to the order that the feature appeared in the original Phatr3 file. A key-value pair file was then created to link the newly formatted metadata with the original versions so that the original metadata could be restored once remapping analysis was completed. The formatted Phatr3 file was then partitioned into multiple GFF files corresponding to the types of features that were

present in the original annotation file and all the corresponding data (CDS, chromosome, direct repeat, exon, 5'-UTR, gene, intron, lncRNA, mRNA, ncRNA, ncRNA gene, pseudogene, pseudogenic transcript, region, repeat region, rRNA, sequence feature, snoRNA, snRNA, supercontig, 3'-UTR, tRNA).

3.2.2 Sequence Extraction and Prefiltering

Sequence information from each of the features in the GFF files was extracted using bedtools getfasta (v2.30.0) using the Diatom Consortium genome assembly as the input FASTA file and using the -bed flag for the GFF files. Additionally, the -name flag was used to append the feature name and coordinates to the FASTA header and the -s flag was used to ensure that the reverse complement of the sequence will be extracted if the feature lies on the antisense strand of its chromosome or scaffold. Using a custom script, extracted sequences were analyzed and removed from downstream analysis if any non-resolved bases (N's) were present in their sequences. The remaining sequences were removed from downstream analysis if any features were shorter than 20 bp using another custom script.

3.2.3 Local and Global Alignments

Phatr3 features that were retained following preliminary filtering were searched against the telomere-to-telomere genome assembly using the blastn application of the command line blast package (v2.10.1). A BLAST database consisting of the chromosome sequences from the *P. tricornutum* T2T genome assembly and the mitochondrial and chloroplast assemblies was first constructed using the makeblastdb application of the command line blast package. A custom output format was specified for all the blastn queries to ensure that strand-specific information was retained in the output (flag -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send eval bitscore sstrand"). Row numbers (0-indexed) were also appended to BLAST outputs for the

downstream analyses to work. Sequence information of all of the blast outputs was then extracted by converting all the outputted blast outputs into BED format using a custom script and then by using bedtools getfasta (v2.30.0) with the telomere-to-telomere genome assembly as the input FASTA. Global alignments were performed using a custom script utilizing EMBOSS Needle (v6.3.0). Briefly, this script takes in FASTA files containing the sequence information for the BLAST outputs, as well as the FASTA files used as BLAST inputs. Sequences from BLAST outputs are searched against a dictionary of BLAST inputs. If the header names of the sequence in the BLAST outputs and input FASTA files match, a global alignment between the two features is performed using EMBOSS Needle. Following global alignment, the script checks to see how many of the first and last 10 bp between the two sequences match. The global alignment score, as well as the first and last 10 bp matching information for each feature are appended to the BLAST output file.

3.2.4 Assembly to Assembly Mapping

The old *P. tricornutum* Diatom Consortium genome assembly was mapped to the T2T genome assembly using minimap2 (v2.22-r1101, using the -x asm5 flag to optimize to optimize aligner for full genome/assembly alignment and outputted in PAF format). A custom script was used to remove alignments that had low mapping scores ($\text{MAPQ} < 20$) and small alignments (alignment < 450 bp). A text file containing chromosome sizes of the T2T genome assembly was also made for the following filtering to work. Updating the coordinates of the mappings from the Diatom Consortium assembly to the telomere-to-telomere assembly to remove overlapping alignments was performed with a custom script. Briefly, this script first sorts the mappings from largest to smallest. Then starting with the largest mapping, the region of the telomere-to-telomere genome that the mapping maps to is designated as a region that has a mapping and becomes a restricted region. The next

largest mapping is then considered. If part of the mapping falls within a restricted region, then that mapping is considered illegitimate and removed. To mitigate over-filtering, if the region of overlap is 20% or less than the size of the mapping itself, then that mapping is allowed to map and the region that overlaps with the restricted region is marked. This process is repeated until all the mappings have been processed. The remaining maps are then looped through again to remove any overlaps that they may have with other mappings. The start and end coordinates are updated on the target region (the region on the T2T assembly) and the query region (the region the mapping derived from on the Diatom Consortium assembly) of the mapping to remove overlaps. Which coordinates are updated is dictated by the location of the overlap and the strand of the mapping (Figure A-7). The final assembly-to-assembly mapping consists of the largest mappings with mostly unique mapping regions to the new genome with no overlaps. The GFF file containing the Phatr3 annotations is then looped through, and each feature is checked to see if its coordinates fall within a region of the Diatom Consortium assembly that got remapped to the telomere-to-telomere assembly. Coordinate and strand data related to how the mapping that the feature pertains to maps to the new assembly is then appended to the Phatr3 annotation file.

3.2.5 Filtering and Calling Reciprocal Best Hits

Prior to filtering, all the data pertaining to the assembly-to-assembly remapping and the feature type and size information was appended to the BLAST output. Preliminary filtering based on end alignments and feature sizes was then performed. The output was further filtered to maintain features that had BLAST outputs consistent with the assembly-to-assembly mapping results. Reciprocal best hits were then called. All of these steps were performed using custom scripts.

3.2.6 Updating Coordinates

The file containing all the reciprocal best hits was formatted into a GFF file. The file was then grouped by chromosome number and then sorted in ascending order by feature start coordinates. Annotation data from the mitochondrial and chloroplast genomes were then appended to this file. Finally, the custom metadata was replaced with the original metadata for the remapped features. All of these steps were performed using custom scripts.

3.2.7 Calculating Success Rate Chromosome by Chromosome

For each feature in the original Phatr3 file, columns were added to append the start and end coordinates of where the portion of the Diatom Consortium assembly harboring the feature mapped to on the T2T assembly, as well as the chromosome from the T2T assembly that it was mapped to. Features that were not captured within a mapped fragment in the assembly-to-assembly mapping were excluded from analysis. The number of features estimated to map to each chromosome in the T2T assembly based off the assembly-to-assembly mapping was then compared to the actual number of features remapped for each chromosome and computed as a success percentage. All steps were performed using custom scripts.

3.2.8 Data availability

All files used in this study (including assembly files, annotation files, and scripts used) can be found https://github.com/mpampuch/pt_genome_remapping along with detailed instructions for how to perform each step of the process. Remapped and unremapped annotation data is hosted on the University of California Santa Cruz Genome Browser portal and can be accessed through https://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=1828134926_AOVBj4mafegeRJoItfGJ2k2cDR0R alongside an interactive genome browser of the T2T assembly with the remapped annotation data.

3.3 Results

The strategy implemented to remap the annotation data from the Diatom Consortium assembly to the T2T assembly was a reciprocal best hit approach, a technique often used to find orthologous genes across different species⁷. For the purposes of remapping genomic data, this approach was employed by first performing a local alignment (subsequence alignment) of a genomic feature from the old genome assembly to the new genome assembly using BLAST⁸. All the subject sequences produced by the alignment were then globally aligned (end-to-end alignment) back to the original genomic feature using EMBOSS Needle⁹. Subject sequences that had the highest global alignment score were then candidates for successfully remapped genomic features. This was done on all features from the old assembly except those that contained any “N” symbols in their sequences and that had a feature size less than 20 bp, as both of these could cause problems for the alignment softwares or filtering steps. Before a feature was deemed successfully remapped, a few additional filtering steps were employed to mitigate false-positive remappings: (1) The first and last base pair of the feature from the old assembly and the new assembly must be identical. This was done to ensure that the entire sequence of the original feature was found in the local alignment as alignment softwares tend to offer the least accuracy towards the ends of the query sequences. (2) At least nine of the first and last ten base pairs of the feature from the old assembly and the new assembly must be identical. Since there is a real possibility that the very first and last base pair of a query and subject sequence can match purely by random chance, this filtering step was employed as an extra precaution to ensure the entire sequence is found in the local alignment output. (3) The first three and last three base pairs of the feature from the old assembly and the new assembly must be identical if the feature is annotated as a gene, transcript, CDS, or exon.

This was done to mitigate any frameshifted sequences that may have passed the first two filtering steps from remapping to the new assembly. (4) The length of the feature from the old assembly must not be 50% longer or shorter than the feature in the new assembly. (5) The length of the feature from the old assembly must not be 10% longer or shorter than the feature in the new assembly if the gene, transcript, CDS, or exon. These features were thought to be more conserved than other features in the genome, therefore a more stringent filtering cut-off was applied. (6) The feature from the old chromosome maps to an expected region on the new assembly. (7) The feature from the old chromosome maps to the expected strand on the new assembly. These last two filters were applied because based on where and on which strands the chromosomes and DNA scaffolds in the old assembly align to the chromosomes in the new assembly, the general vicinity and on which strands the genes and genome features pertaining to those chromosomes and scaffolds will map to can be predicted as well. Therefore, a high-quality assembly-to-assembly alignment was performed from the Diatom Consortium assembly to the T2T assembly. The alignment was performed using minimap2¹⁰, and only assembly-to-assembly mappings with a mapping quality score of 20 or higher, and a size greater than 450 bp were retained (450 bp was chosen as a cut-off because the smallest scaffold in the old assembly that still contains annotated genomic data is 450 bp). The mappings were then sorted in descending order by size, and then analyzed from largest to smallest to check if they overlap with any other mappings, in which case they were subsequently filtered out from the mapping file. To mitigate over filtering and allow for cases where old assembly chromosomes and scaffolds overlap with other ones in the new genome assembly, mappings were allowed to overlap if the size of the overlapping region was not greater than 20% of the size of the mapping.

The last step to ensure the assembly-to-assembly mapping could unambiguously help remap genome features between assemblies was to update the remapping coordinates to remove any overlaps in the final alignment (detailed in Figure A-7). This was done to ensure that no remapping conflicts could occur if two reciprocal best hits came from two regions of the original assembly that overlap with each other on the assembly-to-assembly mapping and map to the same location on the new assembly. After preliminary filtering was performed, reciprocal best hits were called by first grouping the BLAST outputs to their query sequences, sorting the outputs from highest-to-lowest global alignment score, removing any edge cases where two or more alignments had the highest global alignment score, and then filtering out all alignments that did not have the highest global alignment score. A flowchart overviewing the remapping procedure can be found in (detailed in Figure A-6). Of the initial 69,070 annotated genomic data features, 56,624 features remapped successfully, whereas 9250 features did not pass the additional filtering, 1532 features were not found in the following the BLAST alignments, 936 features were smaller than 20 bp, and 48 features had N's in their sequences (Figure 3-1).

P. tricornutum Telomere-to-Telomere Assembly Remapping Summary

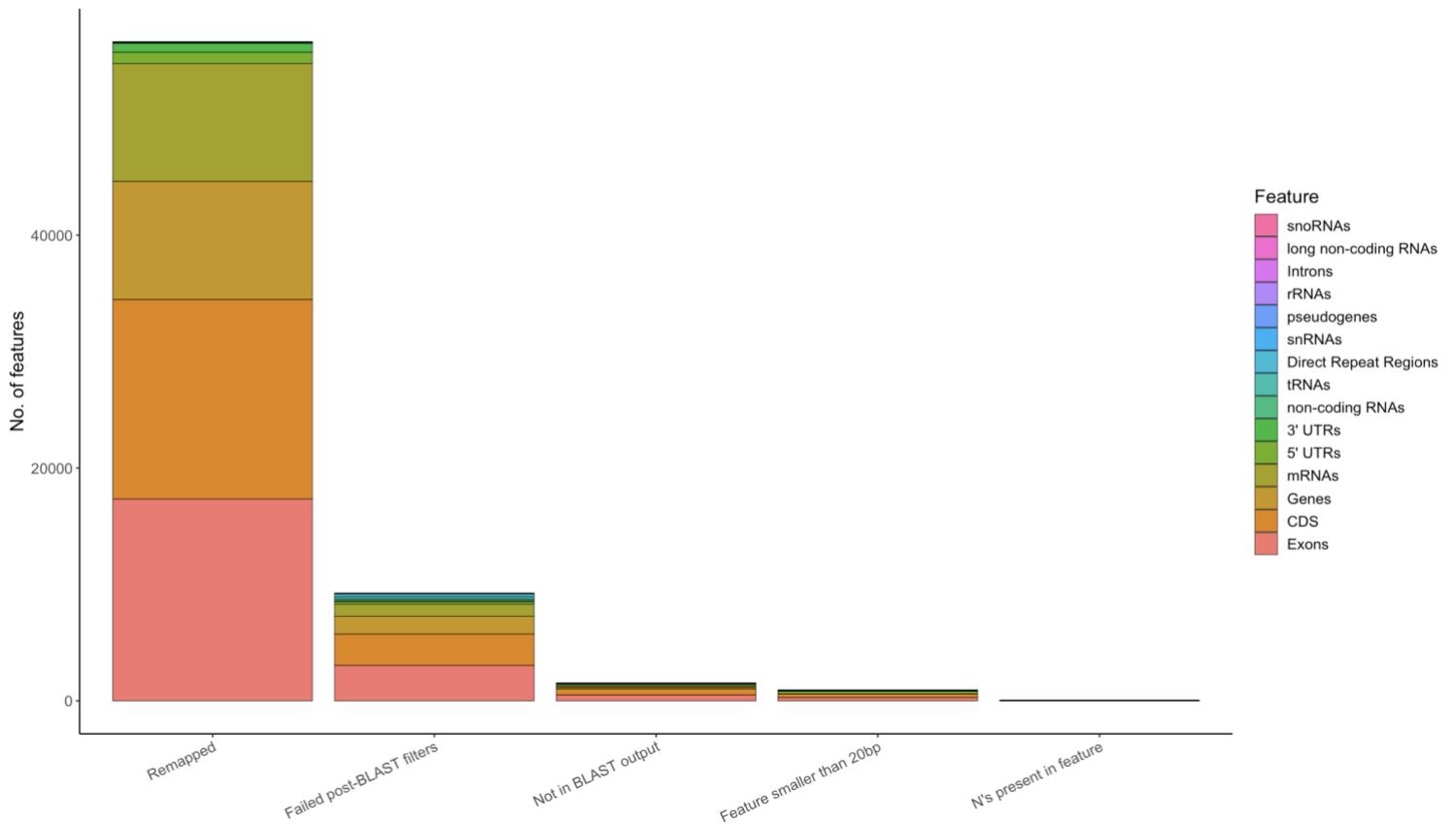
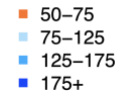
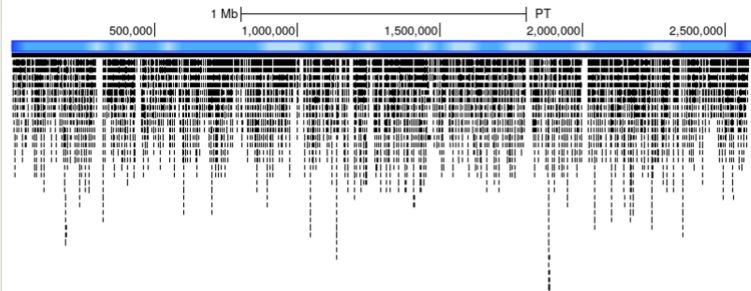
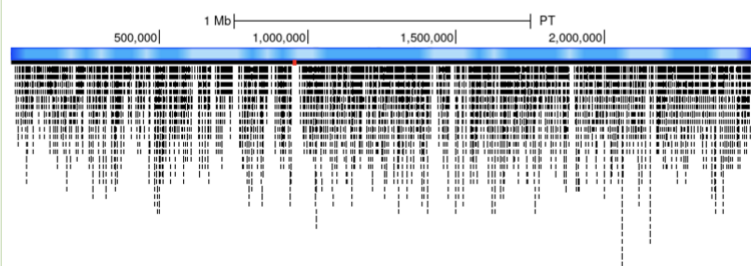
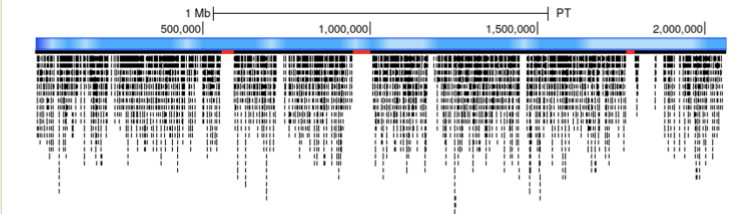
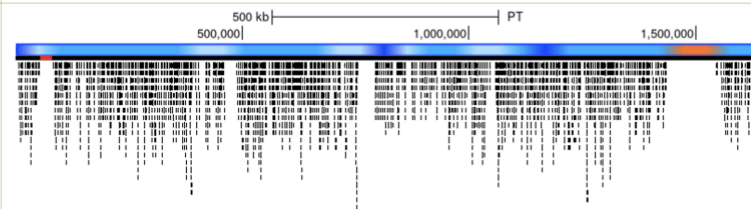
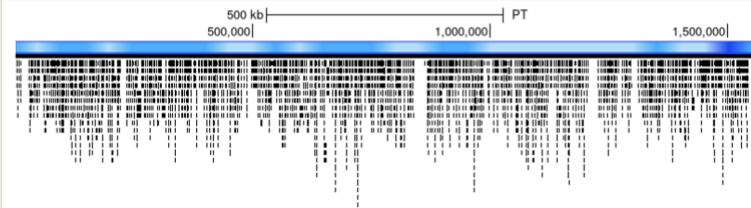


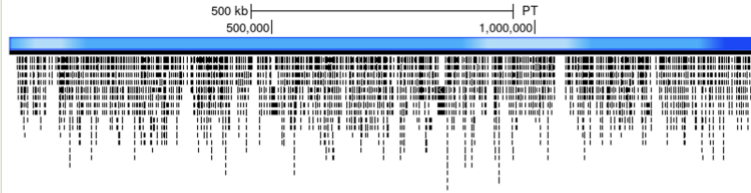
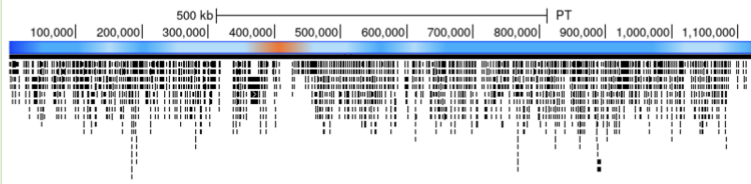
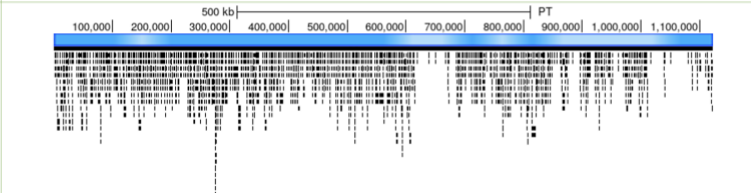
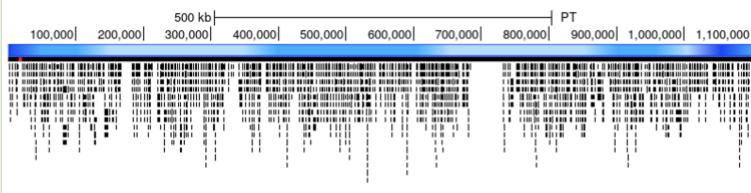
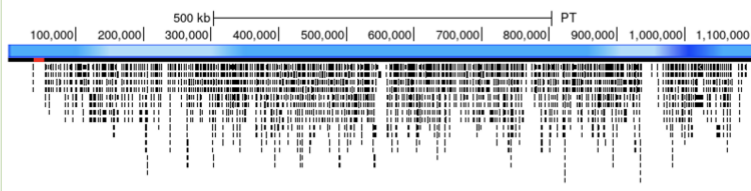
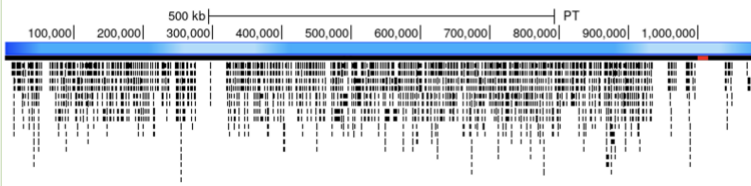
Figure 3-1: Genome-wide summary of remapped features to the telomere-to-telomere assembly of *P. tricornutum*.

Since *P. tricornutum* is a diploid organism that has shown to exhibit different levels of heterozygosity on different chromosomes, the number remapped features were checked chromosome by chromosome to detect if heterozygosity could have had an effect on the remapping success. In order to quantify this, an expected value of how many features were expected to map to each chromosome needed to be established. This was done by isolating all the assembly-to-assembly mappings that mapped to the chromosome of interest and then extracting all the genome features from the old assembly pertaining the regions where the mappings derived from. The success rate per chromosome was calculated to range from 67% to 91.7% (detailed in Figure 3-2).

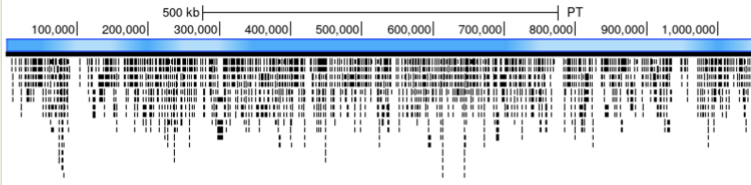
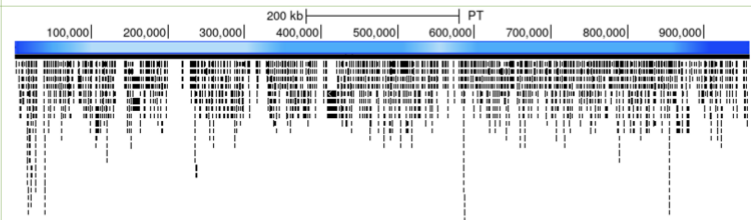
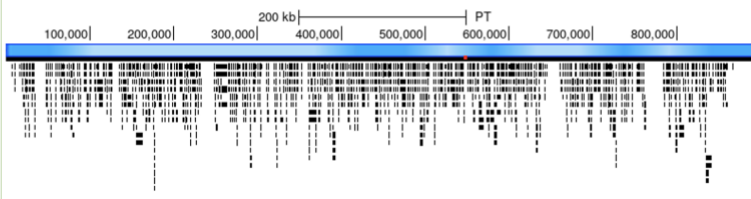
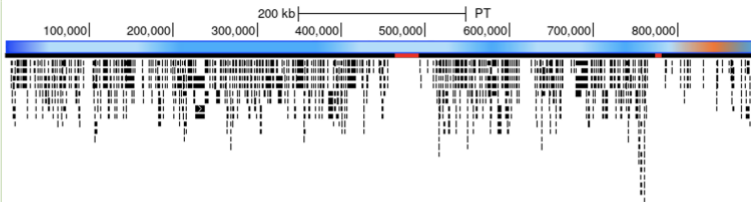
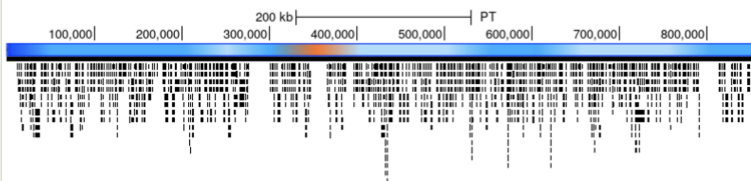
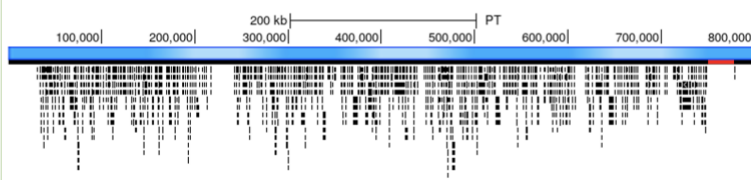
Coverage depth

**Chromosome 1** 2,608,419 bpRemapped features: [6224](#)Expected: [6787](#)Data retained after filtering: [91.7%](#)**Chromosome 2** 2,449,861 bpRemapped features: [5425](#)Expected: [6256](#)Data retained after filtering: [86.7%](#)**Chromosome 3** 2,064,744 bpRemapped features: [4470](#)Expected: [5279](#)Data retained after filtering: [84.6%](#)**Chromosome 4** 1,629,129 bpRemapped features: [3036](#)Expected: [3605](#)Data retained after filtering: [84.2%](#)**Chromosome 5** 1,555,020 bpRemapped features: [3328](#)Expected: [3864](#)Data retained after filtering: [86.1%](#)

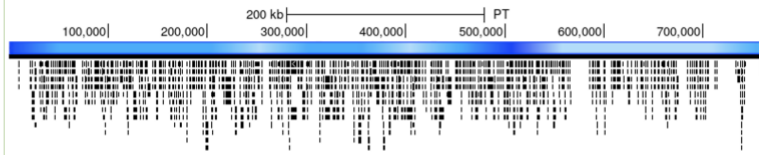
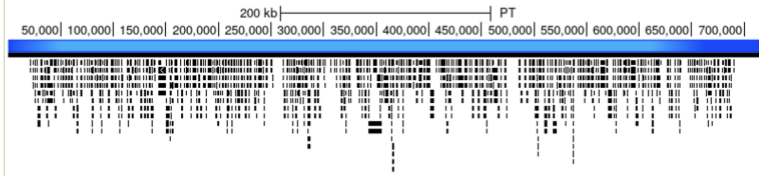
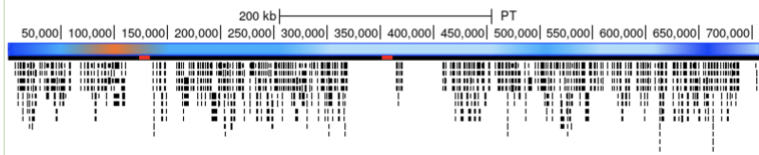
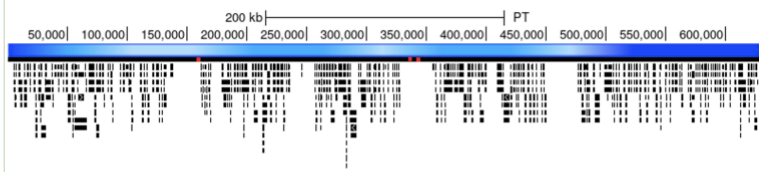
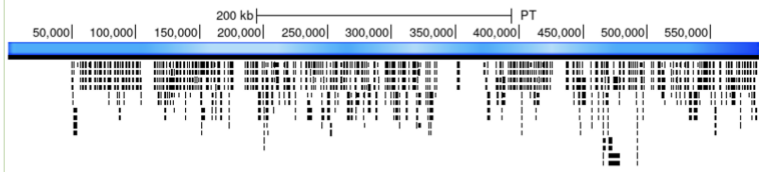
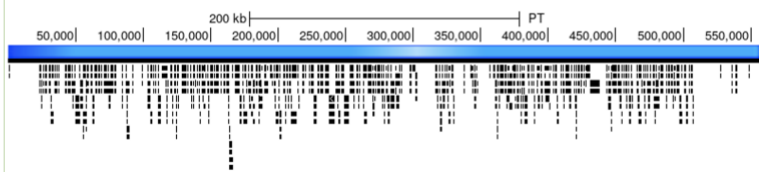
(Figure extends to next page)

Chromosome 6 1,417,157 bpRemapped features: 3110Expected: 3510Data retained after filtering: 88.6%**Chromosome 7** 1,124,623 bpRemapped features: 2180Expected: 2638Data retained after filtering: 82.6%**Chromosome 8** 1,222,386 bpRemapped features: 2194Expected: 2590Data retained after filtering: 84.7%**Chromosome 9** 1,108,211 bpRemapped features: 2382Expected: 2773Data retained after filtering: 85.9%**Chromosome 10** 1,107,389 bpRemapped features: 2417Expected: 2753Data retained after filtering: 87.7%**Chromosome 11** 1,087,446 bpRemapped features: 2190Expected: 2473Data retained after filtering: 88.6%

(Figure extends to next page)

Chromosome 12 1,052,234 bpRemapped features: 2160Expected: 2499Data retained after filtering: 86.4%**Chromosome 13** 959,323 bpRemapped features: 2249Expected: 2579Data retained after filtering: 87.2%**Chromosome 14** 898,576 bpRemapped features: 1821Expected: 2211Data retained after filtering: 82.4%**Chromosome 15** 897,230 bpRemapped features: 1403Expected: 1757Data retained after filtering: 79.9%**Chromosome 16** 860,830 bpRemapped features: 1644Expected: 1997Data retained after filtering: 82.3%**Chromosome 17** 803,256 bpRemapped features: 1552Expected: 1781Data retained after filtering: 87.1%

(Figure extends to next page)

Chromosome 18 759,751 bpRemapped features: 1533Expected: 1839Data retained after filtering: 83.4%**Chromosome 19** 716,929 bpRemapped features: 1482Expected: 1645Data retained after filtering: 90.1%**Chromosome 20** 709,265 bpRemapped features: 1133Expected: 1377Data retained after filtering: 82.3%**Chromosome 21** 629,758 bpRemapped features: 1047Expected: 1563Data retained after filtering: 67%**Chromosome 22** 587,839 bpRemapped features: 1049Expected: 1218Data retained after filtering: 86.1%**Chromosome 23** 557,589 bpRemapped features: 993Expected: 1196Data retained after filtering: 83%

(Figure extends to next page)

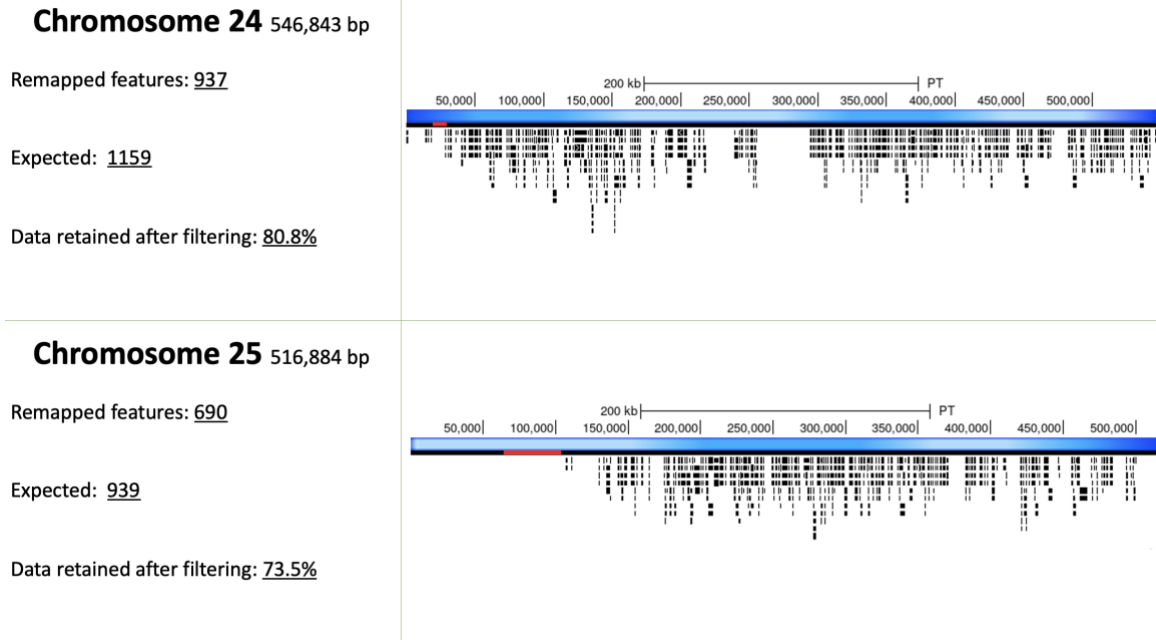


Figure 3-2: Chromosome-wide summary of remapped features to the telomere-to-telomere assembly of *P. tricornutum*. Each black bar indicates a remapped Phatr3 feature and its location relative to the T2T assembly chromosome. Chromosome coverage plots from Giguere et al., (2021) were overlaid for each chromosome. Coverage depth across each chromosome is denoted by blue or orange colouration. Orange colouration indicate regions where sequencing depth was 50 – 75X. Light blue colouration indicate regions where sequencing depth was 75 – 125X. Blue colouration indicate regions where sequencing depth was 125 – 175X. Dark blue colouration indicate regions where sequencing depth was 175X+. Red lines indicate regions where overlaps were removed from the assembly-to-assembly mapping.

3.4 Discussion

The remapping procedure outlined here was able to confidently remap 81.8% of the Phatr3 annotated genome models to the T2T genome assembly. Various bioinformatics tools exist to perform assembly-to-assembly annotation data remapping, however the accuracy of these tools relies heavily on the similarity between the two assemblies and validation of the results is challenging^{11,12}. Although the same strain of *P. tricornutum* was used to assemble the 2008 Diatom Consortium and the T2T assembly (strain CCAP

1055/1), the two genome assemblies differ significantly. *k*-mer completeness is a metric that is used to estimate the completeness of a genome assembly. *k*-mers are (DNA subsequences of length *k* and by comparing the *k*-mers found in high quality sequencing reads to the those found in a *de novo* genome assembly, the portion of the *k*-mers found in the sequencing library that appear in the final reference genome in the be intuited¹³. For both the T2T and Diatom Consortium assembly, the *k*-mer completeness estimate was only 80%⁵. This may be due to *P. tricornutum* having an intrinsically high level of heterozygosity between both homologs for most of its chromosomes that are not captured in either of the genome assemblies. Both the T2T and the Diatom Consortium genome assemblies are collapsed assemblies, meaning that each chromosome has only one reference sequence despite each chromosome having two homologs in the nucleus. Any variation between the two homologs will have been flattened down into a single sequence by the genome assembly algorithm^{5,14}. *P. tricornutum* is also known to have a high degree of intraspecies genomic variation, and genomic changes occur very often between mother and daughter cells¹⁵. Diatoms are a very ecologically successful organisms and occupy a wide array of ecological niches, implying that these organisms are very effective at producing genomic variation in their populations¹⁶. Despite this, sexual reproduction and meiosis are rarely observed in these organisms^{17,18}. Instead, diatoms have been observed to exhibit a large degree of interhomolog mitotic recombination that produce genetic compositions similar to those generated through meiotic recombination¹⁵. Further investigation needs to be performed to determine whether this phenomenon is responsible for the low level of *k*-mer completeness found in this diatom, however as it stands the differences in sequences between the two assemblies may be up to 80%. This number is

strikingly similar to the percentage of Phatr3 features that successfully remapped between the assemblies, therefore homolog specific sequence differences may be a partial explanation for why some Phatr3 features could not be confidently remapped using this methodology. One potential solution for this problem could be to perform “phasing” on the T2T genome assembly, whereby the sequence for each chromosome is parsed into two reference sequences indicative of the two homologs in the nuclear genome^{14,19}. Repeating the remapping process on a “haplotype resolved” assembly could provide a higher success rate than that observed here on a collapsed assembly. Since there may be a large amount of sequence variation between the two assemblies, it would also be valuable to compare how similar the sequences are between the Phatr3 gene models on the Diatom Consortium assembly and the sequences that are derived from the Phatr3 gene models mapped to the T2T assembly. The sequence variability at the gene loci is likely to be much lower to that of intergenic regions, as these regions are more likely much more heavily conserved, however it still needs to be determined whether translating each sequence *in silico* will result in the exact same polypeptide sequence. Such an analysis may elucidate potential errors made in the assembly of either reference genomes or novel gene variants within the species. Finally, chromosome-level analysis revealed that many regions of the T2T chromosomes that contain low levels of remapped Phatr3 annotations correspond to regions of low coverage depth in the T2T genome assembly. This is very well exemplified on chromosome 4 in Figure 3-2, where a lack of annotation data can be observed in the region of high sequencing variability (coloured in orange). It was noted in the T2T assembly study that regions of low coverage during genome assembly are highly correlated with predicted dense regions of long terminal repeat (LTR) retrotransposons⁵. This would

explain the lack of remapped data to those regions as features that exhibit a lot of sequence similarity to other features or many portions of the genome are one of the hardest features to confidently remap. The LTR-rich regions of the *P. tricornutum* genome are one of the most repetitive regions as LTR retrotransposons are predicted to be the most common transposable elements in *P. tricornutum*.

In conclusion, the results obtained in this chapter improve the quality of the *P. tricornutum* reference genome and provide a good reference for future remapping initiatives.

3.5 References

1. Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
2. Venter, J. C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
3. Mao, Y. & Zhang, G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nat. Methods* **19**, 635–638 (2022).
4. Filloramo, G. V., Curtis, B. A., Blanche, E. & Archibald, J. M. Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics* **22**, 379 (2021).
5. Giguere, D. J. *et al.* Telomere-to-telomere genome assembly of *Phaeodactylum tricornutum*. *PeerJ* **10**, e13607 (2022).
6. Rastogi, A. *et al.* Integrative analysis of large-scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci. Rep.* **8**, 4834 (2018).

7. Geiß, M., Stadler, P. F. & Hellmuth, M. Reciprocal best match graphs. *J. Math. Biol.* **80**, 865–953 (2020).
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
9. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
10. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* **34**, 3094–3100 (2018).
11. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinforma. Oxf. Engl.* **30**, 1006–1007 (2014).
12. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinforma. Oxf. Engl.* **37**, 1639–1643 (2021).
13. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
14. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
15. Bulankova, P. *et al.* Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. *Curr. Biol. CB* **31**, 3221–3232.e9 (2021).
16. Butler, T., Kapoore, R. V. & Vaidyanathan, S. *Phaeodactylum tricornutum*: A Diatom Cell Factory. *Trends Biotechnol.* **38**, 606–622 (2020).
17. Mao, Y. *et al.* Sexual reproduction potential implied by functional analysis of SPO11 in *Phaeodactylum tricornutum*. *Gene* **757**, 144929 (2020).

18. Bowler, C. & Falciatore, A. *Phaeodactylum tricornutum*. *Trends Genet. TIG* **35**, 706–707 (2019).
19. Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).

Chapter 4

4 General Discussion

4.1 Transformation and Remapping Data enable Large Scale Genome Engineering Efforts

The data obtained in this thesis improves our understanding of the genetic basic of *P. tricornutum* and opens the door for larger-scale engineering efforts to be performed. After optimizing a method for integrating exogenous DNA into the nuclear genome of *P. tricornutum*, the next objective was to introduce landing pads, or sites where DNA can be recombined deliberately, into the genome. This would allow for the possibility of achieving a major objective of the Synthetic Diatoms project, which is to develop a partial chromosome replacement system through the use of controlled translocations¹. For such a system to be implemented, two primary stipulations need to be met. (i) A method for “swapping in” a large exogenous piece of DNA with the end of a chromosome needs to be achievable, and (ii) knowledge about where essential genes lie needs to be established. Based on the results obtained from this study, a *Cre/lox* system for replacement chromosome fragments can be attempted². Transformation cassettes containing a *loxP* site can electroporated into *P. tricornutum*. If a *loxP* harboring cassette integrates towards the end of a nuclear chromosome, then this site can act as a “landing pad” for edits to be made. If the *Cre/lox* recombination system can be shown to work in *P. tricornutum*, then by transforming modified chromosome fragment harboring a *loxP* site via a conjugative vector, excising it, and then introducing the *Cre* recombinase into the organism, translocations can occur at the *loxP* sites, swapping in the exogenous chromosome fragment with the native one. One caveat is that maintenance and transformation of DNA fragments becomes more difficult the larger the DNA is, hence if modified chromosome

fragments can be reduced in size, this may streamline any genome editing pipelines^{3,4}. To this end, removing any non-essential genes and intergenic regions can be done if these regions are known. Electroporation on a large scale can be an effective way of determining non- or quasi-essential genome regions. Researchers from the J. Craig Venter Institute were able to a strain of *Mycoplasma mycoides* with a minimized genome using a similar strategy of creating a disruption map of randomly integrated cassettes⁵. *P. tricornutum* is a diploid organism and hence performing a gene essentiality experiment is much more difficult than in haploid organisms, but this methodology can still provide of a good indicator of which genes or genome regions are non-essential in a homozygous context. This is even demonstrated through this study as a cell line generated by the heterozygous knockout of the predicted gene PHATRDRAFT_45242 was established and has been shown to be viable. Once essential genes are known, the other piece of the puzzle is knowing where they exist with the genome. To this end, the data obtained through the genome remapping project provides substantial progress towards elucidating this information.

4.2 Conclusions

In conclusion, *P. tricornutum* has a high potential to become leading industrial microbe in a sustainable bioeconomy. The results obtained in this study provide new insights into the underlying biological mechanisms and genome of *P. tricornutum*. The genetic engineering advances made will increase the rate at which basic and applied diatom research can be performed and move us one step closer to performing ambitious synthetic biology scale research on this organism.

4.3 References

1. Pampuch, M., Walker, E. J. L. & Karas, B. J. Towards synthetic diatoms: The *Phaeodactylum tricornutum* Pt-syn 1.0 project. *Curr. Opin. Green Sustain. Chem.* **35**, 100611 (2022).
2. McLellan, M. A., Rosenthal, N. A. & Pinto, A. R. Cre-loxP-Mediated Recombination: General Principles and Experimental Considerations. *Curr. Protoc. Mouse Biol.* **7**, 1–12 (2017).
3. Cochrane, R. R. *et al.* Rapid method for generating designer algal mitochondrial genomes. *Algal Res.* **50**, 102014 (2020).
4. George, J. *et al.* Metabolic Engineering Strategies in Diatoms Reveal Unique Phenotypes and Genetic Configurations With Implications for Algal Genetics and Synthetic Biology. *Front. Bioeng. Biotechnol.* **8**, 513 (2020).
5. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).

Appendices

Appendix A: Supplementary Figures

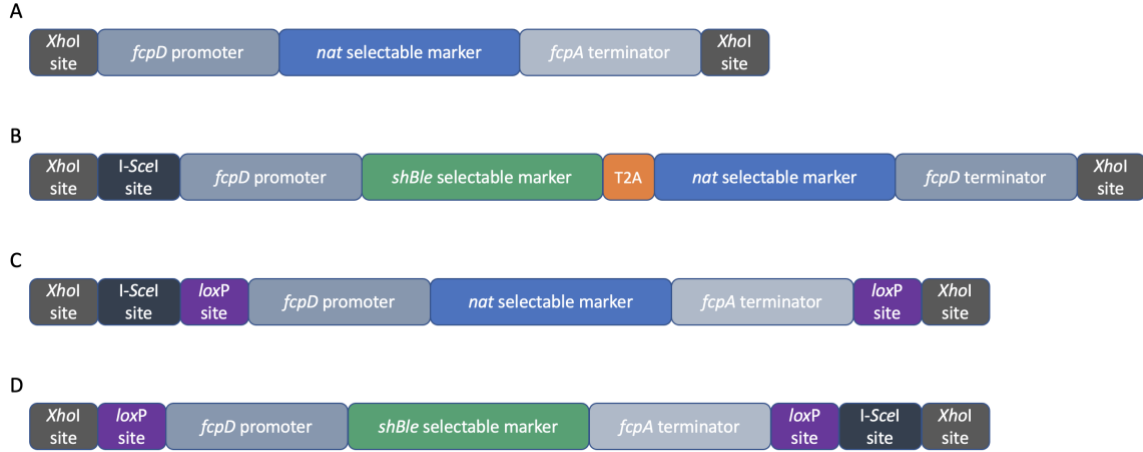


Figure A-1: Schematic of important transformation cassettes used in this study. (A) *nat* single selectable marker cassette. **(B)** *shBle*-T2A-*nat* double selectable marker cassette. **(C)** *nat* single selectable marker cassette with *loxP* sites. **(D)** *shBle* single selectable marker cassette with *loxP* sites.

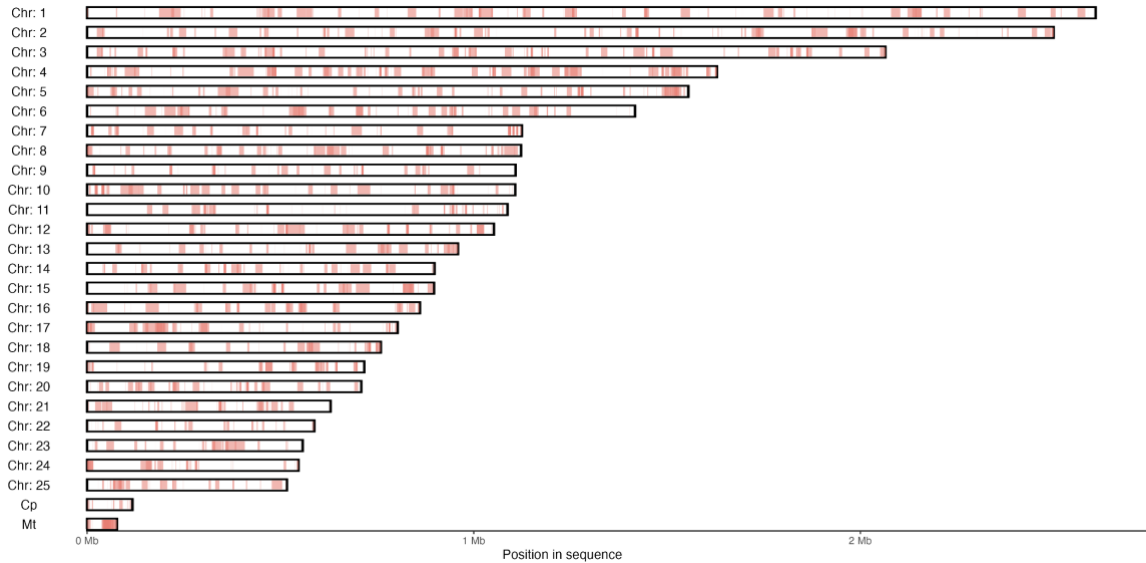


Figure A-2: Potential coverage plot of sequencing reads obtained from transformation line Z-Z1. Red regions indicate where sequencing reads potentially mapped to in the genome, as predicted by minimap2. Cp denotes chloroplast genome (represented linearly). Mt denotes mitochondrion genome (represented linearly). One instance of transformation marker integration was discovered in this cell line.

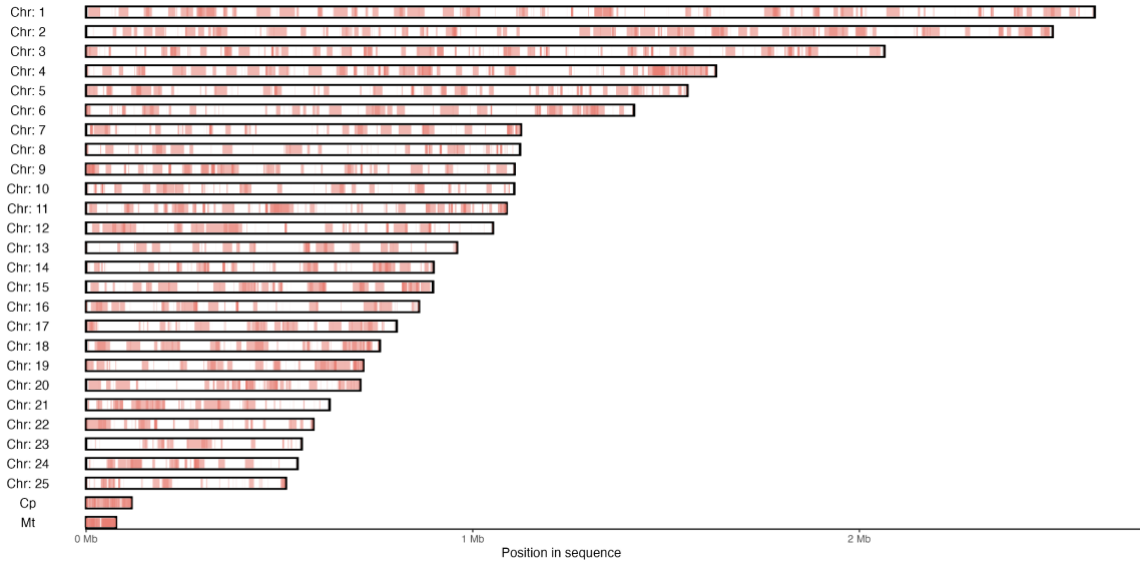


Figure A-3: Potential coverage plot of sequencing reads obtained from transformation line Z-Z₂. Red regions indicate where sequencing reads potentially mapped to in the genome, as predicted by minimap2. Cp denotes chloroplast genome (represented linearly). Mt denotes mitochondrion genome (represented linearly). One instance of transformation marker integration was discovered in this cell line.

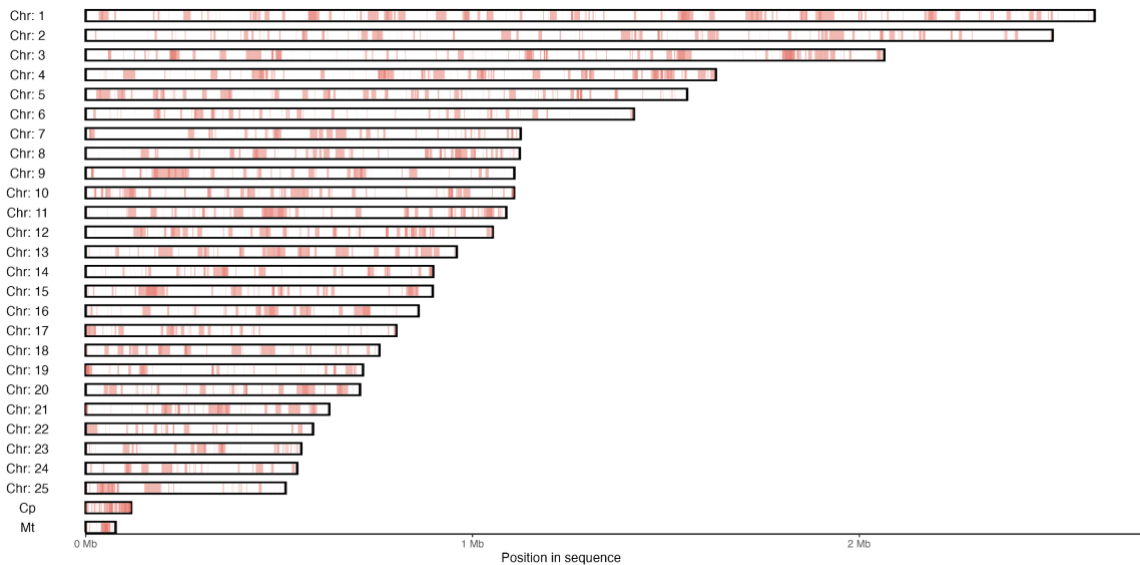


Figure A-4: Potential coverage plot of sequencing reads obtained from transformation line N-N₁. Red regions indicate where sequencing reads potentially mapped to in the genome, as predicted by minimap2. Cp denotes chloroplast genome (represented linearly). Mt denotes mitochondrion genome (represented linearly). No instances of transformation marker integration were discovered in this cell line.

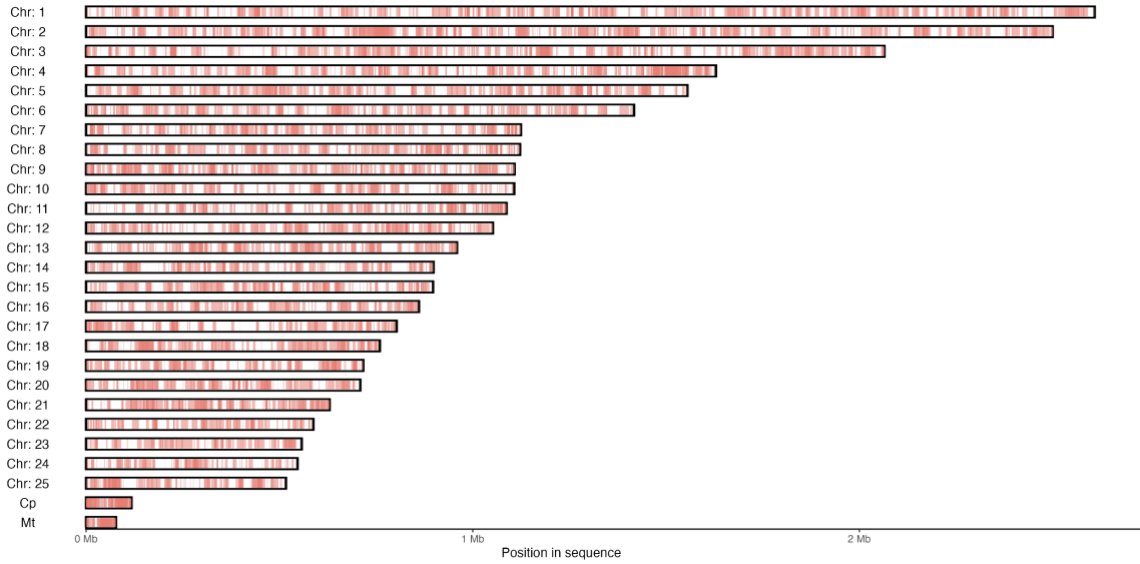


Figure A-5: Potential coverage plot of sequencing reads obtained from transformation line N-N₂. Red regions indicate where sequencing reads potentially mapped to in the genome, as predicted by minimap2. Cp denotes chloroplast genome (represented linearly). Mt denotes mitochondrion genome (represented linearly). No instances of transformation marker integration were discovered in this cell line.

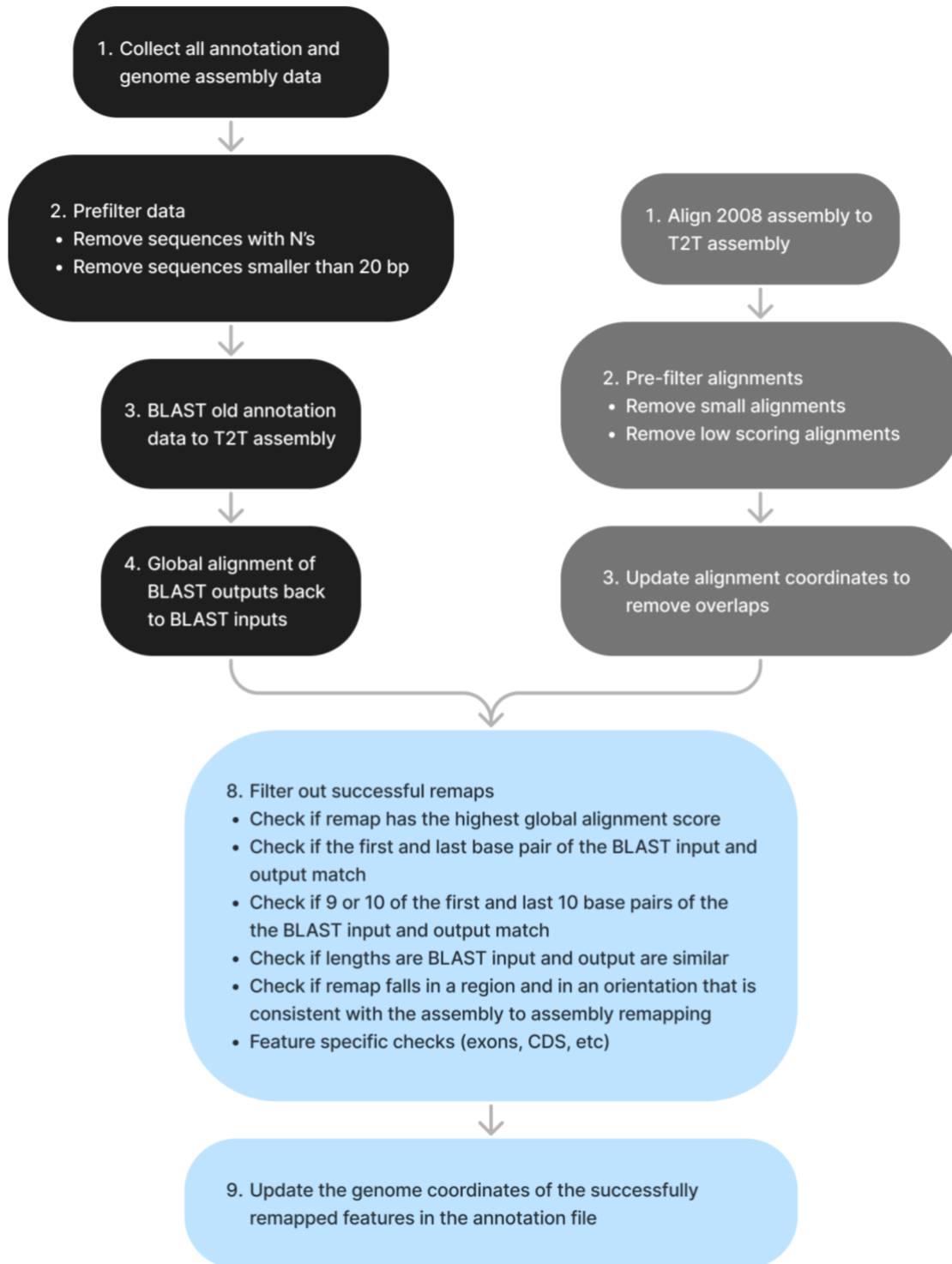


Figure A-6: Flowchart of remapping process.

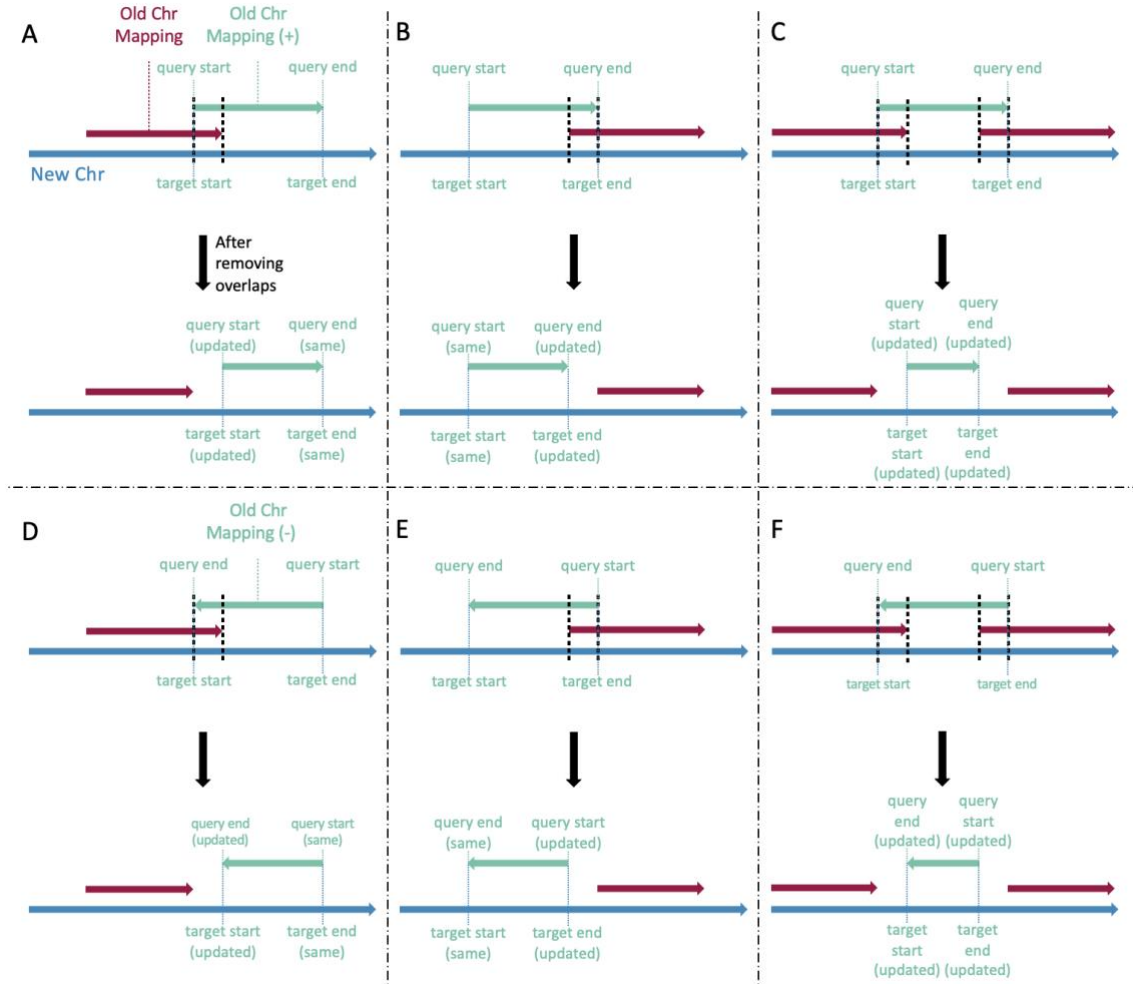


Figure A-7: Diagram describing genome coordinate modifications that need to be made to mappings to remove overlaps from old assembly to new assembly mapping.

(A) If the overlap exists on the left side (relative to the sense strand of the new assembly chromosome) of the old assembly mapping, and the mapping is on the sense (+) strand of the new chromosome, then the query (region on old assembly) start coordinates and target (region on new assembly) start coordinates must be updated. (B) If the overlap exists on the right side (relative to the sense strand of the new assembly chromosome) of the old assembly mapping, and the mapping is on the sense strand of the new chromosome, then the query end coordinates and target end coordinates must be updated. (C) If the overlap exists on the left and right side of the old assembly mapping, and the mapping is on the sense strand of the new chromosome, then all the genomic coordinates need to be updated. (D) If the overlap exists on the left side of the old assembly mapping, and the mapping is on the antisense (-) strand of the new chromosome, then the query start coordinates and target start coordinates must be updated. (E) If the overlap exists on the right side of the old assembly mapping, and the mapping is on the antisense strand of the new chromosome, then the query start coordinates and target end coordinates must be updated. (F) If the overlap exists on the left and right side of the old assembly mapping, and the mapping is on the antisense strand of the new chromosome, then all the genomic coordinates need to be updated.

Appendix B: Primers and Genetic Parts

Table B-1: Oligonucleotides used in this study.

Name	Sequence (5' to 3')	Description
BK1980_F	GAGCTGTAAGTACATCACCGACGAGCAAGGCAAGACGATCAATC AAAAAACCCACCTTTCT	Forward primer: Amplifies <i>oriC</i> region of pAL1 backbone (used for transformation negative control)
BK1980_R	AGGGTTTTCCCAGTCACGACATTAACCTCACTAAAGGGAATTTT ACCCCTTTTATTAAT	Reverse primer: Amplifies <i>oriC</i> region of pAL1 backbone (used for transformation negative control)
BK2510_F	CTCGAGATTGGGATATCTCGCTCGTG	Forward primer: Amplifies <i>nat</i> from pPtGE27 with <i>XhoI</i> sites.
BK2510_R	CTCGAGCCCTGGTTGAGTTCGATAGC	Reverse primer: Amplifies <i>nat</i> from pPtGE27 with <i>XhoI</i> sites.
BK2598_F	GACCAAGGTGTTCCCGA	Forward primer: Genotyping <i>nat</i> CDS (283 bp)
BK2598_R	GTTGACGTTGGTGACCTCC	Reverse primer: Genotyping <i>nat</i> CDS (283 bp)
BK2618_F	ATGGGGTTCACCCTCTGC	Forward primer: Inverse PCR primer inside <i>nat</i> CDS.
BK2618_R	GAAGACGGTGTCGGTGGT	Reverse primer: Inverse PCR primer inside <i>nat</i> CDS.
BK2632_F	ACCTCTACGGGCCAAAGATT	Forward primer: Genotyping <i>fcgA</i> gene on chromosome 2 for construct integration (1453 bp).
BK2632_R	GGCTCATAGTCGGTTTGGGA	Reverse primer: Genotyping <i>fcgA</i> gene on chromosome 2 for construct integration (1453 bp).
BK2790_F (b)	CTCGAGATAACTTCGTATAATGTATGCTATACGAAGTTATTGTGA GCGGATAACAATTTCA	Forward primer: Amplifies <i>shBle</i> with <i>loxP</i> sites, <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2790_R (b)	CTCGAGATAACTTCGTATAATGTATGCTATACGAAGTTATGCTAG TGTTATTCCTGACTG	Reverse primer: Amplifies <i>shBle</i> with <i>loxP</i> sites, <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2791_F (b)	CTCGAGATAACTTCGTATAATGTATGCTATACGAAGTTATACTAG CTTGATTGGGATATC	Forward primer: Amplifies <i>nat</i> with <i>loxP</i> sites, <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2791_R (b)	CTCGAGTAGGGATAACAGGGTAATATAAAGTTCGTATAATGTATG CTATACGAAGTTATAACAATTTACACAGGAAAC	Reverse primer: Amplifies <i>nat</i> with <i>loxP</i> sites, <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2821_F	CTCGAGTAGGGATAACAGGGTAATACTAGCTTGATTGGGATATC TCG	Forward primer: Amplifies <i>shBle-T2A-nat</i> from pPtGE32 with <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2821_R	CTCGAGGACGTTTTCACTCTCGAGCACAGGTTTTTTACTAATTG	Reverse primer: Amplifies <i>shBle-T2A-nat</i> from pPtGE32 with <i>XhoI</i> sites and a I- <i>SceI</i> site.
BK2923_F	GACCAAGGTGTTCCCGA	Forward primer: Genotyping <i>nat</i> CDS (225 bp).
BK2923_R	GTCGCGAGCCCATCAAC	Reverse primer: Genotyping <i>nat</i> CDS (225 bp).
BK2928_F	TGGCCAAGTTGACCAGTGC	Forward primer: Genotyping <i>shBle</i> CDS (150 bp).
BK2928_R	TGATGAACAGGGTCACGTCG	Reverse primer: Genotyping <i>shBle</i> CDS (150 bp).
BK2935_F	CATTTGCTGGACACGGATGC	Forward primer: Genotyping <i>fcgD</i> gene (1928 bp).
BK2935_R	CTGCAAATTGTCCTCCTGGG	Reverse primer: Genotyping <i>fcgD</i> gene (1928 bp).
BK2936_F	ACCGACGAATATCTGACAGTCA	Forward primer: Genotyping <i>fcgC</i> gene (1183 bp).
BK2936_R	GAGGTCCTCAGCGTACGTTA	Reverse primer: Genotyping <i>fcgC</i> gene (1183 bp).

BK2937_F	GAACAATTACAACCCCGGCC	Forward primer: Genotyping <i>fcpB</i> gene (MPX + control; 865 bp).
BK2937_R	TCCTGGTTGAACACGTATCCT	Reverse primer: Genotyping <i>fcpB</i> gene (MPX + control; 865 bp).
BK2938_F	CACCCTGTGCCTTGCTCA	Forward primer: Genotyping <i>fcpA</i> gene (1453 bp).
BK2938_R	TATCTCCACCGCCGAAAC	Reverse primer: Genotyping <i>fcpA</i> gene (1453 bp).
BK2998_F	TCACACCGACATGCATCTCT	Forward primer: Genotyping upstream region of integration event in Z-Z ₁ line (Chr 5) to <i>shBle</i> (~1420 bp).
BK2998_R	TCCCGGAAGTTCGTGGAC	Reverse primer: Genotyping upstream region of integration event in Z-Z ₁ line (Chr 5) to <i>shBle</i> (~1420 bp).
BK3000_F	TCCTTACCACCGACACC	Forward primer: Genotyping downstream region of integration event in Z-Z ₁ line (Chr 3) to <i>nat</i> (~2338 bp).
BK3000_R	CAAGTTGGCTGTCATACGCA	Reverse primer: Genotyping downstream region of integration event in Z-Z ₁ line (Chr 3) to <i>nat</i> (~2338 bp).
BK3003_F	GTCTTTCACACCGACATGCA	Forward primer: Genotyping upstream region (Chr 5) to downstream region (Chr 3) of integration event in Z-Z ₁ line (~4157 bp).
BK3003_R	AGCTTGCAAAACACCATCTG	Reverse primer: Genotyping upstream region (Chr 5) to downstream region (Chr 3) of integration event in Z-Z ₁ line (~4157 bp).
BK3004_F	CGGACAGCAAGCCAGATTG	Forward primer: Genotyping upstream region of integration event in Z-Z ₂ line (Chr 9) to <i>shBle</i> (~1126 bp).
BK3004_R	TCCCGGAAGTTCGTGGAC	Reverse primer: Genotyping upstream region of integration event in Z-Z ₂ line (Chr 9) to <i>shBle</i> (~1126 bp).
BK3006_F	TGACCACTCTTGACGACACG	Forward primer: Genotyping downstream region of integration event in Z-Z ₂ line (Chr 9) to <i>nat</i> (~1580 bp).
BK3006_R	ACAAGAATGGCGCTTCGATC	Reverse primer: Genotyping downstream region of integration event in Z-Z ₂ line (Chr 9) to <i>nat</i> (~1580 bp).
BK3009_F	CGAATCAAAGTCACCCCGGA	Forward primer: Genotyping upstream region (Chr 9) to downstream region (Chr 9) of integration event in Z-Z ₂ line (~2905 bp).
BK3009_R	TTGTCCTTCCTGTTGCCAC	Reverse primer: Genotyping upstream region (Chr 9) to downstream region (Chr 9) of integration event in Z-Z ₂ line (~2905 bp).

Table B-2: Genetic parts used in this study.

Name	Sequence (5' to 3')	Description
<i>fcpD</i> promoter	ACTAGCTTGATTGGGATATCTCGCTCGTGCTTGTGCGGTGCTATGTCTTTAGGGTACTTGA ACCTACGTTTCGTAATTGTATAATATGATCATCGTCGTCATCGTATTATCGTTTTTCATCCGTC CAGCGCAAAATGCATTAGCAGCTAGTCTAGCGTGCAGGCTACCTGTACAGGTGCATGAC GGATGCGTGTCTTAAGTGAGTTTCTAATTAACAGTAACCTCTTTACTTATGTTTCAGTTTGT AAGAAGCGGGATTTCGCTCGTTCGGTTGACATCTGATTGGACTGCGTCGGCACGTGAAAACTA CATTTGTGAAATCTGCTAAAACTCCGGGTATCTCTGACACAAAACGATTTCGCGTCTCAATTTT AACATTACGGTCAAGGCTAACGTATCTTTCTCGGTCAACTTCAGATTACGCCGAGTAAATTG TCGTAGCTTTCAAGGCGTTTTGAGTACTGCGGCAGTTGTTGAACCTGCAAGGAGAAGATCT CGACAACAGAATACAGCGAAAAATGGGTCTCATGCACTAACACTCAGTCTCCCTCATAAT CTCTGTTAGAGTTTACCAACAACACATATATACATTTTCGACAAA	Regulatory element native to chromosome 2 of <i>P. tricornutum</i> .
<i>fcpD</i> terminator	TTTTGTTACATTTACTGACTTCAAGGAGTCGAGGAATCGATACTGCCGTCGTTCCAGGATC CGAGGTTTCATAAACTCTGTTAACGTTATAGAAAACAGACTTACCTCTCCTACGCCATTACAG TAATATTCGCAATATGCTATTCTTCTCTGAAGACCAGGTTTATGTGCTGCCTGAACTATT TCAATAAGTCAGCTGCACTTGCACAGGGTTTCAAGGAAAGCGTGTCTTTTTTCCAACGT AGGCGTCGCTTTCGTCGACTCTTACTCTTACATTCACAGCCAATACTTACAATTAGTAAAA AACCTGTGCTCGAGAGTGAAAACGTC	Regulatory element native to chromosome 2 of <i>P. tricornutum</i> .
<i>fcpA</i> terminator	CCGCAACAACACTACCTCGACTTTGGCTGGGACACTTTCACTGAGGACAAGAAGCTTCAGAAG CGTGCTATCGAACTCAACCAGGGACGTGCGGCACAAATGGGCATCCTTGCTCTCATGGTGC ACGAACAGTTGGGAGTCTCTATCCTTCTTAAAAATTTAATTTTCATTAGTTGCAGTCACTC CGCTTTGGTTT	Regulatory element native to chromosome 2 of <i>P. tricornutum</i> .
<i>nat</i> CDS	ATGACCACTCTTGACGACACGGCTTACCGGTACCGCACCAGTGTCCCGGGGGACGCCGAGG CCATCGAGGCACTGGATGGGTCTTACCACCGACACCGTCTTCCGCGTCACCGCCACCGG GGACGGCTTACCCCTGCGGGAGGTGCCGTTGACCCGCCCTGACCAAGGTGTTCCCGCAG ACGAATCGGACGACGAATCGGACGACGGGAGGACGGCGACCCGGACTCCCGGACGTTT GTCGCGTACGGGGACGACGGCGACCTGACGGGCTTCGTGGTCTGCTACTCCGGGTGGA ACCGCCGGCTGACCGTCGAGGACATCGAGGTGCGCCCGGAGCACCGGGGGACGGGGTTCG GGCGCGCGTTGATGGGGCTCGCGACGGAATTCGCCCCGCGAGCGGGGGCGCCGGGACCTCTG GCTGGAGGTCACCAACGTCAACGCACCGGCGATCCACGCGTACCGGCGGATGGGGTTCACC CTCTGCGGCTGGACACCGCCCTGTACGACGGCACCGCCTCGGACGGCGAGCAGGCGCTCT ACATGAGCATGCCCTGCCCTGA	Coding sequence conferring resistance to nourseothricin.
<i>cat</i> CDS	ATGGAGAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAAC ATTTTGAGGCATTTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTTCAGCTGGATATT ACGCCTTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTTATCCGCTTTTATTCACAT TCTTGCCCGCTGATGAATGCTCATCCGGAATTTTCGTATGGCAATGAAAGACGGTGAGCTG GTGATATGGGATAGTGTTCACCTTGTACACCGTTTTCCATGAGCAAACTGAAACGTTTTTC ATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAAGATG TGGCGTGTTACGGTGAAAACCTGGCTATTTCCCTAAAGGGTTATTGAGAATATGTTTTTTC GTCTCAGCCAATCCCTGGGTGAGTTTACCAAGTTTTGATTAAACGTGGCCAATATGGACAA CTTCTTCGCCCCGTTTTACCATGGGCAATATTATACGCAAGGCGACAAGGTGCTGATGC CGCTGGCGATTACAGTTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAAT GAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCG	Coding sequence conferring resistance to chloramphenicol.
<i>shBle</i> CDS	ATGGCCAAGTTGACCAAGTGCCGTTCCGGTGCTCACCGCGCGACGTCGCCGGAGCGGTGCG AGTTCTGGACCGACCGGCTCGGGTTCTCCCGGCACTTCGTGGAGGACGACTTCGCCGGTGT GGTCCGGGACGACGTGACCTGTTTCATCAGCGCGGTCCAGGACCAGGTGGTCCCGGACAAC ACCCTGGCCTGGGTGTGGGTGCGCGGCTGGACGAGCTGTACGCCGAGTGGTCCGAGGTCG TGTTCCACGAACCTCCGGGACGCTCCGGGCGGCCATGACCGAGATCGCGGAGCAGCCGTG GGGGCGGGAGTTTCGCCCTGCGGACCCGCGGCCAAGTGCCTGCACTTCGTGGCCGAGGAG CAGGAC	Coding sequence conferring resistance to zeocin.
T2A	TCCCGTGAAGCCCGTCACAAGCAGAAGATTGTGCCCCCGTCAAGCAGACCTTGAACCTTCG ACTTGCTGAAGTTGGCCGGAGACGTCAATCCAACCCCGGACCC	T2A self-cleaving peptide linker sequence. Promotes ribosome skipping in <i>P. tricornutum</i> .
<i>Cre</i> CDS	ATGCCCAAGAAGAAGAGGAAGGTGTCCAATTTACTGACCGTACACCAAAATTTGCCTGCAT TACCGGTCGATGCAACGAGTGATGAGGTTTCGCAAGAACCTGATGGACATGTTTCAGGGATCG CCAGGCGTTTTCTGAGCATACCTGGAAAAATGCTTCTGTCCGTTTGCCGGTCTGGGCGGCAT GGTGCAAGTTGAATAACCGGAAATGGTTTCCCGCAGAACCTGAAGATGTTTCGCGATTATCT TGTATATCTTCAGGCGCGCGGTCTGGCAGTAAAACTATCCAGCAACATTTGGGCCAGCTA AACATGCTTCATCGTCGGTCCGGGCTGCCACGACCAAGTGACAGCAATGCTGTTTCACTGG TTATGCGGGCGATCCGAAAAGAAAAACGTTGATGCCGGTGAACGTGCAAAACAGGCTCTAG CGTTTCGAACGCACTGATTTTCGACCAGGTTTCGTTCACTCATGGAAAAATAGCGATCGCTGCCA GGATATACGTAATCTGGCATTCTTGGGGATTGCTTATAACACCCTGTTACGTATAGCCGAAA TTGCCAGGATCAGGGTTAAAGATATCTCAGTACTGACGGTGGGAGAATGTTAATCCATAT TGGCAGAACGAAAACGCTGGTTAGCACCGCAGGTGTAGAGAAGGCACTTAGCCTGGGGGT	Sequence encoding <i>Cre</i> recombinase, codon optimized for <i>P. tricornutum</i> .

	AACTAAACTGGTCGAGCGATGGATTTCCGTCTCTGGTGTAGCTGATGATCCGAATAACTAC CTGTTTTGCCGGGTGAGAAAAAATGGTGTGGCCGCGCCATCTGCCACCAGCCAGCTATCAA CTCGCGCCCTGGAAGGGATTTTTGAAGCAACTCATCGATTGATTTACGGCGCTAAGGATGA CTCTGGTCAGAGATACCTGGCCTGGTCTGGACACAGTGCCCGTGTGCGAGCCGCGCGAGAT ATGGCCCGCGCTGGAGTTTCAATACCGGAGATCATGCAAGCTGGTGGCTGGACCAATGTAA ATATTGTCATGAACTATATCCGTAACCTGGATAGTGAAACAGGGGCAATGGTGCGCCTGCT GGAAGATGGCGATTAG	
<i>loxP</i> site	ATAACTTCGTATAATGTATGCTATACGAAGTTAT	Site of recombination for <i>Cre</i> recombinase
<i>XhoI</i> site	CTCGAG	Restriction site for the <i>XhoI</i> nuclease.
<i>I-SceI</i> site	TAGGGATAACAGGGTAAT	Restriction site for the <i>I-SceI</i> nuclease.
pAL backbone fragment	GAGCTGTAAGTACATCACCGACGAGCAAGGCAAGACGATCAATCAAAAAACCACCTTTCTT ATGAAACCTTGCTTTTCTTATTATAAATAAAGTGTAATTTAAAGTCAAACATAAAATGGTCT TTCTTTTATTTATTTTATTTGATTTTCCACAAATTTTAAATGAATGTTCCCCACAATTA TTGTCCACACATTGTGGATAAAGTTCCACATTTTATTCACAATGTTGATAAGTAGCGTAAG TATATTTAACAGCCTTACAAAGCAAATGATACACTGAAAAGTTATCCACAATTTAATATTTAA AGAACAGCTAAATCAAAAAGTTATCCACAATAATGTGGAAAACCTTTTATTAATTTGTG GTTTCTTATGCTATCATAGTTTTACATAAATTATTAACCTCAGGGAGGCAGTCATGAGTCCAA ACAGCACACTATGGCAGACAATATTACAGGATTTAGAAAACTATACAACGAGGAGACTT ACAACGAGCTATTTCTACCAGTGACTTCTACTTTTAAAGATCAAAACGGATTACTTACAATG GTTGTAGCAAATGAGTTCTTAAAGAATCGTATCAATAAACTATACATCGCAAAAATTAACG AATCTGCTACTAAATATTCAAGTACTCCAGTTAGATTGAAATTCATATCACAGAGGAAGT TATTGAAGAACCAGTAGCGGATCGTAAATTAACCATTGATTATCGTCAAGGTAACCTAAAC TCTACATATACCTTTGACTCTTTTGTGTTGGAATACTAACATGTTTGCTTTTCGTATGGCG ATGAAGGTTGCTGATCAACCTGCAGCAGTAGCAAAACCCCTTCTACATATTTGGTGATGTAG GTTTAGGTAACCACTCTTATGCAAGCAATAGGTAACCTATATATTAGATAATGATGTTGA AAAACGTATATTATATGTTAAAGCTGATAATTTTATTGAAGACTTTGTATCATTATTATCAA GAAACAAAAATAAGACTGAAGAATTCAATGCTAAATATAAAGATATAGATGTTATATTAGT GGATGATATTCAAATTATGGCCAACGCTAGTAAACTCAAATGGAATTTCTTAAACTCTTTG ACTATCTATATTTAAATAATAAGCAAATTGTCATTACGTCTGATAAACCAGCTTCACAATTA ACAAATATTATGCCGCGATTAAACACACGTTTTGAAAGCTGGTCTCTCTGTAGACATACAAAT ACCTGAATTAGAACATAGAATAAGCATTTTAAAGAGAAAAACAGCTACATTAGATGCAAA CTTAGAGGTAGGAGAGGATATCTTAACCTTTATTGCATCTCAATTTGCAGCAATATTAGAG AAATGGAGGGTGCACTCATTCTGTTAATTAGTTATGCACAAACCTTTAATCTAGAAATAAC AATGGATGTTGTTGAAGAAGCACTTGGTGCTGTCTTAAAAACAAAGAAGAAAAACAAATCA ATTAACGAAAAATAACTACGATAAGATCCAAAGTATCGTTGCAGATTACTTCCAAGTGTC TTACCAGACTTAATTGGTAAAAAAGACATGCTAAATTCACATTACCTAGACATATCGCAA TGTATCTTATCAAACCTCAAATTCAATATACCTTATAAAACGATTGGTTCTTTGTTAATGAT AGAGACCACTCTACTGTATTGGCTGCTTGTGAGAAAGTAGAACGCGATATGAGGATGGATT CGAACTTAAAGTTTGCTGTTGACTCAATTGTCAAAAAAATAGATTACCATCATTAAAGTG ATAAATGTTTATAAAAAATGATTAATGTGGTAAAAATAAATGGTAGATGAAGCGATTATTGC TAGTTTCCCACTTTCCACAGACACTAACATAACAAAGAAGAAATAATAATTAATAAAG GGTAAAAATCCCTTTAGTGAGGGTTAATGTCGTGACTGGGAAAACCTT	Region of pAL backbone used as negative control (DNA encoding nothing) in transformation experiments.

Appendix C: Electroporation Data

Table C-1: Data for all electroporations performed in this study.

Asterisks (*) denote a set of transformations where samples that did not grow on initial selection plates but were confirmed to grow in selective media at a later stage, likely owing to a bad batch of media.

Transformation	Type growing	Transformed with	Blunt or digested	conc DNA (ng/uL)	volume DNA added (uL)	carrier DNA	Selectable media	O D730	Initial cell concentration before transformation (cells / mL)	Final cell concentration before transformation (cells/uL)	Voltage (V)	Capacitance (uF)	Resistance (Ohms)	TC (ms)	CFUs (10 ⁴ cells/plate)
1	WT	pPtGE27	digested (<i>Xho</i> I)	51	6	calflymus (40 µg)	½ L1 + NTC 100	n/a	n/a	2x10 ⁶	500	25	400	0.6	0
2	WT	pPtGE27	digested (<i>Xho</i> I)	51	6	calflymus (40 µg)	½ L1 + NTC 100	n/a	n/a	2x10 ⁶	500	25	400	0.5	0
3	WT	n/a (- control)	-	-	-	calflymus (40 µg)	½ L1 + NTC 100	n/a	n/a	2x10 ⁶	500	25	400	0.5	0
4	ΔHIS	HIS 2000 bp	undigested	73	6	calflymus (40 µg)	½ L1	n/a	n/a	2x10 ⁶	500	25	400	0.6	0
5	ΔHIS	HIS 2000 bp	undigested	73	6	calflymus (40 µg)	½ L1	n/a	n/a	2x10 ⁶	500	25	400	0.5	0
6	ΔHIS	n/a (- control)	-	-	-	calflymus (40 µg)	½ L1 + NTC 100	n/a	n/a	2x10 ⁶	500	25	400	0.5	0
7	WT	pPtGE27	digested (<i>Xho</i> I)	54	6	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
8	WT	n/a (- control)	-	-	-	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
9	WT	pPtGE27	digested (<i>Xho</i> I)	65	6	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
10	WT	n/a (- control)	-	-	-	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.6	0
11	WT	pPtGE27	digested (<i>Xho</i> I)	51	6	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
12	WT	n/a (- control)	-	-	-	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
25	WT	pPtGE31	digested (<i>Bmt</i> I-HF)	78	6	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
12	WT	n/a (- control)	-	-	-	ssss DNA (40 µg)	½ L1 + NTC 100	0.23	n/a	2x10 ⁶	500	25	400	0.5	0
13	WT	XhoI-site-fcpD.Promoter-nat-	undigested	101	6	ssss DNA	½ L1 +	0.25	n/a	2x10 ⁶	500	25	400	0.6	0

		fcpA.Terminator-XhoI site				(40 µg)	NTC 100								
14	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 100	0.25	n/a	2x10 ⁶	500	25	400	0.6	0
15	ΔHIS	HIS 2000 bp	undigested	73	6	ssss DN A (40 µg)	½ L1	0.28	n/a	2x10 ⁶	500	25	400	0.6	0
16	ΔHIS	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1	0.28	n/a	2x10 ⁶	500	25	400	0.6	0
17	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	132	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.3	n/a	2x10 ⁶	500	25	400	0.6	0
18	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 100	0.3	n/a	2x10 ⁶	500	25	400	0.6	0
19	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	25	400	0.6	0
20	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	400	28.5	2
21	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	150	12.2	0
22	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	25	400	0.6	0
23	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	400	29.9	2
24	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	150	12.2	0
25	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	25	400	0.6	0
26	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	400	25.3	6
27	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	150	12.2	0
28	WT	pPtGE27	digested (XhoI)	125	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	25	400	0.7	0
29	WT	pPtGE27	digested (XhoI)	125	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	400	26.7	2
30	WT	pPtGE27	digested (XhoI)	125	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	50	150	11.9	7
31	WT	pPtGE31	digested (BmtI-HF)	110	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.26	n/a	2x10 ⁶	500	25	400	0.6	0

32	WT	pPGE31	digested (<i>Bmt</i> I-HF)	110	6	ssss DN A (40 µg)	½ L1 + NTC 100	0.2 6	n/a	2x10 ⁶	500	50	400	26. 3	0
33	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 100	0.2 6	n/a	2x10 ⁶	500	50	400	30. 5	0
34	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 6	n/a	2x10 ⁶	501	25	400	0.5	0
35	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	77	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 6	n/a	2x10 ⁶	500	50	400	0.7	0
36	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 6	n/a	2x10 ⁶	501	25	400	0.3	0
37	ΔHIS	HIS 2000 bp	undigested	110	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	501	25	400	0.5	0
38	ΔHIS	HIS 2000 bp	undigested	110	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	500	50	400	1	0
39	ΔHIS	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	501	25	400	0.7	0
40	ΔURA	URA 2000 bp	undigested	110	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	501	25	400	0.4	0
41	ΔURA	URA 2000 bp	undigested	110	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	500	50	400	0.9	0
42	ΔURA	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 9	n/a	2x10 ⁶	501	25	400	0.4	0
43	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	129	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	486	100	100	8.3 (10 ms pro to col)	0
44	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	129	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	490	100	200	16. 1 (20 ms pro to col)	0
45	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	129	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	495	150	200	29. 4 (30 ms pro to col)	0
46	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	486	100	100	8.3 (10 ms pro to col)	0
47	ΔURA	URA 2000 bp	undigested	177	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	486	100	100	8.3 (10 ms pro to col)	0
48	ΔURA	URA 2000 bp	undigested	177	6	ssss DN A (40 µg)	½ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	492	125	200	20. 2 (20 ms pro	0

														toc ol)	
49	Δ URA	URA 2000 bp	undigested	177	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	492	150	200	29. 5(3 0ms pro toc ol)	0
50	Δ URA	n/a (- control)	-	-	-	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 150	0.2 1	n/a	2x10 ⁶	487	100	100	8.3 (10 ms pro toc ol)	0
51	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	129	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 150	0.3	n/a	2x10 ⁶	500	50	150	0.6	0
52	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	129	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 150	0.3	n/a	2x10 ⁶	500	50	150	0.5	0
53	WT	n/a (- control)	-	-	-	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 150	0.3	n/a	2x10 ⁶	500	50	150	0.8	0
54	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	400	20. 5	1000 +
55	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	400	22. 9	1000 +
56	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	150	10. 9	657
57	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	150	11. 4	84
58	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	486	75	150	11. 4	84
59	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	220	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	486	75	150	11. 4	122
60	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	220	15	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	150	10. 6	389
61	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	undigested	220	15	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	150	10. 9	214
62	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	220	15	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	487	75	150	7.1	67
63	WT	n/a (- control)	-	-	-	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.2 9	n/a	2x10 ⁶	500	50	150	11. 9	3
64	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	140	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.0 8	n/a	2x10 ⁶	500	50	400	25. 4	0
65	WT	XhoI-site- fcpD.Promote r-nat- fcpA.Termina tor-XhoI-site	digested (Xho I)	140	6	ssss DN A (40 μ g)	$\frac{1}{2}$ L1 + NTC 200	0.0 8	n/a	2x10 ⁶	500	50	400	25. 4	0
66	WT	XhoI-site- fcpD.Promote	digested	140	6	ssss DN	$\frac{1}{2}$ L1 +	0.0 8	n/a	2x10 ⁶	500	50	150	12	0

		r-nat-fcpA.Terminator-XhoI site	(Xho I)			A (40 µg)	NTC 200								
67	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	digested (Xho I)	140	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.08	n/a	2x10 ⁶	500	50	150	12	0
68	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 200	0.08	n/a	2x10 ⁶	500	50	150	12.1	0
69	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	140	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	16.1	lawn
70	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	235	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	16.7	55
71	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	194	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	16.8	3
72	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	15	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	15.9	lawn
73	WT	XhoI site-fcpD.Promoter-nat-fcpA.Terminator-XhoI site	digested (Xho I)	26	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	14.7	lawn
74	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	15.8	lawn
75	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	362	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	14.1	1
76	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	362	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	17.7	49
77	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	362	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	18	37
78	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	72	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	18.1	11
79	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	72	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	18.4	10
80	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	72	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	15.3	5
81	WT	2 kbp pAL backbone (- control)	undigested	15	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.19	n/a	2x10 ⁶	500	50	400	18	0
82	WT	XhoI site-fcpD.Promoter-shBle-T2A-nat-	undigested	152	6	ssss DN A	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	17.5	0

		fcpD.Terminator-XhoI site				(40 µg)									
83	WT	XhoI site-fcpD.Promoter-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	362	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	20.4	0
84	WT	XhoI site-Truncated.fcpD.Promoter-nat-fcpA.Terminator-XhoI site	undigested	139	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	19.6	0
85	WT	2 kbp pAL backbone (- control)	undigested	232	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	19.4	0
86	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	digested (Xho I)	52	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	17.6	0
87	WT	XhoI site-fcpD.Promoter-r-cat-T2A-nat-fcpA.Terminator-XhoI site	digested (Xho I)	45	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	20.2	0
88	WT	XhoI site-Truncated.fcpD.Promoter-nat-fcpA.Terminator-XhoI site	digested (Xho I)	30	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	19.8	0
89	WT	-	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 200	0.27	n/a	2x10 ⁶	500	50	400	20.6	0
90	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	151	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	23.8	25
91	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	151	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	21.1	28
92	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	151	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	22.2	18
93	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	digested (Xho I)	40	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	25.2	5
94	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-fcpD.Terminator-XhoI site	digested (Xho I)	40	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	22.3	6
95	WT	2 kbp pAL backbone (- control)	undigested	73	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.28	n/a	2x10 ⁶	500	50	400	23.6	0
96	WT	XhoI site-fcpD.Promoter-r-nat-fcpA.Terminator-XhoI site	undigested	437	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.22	n/a	1.35x10 ⁶	500	50	400	15.2	2
97	WT	XhoI site-fcpD.Promoter-r-cat-T2A-nat-fcpA.Terminator-XhoI site	undigested	98	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.22	n/a	1.35x10 ⁶	500	50	400	18.4	38
98	WT	XhoI site-fcpD.Promoter-r-shBle-T2A-nat-	undigested	213	6	ssss DN A	½ L1 + NTC 200	0.22	n/a	1.35x10 ⁶	500	50	400	14.5	120

		fcpD.Terminator-XhoISite				(40 µg)									
99	WT	2 kbp pAL backbone (- control)	undigest	230	6	ssss DN A (40 µg)	½ L1 + NTC 200	0.22	n/a	1.35x10 ⁶	500	50	400	15.6	0
100	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	453	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 100	0.35	n/a	2x10 ⁶	500	50	400	21.9	790 (NTC), 47 (ZEO)
101	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	453	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 100	0.35	n/a	2x10 ⁶	500	50	400	23.1	333 (NTC), 20 (ZEO)
102	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	453	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 100	0.35	n/a	2x10 ⁶	500	50	400	22	306 (NTC), 18 (ZEO)
103	WT	2 kbp pAL backbone (- control)	undigest	453	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 100	0.35	n/a	2x10 ⁶	500	50	400	23.9	0 (NTC), 0 (ZEO)
104	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	400	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.29	n/a	2x10 ⁶	500	50	400	24.1	1322 (NTC), 195 (ZEO)
105	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	400	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.29	n/a	2x10 ⁶	500	50	400	27.1	782 (NTC), 107 (ZEO)
106	WT	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	400	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.29	n/a	2x10 ⁶	500	50	400	n/a	0 (NTC), 0 (ZEO)
107	WT	2 kbp pAL backbone (- control)	undigest	230	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.29	n/a	2x10 ⁶	500	50	400	25.6	0 (NTC), 0 (ZEO)
108	WT (thawed from -80)	XhoISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	400	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.35	n/a	2x10 ⁶	500	50	400	8.8	0 (NTC), 0 (ZEO)
109	WT	XhoISite-ScISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	289	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 50	n/a	4.88x10 ⁶	2x10 ⁶	500	50	400	21.2	547 (NTC), 82 (ZEO)
110	WT	XhoISite-ScISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigest	289	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1	n/a	4.88x10 ⁶	2x10 ⁶	500	50	400	22.3	614 (NTC), 93 (ZEO)

		fcpD.Terminator-XhoISite					+ ZEO 50								
111	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	289	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 50	n/a	4.88x10 ⁶	2x10 ⁶	500	50	400	23.7	479 (NTC), 53 (ZEO)
112	WT	2 kbp pAL backbone (- control)	undigeste d	289	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 50	n/a	4.88x10 ⁶	2x10 ⁶	500	50	400	22.9	0 (NTC), 0 (ZEO)
113	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.2x10 ⁶	2x10 ⁶	500	50	400	25	218 (NTC), 73 (ZEO)
114	WT	2 kbp pAL backbone (- control)	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.2x10 ⁶	2x10 ⁶	500	50	400	23.2	0 (NTC), 0 (ZEO)
115	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4x10 ⁶	2x10 ⁶	500	50	400	23.3	95 (NTC), 12 (ZEO)
116	WT	2 kbp pAL backbone (- control)	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4x10 ⁶	2x10 ⁶	500	50	400	24	0 (NTC), 0 (ZEO)
117	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.17x10 ⁶	2x10 ⁶	500	50	400	24	56 (NTC), 14 (ZEO)
118	WT	2 kbp pAL backbone (- control)	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.17x10 ⁶	2x10 ⁶	500	50	400	23.5	0 (NTC), 0 (ZEO)
119	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5x10 ⁶	2x10 ⁶	500	50	400	23.2	53 (NTC), 7 (ZEO)
120	WT	2 kbp pAL backbone (- control)	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5x10 ⁶	2x10 ⁶	500	50	400	24.6	0 (NTC), 0 (ZEO)
121	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.39x10 ⁶	2x10 ⁶	500	50	400	25.8	52 (NTC), 16 (ZEO)

122	WT	2 kbp pAL backbone (- control)	undigested	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.39X10 ⁶	2x10 ⁶	500	50	400	26.1	0 (NTC), 0 (ZEO)
123	WT	XhoI site-SceI site-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.21X10 ⁶	2x10 ⁶	500	50	400	25.6	141 (NTC), 13 (ZEO)
124	WT	2 kbp pAL backbone (- control)	undigested	160	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	5.21X10 ⁶	2x10 ⁶	500	50	400	26.4	0 (NTC), 0 (ZEO)
125	WT (thawed from -80° C)	XhoI site-SceI site-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	230	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	0.35	n/a	2x10 ⁶	500	50	400	13.2	0 (NTC), 0 (ZEO)
126	WT	pDMI2 with HASP1 promoter, signal peptide, gene (w/ introns) and terminator	circular	147	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	16.5	9
127	WT	pDMI2 with HASP1 promoter, signal peptide, gene (w/ introns) and terminator	circular	152	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	17.6	1
128	WT	pDMI2 with HASP1 promoter, signal peptide, gene (cDNA) and terminator	circular	168	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	17.7	12
129	WT	pDMI2 with HASP1 promoter, signal peptide, gene (cDNA) and terminator	circular	201	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	15.8	3
130	WT	XhoI site-SceI site-loxP-fcpD.Promoter-r-nat-fcpA.Terminator-loxP-XhoI site	undigested	180	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	20.1	214
131	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	20.2	5
132	WT	XhoI site-loxP-fcpD.Promoter-shBle-fcpA.Terminator-loxP-SceI site-XhoI site	undigested	180	6	ssss DN A (40 µg)	½ L1 + ZEO 25	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	21.8	16
133	WT	n/a (- control)	-	-	-	ssss DN A (40 µg)	½ L1 + ZEO 25	n/a	2.43x10 ⁷	2x10 ⁶	500	50	400	20.2	0
134	WT	XhoI site-SceI site-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 +	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.7	0*

							ZEO 25								
135	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21.7	0*
136	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21.5	0*
137	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23	0*
138	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21	0*
139	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.3	0*
140	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.5	0*
141	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21.5	0*
142	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22	0*
143	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.9	0*
144	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.9	0*
145	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21.7	0*

146	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24	0*
147	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21. 7	0*
148	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24	0*
149	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21. 1	0*
150	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24	0*
151	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 2	0*
152	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 3	0*
153	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 7	0*
154	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23. 4	0*
155	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 5	0*
156	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23	0*
157	WT	XhoI site- SceI site- fcpD.Promote r-shBle-T2A-	undigested	112	6	ssss DN A	½ L1 + NTC 200,	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 2	0*

		nat-fcpD.Terminator-XhoISite				(40 µg)	½ L1 + ZEO 25								
158	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.9	0*
159	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.9	0*
160	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.3	0*
161	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.3	0*
162	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	20.9	0*
163	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.6	0*
164	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22.4	0*
165	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23	0*
166	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.5	0*
167	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.8	0*
168	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.8	0*

169	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 8	0*
170	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24	0*
171	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 2	0*
172	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 1	0*
173	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	19. 3	0*
174	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 4	0*
175	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 8	0*
176	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	20. 4	0*
177	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21. 3	0*
178	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 5	0*
179	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23. 8	0*
180	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A-	undi geste d	112	6	ssss DN A	½ L1 + NTC 200,	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21. 5	0*

		nat-fcpD.Terminator-XhoISite				(40 µg)	½ L1 + ZEO 25								
181	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23	0*
182	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.8	0*
183	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	25	0*
184	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.3	0*
185	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.4	0*
186	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.4	0*
187	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.9	0*
188	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.6	0*
189	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.2	0*
190	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.2	0*
191	WT	XhoISite-SceISite-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoISite	undigeste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	26.5	0*

192	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	26. 3	0*
193	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 2	0*
194	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	26. 2	0*
195	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	26. 6	0*
196	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	26. 2	0*
197	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	21. 7	0*
198	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24. 5	0*
199	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23. 6	0*
200	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23. 3	0*
201	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	25	0*
202	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A- nat- fcpD.Termina tor-XhoISite	undi geste d	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24	0*
203	WT	XhoISite- SceISite- fcpD.Promote r-shBle-T2A-	undi geste d	112	6	ssss DN A	½ L1 + NTC 200,	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	22. 6	0*

		nat-fcpD.Terminator-XhoI site				(40 µg)	½ L1 + ZEO 25								
204	WT	XhoI site-SceI site-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	24.9	0*
205	WT	XhoI site-SceI site-fcpD.Promoter-shBle-T2A-nat-fcpD.Terminator-XhoI site	undigested	112	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	25.9	0*
206	WT	n/a (- control)	-	-	6	ssss DN A (40 µg)	½ L1 + NTC 200, ½ L1 + ZEO 25	n/a	4.26x10 ⁶	2x10 ⁶	500	50	400	23.2	0
207	WT	PtMt 14.1.1	circular	141.6	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.13x10 ⁷	2x10 ⁶	500	50	400	20.4	0
208	WT	ptMt MinV3 C3.1	circular	448	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.13x10 ⁷	2x10 ⁶	500	50	400	22.5	0
209	WT	pDMI2	circular	676	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.13x10 ⁷	2x10 ⁶	500	50	400	22.1	0
210	WT	n/a (- control)	-	-	6	ssss DN A (40 µg)	½ L1 + NTC 100	n/a	2.13x10 ⁷	2x10 ⁶	500	50	400	23.4	0

Curriculum Vitae

Name: Mark Pampuch

Post-secondary Education and Degrees: The University of Western Ontario, London, ON, Canada
2021–2023 MSc.
Department of Biochemistry
Supervisor: Dr. Bogumil Karas

The University of Western Ontario, London, ON, Canada
2016–2020 BSc.
Honours Specialization in Genetics

Courses Taken: BIOCHEM 9545Q: Bioinformatics I
BIOCHEM 9546R: Bioinformatics II
BIOCHEM 9533: Ideas to innovation

Honours and Awards: Chair's Travel Award
\$500 – 2022

NSERC Canada Graduate Scholarships – Master's Program
\$17,500 – 2022

Ontario Graduate Scholarship
\$15,000 – 2022 (Declined due to NSERC conflict)

Ontario Graduate Scholarship
\$15,000 – 2021

Dean's Honor List
2017 – 2020

Western Scholarship of Distinction (Entrance Scholarship)
\$1000 - 2016

Publications: Walker, E.J.L., **Pampuch, M.**, Chang, N., Cochrane, R.R., and Karas, B.J. (2023) Design and assembly of the 117-kb *Phaeodactylum tricornutum* chloroplast genome. Plant Physiology. Submitted, manuscript number: PP2023-RR-00842

Pampuch, M., Walker, E.J.L., and Karas, B.J. (2021) Towards synthetic diatoms: The *Phaeodactylum tricornutum* Pt-syn 1.0 project. Current Opinion in Green and Sustainable Chemistry. <https://doi.org/10.1016/j.cogsc.2022.100611>

In Preparation

Pampuch, M., remaining authors TBD, and Karas, B.J. Advances in *Phaeodactylum tricornutum* nuclear genome engineering (temporary title).

Presentations:

(Talk) International Conference on Algal Biomass, Biofuels, and Bioproducts (AlgalBBB)

Title: Towards Synthetic Diatoms: Advances in *Phaeodactylum tricornutum* nuclear genome engineering

Waikōloa Beach, Hawaii, USA

June 2023

(Talk) SynDiatoms 2023 Workshop

Title: Towards Synthetic Diatoms: Advances in *Phaeodactylum tricornutum* nuclear genome engineering

Virtual Meeting

May 2023

(Talk) Graduate Research Spring Symposium 2023

Title: Towards Synthetic Diatoms: Advances in *Phaeodactylum tricornutum* nuclear genome engineering

University of Western Ontario, London, ON, Canada.

May 2023

(Talk) SynDiatoms 2022 Workshop

Title: Optimizing transformation methods for *Phaeodactylum tricornutum*

Virtual Meeting

December 2022

(Talk and Poster) Canada SynBio 2022

Title: Towards Synthetic Diatoms: The Pt-syn1.0 project

Toronto, ON, Canada

May 2022

(Poster) Graduate Research Winter Symposium 2022

Title: Towards Synthetic Diatoms

University of Western Ontario, London, ON, Canada.

January 2022

**Workshops
Attended:**

An introduction to machine learning for Oxford Nanopore data
Nanopore Community Meeting – Manhattan, NY, USA
December 2022

**Mentorship
Experience:**

Ayagiysan Kaneshan – Work study student
September 2022 – April 2023

Garvin Tran – 4th year Biochemistry thesis student
May 2022 – April 2023

Megan Demers – Volunteer
May 2022 – August 2022

Danish Zahid – Volunteer
January 2022 – August 2022

**Leadership and
Extracurricular
Activities:**

Seed your Startup 2023 Semi-Finalist
Morrisette Institute for Entrepreneurship, London, ON, Canada
March 2023

2023 Ivey Business Plan Competition Semi-Finalist
Ivey Business School, London, ON, Canada
November 2022 – January 2023

Skiing and Snowboarding Instructor
London Track 3 Adaptive Snow School, London, ON, Canada
December 2021 – March 2022, December 2022 – March 2023

Science Rendezvous Booth Organizer
London, ON, Canada
May 2022, May 2023

Biochemistry Graduate Student Association Member
Western University, London, ON, Canada
September 2021 - Present

cGEM Competition Judge
Online
October 2021

**Relevant Work
Experience:**

Summer Research Assistant
Karas Lab, Western University, London, ON, Canada
April 2021 – August 2021

Undergraduate Researcher
Agriculture and Agri-Food Canada, London, ON, Canada
June 2019 – August 2020

Internship
MolecuLight Inc., Toronto, ON, Canada
May 2018 – August 2018