New tools, targets and approaches for gene, genome and metabolic engineering

Blake Tyler Hovde

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Raymond J. Monnat Jr, Chair

Rose Ann Cattolico

Andrew Scharenberg

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

New tools, targets and approaches for gene, genome and metabolic engineering

Blake Tyler Hovde

Chair of the Supervisory Committee:
Dr. Raymond J Monnat Jr. MD
Genome Sciences/Pathology

Genome engineering tools and DNA sequencing technologies have improved dramatically over the past decade. These technologies are transforming many areas of research ranging from evolutionary biology to human disease. My thesis research has been focused on developing and applying DNA sequence-guided approaches to genome engineering. The specific projects described in this dissertation and the associated manuscripts from this work have focused on characterizing the sequence of a new algal genome from *Chrysochromulina tobin,* an organism of considerable biotechnology interest for metabolic engineering and evolutionary studies. I describe the sequencing, assembly and annotation of the *Chrysochromulina* organellar and nuclear genomes, and transcript profiling over the light-dark cycle that entrains metabolism, lipid synthesis and degradation and cell division. Novel findings included the unique repeat structure of the organellar genomes and discovery of novel biochemical pathways never before seen in algae.

I also worked on genome engineering application to human disease. This work included the engineering of DNA target site-specific nucleases to serve as a gene drive system designed to control fertility and thus the population size of *Anopheles gambiae*, the vector for transmission of malaria. This work was part of the first demonstration of homing as a gene drive system in metazoans. In addition, I developed the molecular tools and protocols to diagnose disease-causing mutations in Shwachman-

Diamond syndrome (SDS) and to target the correction of these mutations in iPS cells derived from SDS patients. An alternative approach makes use of a complementing SDS gene into a human genome safe harbor site.

Acknowledgements

I would like to thank Ray Monnat for his guidance and support, and allowing me to pursue many projects of interest during my time in the lab. I also want to thanks Rose Ann Cattolico for the countless hours of discussion and help in all things algae.

Many thanks go out to all members of both the Monnat and Cattolico labs for the great friendships and scientific support they provided during the past few years.

I also want to thank my family and friends locally and abroad for fun and support.

Many thanks to those who helped get me where I am today by sparking interest and mentoring me in scientific research, particularly Amy Lawson, Matthew Smith and Zane Kraft.

Table of Contents

List of Figures 3
List of Tables 4
Introduction 5
Chapter 1: Genome engineering tools 6
Nuclease double strand break repair enables genome engineering 6
Homing endonucleases 7
TALENs 8
CRISPR/Cas9 9
ZFNs and others 10
Chapter 2: A novel algal species for metabolic engineering 12
Introduction 12
Requirement of sequencing 14
Results 15
Genome Sequencing 15
Organellar genomes 15
Mitochondrial genome 15
Mitochondrial repeats 18
Chloroplast genome 19
Ribosomal spacers 24
Repeat analysis 25
Genes of interest 29
Nuclear genome 35
Transcriptomics 36
Lipid biogenesis 41
Meiosis and DNA repair 43
Polyketide synthesis 49
Transformation of *C. tobin* 48
Conclusions 48
Materials and Methods 49
Chapter 3: Genetic engineering of mosquitoes for malaria prevention 57
Background 57
Homing endonucleases as ideal tools 57
Engineering challenges 58
Single base pair variation 59
Pairwise combinatorial engineering 61
Central 4 information 65
Generation of *Anopheles gambiae* fertility gene-specific variants 65
Conclusions 73
Materials and Methods 73
Chapter 4: New human chromosomal safe harbor sites for targeted genome engineering 76
Background 76
Target finding strategy 77
Targeting tool methods 80
Validation of target integrations 84
Discussion 92
Chapter 5: Targeted correction of human disease-causing mutations 96
Background 96
Target analysis 97
Repair template design 101
Prospects and next steps 102
Chapter 6: Conclusions 104

References 106
Appendix 1: Supplementary tables 116
Appendix 2: Transformation of *C. tobin* 119
Appendix 3: Targeted epigenetic modification tools 128

List of Figures:

Figure 1.1: DNA repair outcomes 7
Figure 1.2: Homing endonuclease transmission 8

Figure 2.1: *Chrysochromulina tobin* 24 hour photocycle 13
Figure 2.2: *Chrysochromulina tobin* mitochondrial genome map 17
Figure 2.3: *Chrysochromulina tobin* chloroplast genome map 22
Figure 2.4: Ribosomal operon repeat structure of haptophytes 24
Figure 2.5: clpC associated inverted repeats 27
Figure 2.6: Analysis of small repeats 29
Figure 2.7: Phylogenetic analysis of RPL36 proteins 31
Figure 2.8: Cell division and lipid accumulation 37
Figure 2.9: Transcript heatmap 38
Figure 2.10: Fatty acid synthesis genes and transcripts 42
Figure 2.11: Fatty acid desaturases 43
Figure 2.12: Polyketide synthesis gene 47

Figure 3.1: I-CreI crystal structure and DNA recognition sequence 59
Figure 3.2: I-CreI single base pair degeneracy 60
Figure 3.3: Multiple target site substitution analysis 62
Figure 3.4: Pairwise combinations heatmap 64
Figure 3.5: *Anopheles* gene target selection criteria 66
Figure 3.6: Homing endonuclease design and testing workflow 67
Figure 3.7: Competitive cleavage assay 69
Figure 3.8: In-vitro combined degeneracy test 70
Figure 3.9: In-vitro full target site testing with engineered mCre 71
Figure 3.10: In-vivo pDR-GFP recombination assay 72

Figure 4.1: I-CreI sites in the human genome 78
Figure 4.2: Safe harbor site targeting and modification 85
Figure 4.3: Safe harbor site insertions 91
Figure 4.4: Workflow of human safe harbor site identification 92

Figure 5.1: *SBDS* gene and pseudogene map 97
Figure 5.2: *SBDS* gene targeting with mCre 97
Figure 5.3: In-vitro combined degeneracy test of *SBDS* target 1 99
Figure 5.4: In-vitro partial design test of *SBDS* target 1 100
Figure 5.5: Restriction enzyme/PCR diagnostic of *SBDS* mutations 101
Figure 5.6: TALEN platform targeting of SBDS 101

List of Tables:

Table 1.1: Genome engineering tools comparison 11

Table 2.1: Haptophyte mitochondrial genome comparison.....16
Table 2.2: Haptophyte chloroplast genome comparison 21
Table 2.3: *C. tobin* nuclear genome statistics 35
Table 2.4: *C. tobin* nuclear gene calling and annotation 36
Table 2.5: Meiosis and DNA repair genes.....45

Table 3.1: *Anopheles* gene targets 67
Table 3.2: Engineering solutions for two *Anopheles* gene targets 70
Table 3.3: pDR-GFP recombination quantification 72
Table 3.4: Two *Anopheles* gene target summary.....73

Table 4.1: Human genome safe harbor site criteria 80
Table 4.2: Human I-CreI safe harbor scores 83

Introduction

Genome engineering and genomics, the focus for my thesis research, are two rapidly evolving fields at the intersection of technology and biology. Genome engineering targets specific genes or genomic locations for insertion, deletion or modification. Genomics enabled by the continued rapid drop in cost of DNA sequencing has provided significant new data on the genetics, genome structure and metabolism of both model and non-model organisms. My thesis research provided the opportunity to gain expertise in both genome engineering and genomics as applied to specific biological problems. My interest in genome engineering stemmed from curiosity about the ways in which homing endonucleases could be adapted to enable genome engineering in a wide range of organisms. Over the course of my thesis research I developed approaches to engineer gene- and target site-specific variants of homing endonucleases and other genome engineering nucleases, and applied these to the development of a novel gene drive system aimed at combatting malaria transmission. I also developed the genome engineering tools for the targeted correction or functional complementation of disease-causing mutations in a human inherited bone marrow failure syndrome, Shwachman-Diamond syndrome. My genomics efforts have focused on determining the genome sequence of a novel algal species with high potential for biotechnology applications. Each of these projects is discussed in the chapters that follow, together with a description of manuscripts that resulted from each project and the most important work that remains to be done in each project.

Chapter 1: Genome engineering tools

When DNA single or double strand breaks occur, the cell recruits DNA repair machinery to prevent accumulation of mutations. DNA damage, often incurred by UV light, other radiation sources or mutagenic chemicals, is common and repair mechanisms are required to maintain the viability of genome is order to prevent the accumulation of harmful mutations. In general, non-homologous end joining and homology driven repair are both accurate repair mechanism for correct DNA repair and relegation. These mechanisms can be taken advantage of however, for genome editing purposes (Figure 1.1). By transforming a site specific genome engineering tool into a cell for example, one can target a genomic locus of interest. DNA cleavage would normally be repaired as discussed above, but because a locus specific enzyme is used, DNA sequence repaired correctly will be subsequently cleaved again while the nuclease is present. Therefore, there are two potential outcomes for this cycle of DNA cleavage and repair.

One outcome, via the non-homologous end joining (NHEJ) pathway, can mutagenize the target being cleaved due to NHEJ being an often imperfect repair system. DNA cleavage will continue to take place until a mutation occurs via a repair mistake caused by NHEJ, and cleavage sensitivity of the nuclease is subsequently ablated by the mutant sequence. Because NHEJ is mutagenic, NHEJ can be used for gene perturbation. NHEJ process can be accelerated by including an exonuclease along with the nuclease. By attaching an exonuclease such as Trex2, which removes DNA overhangs from the DNA cleavage event, mutagenesis is greatly enhanced (Certo et al., 2012).

The second outcome takes advantage of the cell's ability to perform homology driven repair (HDR). By providing exogenous DNA with partial homology to a cleaved genomic locus of interest, HDR can drive gene insertion or repair. The added DNA template can contain any DNA sequence of choice, as long as the sequence is flanked by regions of exact homology to both sides of the nuclease's genomic target. These two options provide powerful genetic modification methods that can be used in a variety of research fields, from gene modification and knockout, to the repair of disease causing DNA mutations as a therapy. Genome engineering tools are now becoming more mainstream as the success of newest genome editing tools provides a relatively low barrier to entry. Many of these tools are

discussed briefly below, though the much of the work presented in this thesis utilizes homing endonucleases.

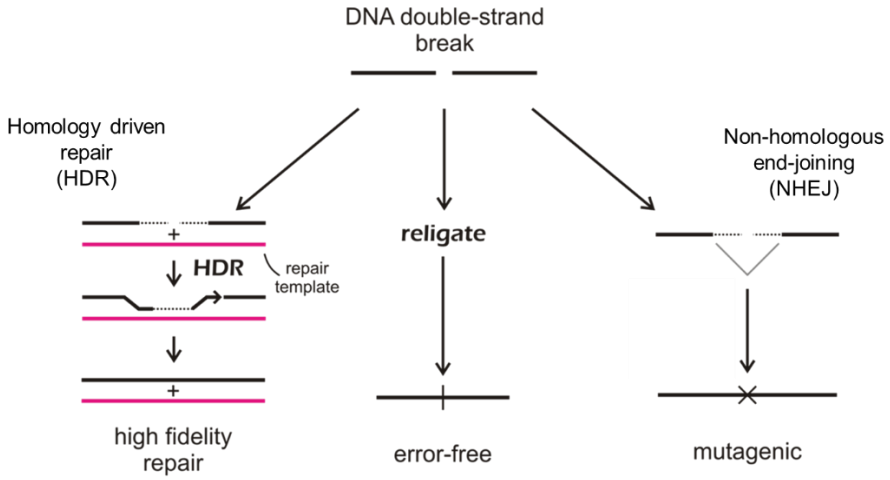


Figure 1.1: DNA double strand breaks mediated by any genome editing tool of interest yield multiple outcomes. Re-ligation by normal DNA repair machinery is common. However, continuous DNA cleavage will eventually yield mutagenesis via non-homologues end joining. By providing a repair template with homology to the native DNA, one has the ability to make insertions or point mutants as well.

Homing Endonucleases (Meganucleases):

Homing endonucleases (HEs) represent an evolutionary iteration of the selfish genetic element. These elements are by far the most compact engineering tool and represent "all in one" enzymes – meaning the DNA targeting and cleavage components are contained in a single molecular construct. The biggest drawback thus far has been the difficulty in re-engineering the DNA binding/recognition sites of homing endonucleases to take advantage of the useful properties of these enzymes. Homing endonucleases are a family of parasitic DNA elements that are found in all kingdoms of life (Stoddard, 2005). As the name implies, HEs are nucleases, yet they are often encoded as open reading frames in mobile introns or inteins. By recognizing and cleaving a target site in an allele not containing the intronic sequence, homologous recombination is initiated to copy the HE coding sequence into the cleaved allele (Belfort and Perlman, 1995). HEs recognize and bind long DNA sequences (18-24 bp) with high specificity and cleave the target site with a double strand break (DSB). The specificity of homing endonucleases allows them to be genome specific in most cases, meaning that they will recognize one site in an entire genome.

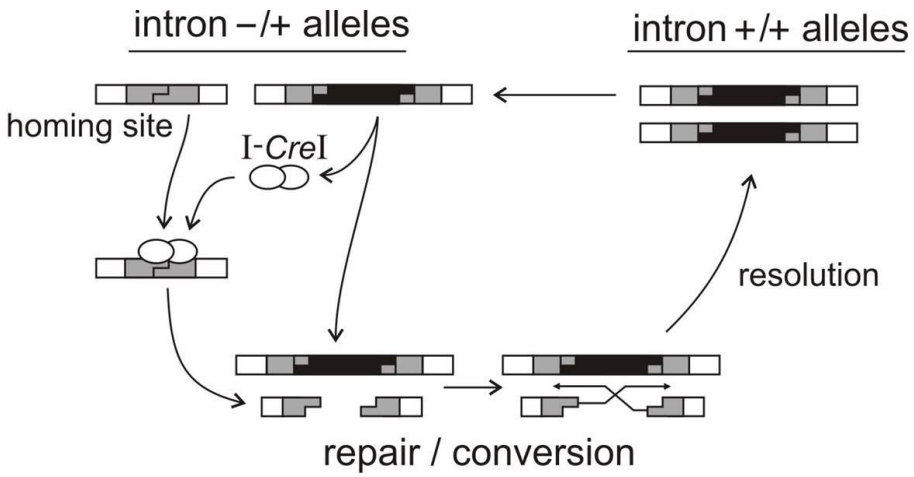


Figure 1.2: Homing endonucleases are naturally occurring self-propagating genetic elements. Found coded in introns, HE transcription and translation will cleave homing sites of the correct sequence and allow the genetic element to be copied into the allele lacking the HEG (Eklund 2005). Homologous recombination can use a copy of the HE coding allele for DSB repair in which gene correction with allele transfer occurs. Alternatively, without a correction template, gene disruption can occur via non-homologous end joining (NHEJ). NHEJ is error prone, and can cause deletions leading to frame shift mutations or codon deletions (Brugmans et al., 2007).

Much effort has been put into altering homing endonucleases to recognize novel target sites instead of being limited to the enzyme's native recognition site. Moreover, the identification of new homing endonucleases continues to grow (Grishin et al., 2010). The LAGLIDAG family of homing endonucleases share a common amino acid sequence motif designated by their family name. Generally, the enzyme becomes functional as a heterodimer, with dimerization or folding mediated by the LAGLIDAG motif. As discovery of new enzymes in this family provides an increasingly robust library of HEs, efforts to create novel HEs have also been successful. Monomers from two different homing endonucleases have been combined to form functional chimeric dimers (Chevalier et al., 2002). As our ability to engineer these enzymes improves and the library of variant enzymes grows, the ability to target genetic sites of interest for homologous recombination or NHEJ events will improve.

Transcription activator-like effector nuclease (TALEN):

TALENs have become popular over the past few years due to the ease of engineering. TALENs are composed of two domain proteins that contain a DNA recognizing transcription activator-like

effector (TALE) component fused to a nuclease component (N). The TALE binding domain is derived from a bacterial protein which has the ability to bind plant promoter DNA sequences, to prevent expression of plant antibacterial genes. It was discovered that the DNA recognition sequence was easily changed by modification of 2 adjacent amino acids in the TAL effector central domain repeat region (Boch et al., 2009). Each 34 amino acid repeat makes up a domain that recognizes and binds a single DNA base pair. By combining multiple repeat domains in sequence, each domain coding recognition for an individual base pair, TAL effector domains that can easily be engineered to recognize any DNA sequence (Miller et al., 2011). When fused to a sequence independent FokI nuclease domain, pairs of TALENs can be used to cleave DNA targets of interest with high specificity (Cermak et al., 2011). Though TALENs are easy to engineer, there are also drawbacks associated with them. First, TALEN constructs are made of large proteins (34 amino acids per DNA base pair). This size constraint pushes TALENs to the upper limits of viral packaging making alternative (non-viral) methods of TALEN delivery necessary. Second, because the amino acid domains required for DNA recognition are repeated many times in a row, the stability of TALEN constructs can be compromised during replication.

CRISPR/Cas9:

The CRISPR/Cas9 system was discovered in bacteria. This complex, which provides an acquired immunity like mechanism in prokaryotes, has now been harnessed for genome engineering in eukaryotic cells The mechanism of CRISPR/Cas9 is similar to that of RNAi in eukaryotes in which an RNA guide sequence is recognized by the CRISPR/Cas9 complex to target a genomic locus of interest. Because the guide RNAs used to cause genome targeting are easy and inexpensive to obtain commercially, multiplexing with this system is more cost effective than in any other system. Off target effects are still being determined, but the ease of use and low barrier to entry of this system makes it very appealing (Cong et al., 2013; Ran et al., 2013).

Zinc finger nucleases (ZFN):

Zinc finger nucleases are part of the first generation of genome editing tools. These enzymes are similar to TALENs, in that they are two domain derived proteins. The first domain is that of a DNA

recognition motif. This motif is derived from a DNA recognition zinc finger. Zinc finger DNA recognition can be programmed using a modular approach to construction (Miller et al., 2007). Three base pair recognition motifs are made of a single zinc finger. Targeting a specific DNA sequence within a genome is possible using a pair ZFNs comprised of 3-4 domains zinc finger domains together with a linker and FokI nuclease domain. Zinc finger nucleases have been reported to have off targeting effects due to the separate non-specific FokI domain (Cheng et al., 2011).

Mega-TALs:

Mega-TALs represent a fusion of a homing endonuclease to a TAL DNA binding domain. The TAL domain is used hold the fused homing endonuclease in place to allow for hyper specific genome editing (Boissel et al., 2014). Extremely low off target cleavage is a benefit of using an endonuclease fusion with built in specificity rather than a non-specific nuclease domain seen in TALEN and ZFN platforms. The drawbacks of this system include the high engineering effort required due to the homing endonuclease fusion section and the large size of the resulting fusion protein.

This growing assortment of tools is representative of how quickly the field of genome engineering has advanced. Each genome engineering tool in this list has specific benefits which depend on the application of interest that will be used. For therapeutic applications for example, nucleases with off target effects should be used with caution. Table 1.1 briefly describes some of the benefits and drawbacks of each platform.

Genome Engineering Tool	Pros	Cons
Homing Endonucleases (Meganucleases)	Compact, single domain for cleavage and DNA specificity	Difficult to alter DNA recognition sequence
Zinc Finger Nucleases	Ease of engineering	Off-target effects, separate DNA Recognition and cleavage domains
Tale effector Nucleases (TALENs)	Ease of engineering	Large size, repeated sequence makes applications difficult
CRISPR/Cas9	Ease of engineering, Ease of multiplexing	Potential off target effects
Mega-TAL fusion	Very high locus specificity, separate cleavage domains are also sequence specific to reduce off target effects	High engineering effort required, large size

Table 1.1: A brief summary of the current genome engineering tools with corresponding benefits and drawbacks.

With the wide variety of genome engineering tools available, there is no shortage of scientific hypotheses and problems one could test. The ability of genome engineering has also been greatly augmented by significant advances DNA sequencing. Costs are very low for generating data for new genomes as well as deep sequencing of known genomes while providing high accuracy. Assembly of new genomes for the identification and annotation of individual genes as described in Chapter 2 is made possible by these advances of sequencing and improvement of bioinformatics tools. As shown, is it now possible for one or a few individuals to produce a draft genome sequence *de novo*. While discussion of current sequencing technologies reaches beyond the scope of this thesis, the benefits of these technologies for genome engineering are highly enabling.

The scientific contributions presented in this dissertation represent various aspects of genome engineering. Ranging from sequencing and annotating a new algal genome; engineering homing endonucleases to target genes in mosquitoes for disease prevention; as well as work to identify and develop reagents to safely insert therapeutic transgenes or cleave and repair endogenous genes and disease-associated mutations.

Chapter 2: A novel algal species for metabolic engineering

Background

The development of new genome engineering tools provides opportunities in both model and non-model organisms. The opportunity to apply genome engineering tools to a new model arose with a new algal species. The diminishing supply of fossil fuels has driven research into more renewable and economically sustainable energy sources, such as biofuels. Thus far, land plants have been the main source of bio-fuel production such as ethanol production of corn. However, studies show that the amount of land required to produce these crops for large-scale fuel production would hinder food production by infringing on available farmland (Singh et al., 2011). As an alternative, algal sources of bio-fuel place fewer demands on land resources and have the potential to be more economically sustainable. However, for this potential to be met, we need to maximize the efficiency of bio-fuel production in algae. Production diesel products from plant or algae material, relies on the production and accumulation of lipids, which can be harvested and transesterified for diesel fuel. Lipid content and quality vary widely in algae, and identification of an alga with all properties ideal for efficient fuel production has been unsuccessful thus far. Use of genetic modification is now shown as tractable in algae (Radakovits et al., 2010; Trentacoste et al., 2013). Therefore, a better understanding of how high energy products such as lipids are made, stored and utilized will allow for identification of potential targets for genetic modification. The organelles associated with lipid storage are called lipid bodies or lipid droplets. The biogenesis of these organelles is poorly understood in algae, but their production is crucial for storage of high value products. The freshwater alga, *Chrysochromulina tobin* represents a potential model for investigating lipid body properties and could be amenable to genetic modifications to study other characteristics of lipid or other high value biological products.

A practical problem in further investigating and engineering the intriguing biology of *Chrysochromulina tobin* has been the lack of a genome sequence. In order to address this issue, I used two different genome sequencing methods together with the sequencing of expressed RNA to generate and annotate a high quality de novo genome sequence of the *Chrysochromulina tobin* nuclear and

organellar genomes. In addition, genomic information is a prerequisite for identification of genome engineering targets of interest.

Chrysochromulina tobin is a member of the Prymnesiophyte class of Haptophyta and represents a model for lipid body studies. This alga also has a combination of unique properties that make it a desirable starting point as a potential biofuel producer. It forms two large lipid bodies which contain both medium and long chain fatty acids (Bigelow et al., 2011) which can be used in bio-diesel production through transesterification. In a 24 hour period, the size of the lipid bodies are largest at the end of a 12h light cycle, and smallest and the end of a 12h dark cycle (Figure 2.1) (Bigelow et al., 2011). Many potential biofuel algae currently require expensive media to be grown (>$1 per L) making large scale culture cost a limiting factor. *Chrysochromulina tobin* can grow in supplemented waste water which is important when considering cost feasibility. *Chrysochromulina tobin* also thrives in high pH environments (~ 9 pH), eliminating many potential microbial competitors, allowing for easier axenic large scale growth conditions.

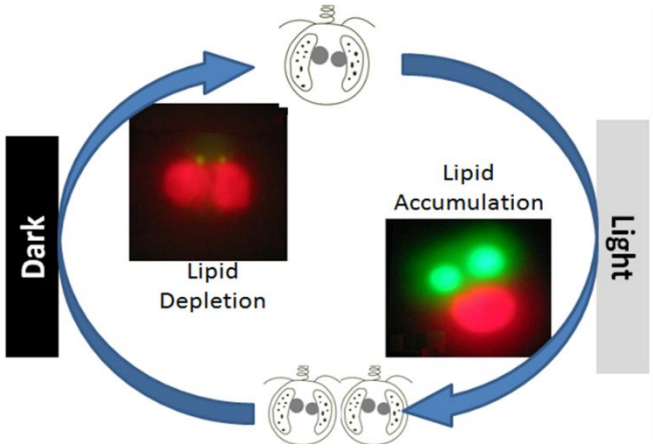


Figure 2.1 Cell division and lipid accumulation is regulated within a 24 hour light dark cycle.

In addition to being an interesting bioproduct generating organism, haptophytes such as *C. tobin* are environmentally significant. Globally, primary producers fix ~100 gigatons of carbon each year. This production is equally distributed between terrestrial and aquatic ecosystems (Field et al., 1998). Haptophytes are globally abundant and important photosynthetic microalgae found in both

marine and freshwater environments. Recent estimates indicate that haptophytes alone may represent "...30-50% of total photosynthetic standing stock in the world's oceans" (Liu et al., 2009), where they play a key role in carbon fixation. Some haptophyte species are photosynthetic as well as mixotrophic, in that they actively prey on bacteria, and thus can live in dysphotic zones where light levels are too low to support photosynthesis (Keeling, 2004). This metabolic versatility may contribute to fitness, and help explain haptophyte prevalence within global algal populations.

Despite their important ecological roles and interesting evolutionary history, there has been little genomic characterization of diverse haptophyte species. Two classes define Haptophytes. The monophyletic Pavlovophyceae display minimal diversity, being described by 4 orders. In contrast, the polyphyletic and globally abundant Prymnesiophyceae encompass 6 orders, of which the B2 clade seems most dominant in marine and fresh water ecosystems (Bittner et al., 2013). Of this vast assemblage of haptophytes, the organellar genomes of only one representative of the Pavlovophyceae, (*Pavlova lutheri*: Pavlovales), and three of the Prymnesiophyceae (*Emiliania huxleyi*: Isochrysidales; *Phaeocystis antarctica* and *Phaeocystis globosa*: Phaeocystales) have been sequenced. The large and complex Prymnesiales that encompass the B1 to B5 clades (Edvardsen et al., 2000; Medlin et al., 2008), lack a sequenced representative. This omission is surprising given reports demonstrating that >55% of all haptophyte sequences in a Mediterranean location belong to this taxonomic assemblage (Bittner et al., 2013; McDonald et al., 2007), and that members of this clade can dominate fresh water ecosystems (Simon et al., 2013). We reasoned that determining the genomic sequence of a B2 representative in the Prymnesiales would provide new information on haptophyte evolutionary origins and ecosystem roles.

Therefore sequencing this genome provides required genome information that can be used for genetic targeting, but also evolutionary history and lipid metabolic pathways such as fatty acid synthesis and lipid body formation. Initially, the organellar genomes of the mitochondria and chloroplast were isolated and annotated, followed by assembly and annotation of the nuclear genome.

Results

Genome sequencing:

Purified total genomic DNA was used to prepare libraries for both the 454 pyrosequencing and Illumina platforms. A total of 4.7 million reads and 79 million reads were generated on the 454 and Illumina platforms respectively, and then assembled using Newbler (Margulies et al., 2005) and Velvet (Zerbino and Birney, 2008) (see Materials and Methods). The resulting draft assembly included 3,472 contigs with an average length of ~17 kb (NCBI Bioproject:SAMN03102970). A total of 59 Mb of sequence was assembled representing coverage of over 100x for the nuclear genome. Contigs containing the portions of the organellar genomes were identified as well. A single contig of 25,263 bp represented 74 % of the mitochondrial genome, but no other assembled contigs contained remaining known mitochondrial sequence, likely due to the presence of a large repeat structure. This repeat structure required PCR amplification and sequencing to resolve the final circular draft. The chloroplast genome was contained in two assembled contigs that totaled 101,192 bp in length. Due to the ribosomal inverted repeat, PCR followed by Sanger sequencing of the amplified products was used to join the two sequences and form a complete, circular mapping chloroplast assembly.

Mitochondrial gene content:

The *Chrysochromulina tobin* mitochondrial genome [GenBank:KJ201908] is 34,288 bp in size, has a GC content of 31.4%. The genome encodes 48 genes, including 25 tRNAs, 21 protein coding genes and a split ribosomal operon comprising the 16S and 23S rRNA genes (Figure 2.2). The mitochondrial 21 protein coding gene complement includes a single novel open reading frame (orf457) that encodes a 457 amino acid protein that lacks strong homology to any other protein within the NCBI database. As in other sequenced haptophytes, NCBI translation table 4 was used, where UGA codes tryptophan rather than a termination codon. Comparison of the genomic content among all published haptophyte genomes (*E. huxleyi* (Sánchez Puerta et al., 2004; Smith and Keeling, 2012): [GenBank:AY342361, JN022704]; *P. antarctica* (Smith et al., 2014): [GenBank:JN131834, JN131835]; *P. globosa*: [GenBank:KC967226]; *P. lutheri*: [GenBank:HQ908424]) indicate that 14 energy and metabolism genes

are conserved in all examined taxa. All haptophyte genomes also contain an identical complement of five ribosomal proteins (*rps*3, *rps*8, *rps*12, *rps*14, and *rpl16*) except for *P. antarctica* and *P. globosa* which are missing the *rps*8 or the *rps*8 and *rps*14 genes respectively (Table 2.1). Most notably, *nad*7, *nad*9 and *nad*11 are consistently missing from all haptophyte and rhodophyte mitochondrial genomes sequenced to date. Interestingly, these three genes are present in all cryptophyte and stramenopile mitochondrial genomes. An estimated 800 mtDNA copies are present per cell.

Table 2.1: Comparison of haptophyte mitochondrial genomes

	Chrysochromulina tobin UWC 291	*Phaeocystis antarctica* (Partial) CCMP 1374	*Phaeocystis globosa* (Partial) Pg-G(A)	*Emiliania huxleyi* CCMP 1516	*Emiliania huxleyi* CCMP 373	*Pavlova lutheri* CCMP 1325 (Partial)
Genome Size (bp)	34288	27547	24477	28660	29013	34086
GC%	31.4	29.7	30.5	28.5	28.3	37.3
Protein-coding genes	21	19	18	20	20	22
Respiratory coding proteins	15	15	15	14	14	15
Ribosomal proteins	5	4	3	5	5	5
Unique gene content	ORF457	-	-	*dam*	*dam*	ORF636 (*dam*), ORF105
Missing genes found in other haptophytes	-	*rps8*	*rps8*, *rps14*	*atp8* (partial only)	*atp8* (partial only)	-
RNA-coding genes						
tRNAs	25	26	25	25	25	24
rRNA content	1 (split operon)	1 (split operon)	1 (split operon)	1 (intact operon)	1 (intact operon)	1 (split operon)
Repeat elements						
Tandem repeats	3	27	4	5	7	27
Inverted repeats	1	4	6	3	1	1
Large repeat regions	1	2	2	1	1	1

Note: No introns were found within any of the listed genomes

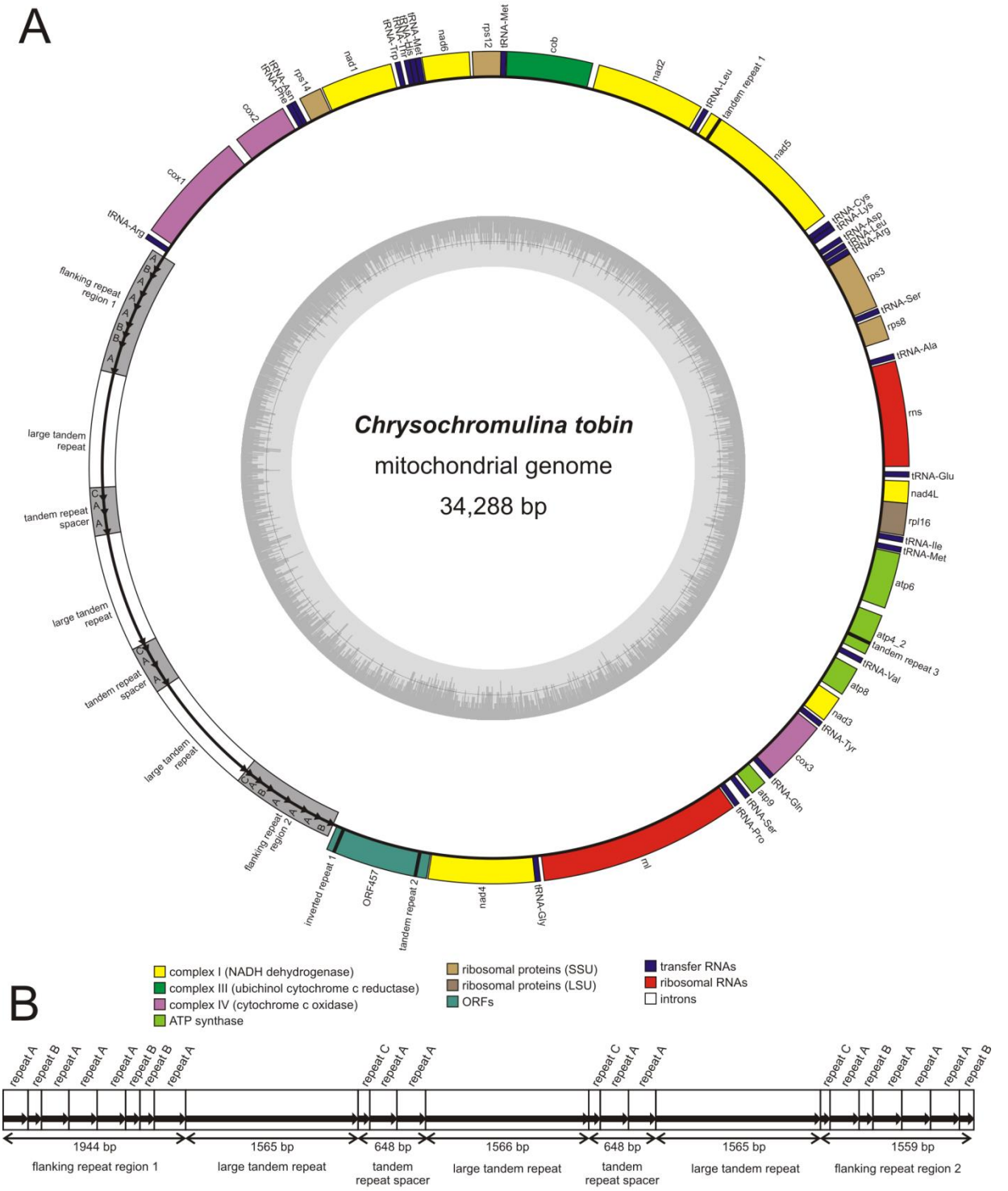


Figure 2.2: *Chrysochromulina tobin* mitochondrial genome map. All genes are transcribed in the same direction (counterclockwise). A split ribosomal operon is present. The large repeat region of 9495 bp represents a significant portion (~28%) of the genome. (B) Detailed representation of the complex repeat region found in the mitochondrial genome. Three large tandem repeat regions are flanked by two sections containing small tandem repeats designated A, B and C. These repeat subunits also make up the regions separating the large tandem repeats from each other. Blocks A, B and C have strong, but rarely perfect, sequence identity.

Mitochondrial repeats:

The *Chrysochromulina tobin* mitochondrial genome contains a large repeat region 9.3 kb (Figure 1B). This region features three large tandem repeats, each ~1.5 kb, that are flanked by two regions consisting of additional small tandem repeats. These small tandem repeat regions are composed of three subunits, arbitrarily classified A, B and C, based on sequence homology (though all sequences within each subunit class are not 100% identical). Repeat unit A is 290 bp. Unit B is 156 bp, of which ~84 bp exhibit significant sequence identity to unit A. Unit C is 85 bp. Although the flanking repeat regions are not identical in size (regions 1 and 2 are 1896 bp and 1558 bp, respectively), a consistent pattern of B-A-A-A-B is found within these two flanking domains. The three large tandem repeats are separated from each other within the repeat region by spacers (consisting of a C-A-A pattern). Interestingly, this direct repeat arrangement is strikingly similar to the larger (35kb) repeat structure found in the diatom *Phaeodactylum tricornutum* (Oudot-Le Secq and Green, 2011). The cryptophytes, *Hemiselmis andersenii* and *Rhodomonas salina*, and the chlorophytes *Pedinomonas minor* and *Acutodesmus obliquus* also contain large tandem repeat regions (>4 kb) that differ from the minimal repeat structure seen in most mitochondrial genomes of other algae.

Not surprisingly, the complexity of this repeat caused assembly challenges. The fact that *P. antarctica, P. globosa,* and *P. lutheri* mitochondrial genomes remain incomplete is likely due to the presence of one or more large repeat structures. For example, Smith et. al. (2014) reported unresolved repeats within two repeat regions in *P. antarctica* and *P. globosa*. Unfortunately, the use of short read, high throughput sequencing techniques does not easily facilitate solving these complex repeat structures. The first *E. huxleyi* mitochondrial genome published in 2004 (Sánchez Puerta et al., 2004), the stramenopile, *Heterosigma akashiwo* (Karol et al.), as well as the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Oudot-Le Secq and Green, 2011) utilized fosmid sequencing, that supported assembly and primer walking for the resolution of longer repeats.

Chloroplast gene content:

The *Chrysochromulina tobin* chloroplast genome [GenBank:KJ201907] is 104,518 bp, circular, and has a GC content of 36.3%. The genome encodes 145 genes (Figure 2.3) including 27 tRNA coding genes, 112 protein coding genes and an inverted repeat, with each repeat copy containing the 23S, 16S and 5S rRNA genes. The tRNAs present in this genome represent all potential amino acid anticodons, including a start methionine. Within the chloroplast genome, codon usage is standard for plastid genomes, and protein alignments suggest that codon GTG (valine) may serve as the translation initiation codon for *ycf*55, *rps*3, *psb*E, *ycf*65 and *psb*C. Such alternative translational start codons have been reported in algal chloroplast genomes of wide taxonomic divergence (e.g., *Cyanidium caldarium, Odontella sinensis, Heterosigma akashiwo,* and *Emiliania huxleyi*) (Sánchez Puerta et al., 2005; Starkenburg et al., 2014; Wang et al., 2013) although not for the same genes established in *C. tobin*. Approximately 250 copies of a chloroplast genome are found per cell.

The *C. tobin* chloroplast gene complement is similar to other sequenced haptophyte chloroplast genomes: (Table 2; *E. huxleyi*: [Genbank:AY741371, JN022705]; *P. antarctica*: [GenBank:NC_016703]; *P. globosa*: [GenBank:NC_021637] and *P. lutheri*: [GenBank: NC_020371]). Additionally, an "uncultured prymnesiophyte C19847" (derived from metagenomic data of oceanic samples collected from the North Atlantic [GenBank:HM565909] (Cuvelier et al., 2010)) was included in this analysis. All haptophyte genomes are relatively small in size when compared to other microalgal species (Additional File 2). Gene content comparison shows *E. huxleyi* (113 protein-coding genes) contains *dfr* (a two component signaling protein) that is absent in *C. tobin*. Unlike *C. tobin, P. antarctica* and *P. globosa* chloroplast genomes (both having 108 genes) are missing ORF132 (unknown function), *ycf20* (unknown function), *thi*G, and *thi*S (thiamine biosynthesis protein G and S respectively). A conserved coding region (ORF154), found uniquely in *C. tobin* (154 amino acids) and *E. huxleyi* (132 amino acids), is located adjacent to *psb*V in both genomes. The amino acid identity of these two hypothetical genes is low, suggesting remnants of functional proteins. The chloroplast genome of the phylogenetically distant haptophyte, *P. lutheri*, is missing 7 genes, and contains an additional 9 genes that are not found in

available haptophyte chloroplast genomes (Table 2.2). While gene content is similar among all of the haptophytes analyzed, the freshwater *C. tobin* actually has the highest sequence identity to the marine uncultured prymnesiophyte C19847. This result is not too surprising given recent studies that document the occurrence of multiple, independent freshwater colonizations by haptophytes (Shalchian-Tabrizi et al., 2011). Our 18s rDNA based phylogenetic analyses (Deodato, Barlow et. al., in prep) show *C. tobin* to cluster with species isolated from fresh water lakes in France (Simon et al., 2013).

Co-linearity in gene placement among haptophyte chloroplast genomes was assessed. Unlike in diatoms (Oudot-Le Secq et al., 2007), gene clusters have been exchanged between the large and small single copy regions within these haptophyte chloroplast genomes. When comparing *E. huxleyi* to *C. tobin*, and *E. huxleyi* to *P. antarctica*, 17 and 13 gene clusters were conserved, respectively. A highly conserved region of 20,610 bp encompassing *ccs*1 through *atp*A (18 genes) was identified. This region contains a single inversion in the *C. tobin rps*2 and *rps*4 coding region, and is more highly conserved between *E. huxleyi* and *P. antarctica* – expanding to a ~30,000 bp region that initiates with *pet*L (cytochrome b6/f complex component) and ends with *psb*E (photosystem II protein). Another large gene cluster conserved in all three species consists of ~15,000 bp that contains the commonly preserved 24 ribosomal protein gene operon and the *dna*K heat shock protein. GRIMM (Tesler, 2002) analysis was used to quantify the degree of gene rearrangement among the three completed haptophyte chloroplast genomes above. The most parsimonious result found 11 genome rearrangements occurring between *E. huxleyi* and *P. antarctica*, 10 rearrangements between *P. antarctica* and *C. tobin*, and 19 rearrangements between *E. huxleyi* and *C. tobin*. The scrambled placement of genes among these haptophytes yields no clear insight into relatedness. Four genera of haptophytes now have sequenced chloroplast genomes and the contents are compared in the table below. In addition to known haptophytes, a metagenomic sample contained another likely haptophyte sequence (Prymnesiophyte C19847) which has the strongest sequence homology in the NCBI database to that of *C. tobin*.

Table 2.2 Comparison of haptophyte chloroplast genomes

	Chrysochromulina tobin UWC 291	*Phaeocystis antarctica* CCMP 1374	*Phaeocystis globosa* Pg-G(A)	*Emiliania huxleyi* CCMP373	*Emiliania huxleyi* CCMP1516	*Pavlova lutheri* CCMP 1325	Uncultured Prymnesiophyte C19847 (Partial genome)
Genome Size (bp)	104518	105651	107461	105309	105297	95281	45567
GC%	36.3	35.5	35.4	36.8	36.8	35.6	37.2
Protein-coding Genes	112	108	108	119	113	111	45
Unique gene content	ORF154, *ycf20*, *thiG*, *thiS*			ORF132, *ycf20*, *dfr*, *thiG*, *thiS*		ORF140, *ycf66*,ORF 208, ORF66, rpoZ, ORF84, RF489,OR F70, ORF69	N/A
Missing genes found in other haptophytes						*ycf55, ycf47, ycf80, ycf65, rpl34,ycf4 6, ycf35*	N/A
RNA-coding genes							
tRNAs	27	27	27	28	30	27	18
Ribosomal operons (23S, 16S, 5S)	2 (inverted repeat)	2 (inverted repeat)	2 (inverted repeat)	2 (inverted repeat)	2 (inverted repeat)	1	1
Repeat elements							
Inverted repeats	16	7	10	15	16	6	11
Tandem repeats	1	6	2	1	1	2	1

Note: No introns were found within any of the listed genomes

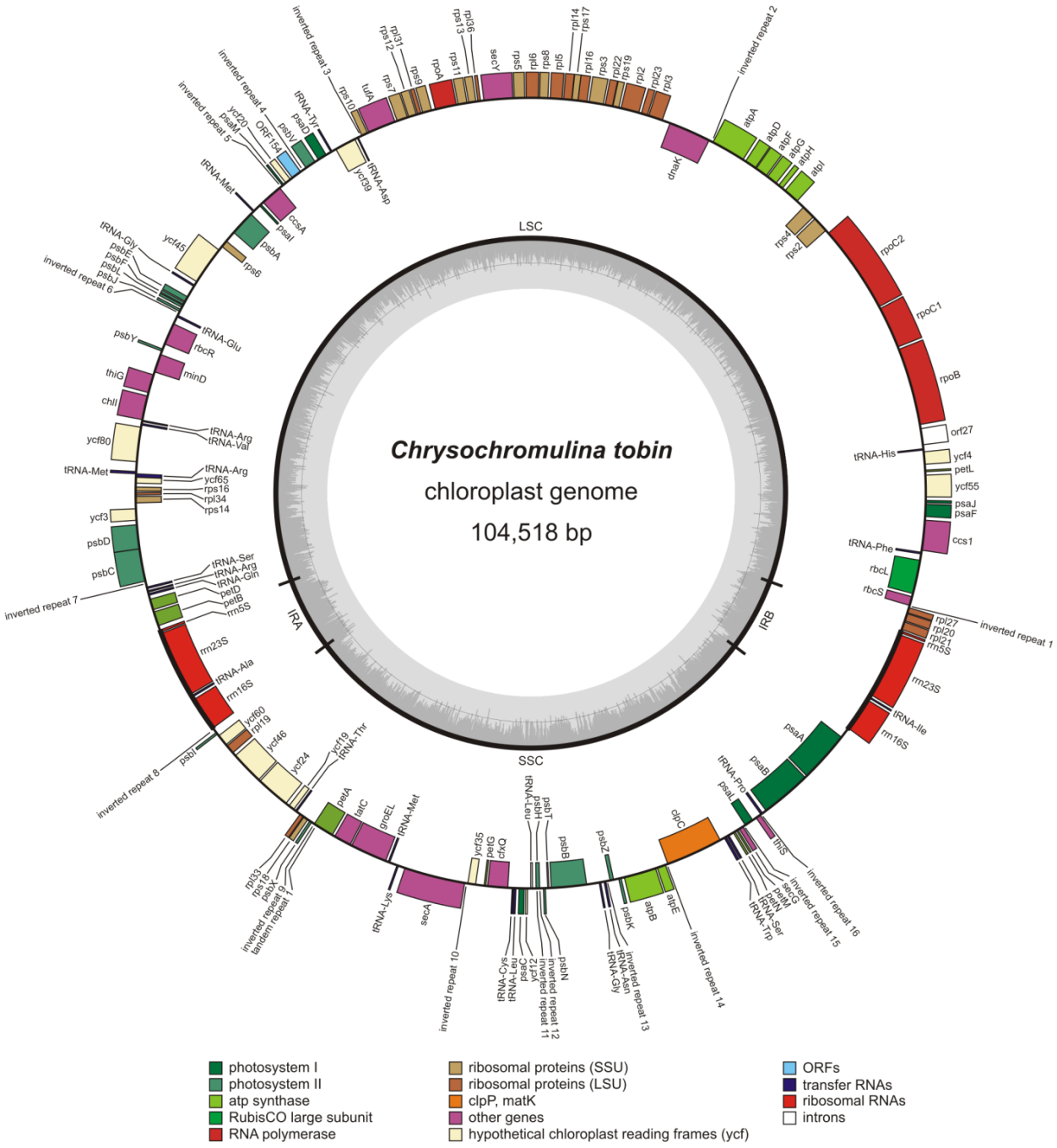


Figure 2.3: *Chrysochromulina tobin* chloroplast genome map. Genes facing outside are transcribed in the counter-clockwise direction and genes facing inside are transcribed in a clockwise direction. Two copies of the ribosomal operon are inverted and the repeat region contains no other genes beyond the ribosomal subunits. The small single copy (SSC) and large single copy (LSC) regions are labeled. Inverted and tandem repeats are also designated.

Chloroplast repeats and evolution:

Haptophyte evolutionary history remains enigmatic. Based on fossil records and 18S rDNA phylogenetic analyses (Keeling et al., 2005; Medlin et al., 1997), it is estimated that these algae are an ancient lineage, arising over 1.2 billion years ago. Phylogenomic analyses of the plastids of cryptophytes, dinoflagellates (alveolates), stramenopiles and haptophytes, show that the plastids of these four groups, collectively termed "CASH" (Baurain et al., 2010), form a monophyletic grouping descendent from red algal plastids. However, the relationships among CASH plastids remains controversial, as differing topologies are recovered in phylogenetic analyses of chloroplast genes using various methods and loci (Green, 2011). Aside from the plastid lineage controversy, the haptophyte host lineage may be affiliated with the stramenopile-alveolate-rhizaria (SAR) group (Burki et al., 2008).

Large inverted repeat:

The *Chrysochromulina tobin* ribosomal repeat region is structurally unique when compared to those found in both land plants and all algal chloroplasts sequenced to date. Eighty two percent (209/256) of all chloroplast genomes (including non-algal species) examined at the genus level contain a large inverted repeat. The conventional structure of this conserved operon includes the 16S ribosomal gene, an intergenic spacer region (ISR) that encodes the tRNA-isoleucine (anticodon GAU), and the tRNA-alanine (anticodon UGC). The ISR is followed by the 23S ribosomal subunit gene and the 5S ribosomal gene (Figure 2.4). In land plants and chlorophytic algae, and less often in rhodophytes and CASH members, the repeat region expands to include additional genes that flank the ribosomal gene operon, making the inverted repeat in chlorophytes larger on average (Appendix Table A1.1). The ribosomal inverted repeat structures in *C. tobin* have non-identical tRNA coding sequences within each ribosomal intergenic spacer region. This domain normally contains two identical tRNA coding regions in each inverted repeat. However, *C. tobin* has only tRNA alanine in inverted repeat A, and tRNA isoleucine in inverted repeat B (Figure 2.4). This pattern is also likely present in the uncultured prymnesiophyte C19847. Only a single operon was assembled for this organism, but that operon solely

contained tRNA-Ala within the ISR. Interestingly, the C19847 ribosomal operon that is adjacent to the *pet*B gene contains a tRNA-Ala, while the ribosomal operon adjacent to *pet*B in *C. tobin* contains tRNA-Ile. Intramolecular recombination within chloroplast genomes having an inverted repeat is well documented for several algae as well as in land plants (Aldrich et al., 1985; Cattolico et al., 2008; Palmer and Thompson, 1982). This process generates genomic isomers that differ solely in the orientation of their single copy domains. Surprisingly, using long PCR to bridge both *C. tobin* repeat regions, no evidence of such flip-flop recombination was found.

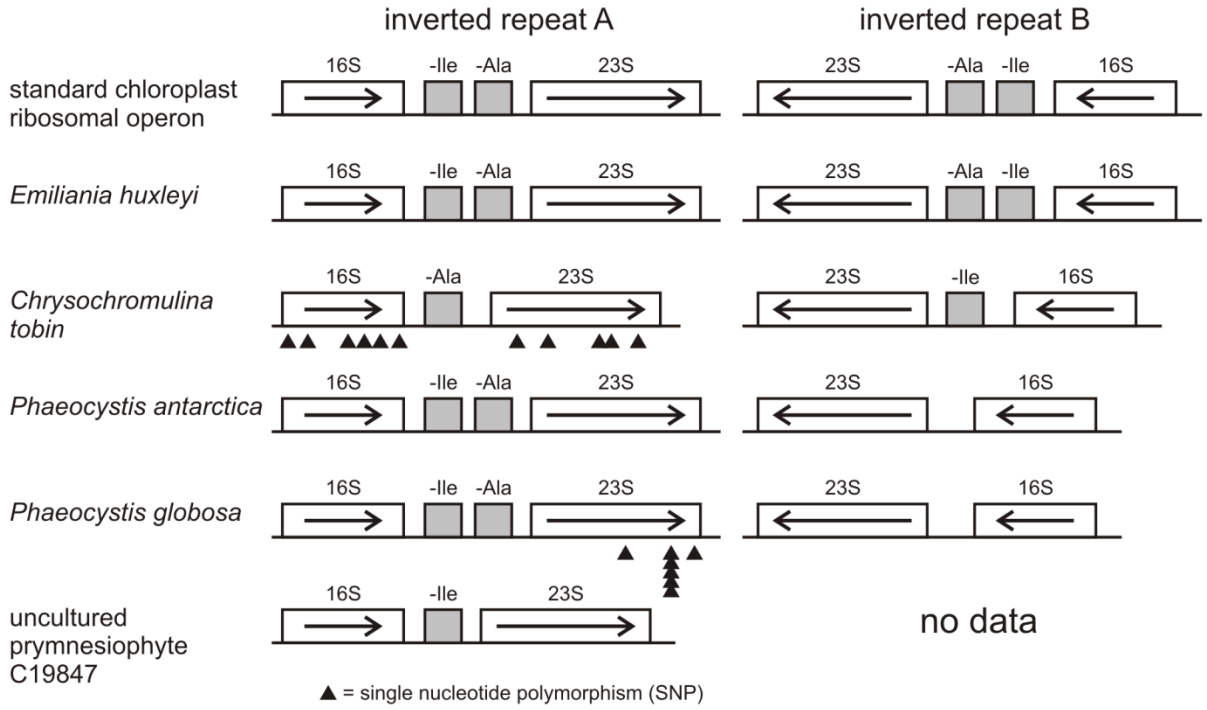


Figure 2.4: Ribosomal operon repeats of *C. tobin* and other haptophyte chloroplast genomes. The "standard" chloroplast ribosomal operon contains tRNA-Ile and tRNA-Ala in the ribosomal intergenic spacer regions (ISR). *Chrysochromulina tobin* contains tRNA-Ala in operon A and tRNA-Ile in operon B. Multiple SNPs (black triangles) are present between the *C. tobin* large and small ribosomal subunits coding regions to each other. One copy of the ribosomal operon contains both tRNAs while the other operon lacks both tRNAs in *P. antarctica* and *P. globosa* plastid genomes. A metagenomic sample, uncultured prymnesiophyte C19847 [28], contains a single tRNA-Ala; the second ribosomal copy was not assembled and is therefore unknown in structure.

Within haptophytes, *E. huxleyi* has a canonical ribosomal operon structure, but *P. antarctica* and *P. globosa* do not. In both *P. antarctica* and *P. globosa*, one copy of the ribosomal operon contains both tRNAs in the intergenic spacer region (conventional arrangement), while the second copy lacks

both tRNAs in the intergenic spacer domain (non-conventional). As also seen in Figure 2.4, copies of the ribosomal gene sequences encoded within the *C. tobin* repeat, contain single nucleotide polymorphisms in the 16S (6 SNPs) and 23S (5 SNPs) ribosomal genes. Notably, every cryptophyte and rhodophyte chloroplast genome examined that encodes a repeat structure also shows the presence of SNPs between replicated ribosomal genes (Appendix Table A1.1). In contrast, only one alga in the chlorophyte lineage, *Ostreococcus tauri*, contains a SNP. Additionally, no land plant species queried show SNPs in either 16S or 23S genes. Though speculative, the presence of alternative operon structure for the ribosomal genes, combined with the elevated appearance of SNPs suggest that the well-recognized "copy correction" mechanism (Newman et al., 1990) may be more effective in some "green" algal lineages (chlorophytic algae and algae with chlorophytic algal symbionts), than in the "red" lineage of autotrophs (rhodophytes and CASH taxa). This suggests that one method of repeat loss within an algal chloroplast genome may be driven by the accumulation of SNPs and eventual disintegration of operon integrity.

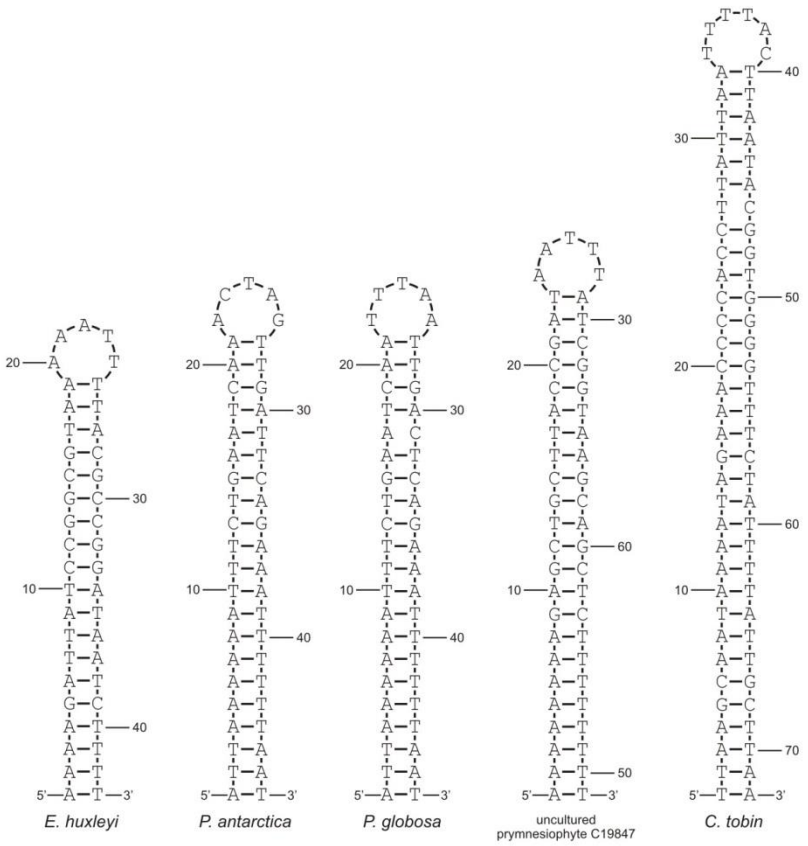
Small repeat function in chloroplast genomes:

Chloroplast genomes are consistently embellished with small repeats that are either tandem or inverted in orientation. *Chrysochromulina tobin* is no exception having 16 inverted repeats with an average length of 25.4 ±5.2 bp in stem length with loop domains averaging 6.1 ± 3.0 bp in size. A single tandem repeat comprised of a duplicated 15 bp sequence was also found. Similar to observations made for other chloroplast genomes (Cai et al., 2008; Haberle et al., 2008) and bacterial genomes (Delihas, 2011; Emmanuel D Ladoukakis, 2008), most *C. tobin* chloroplast repeats occur within the intergenic space, at the termini of genes located on opposite coding strands (Figure 2.3).

The conservation of repeats within the chloroplast genomes of all algal taxa suggests a functional constraint for these structures. Insight into the contribution of repeats to chloroplast metabolic processes was accomplished by elegant studies with *Chlamydomonas reinhardtii* (Drager et al., 1996; Rott et al., 1998) that exploited the early ability to transform this organism. Though these

studies were predominantly focused on inverted repeats and confined to a limited gene set (e.g., *atp*B and *rbc*L), observations reveal a multifunctional role for repeat structures (Adams and Stern, 1990). Data show that the presence of a repeat at the terminus of a gene is necessary for proper mRNA processing by exo- and endonucleases, maintaining RNA stability and supporting protein translation.

Given the established functional contribution of repeats, we asked whether specific genes or conserved chloroplast gene clusters were targeted for repeat association. Significant gene-repeat association was observed. For example, an inverted repeat is found at the terminus (*rps*10) of the ribosomal protein operon that encompasses 24 genes; inverted repeats are often present after the RuBisCO operon (*rbc*L/*rbc*S), as well as following the photosystem II gene pair (*psb*C/*psb*D). Each of these targeted genes has no spatial relationship to one another with respect to in-chromosome placement; no concurrence exists among these genes in repeat type (tandem, inverted) and no similarity in sequence identity is seen in the repeat structures that are associated with the targeted gene. Despite these facts, selected gene-repeat associations (e.g., *rbc*L/*rbc*S) are conserved in chloroplast genomes as taxonomically disparate as *C. tobin and E. huxleyi* (Haptophyta), *Ectocarpus siliculosus* (Stramenopila: Phaeophyceae), *Cyanidium caldarium* (Rhodophyta) and *Rhodomonas salina* (Cryptophyta) - a list that is by no means exhaustive. Single genes that are not associated with operons can also be targeted for repeat tagging. A good example is *clp*C that encodes an ATPase-dependent protease. This gene is found in different locations with many dissimilar up and downstream gene neighbors among CASH taxa. *clp*C is tagged with a repeat in all haptophytes except *P. lutheri*. A repeat is also found next to *clpC* in 19 of 44 (43.2%) CASH plastid genomes analyzed to date. As shown in Figure 2.5, even within haptophytes, the repeat is conserved *only* in gene association. Neither the size, sequence, nor stem loop structure formed by these repeats is conserved.



C. tobin KJ201909	TTAAGCAATAAAATAGAAACCCCACCTTATTAATTTTACTTAATACGGTGGGGTTTCTATTTTATTGCTTAA
E. huxleyi JN022705	TAGAAAAGATTATCCGGCGTAAAAATTTTACGCCGGATAATCTTTTTTA
E. huxleyi NC_007288	TAGAAAAGATTATCCGGCGTAAAAATTTTACGCCGGATAATCTTTTTTA
P. antarctica JN117275	ATTAAAAAATTTCTGAATCAAACTAGTTGATTCAGAAATTTTTTAAT
P. antarctica NC_016703	ATTAAAAAATTTCTGAATCAAACTAGTTGATTCAGAAATTTTTTAAT
P. globosa NC_021637	ATTAAAAAATTTCTGAATCAATTTAATTGACTCAGAAATTTTTTAAT
uncultured prymnesiophyte C19847 HM56909	AAAAAAAAGAGCTGCTTACCGATAAATTATCGGTAAGCAGCTCTTTTTTTT

Figure 2.5: Conserved inverted repeats found adjacent to haptophyte clpC genes.

Given that repeats appear to have a functional significance, it was also of interest to determine whether a pattern in repeat acquisition exists among evolutionary diverse algae. Repeat properties were queried across three groups of algae: rhodophytes (red algae), the 'green' algal lineage (green algae and those algae derived from the secondary endosymbiotic uptake of a chlorophyte [i.e. euglenids]) and the CASH grouping of algal species (derived from secondary or higher order endosymbioses of a rhodophytic plastid). Data clearly show that the number of repeats found in a chloroplast genome varies when different algal groups are compared. Rhodophytes appear to have few repeats (10 to 16 when excluding *Cyanidioschyzon merolae,* n=79), and the CASH taxa have a moderate

number (4 to 49 repeats [average = 25]). In contrast, the green plastid lineage has on average 80 repeats per genome, though representatives have as many as 281 (*Dunaliella salina*) to 435 (*Chara vulgaris*) (Additional File 2). Repeat type is also group dependent. CASH algae have a greater number of inverted repeats in their chloroplast genomes, whereas the green lineages have significantly more tandem repeats (Figure 2.6A). Attempts to assess whether differences in chloroplast size and intergenic distance influenced the number and size of repeat structures show both parameters to be positively correlated with an increase in repeat number for green and CASH plastid lineages (Figure 2.6B). However, there appears to be a limit on repeat size in the CASH plastid lineage, for even as genomes become larger and/or intergenic distances increase, repeat size does not exceed ~65 bp. This result significantly contrasts with that seen in the green lineage. A strong correlation exists between increased repeat size, and either an increased genome size or an increased intergenic distance. The fact that repeat embellishment occurs in every algal chloroplast genome analyzed to date, that repeats often are conserved near specific genes, and repeats contribute to chloroplast gene expression, suggest that future research analyzing chloroplast intergenic regions is warranted.

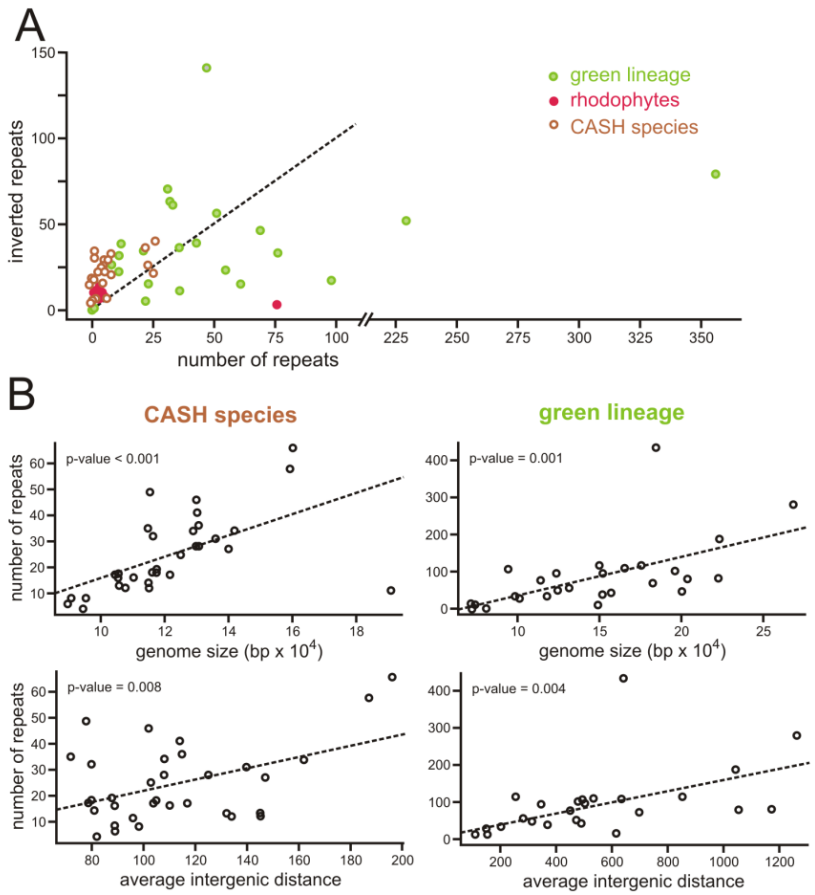


Figure 2.6: Small repeat analysis across algal groups. (A) Tandem and Inverted repeat complement across CASH, rhodophyte and "green" algal species. The dotted line represents a 1:1 ratio of tandem and inverted repeat counts. (B) Linear association of repeat number versus genome size and average intergenic distance.

Chloroplast protein characterization:

Analysis of select genes can provide unique insight into chloroplast genome evolution and function. In this context, several genes within the *Chrysochromulina tobin* chloroplast genome are described below.

Ribosomal protein RPL36:

It is now known that some ribosomal proteins are multifunctional. Not only do these proteins serve as architectural components in the ribosome itself, but may also have additional extra-ribosomal functions that help maintain cellular homeostasis (Warner and McIntosh, 2009). As shown in Figure 2.7B, the 'conventional' (C^+ motif) RPL36 protein encoded by chlorophytes and rhodophytes has a

highly conserved zinc finger motif of the cysteine-cysteine-cysteine-histidine (CCCH) type (indicated by arrows). The haptophyte and cryptophyte RPL36 (C^- motif) proteins lack the conserved zinc finger domain. In both haptophytes and cryptophytes the first cysteine is replaced by a serine, and the terminal histidine of the zinc finger motif is replaced by a leucine. Therefore it is very unlikely that the haptophyte/cryptophyte RPL36 C^- retains zinc finger protein function.

Figure 2.7: Phylogenetic analysis of RPL36 proteins. Bayesian majority rule consensus tree of 85 RPL36 proteins from haptophytes, cryptophytes, a haptophyte plastid-containing dinoflagellate, stramenopiles, rhodophytes, *Paulinella chromatophora*, and select bacteria and cyanobacteria. Taxa are colored according to the legend. Bayesian posterior probabilities and Maximum Likelihood bootstrap support are shown at nodes. Scale bar shows amino acid substitutions per site (A). Logo plot consensus sequences for the C^- and C^+ RPL36 protein (B). The zinc finger residues are completely conserved in the C^+ genotype, while 2 residues are absent from the C^- clade.

Although the zinc finger function was not recognized, earlier studies used the unique RPL36 C^- sequence observed in haptophyte and cryptophyte chloroplast genomes to argue for a sister relationship between the plastids of these taxa (Rice and Palmer, 2006). Sanchez-Puerta and Delwiche attempted to reconcile the presence of an *rpl*36 C^- gene (likely derived from a bacterium via lateral gene transfer) in the cryptophyte and haptophyte plastids with the presence of the ancestral *rpl*36 C^+ gene in stramenopiles by positing that, for a time, two chloroplast genomes co-existed in the haptophytes and cryptophytes, some genomes containing the *rpl*36 C^+ gene and others with the C^- gene. One or the other genome was then fixed in particular lineages. This hypothesis predicts the discovery of an *rpl*36 C^+ gene in the chloroplast genomes of some haptophytes or cryptophytes. To better test this hypothesis, we infer a new *rpl36* phylogeny including five additional haptophyte genera (seven species), 24 additional stramenopiles, four additional rhodophytes, as well as a representative of the recently recognized algal lineage *Paulinella*.

Mining NCBI as well as our publically available chloroplast genome database (http://chloroplast.ocean.washington.edu), a total of 462 non-redundant RPL36 amino acid sequences were recovered for phylogenetic analysis. In figure 2.7, this large dataset is condensed and re-inferred to include sequences from a limited number of bacterial and cyanobacterial representatives; rhodophytic as well as CASH plastids. We confirm the RPL36 C^- identity of all haptophyte and cryptophyte algae to the exclusion of all other eukaryotic algal taxa sampled, including stramenopiles and rhodophytes (Rice and Palmer, 2006), thus Sanchez-Puerta and Delwiche's hypothesis as described above was not confirmed. Furthermore, even though our *rpl*36 phylogeny increases bacterial sampling beyond that used for prior analyses, results do not support a planctomycete-origin of the laterally transferred *rpl*36 as previously suggested by Rice and Palmer. No bacterial clade is strongly supported as the donor of the C^- gene. Determination of the donor lineage is made difficult by the short length of the *rpl*36 protein (C^- 49 amino acids, C^+ 38 amino acids) and the ancient nature of the lateral gene transfer event.

The duality in chloroplast encoded *rpl*36 genes poses questions concerning the contribution of each alternative protein type to different algal lineages. The functional contribution is most likely multifaceted. Certainly, the RPL36 protein, whether of the C^+ or C^- type, contributes to ribosomal structure (Ban et al., 2000). However, bacteria that contain both paralogs of the RPL36 protein differentially express *rpl*36 C^- and *rpl*36 C^+ when subject to zinc stress (Panina et al., 2003). The *rpl*36 C^- gene is up-regulated under limiting conditions. Since the zinc finger domain of RPL36 not only binds zinc, but also bind other cationic species (Boysen and Hearn, 2001), one might speculate that in algal cells an increased metal ion binding potential might provide a competitive advantage when living in ecosystems where particular cofactors are in short supply. The fact that haptophyte/cryptophyte RPL36 C^- proteins have an extended 7 to 9 amino acid C terminus that is enriched with positively charged as well as hydrophobic moieties supports the possibility that this small molecule also has a regulatory function, similar to that established for other ribosomal zinc finger proteins (Matthews and Sunde, 2002).

Two component signal transduction systems:

Two-component regulatory systems are key mechanisms through which many organisms (bacteria, archaea, and eukaryotes) control responses to fluctuating environmental conditions (Mascher et al., 2006; Stock et al., 2000). Numerous two component regulatory systems exist. In its simplest form, when cued by an external stimulus, a phosphoryl group from a conserved histidine residue within a sensor kinase protein is transferred to an aspartic acid in the receiver domain of a response regulator protein. Phosphorylation of the response regulator protein activates an effector domain (usually through a conformation change) to propagate the intended regulatory effect.

The *Chrysochromulina tobin* chloroplast genome encodes two potential response regulator proteins but is devoid of sensor kinase genes. The first response regulator protein, *orf*27, encodes a protein similar to the TRG1 response regulator described for the stramenopile *Heterosigma akashiwo* (Jacobs et al., 1999) and is also found in other CASH species, rhodophytes, and cyanobacteria. The

second response regulator, *ycf*55, is likely a member of a new subclass of response regulators evolved from the cyanobacterial type "CheY-like" response regulator proteins. Many of the cyanobacterial CheY-like homologs are comprised of approximately 550 amino acids and contain two domains; the aforementioned receiver domain and a conserved domain of unknown function (DUF3685), hypothesized to be the effector. Intriguingly, the *che*Y-like homolog (*ycf*55) found in *C. tobin* is comprised of only 314 amino acids. Multi-sequence protein alignments of a variety of response regulators from cyanobacteria, algae, and *Arabidopsis* revealed that the C-terminus of the *C. tobin ycf*55 is most similar to the cyanobacterial type CheY-like proteins, as both contain the terminal DUF3685 domain. In contrast, the N-terminus of the *C. tobin ycf*55 is divergent from both cyanobacterial CheY-like response regulators and the plant type response regulators (i.e., ARR1-14), including loss of the canonical site of phosphorylation. Nevertheless, sequences that resemble the *C. tobin* type of *ycf*55 are conserved in rhodophytes (*Chondrus crispus*, *Calliarthron tuberculosum*, *Gracilaria tenuistipitata*, *Porphyra purpurea*, and two *Pyropias* species), other haptophytes (except *P. lutheri*), and some cyanobacteria (classified as 'RRI-other') indicating that this protein still provides an important function. Within this divergent subclass of *ycf*55 proteins an aspartic acid residue (D43) just upstream of the canonical position is conserved, suggesting that this residue could replace the canonical site of phosphorylation by an as yet unknown sensor kinase.

Organellar genomes summary: The complete sequence of *Chrysochromulina tobin* mitochondrial and chloroplast genomes, representative of the ecologically important haptophyte prymnesiales B2 clade, have been determined and annotated. Within the mitochondrial genome, a large repeat structure consisting of ~9 kb was found along with a novel 457 amino acid open reading frame of unknown function. The large inverted repeats in the chloroplast genome contain a combination of novel of intergenic spacer region structures and SNP variants when rDNA-containing domains are compared, indicating the possible loss of a copy correction mechanism. Notably, no recombined structural isomers of the *C. tobin* chloroplast genome were found. Small repeats within intergenic regions of the chloroplast genome have taxon-specific evolutionary features. The consistent

association of specific genes with small repeat sequences of specific genes by repeats argues for a functional role in metabolism for these structures. Several genes found in *C. tobin* chloroplast remain uncharacterized, yet conserved in other algal species. They include: a ribosomal protein *rpl*36; a new two component signal transduction protein; the potential NmrA-like NADP-binding nitrogen regulator, and two chloroplast protein import genes.

***C. tobin* nuclear genome:**

The *C. tobin* genome (59 Mb) is significantly smaller than other haptophyte genomes previously studied. The only other haptophyte genome that has a draft nuclear genome is that of *Emiliania huxleyi* which has ~140 Mb in the genome (Read et al., 2013). Other haptophyte genomes (haploid size) have been predicted to range from ~117 Mb (*Phaeocystis antarctica*) to ~230 Mb (*Prymnesium polylepis*) (John et al., 2010). Propidium iodide staining with flow cytometry estimated *C. tobin*'s DNA content to be approximately 55 Mb. Because this genome size corresponds very closely with the size of the draft genome assembly presented in this work, we suggest that *C. tobin* is haploid. A diploid state is likely however even though it has not been observed, due to the presence of a full complement of meiosis related genes (Deodato et. al., in prep) that are found in the other sequenced haptophyte to date, *Emiliania huxleyi*, which has been shown to have haploid and diploid phases (Mausz and Pohnert, 2014).

Table 2.3: *C. tobin* nuclear genome statistics

Assembled genome size	59 Mb
Sequencing coverage	111x
Assembled contigs	3472
Average contig size	~17 kb
N50 / L50	24114 bp / 798 contigs
Contigs > 75kb	13
GC content	63.4%

Gene annotations were performed using MAKER2 (version 2.10) (Holt and Yandell, 2011) with a variety of *ab initio* and trained gene models. The annotation was also aided by assembled transcriptomic data from Cufflinks (version 2.0.2) (Trapnell et al., 2010). Gene functional annotation

was carried out by Blast2GO searches against NCBI databases with an e value cutoff of $1e^{-6}$. Interpro and Pfam protein databases were also queried using Blast2GO (Conesa et al., 2005). In total, 61.4% of genes called had BLAST homologs identified.

Table 2.4: *C. tobin* nuclear gene calling and annotation

Protein coding genes (Nuclear genome)	16777
Genes supported by BLAST homology (Blast expect value < $1e^{-6}$)	10293 (61.4%)
Genes not supported by BLAST homology	6484 (38.6%)
Average gene length	1405 bp
Average exon length	445 bp
Average exons per gene	2.28
Average intron length	297 bp
Average introns per gene	1.28

Light-Dark transcript profiling:

The *Chrysochromulina tobin* life cycle is highly reliant on a light dark photoperiod given that poor growth and culture failure is observed when this alga is maintained in continuous light. During logarithmic phase growth, approximately one cell division is seen over the photoperiod, generally during the dark period. Lipid body characterization by the use of neutral lipid stain BODIPY 500/515 also defines dramatic changes in the size of the lipid body (lipid droplet) organelle over the 24 hour cell cycle. The largest lipid bodies are observed at the end of the light photoperiod, and lipid bodies are depleted at the end of the dark cycle, suggesting photoperiod tightly controls metabolic processes and cell division in this organism (Figure 2.8). Photoperiod sensitivity is also supported by the expression of cell cycle specific genes that were identified in the genome annotation and transcriptomic data collection. The expression levels of these genes are highly up-regulated at the end of the light period and during the dark period, when cell division is measured.

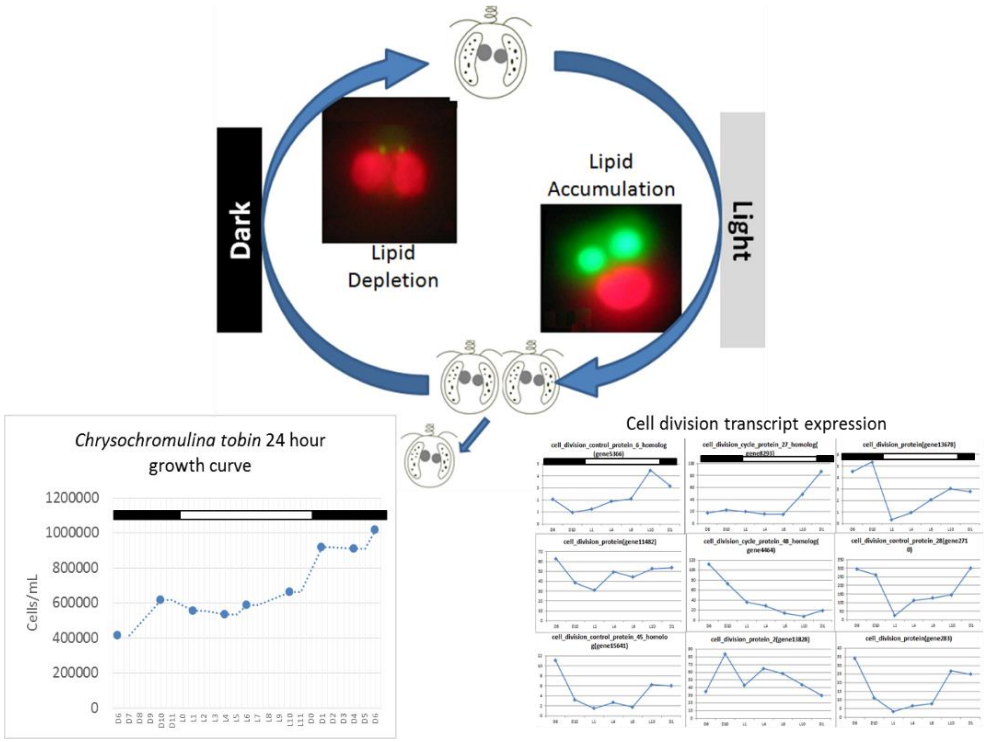


Figure 2.8: *C. tobin* has highly programmed metabolic and cell division response in the context of photoperiod cues. Lipid body organelles were observed to change size dramatically over the 24 hour photoperiod. Cell division also is observed to be regulated by photoperiod, where cell division seems to occur during the dark photoperiod.

Global transcriptomics:

RNA was collected at 7 time points during the 24 hour (12 hour light:12 hour dark) diurnal cycle. Poly-A selection was performed, therefore mitochondrial and chloroplast coded proteins are not found in the transcriptomic dataset. To determine transcripts important to diurnal cycle and cellular division, the top expressing genes with large expression changes were identified and binned into 7 groups for expression analysis (Figure 2.9). Using Fisher's Exact Test, for the analysis of enrichment of gene ontology (GO) terms, several GO terms were identified as overrepresented in each group.

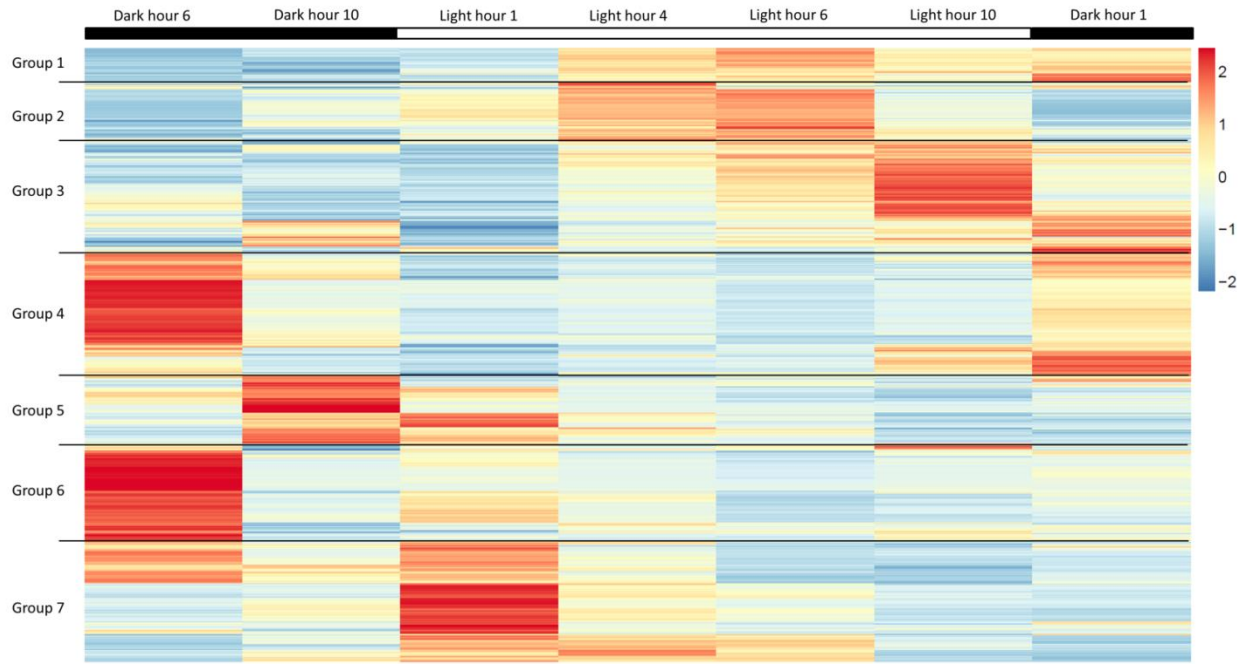


Figure 2.9: Heatmap of expression of genes that are highly expressed and have large differential expression over 7 time points. 1000 gene transcripts were analyzes in this heatmap selected by expression level and large variance between time points.

Group 1 includes genes that are highly expressed from L4 to D1. GO terms over represented include genes linked to post transcriptional regulation of gene expression, negative regulation of cellular processes, response to biotic stimuli and to other organisms, and defense responses.

Group 2 genes that are upregulated from L1 to L6 include the GO terms photosystem I and II, chloroplast thylakoid membrane, cell wall, defense response to fungus, response to heat, ATP binding.

Group 3 has significant over representation of ribosomal subunit gene expression. Upregulation of these genes occurs during the middle to the end of the day, with the majority of members showing maximum expression at hour 10 of the light cycle. Though ribosomal sequences represent < 0.5% of genes annotated, over 38% of genes identified in this group are structural constituents of ribosomes. It is established that ribosomal accumulation usually occurs during the G1 phase of the cell cycle, thereby fostering the production of new biomass, ultimately needed to support cell division.

Group 4, which represents genes that peak in expression at ~D6, show GO term enrichment associated with cell division processes, such as mitosis, cytoskeleton, and cell projection (microtubular and flagellar components) terms. This finding correlates with the timing of cell division that initiates at the end of the light period but is most robust in response during the dark (Figure 2.8). In addition, Group 4 shows highly overrepresented mitochondrial related transcript expression.

Group 5 displays genes that are highly expressed at the end of the dark cycle. An overrepresentation of several gene categories are noted including genes encoding proteins responsible for metal ion binding, as well as a variety of mitochondrial genes that encode both transporters and ion exchange machinery (e.g., ferredoxin and rubredoxin). Transcription during this time period also seems to be favoring the onset of energy molecule and carbohydrate synthesis, as three major polysaccharide pathway member transcripts (glyceraldehyde-3-phosphate, glycosyltranferase and triose-phosphate isomerase) are represented in this group. Nitrite and phosphate transporters are also seen in group 5, likely supporting the acquisition of those precursors needed for the production of both sugar and fatty acid products.

Group 6 genes that are highly expressed in the middle of the dark period include genes having associated GO terms: ribonucleoside bindings (i.e., purine ribonucleoside binding), GTPase activity, microtubule based movement, GTP binding, ATP binding, anchored to plasma membrane, and tubulin.

Group 7 gene expression is highest at the beginning of the light period. Thylakoid, chloroplast, photosynthesis, metabolic processes, glycolysis, and fatty acid biosynthetic process are prominent in the GO term list.

Light serves as a critical abiotic factor in the regulation of algal growth. Our unpublished studies show *Chrysochromulina* cultures grow best when subject to a light/dark photoperiod that has at least 8 hours of darkness. As seen from the information presented above, a circadian photoperiod

drives a wide range of gene expression rhythms. The predictability of these temporal programs provides several excellent metabolic targets for the genetic engineering and may especially be of interest to commercial growers who are dependent on seasonal light availability to serve upscale algal production efforts.

Genomes to pathways - lipid biosynthesis and expression:

Chrysochromulina tobin represents a new model organism for the study of lipid body organelles (lipid droplets) and may be a commercially viable producer of high value lipid products. Production of polyunsaturated fatty acid (PUFA) products are useful nutraceuticals (omega 3 and 6 fatty acids, EPA and DHA) and a source of energy (Hu et al., 2008; Radakovits et al., 2010). To assess lipid synthesis in this oleaginous alga, the genome was queried for the presence of known lipid metabolic pathway genes.

The following genes form the fatty acid biosynthesis pathways were found as adapted from Radakovits et al (2010) Figure 2.10. Transcriptomic data was also collected at 7 time points over a 24 hour photoperiod cycle starting from 6 hours into the night (D6) and continuing through the entire day until the next night (D1) consisting of a 19 hour time window. These data reveal strong regulation of fatty acid synthesis during the light hours and enzymatic degradation during the dark cycle which is to be expected as previous observations have shown lipid synthesis during the lights on periods and lipid metabolism during night hours. Surprisingly there were two major trends identified in the lipid synthesis pathway. First, all chloroplast localized processes whose pathway converts pyruvate to free fatty acids show a strong up-regulation of 3 to 500 fold increase in transcript reads (as measured by fragments per kilobase of exon per million mapped reads (FPKM)) from the end of the dark photoperiod (D10) through the earlier hours of the light photoperiod (L4) with transcripts peaking at L1. The genes identified in this group are the following: acetyl coenzyme a carboxylase (ACCase), 3-ketoacyl-acyl carrier protein synthase (KAS), 3-ketoacyl-acyl carrier protein reductase (KAR), 3-hydroxyacyl-acyl carrier protein (HD), and enoyl-acyl carrier protein reductase (ENR). It is interesting to note the

likelihood of circadian like signaling is high due to the up-regulation of these fatty acid synthesis genes before the presence of light actually occurs.

The second trend is associated with endoplasmic reticulum localized processes once free fatty acid has been transported out of the chloroplast. From hour 4 through hour 10 in the light cycle a gradual upregulation of genes converting Acyl-CoA to triacylglycerol (TAG) takes place. Triacylglyceol is a high value energy product. Genes in this pathway include glycerol 3 phosphate o-acyltransferase (GPAT), lysocardiolipin acyltransferase (LPAAT), lysophosphatidic acid phosphatase; phospholipid glycerol acyltransferase (LPAT), and diacylglycerol acyltransferase (DAGAT).

It is also observed that there is a significant difference between the overall transcript levels of the two groups described above. TAG formation genes (chloroplast) are seen as very high transcript abundance genes; with maximum FPKMs ranging from 70 to almost 1500. Compared to the ER related processes that have maximum FPKM transcript levels about 8 to 35. This differential suggests that the ER related processes may be the rate limiting steps in terms of TAG production potential and therefore potential candidates for genetic modification experiments with overexpression of those gene products. Additionally, it would be useful to compare similar transcriptomic timecourse experiments across multiple species of oil producing algae to see if this trend is consistent.

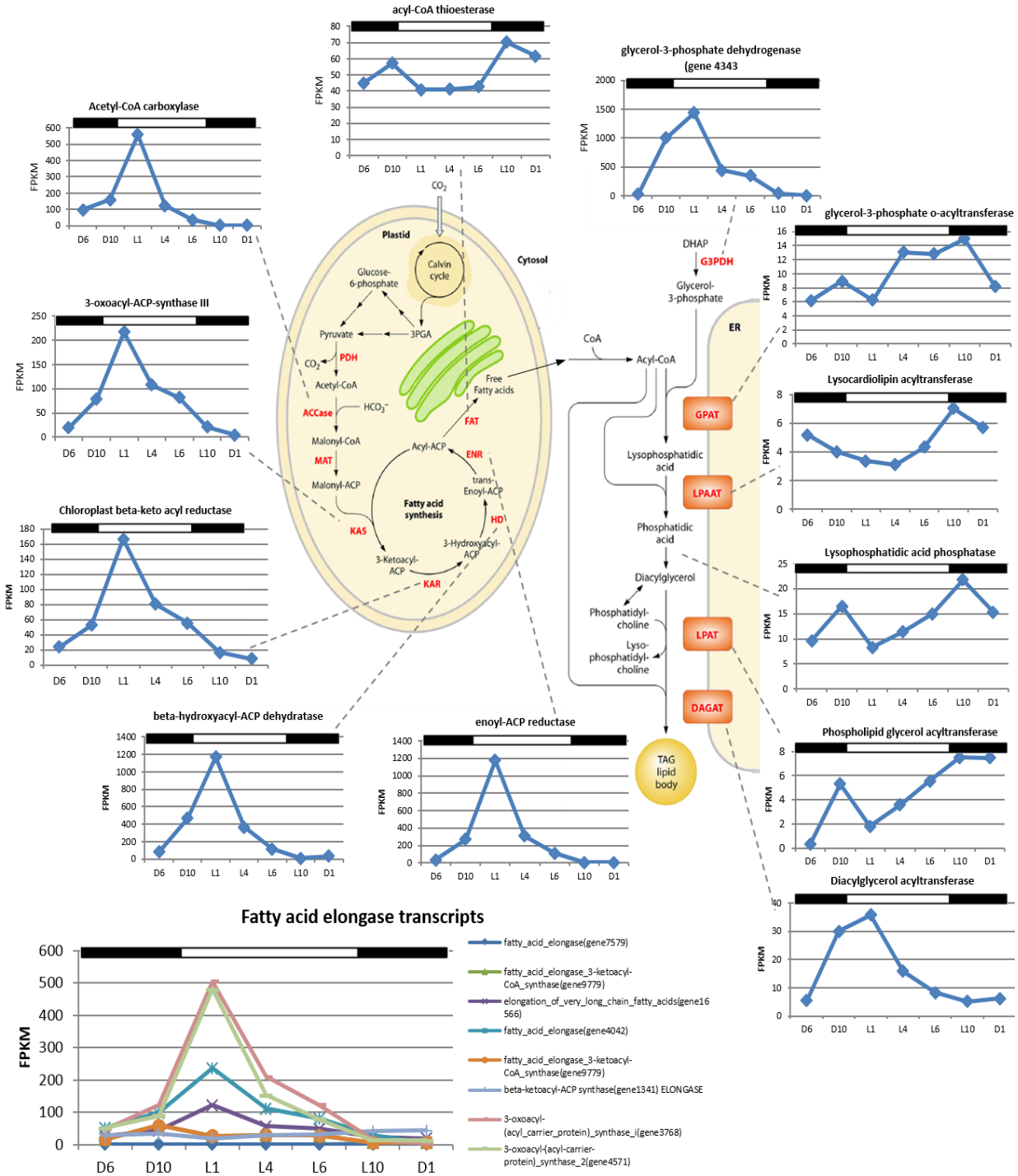


Figure 2.10: The identification of fatty acid synthesis genes (pathway adapted from Radakovits et. al. 2010) of *Chrysochromulina tobin* and corresponding RNA transcript FPKM over a 12:12 light dark photoperiod The genes identified as fatty acid synthesis genes represent a full complement. Gene expression from RNA is shown over 7 time points in the 24 hour light dark cycle. Additionally, elongase genes expression is also shown in the bottom left.

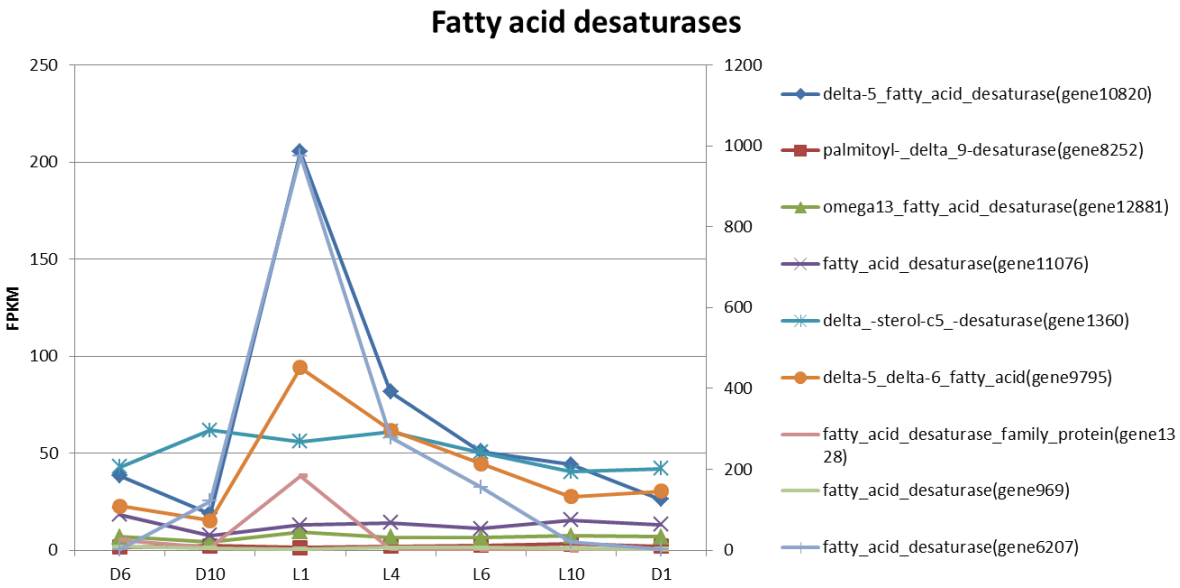


Figure 2.11: Fatty acid desaturase gene transcription level over a 24 hour light dark cycle.

Fatty acid elongases and desaturases also appear to follow the trend of the chloroplast based fatty acid biosynthetic pathway, in that expression is consistently peaking at the L1 time point (Figure 2.10 and 2.11). However, there is great variance between overall transcript levels for each gene queried. It would be of interest to query these gene transcripts across species. We have analyzed a variety of haptophytes using a high throughput GS/MS method for quantification and qualification (Bigelow et. al., 2011). The results of these haptophyte lipid measures (data not shown) would provide good candidates for further study given the ability to use *Chrysochromulina* genome to provide gene homology for de novo transcript sequencing.

Sexual cycle and recombination:

The 59 Mb *Chrysochromulina* genome is significantly smaller than genomes reported for the sequenced *Emiliania huxleyi* genome (140 Mb), and also less than the flow cytometric assessment of DNA complement in various haptophyte strains (ranging from ~117 Mb (*Phaeocystis antarctica*) to ~230 Mb (*Prymnesium polylepis*). Low occurrences of polymorphic loci and genome size concordance with propidium iodide staining using flow cytometric analysis (55 Mb) both strongly support the conclusion that the *Chrysochromulina* genome is haploid.

In order to perform genome engineering using the tools discussed in Chapter 1, homology dependent repair must be feasible. DNA repair mechanism genes relate directly to meiosis, as DNA recombination processes take place during meiosis. Though no properties of sexual reproduction have been observed using flow cytometric analysis of DNA content to date, we queried the genome assembly both for the presence of meiosis related genes and genes associated with gene repair mechanisms. Table 2.5 shows the results of each of these sets of gene queries. The fact that many DNA repair orthologs are found in *Chrysochromulina tobin* suggest that genetic modification methods that rely on homologous recombination are likely viable in this alga and genome editing is likely feasible in this organism. The complement of meiosis genes that are present in *C. tobin* also parallel those found in the other sequenced haptophyte *Emiliania huxleyi*. *E. huxleyi*, and other haptophytes, have been shown to transition between both between haploid and diploid life cycles. *E. huxleyi,* for instance, produces calcified scales while in the diploid life cycle (e.g., Houdan 2003; Edvardsen 1996). Because *E. huxleyi* is known to undergo meiosis, this strongly suggests that *C. tobin* also has the ability to undergo meiosis, though the process has not been observed. This information is useful if classical genetics approaches are to be used in this organism.

Table 2.5: Core genes associated with meiosis and DNA repair in eukaryotic cells

	Chrysochromulina tobin	*Emiliania huxleyi*	*Chondrus crispus*	*Ectocarpus siliculosus*	*Chlamydomonas reinhardtii*	*Volvox carteri*	*Arabidopsis thaliana*	Function
Core meiotic genes								
*spo*11	+ ($2e^{-33}$)	+ ($5e^{-18}$)	+	+	+	+	+	Transesterase
*hop*1	+ ($3e^{-21}$)	+ ($3e^{-7}$)	-	+	+	+	+	DNA DSB binding
*hop*2	-	-	+	-	-	+	+	Associated with MND1, homology searching
*mnd*1	+ ($2e^{-12}$)	+ ($1e^{-13}$)	+	+	+	+	+	DNA heteroduplex formation
*rec*8	-	-	+	-	-	-	+	Sister chromatid binding
*dmc*1	-	-	+	+	+	+	+	Inter-homolog recombination
*rad*51	+ ($1e^{-118}$)	+ ($6e^{-58}$)	+	+	+	+	+	Homologous DNA pairing
*msh*4	+ ($8e^{-22}$)	+ ($5e^{-40}$)	+	+	+	-	+	Holliday junction resolution w/ MSH5
*msh*5	+ ($6e^{-36}$)	+ ($6e^{-32}$)	+	+	+	+	+	Holliday junction resolution w/ MSH4
*mer*3	+ ($2e^{-122}$)	+ ($2e^{-109}$)	+	+	+	+	+	Holliday junction resolution
DNA repair and recombination genes								
*mre*11	+ ($2e^{-99}$)	+ ($2e^{-32}$)	+	+	+	+	+	dsDNA exonuclease/ssDNA endonuclease
*rad*50	+ ($2e^{-73}$)	+ ($2e^{-47}$)	+	+	+	+	+	DNA binding, holds broken DNA ends
*rad*1	+ ($6e^{-118}$)	+ ($1e^{-23}$)	+	+	+	+	+	5'-3' endonuclease for nucleotide excision repair
r*ad*52	-	+ ($4e^{-07}$)	-	+	-	-	+	DSB repair by homologous recombination
*msh*2	+ ($4e^{-157}$)	+ ($4e^{-65}$)	+	+	+	+	+	Binds base-base mismatches with MSH6
*msh*6	+ ($9e^{-161}$)	+ ($2e^{-59}$)	+	+	+	+	+	Binds base-base mismatches with MSH2
*mlh*1	+ ($4e^{-87}$)	+ ($1e^{-41}$)	+	+	+	+	+	Di- and tri-nucleotide mismatch repair
*mlh*2	+ ($7e^{-11}$)	+ ($2e^{-14}$)	+	+	-	+	+	Removal of cisplatin adducts
*mlh*3	+ ($2e^{-26}$)	+ ($2e^{-19}$)	-	+	+	+	+	Frameshift repair
*pms*1	+ ($8e^{-40}$)	+ ($4e^{-7}$)	+	+	-	-	+	DNA mismatch repair
*smc*1	+ ($2e^{-56}$)	+ ($2e^{-32}$)	+	+	+	+	+	Sister chromatid cohesion subunit w/ SMC3
*smc*2	+ ($3e^{-92}$)	+ ($5e^{-74}$)	+	+	+	+	+	Chromosome assembly and segregation
*smc*3	+ ($1e^{-59}$)	+ ($1e^{-14}$)	+	+	+	+	+	Sister chromatid cohesion subunit w/ SMC1
*smc*4	+ ($5e^{-69}$)	+ ($2e^{-66}$)	+	+	+	+	+	Chromosome assembly and segregation
*smc*5	+ ($1e^{-34}$)	+ ($4e^{-40}$)	+	+	+	+	+	DNA repair
*smc*6	+ ($7e^{-39}$)	+ ($6e^{-25}$)	+	+	+	+	+	Post replication DNA repair w/ SMC5
*rad*21	+ ($1e^{-17}$)	+ ($4e^{-20}$)	-	+	-	-	+	Sister chromatid binding
*scc*3	-	+ ($8e^{-30}$)	+	+	+	+	+	Sister chromatid binding
*pds*5	-	-	+	+	-	-	-	Sister chromatid binding in late prophase

Core meiotic gene and DNA repair and recombination gene survey in *Chrysochromulina* CCMP291$_{RAC}$ and other haptophytes and eukaryotes. Genes were selected based on the meiotic gene survey of *Trichomonas vaginalis* (Malik 2008). The "+/-" represents presence or absence of an orthologous gene on the basis of TBLASTN output of less than $1e^{-6}$ E-value. Numbers under the "+" represent the E-value of the orthologous hit.

Polyketide synthesis genes:

Polyketides are biologically active secondary metabolites that are synthesized from the simple fatty acids acetyl- or manoyl-CoA. A variety of biologically active compounds are made by polyketide synthesis (PKS) pathways ranging from antibiotics to toxins. PKS pathways are found to occur in both prokaryotic and eukaryotic organisms. We queried the genome for the presence of polyketide synthesis genes to identify *C. tobin* specific products such as toxins, which may be responsible for harmful algal blooms (HABs) associated with *Chrysochromulina* species. Similar to findings in the previously sequenced *E. huxleyi*, *Chrysochromulina tobin* encodes polyketide synthase genes. In *C. tobin,* I have identified examples of "Type I" PKSs (Staunton and Weissman, 2001)(Figure 2.12). Type I PKSs are large genes (from 5-25 kb) that are comprised of multiple "modules". Each module contains multiple protein active sites, each of which performs a polyketide chain modification. Type I modules generally contain, at the minimum, a beta-ketosynthase catalytic domain (KS), an acyl transferase domain (AT), and an acyl carrier domain (ACP) though the domains found in *C. tobin* PKSs are more complex.

In addition to polyketides, many organisms, including algae, also use nonribosomal peptide synthetases (NRPS) to produce non-ribosomal peptide products. These peptide products are synthesized independently of ribosome protein production pathways and encompass a variety of biological compounds such as antibiotics or toxic compounds. These NRPS genes are similar to that of PKS genes in that they produce a large protein with a multi-functional domain structure. Like an assembly line, a complex polypeptide product is synthesized and modified as it is passed along the NRPS catalytic domains. While we did not find standalone NRPS pathways, we did observe the presence of NRPS domains associated with PKS domains, allowing for potentially novel bioproduct production. As seen in Module 3 (Fig 2.12), a polypeptide adenylation (A) domain is present adjacent to an acyl carrier protein (ACP) arm and a PKS domain. Additional NRPS specific domains are found upstream of the (A) domain as well, including an NRPS condensation domain (CD) and the HxxPF repeat domain. Surprisingly amino transferase (AT) domains are found on both termini of Module 3. A similar AT chain is also found at the C terminus of Module 2. Though speculative, this arrangement suggests that the three Modules may

interact to create a single product. To our knowledge, this is the first time a hybrid PKS-NRPS has been described in an algal species. Hybrid PKS-NRPS pathways have been identified previously in bacteria, cyanobacteria as well as fungi (Fisch, 2013). Such observations are significant, given the extensive interest in identifying new therapeutic compounds that have been shown to be produced by either PKS or NRPS hybrid pathways. For example, fungal products produced by PKS-NRPS hybrid pathways include Fusarin C, a toxin which has been shown to be an estrogen agonist (Sondergaard et al., 2011) and carcinogenic (Gelderblom et al., 1986), as well as Pseurotin A (Maiya et al., 2007) a chitin synthase inhibitor.

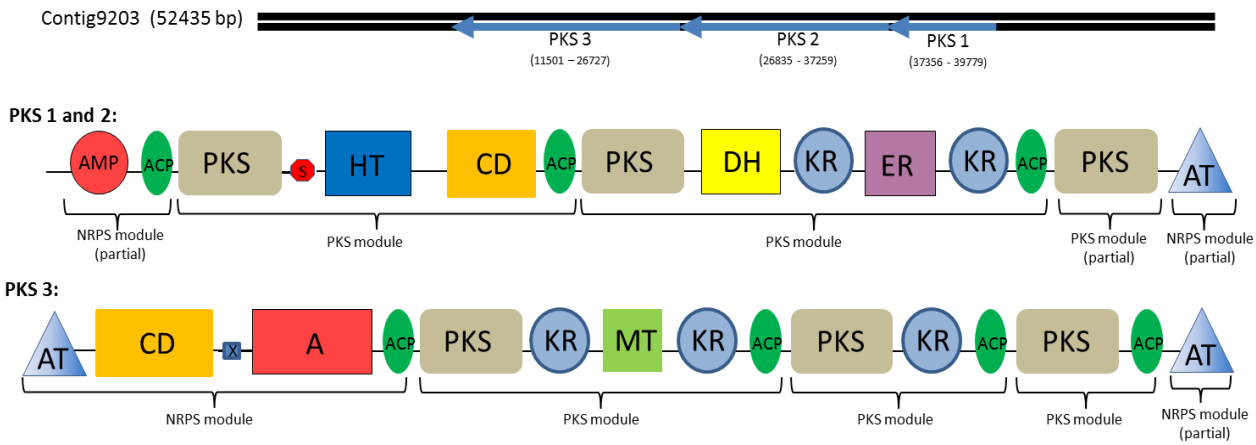


Figure 2.12: Type I Polyketide synthases and PKS-NRPS hybrid domains found in the *C. tobin* genome. These gene clusters each represent one of three Polyketide synthesis domains found on a single assembled contig of over 25000 base pairs. Domains 1 and 2 are separated by a single stop codon. Domain 3 is separated by a stop codon and is frame shifted. Brackets indicate the NRPS specific domains. Modules contain the following components: A: Adenylation domain; AMP: AMP binding domain; ACP: Acyl carrier proteins; AT: acyltransferase; CD: Condensation domain; DH: dehydratase; ER: enoyl reductase; HT: Crotonase/Enoyl-Coenzyme A (CoA) hydratase family; KS: ketosynthase; MT: Methyltransferase; S: Stop codon; X: HxxPF repeat domain

As seen in PKS 3, a polypeptide adenylation domain is located adjacent to an ACP arm and PKS domain. Additional NRPS domains are found upstream of the adenylation domain as well, including an NRPS specific condensation domain and HxxPF repeat domain. Surprisingly, both termini of PKS 3 are amino transferase (AT) domains, which facilitate the transfer of an amine to growing acyl chains (Aron et al., 2005). A similar AT chain is found at the C terminus of PKS 2 as well, which allows speculation that all three domains may interact to create a single product. Additional investigation of these

pathways is warranted and more careful searches of published are in progress. Algal genomes should also be queried to identify additional PKS-NRPS hybrids in other non-fungal eukaryotes.

Transformation systems:

Targeted genetic modification in *Chrysochromulina tobin* was the original objective for completing the genome. In order for this to occur, a working transformation system is required. To date, some successful methods of transformation have been performed in various algae to date (Radakovits). Unfortunately, a successful method of transforming *C. tobin* has not been shown by my research yet. This organism should be amenable to transformation, simply because *C. tobin* cells have no cell wall. I have compiled the transformation materials, methods and results attempted and obtained thus far in Appendix B of this dissertation.

Genome sequencing summary:

Here I report the sequencing and annotation of the complete *Chrysochromulina tobin* mitochondrial and chloroplast and nuclear genomes. *Chrysochromulina tobin* represents only the second haptophyte genome sequenced. *C. tobin* is of interest due to its potential as a model organism for lipid production and lipid body organellar regulation as these metabolic pathways are regulated by a 24 hour light/dark cycle. In addition, we have collected and analyzed transcriptomic data over the 24 hour cycle which reveals both cell cycle regulation and fatty acid metabolic pathways are indeed tightly linked to the diurnal cycle. Targets for future metabolic engineering work have also been identified. Data reported here show the mitochondrial genome to contain a large and complex repeat comprising 28% of the mitochondrial sequence, and to have lost several *nad* genes (*nad*7, 9 and 11). The *C. tobin* chloroplast genome contains a novel intergenic ribosomal spacer region, and multiple SNPs between rDNA copies within the inverted ribosomal repeat regions. Analyses of chloroplast tandem and inverted repeats demonstrate gene-specific associations, regardless of algal species. Features of several genes provide new insight into aspects of chloroplast genome evolution including

lateral gene transfer, gene retention, novel functional rolls and putative regulatory structures localized within intergenic regions.

Materials and Methods

Culture maintenance:

Chrysochromulina strain CCMP291, acquired from The National Center for Marine Algae (NCMA) by the Cattolico laboratory in 2006, was designated as P3. These cultures were maintained in 250 mL Erlenmeyer flasks containing 100 ml of RAC-1, a proprietary fresh water medium. Flasks were plugged with silicone sponge stoppers (Bellco Glass, Vineland, NJ) and capped with a sterilizer bag (Propper Manufacturing, Long Island City, NY). Large volume experimental cultures for genomic DNA harvesting were maintained in 1.0 L of RAC-1 medium contained in 2.8 L large-mouth Fernbach flasks. These flasks were plugged with hand-rolled, #50 cheese cloth-covered cotton stoppers and covered with a #2 size Kraft bag (Paper Mart, Orange, CA). All cultures were maintained at 20°C on a 12 hour light:12 hour dark photoperiod under 100 $\mu Em^{-2}s^{-1}$ light intensity using full spectrum T12 fluorescent light bulbs (Philips Electronics, Stamford, CT). No CO_2 was provided and cultures were not agitated.

Bacterized cultures were treated in the following manner to minimize bacterial contamination: P3 cultures were subject to iterative cell sorting using flow cytometry. *Chrysochromulina tobin* cells were stained for identification using BODIPY 505/515 (4,4-difluoro-1,3,5,7-tetramethyl-4-bora-3a,4a-diaza-s-indacene; Invitrogen, Carlsbad, CA), a neutral lipid binding fluorophore. Approximately 10 stained cells were sorted into a single well of a 96 well plate containing 100 µl RAC-1 medium. Due to poor growth in the 96 well plate, well contents were transferred to 10 ml of RAC-1 medium in 50 ml plastic tissue culture flasks (Nunc, Roskilde, Denmark). This cell sorting process was carried out 4 times with the resulting culture being designated as P4. Cells obtained from reiterative flow cytometric selection (P4) were then grown in RAC-1 medium that contained either streptomycin (resulting in culture P5.5) or hygromycin (P5.6). Treatment with these two antibiotics were identical: cells were exposed to a final concentration of 400 µg/ml antibiotic for 18 hours before 5 mL of treated cultures

were transferred to 100mL of antibiotic free RAC-1 medium. Cultures P5.5 and P5.6 were periodically tested for bacterial contamination using by monitoring the presence of bacterial growth in RAC-1 + LB liquid culture. Sequencing data and a cultured isolate has shown that one bacterial contaminant is still present in the antibiotic treated cultures (data not shown).

Genomic DNA isolation:

Total genomic DNA was collected from each of the P5.5 and P5.6 cultures using the Qiagen Genomic-tip Maxi DNA extraction protocol (Germantown, MD) with the following changes to the standard protocol (Cattolico et al., 2008): 1.5 x 10^8 cells were harvested by centrifugation (Beckman-Coulter JA-10 Rotor at 7000 rpm (5378 x g) for 20 minutes) and resuspended in lysis buffer (20 mM EDTA, pH 8.0; 10 mM Tris-base, pH 8.0; 1% Triton X; 500 mM Guanidine; 200 mM NaCl) with 1.0 hour incubation at 37°C. RNAse A was added to a final concentration of 200 µg/ml and incubated for 30 minutes at 37°C. Next, 600 µl of Proteinase K (20 mg/ml) (Sigma-Aldrich) was added to the tube and incubated at 50°C for 2.0 hours, mixing every 30 minutes by swirling. The Qiagen DNA binding tip (Maxi size) was equilibrated using the manufacturer's instructions. DNA preparation was transferred into the tip and allowed to pass using gravity at room temperature. The tip was washed twice using Qiagen buffer QC. 15 ml of Buffer QF (at 37°C) was added to the tip to elute the DNA. DNA was precipitated by the addition of 10.5 ml of 100% room temperature isopropanol followed by centrifugation (12,000 rpm (11,220 x g) for 20 min, 4°C using a JA-20 rotor). The pellet was washed in 4 ml of 4°C 70% ethanol and centrifuged again using the same conditions. The DNA pellet was air dried for 5 min and resuspended in warmed Qiagen buffer EB (50°C) and incubated at 50°C for 2.0 hours. DNA solution was quantitated using a spectrophotometer and subsequently transferred to 1.7 ml Eppendorf tubes and stored at -80°C.

Genome sequencing, assembly and annotation:

The *Chrysochromulina tobin* chloroplast and mitochondrial genomes were sequenced using a combination of Illumina (Bennett, 2004) and 454 sequencing technologies (Margulies et al., 2005). Two

shotgun libraries (2 X 100 and 1 x 150 base pair) were prepared using standard TruSeq protocols and sequenced from bulk *C. tobin* genomic DNA on an Illumina HiSeq2000 sequencer. Additional shotgun single-end and paired-end (10 kb insert) DNA libraries were prepared for sequencing on the 454 Titanium platform generating 1.2 million and 3.5 million reads, respectively. The 454 single-end data and the 454 paired end data (insert size 8180 +/- 1495 bp) were assembled together using Newbler, version 2.3 (release 091027_1459). The Illumina-generated sequences were assembled separately with VELVET, version 1.0.13 (Zerbino and Birney, 2008). The resulting consensus sequences from both the VELVET and Newbler assemblies were computationally shredded into 10 kb fragments and were re-assembled with reads from the 454 paired end library using parallel Phrap, version 1.080812 (High Performance Software, LLC). Based on homologous BLAST (Altschul et al., 1997) searches against other chloroplast and mitochondrial genomes, the mitochondrial genome was identified as a single contig of 25,263 bp with one gap and the chloroplast genome was comprised of two contigs that totaled a combined 101,192 bp. Most mis-assemblies in the contigs of the mitochondrial and chloroplast genomes were corrected using gapResolution (Cliff Han, personal communication, Los Alamos National Laboratory) or Dupfinisher (Han and Chain, 2006). However, due to the large ribosomal inverted repeat in the chloroplast, PCR amplification anchored by priming of unique regions flanking and within the repeat sections was used as sequence template to resolve the final circular representation of the chloroplast genome structure. Similarly, a large tandem repeat structure identified in the mitochondrial genome prevented automated closure of the remaining gap. De-convolution of this repeat was completed by PCR amplification and cloning of multiple products into the pGem T-easy vector (Promega, Madison, WI) followed by capillary sequencing. The presence of chloroplast genomes containing flip-flop recombined isoforms was queried using all combinations of single copy region primers (*petB*, *ycf60*, *psa* and *rpl21* primers). Only the expected primer pairs, *petB-ycf60* and *psa-rpl21* pairs, produced PCR products.

The final, fully assembled chloroplast and mitochondrial genomes were supported by > 500x average coverage from the combined sequencing platforms. Each assembled genome was verified by

aligning the original Illumina reads to the final draft using the Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009). Continuous coverage without gaps or missing reads was verified using Tablet alignment (Milne et al., 2013) which identified >10 single nucleotide mismatches (both SNPs and indels) in the chloroplast draft assembly, which were corrected in the final assembly.

Annotation of organellar genomes was accomplished by GLIMMER (Delcher et al., 1999) for initial gene calling. BLAST homology searches to CpBase (http://chloroplast.ocean.washington.edu), a curated chloroplast database housed by the University of Washington Department of Oceanography were used for final gene identification and recovery of small or missing genes that were overlooked by automated annotation. Manual examination of BLAST homology searches was performed for each protein coding gene to determine correct start codons and gene length. An identical approach was used to assemble the mitochondrial genome.

The 454 and Illumina data assembled with Newbler and Velvet produced over three thousand contigs. Gene annotation was carried out a combination of processed all compiled into a MAKER2 (Holt and Yandell, 2011) training and annotation pipeline. After masking repeated genomic elements using Repeatmasker (Smit et al., 2010), genes were modeled by combining several methods in MAKER2: 1) Aligning by *C. tobin* transcriptomic BLASTn hits as EST evidence. 2) Aligning *Emiliania huxleyi* EST using tBLASTx. 3) Using the assembled RNAseq data for gene prediction with Tophat and Cufflinks. 4) Aligning all CEGMA (Core eukaryotic genes) genes to the *C. tobin* contigs using BLASTx. 5) Augustus for *ab initio* models trained on the gene structures of *Chlamydomonas reinhardtii*. 6) SNAP for *ab initio* models trained on Hidden Markov Models (HMMs) of the genes called by cufflinks and Tophat models. 7) GenemarkES for *ab initio* gene models. A total of 16777 genes were annotated using the above method. Of these, 10293 were supported by BLAST homology using BLAST2GO and 6484 are considered novel gene calls.

BLAST2GO was used to attach functional annotation to gene call predictions. First, BLASTP was used searching the Non-redundant protein database (nr) with a Blast Expect Value cutoff of 1e-6. Blast2Go Mapping was performed followed by Annotation using E-Value-Hit-Filter: 1e-6, Annotation cutoff of 55 and GO weight of 5.

RNA Sequencing:

$1.5x10^8$ cells were collected per RNA isolation sample. *Chrysochromulina tobin* cells were centrifuged down at 7000 rpm in a Sorvall JA-17 rotor for 20 minutes. The supernatant was decanted and 5 ml of TRIZOL reagent (Invitrogen) was added to the cell pellet. Cells were resuspended by pipetting and vortexing for ~1 minute. The homogenate was transferred equally into four microcentrifuge tubes. To each tube, 250 µl of chloroform was added. Each tube was shaken by hand vigorously for 15 seconds and subsequently centrifuged for 15 minutes at 12,000 x g at 4°C. The mixing and centrifugation was repeated once. After the second centrifugation, the top aqueous phase was transferred to a new microcentrifuge tube being sure not to disturb the lower phenol/chloroform phase. 625 µl of ice cold isopropanol was added to each of the 4 new tubes containing the aqueous phase and incubated at -20°C overnight. The samples were then centrifuged at 12,000 x g for 10 minutes at 4°C. The supernatant was removed and washed with 1.25 mL of 75% ice cold Ethanol followed by 5 minute centrifugation at 7400 x g. The ethanol wash and centrifugation step was repeated one time. The pellet was dried for 10 minutes and resuspended in 30 µl RNAse free water (Qiagen). Four samples were combined into a single tube and treated with RNAse free DNAse for 90 minutes at 37°C. Samples were then cleaned using a Qiagen RNAeasy clean up protocol as specified by the manufacturer's instructions. Samples were stored at -80°C

Poly-A selection was carried out followed by library preparation using a TruSeq library kit Illumina). Sequencing was done on the Illumina high-seq and generated 100 bp paired reads. For each time point collected, 15-30 million reads were generated. Reads were trimmed and groomed using FASTQ trimmer (version 1.0.0) and FASTQ groomer (version 1.0.4) (Blankenberg et al., 2010). Tophat (version 1.5) (Trapnell et al., 2010) was used to assemble the sequences using the *Chrysochromulina*

draft genome as a reference. Cufflinks (v2.1.1) was used to estimate FPKM (fragments per kilobase of exon per million mapped reads) for each transcript at each time point.

Transcript heatmap generation:

To determine which subset of the transcriptomic data to include in the global analysis a high expression and high variance selection method was implemented to determine genes that were 1) highly expressed and 2) had great differences in expression between 2 or more time points. This was done using a formula we derived called "MeanNeighbor" which takes each time point and compares the expression level. FPKM values for each transcript were compared using the following function in R (R Core Team, 2014): MeanNeighbor = function(x) {mean(abs(x[2:length(x)]-x[1:length(x)-1)])). The top 1000 genes as scored by MeanNeighbor value were then plotted as a heatmap using a normalization constant so that all values of gene expression FPKM could be represented by a relative level between -3 and +3. First, each individual FPKM value was subtracted from each transcript's average FPKM value of all 7 time points. The resulting value at each time point was then divided by the standard deviation, giving a relative expression level to be used in the generation of the heatmap. The global heatmap was generated using the R library "pheatmap" using the "ward" clustering method. Fisher's exact test was used in Blast2GO (Conesa et al., 2005) to determine GO term overrepresentation in each group based on the annotation of GO terms by Blast2GO. The p-value cutoff for this was set at 0.05. For group 3, over 100 members were obtained so the p-value cutoff was lowered to $1e^{-7}$ to generate the graphs.

Comparative organellar genomic analyses:

Small repeat analysis: Small (<200 bp) inverted repeats were identified using Einverted from EMBOSS (Rice et al., 2000). Tandem repeats were identified using Tandem Repeats Finder (Benson, 1999). Small repeats located next to genes (*clp*C, *psa*B, *rpo*C, *atp*A, *rps*10, *rbc*S, and *psb*C) appearing to be conserved across chloroplast genomes in multiple species from manual inspection were quantified across all CASH taxa available in CpBase (Rocap and McKay). To refine the repeat list for the gene/repeat association analysis, the presence of tandem or inverted repeats adjacent to genes or

gene clusters was queried using the "Repeat Finder" tool in CpBase. Parameters used were: Search distance: 300 bp, End: "Both", Boundary: "Both". Additionally, no other feature (tRNA or protein coding gene) separating the gene and inverted repeat was counted in this analysis. Small inverted repeat physical structure was determined by inputting the sequence into M-fold (Zuker, 2003) using default parameters and a loop size maximum of 30 bp.

Large inverted repeat homology: The size of large inverted repeats which contain at least the ribosomal 16S-23S operon was determined by 2 sequence BLAST comparison of the two halves of a genome, each half containing one of the ribosomal repeats. BLASTN homology using the default settings was used to determine the borders of the repeat regions. To determine the sequence homology between two copies of 16S or 23S, BLASTN was used with default parameters. Each SNP and single nucleotide insertion or deletion was counted separately. If only one copy of the ribosomal operon was present, no homology search was performed.

Small repeat statistical analysis: A Fisher's Exact test was implemented to compare the proportion of chloroplast with more tandem repeats than inverted repeats in CASH taxa, rhodophytes, and the "green lineage". Linear regressions were used to test for an association between small repeats, genome size, and intergenic length with average repeat size as well as total small repeats with genome size and intergenic length. This analysis was repeated on the green lineage, CASH algae, and rhodophytes.

Phylogenetic analysis of *rpl*36:

By mining CpBase as well as NCBI, a total of 462 non-redundant RPL36 amino acid sequences were recovered for phylogenetic analysis. These sequences were aligned in MUSCLE (Edgar, 2004) and any C-terminal extensions were trimmed to create a 41 amino acid alignment with two gaps in the RPL36 C^- proteins such that functional motifs of the zinc finger domain (and their substituted amino acids in the C^- proteins) were aligned. Protein matrices available in the CIPRES Science Gateway (Moret

and Warnow) MrBayes 3.2.2 tool were evaluated for appropriateness using ProtTest 2.4 (Abascal et al., 2005). The cpREV +I + Γ model of protein sequence evolution was found to best suit the data. Gene trees were inferred with RAxML 7.6.3 (Stamatakis, 2006) with 1000 bootstraps, as well as with MrBayes v3.2.2 (Ronquist and Huelsenbeck, 2003) with two runs each of four chains, 10 million generations, and 25% burn-in. Stationarity and convergence of the Bayesian analysis were assessed with Tracer v1.5 (Drummond et al., 2012). To best represent the data, 85 select taxa were chosen for Figure 2.7 and the phylogeny re-inferred with the same parameters but only 5 million generations in the Bayesian analysis.

Chapter 3: Engineering mosquitoes for malaria prevention

Background

Malaria was estimated to kill approximately 627,000 people in 2010, with the majority of deaths in Africa (World Health Organization – Malaria fact sheet). This disease is generally transmitted by the mosquito (*Anopheles*). Historical efforts to eradicate malaria have been mixed in success and generally relied on widespread chemical treatments to eliminate mosquito populations. However, the environmental impacts of those approaches were extremely harsh. Much effort has been put into vaccine development, which would likely be highly effective in disease prevention, however vaccines with substantial efficacy have not been developed at this point. Alternative methods for mosquito control have been proposed including using a genetic drive element of homing endonucleases to effectively crash a population of mosquitoes to extinction.

Driving a species to extinction seems precarious in terms of the effects to ecosystems that contain the mosquito species to be eliminated. However, one must consider that only certain species of mosquito harbor the disease causing parasite *Plasmodium falciparum*. For this reason it is suggested that the only species targeted for this genetic drive depletion system would be that of *Anopheles gambiae*- the most deadly vector for malaria transmission (Windbichler et al., 2007). If population level elimination of this species is carried out, it is predicted that the numerous other mosquito populations would readily fill the holes in the ecosystem left by *A. gambiae*.

Homing endonuclease based genome targeting is an ideal genome editing tool for use for this purpose due to the small size and the requirement of intra-generation genetic transmission of this platform. However, the engineering challenge of altering a homing endonuclease targeting sequence remains. There is a growing list of homing endonucleases with different target site specificities to use in genetic engineering applications. These come from both an increased discovery of novel homing endonucleases (Jacoby et al., 2012) as well as the ability of novel hybrid homing endonucleases to be

engineered, generally with the fusion of two discrete halves of non-identical homing endonucleases (Chevalier et al., 2002). As the library of homing endonucleases increases, and knowledge on how to engineer these enzymes to recognize new DNA binding sites improves, engineering with these molecules in specific cases will improve.

Engineering challenges: target definition

In order to perform the desired outcome, control of *Anopheles* populations, one of the most important aspects is identification of an appropriate gene target that is cleavable by a homing endonuclease. The homing endonuclease enzyme must have both high enzymatic activity and high specificity to a target to reduce off target genome cleavage events. In order to determine which genomic loci are best possible candidates for a homing endonuclease target, a two front approach is required: 1) identify homing endonuclease DNA recognition sequences that have high enough enzymatic activity to cause the desired effect yet limit off target effects and 2) find a gene target that will provide the desired phenotype (i.e. sterilization or vector incompatibility).

Homing endonuclease target sites:

For this project, we used a monomerized version of the I-CreI homing endonuclease deemed mCre (Li et al., 2009) to investigate potential targets of interest in the Anopheles genome. The use of this enzyme is based on previous investigation of this molecule by our lab and others (Arnould et al., 2011; Ulge et al., 2011). Due to some underlying knowledge of both engineering and enzyme tolerance for DNA recognition single base pair change effects, we were able to generate best hits for potential target sites in the *Anopheles* genome.

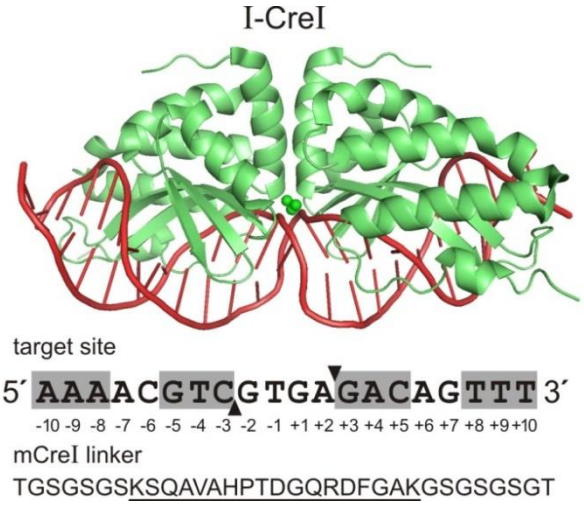


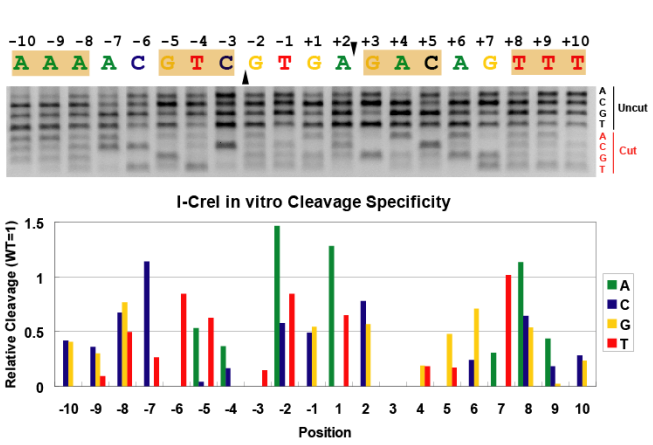
Figure 3.1: Crystal structure representation of I-CreI and the enzyme bound to DNA. The native 20 base pair recognition sequence is also shown as well as the amino acid sequence used to monomerize the mCre version of the enzyme.

Engineering Knowledge:

I-CreI has been characterized in a number of ways as described below. These properties and modifications of I-CreI can be utilized to increase the range of targets that can be accessed by I-CreI or its derivatives, such monomerized version of I-CreI, mCre.

Single base pair degeneracy:

Single base pair degeneracy data in I-CreI has been generated over time by a number of researchers (e.g. Li et al., 2012). This information was generated by testing the effects of DNA cleavage ability of a DNA molecule that had a single base pair change at a position across the 20 bp DNA binding domain mCre targets. Collecting this single base pair degeneracy data allows formation of a position specific scoring matrix (PSSM) for an enzyme (Figure 3.2). A first pass of target site identification is to simply identify target sites that 1) represent the fewest base pair changes between the DNA sequence of interest and the native 20 base pair homing endonuclease gene recognition sequence and 2) Identify the DNA sequences in the genomic target that contain the best tolerated base pair changes based on known degeneracy data of the homing endonuclease. Relatively quick methods have been developed using high throughput competition assays to determine the binding recognition sequence (Zhao and Stoddard, 2014).



	A	C	G	T
-10	1	0.36	0.28	0
-9	1	0.20	0.25	0.09
-8	1	0.58	0.56	0.30
-7	1	1.03	0	0.22
-6	0	1	0	1.05
-5	0.69	0.32	1	0.61
-4	0.31	0.16	0	1
-3	0	1	0	0.17
-2	1.21	0.67	1	0.68
-1	0	0.59	0.52	1
+1	1.20	0	1	0.58
+2	1	0.99	0.79	0
+3	0	0	1	0
+4	1	0	0.29	0.22
+5	0	1	0.28	0.21
+6	1	0.25	0.81	0
+7	0.40	0	1	0.96
+8	0.94	0.23	0.27	1
+9	0.44	0.19	0.02	1
+10	0	0.08	0.32	1

Figure 3.2: Degeneracy activity of single base pair changes in the native I-CreI DNA recognition sequences. A position specific scoring matrix (PSSM) can be generated by combining all collected data. Green highlights in panel two represent the native base pair sequence. Orange highlight positions represent base pair degeneracy that is highly tolerated (maintains 90% or greater enzymatic activity) by the native I-CreI enzyme.

Engineering Designs:

Also of use, is a small library of engineering designs that were previously generated (Ulge et al., 2011). This library contains a list of amino acid residues in the mCre enzyme that can be modified for a desired target sequence shift. For example, modification of residue 24 from I to K and residue 68 from R to T allows the modified enzyme to recognize a cytosine at the -5 position of DNA target sequence rather than the canonical guanine. Using this library of designs, we can also base target site searches to include sequences that are amenable to design solutions.

To combine the current knowledge base of homing endonuclease information, the Stoddard lab at the Fred Hutchinson Cancer Research Center developed and maintains an online server for homing endonuclease information on both degeneracy and engineering solutions (Taylor et al., 2012). This website is www.homingendonuclease.net – The LAGLIDADG homing endonuclease database and engineering server. By combining this information, user friendly search tools are available which use a DNA sequence input (i.e. gene or genome target of interest). The user options include which homing endonuclease enzyme knowledge bases you want to include in each search. Below, I describe additional

mCre specific knowledge obtained through experimentation which can be applied to gene target selections.

Larger numbers of base pair changes are not consistently additive in effect:

More base pair changes away from the native mCre recognition sequence are expected to generate less successful targeting via cleavage assay. Because our knowledge of combining many changes together is minimal, we wanted to combine collected data on combinations of degeneracy data.

We tested native mCre on 35 target sites which had 3 to 9 changes in total across the 20 bp canonical target site AAAACGTCGTGAGACAGTTT. This set of 35 sequences all had individual base pair changes corresponding from .2 to <1 relative cleavage activity as generated by *in vitro* cleavage profiling. Though we have data representing only single base pair change effects on the efficacy of DNA cleavage by the mCre molecule, we have not systematically analyzed the effects of multiple base pair changes and if the effects of these changes are additive.

From this set of 35 alternative mCre target sequences, we observed 31 targets (89%) which exhibited cleavage activity against the modified target site in the plasmid based in vitro cleavage assay. Of these 31, 11 target sites (31%) exhibited at or near mCre native target activity which corresponded to targets ranging from 3 – 7 individual changes across the target site. Adjacent target site substitutions, up to 3 in a row, each with individual relative cleavage activities >.5 were found to be well tolerated. Target sites with 4-5 contiguous substitutions had lower activity, particularly when one or more of the bases has individual relative cleavage activity <.5 of the native mCre target site position.

# of Substitutions in mCre Target	3	4	5	6	7	9
Average Score (0-3)	2	2.3	1.88	1.33	1.14	0

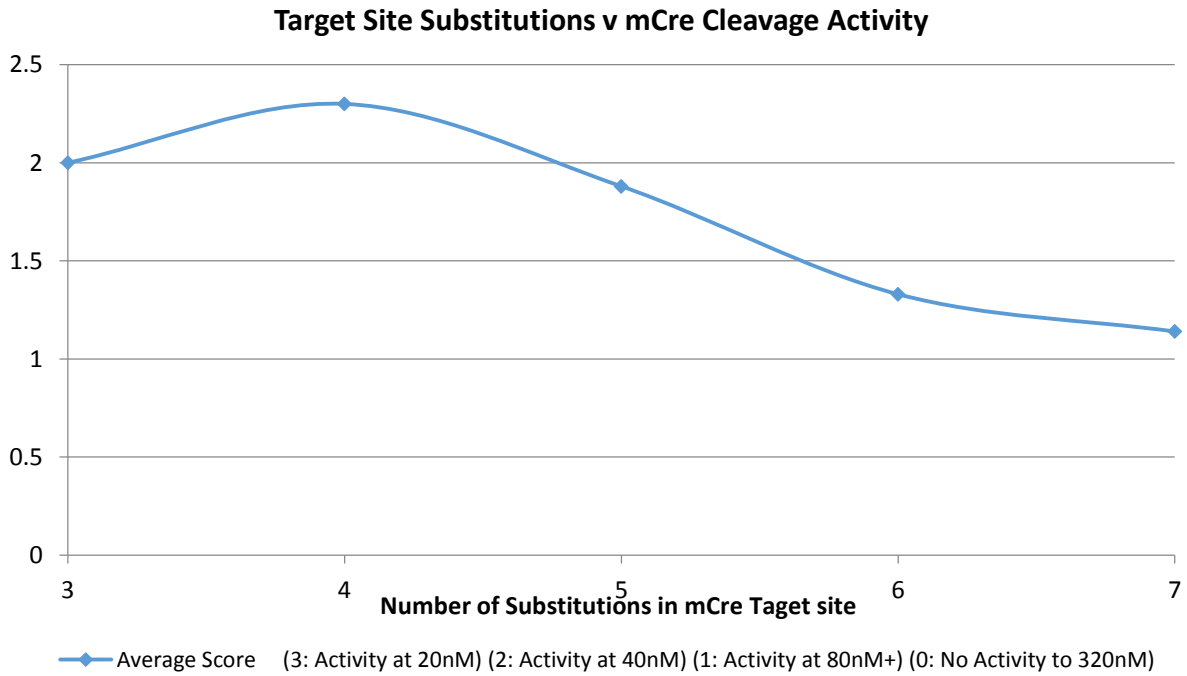


Figure 3.3: Average activity level of WT mCre based on number of degeneracy sites tested. A score of 3 represents WT mCre activity level on the native mCre target site.

Pairwise combination testing:

Previous observations show that some base pair degeneracy changes that exhibit low or no enzymatic activity are actually tolerated when in conjunction with additional base pair changes. This is not too surprising, as multiple base pair changes in proximity to one another may allow tolerance of additional changes based on the altered geometry and charge of the DNA-enzyme binding interface. Pairwise combinations that would be of high interest would include DNA target base pair changes that were predicted to have low or no DNA cleavage activity by mCre according to degeneracy data

available alone, but appear to have DNA cleavage activity when combined with other base pair changes.

In order to identify some of the more interesting base pair changes, I utilized two sets of data. The first dataset is the aforementioned single base pair degeneracy data. The second is a list of sequences that appear to be cleaved by mCre as determined by a randomized sequence selection experiment. This list of cleavage sensitive sequences identified by use of a neural network analysis of DNA cleavage sensitive sequences whose cleavage sensitivity differed from the prediction from single base pair additive data. From this list of the 2900 top scoring sequences that were predicted to have low or ablated cleavage activity according to the PSSM, all pair combinations were counted to observe which combinations gave “non-obvious” cleavage activity (Figure 3.3)

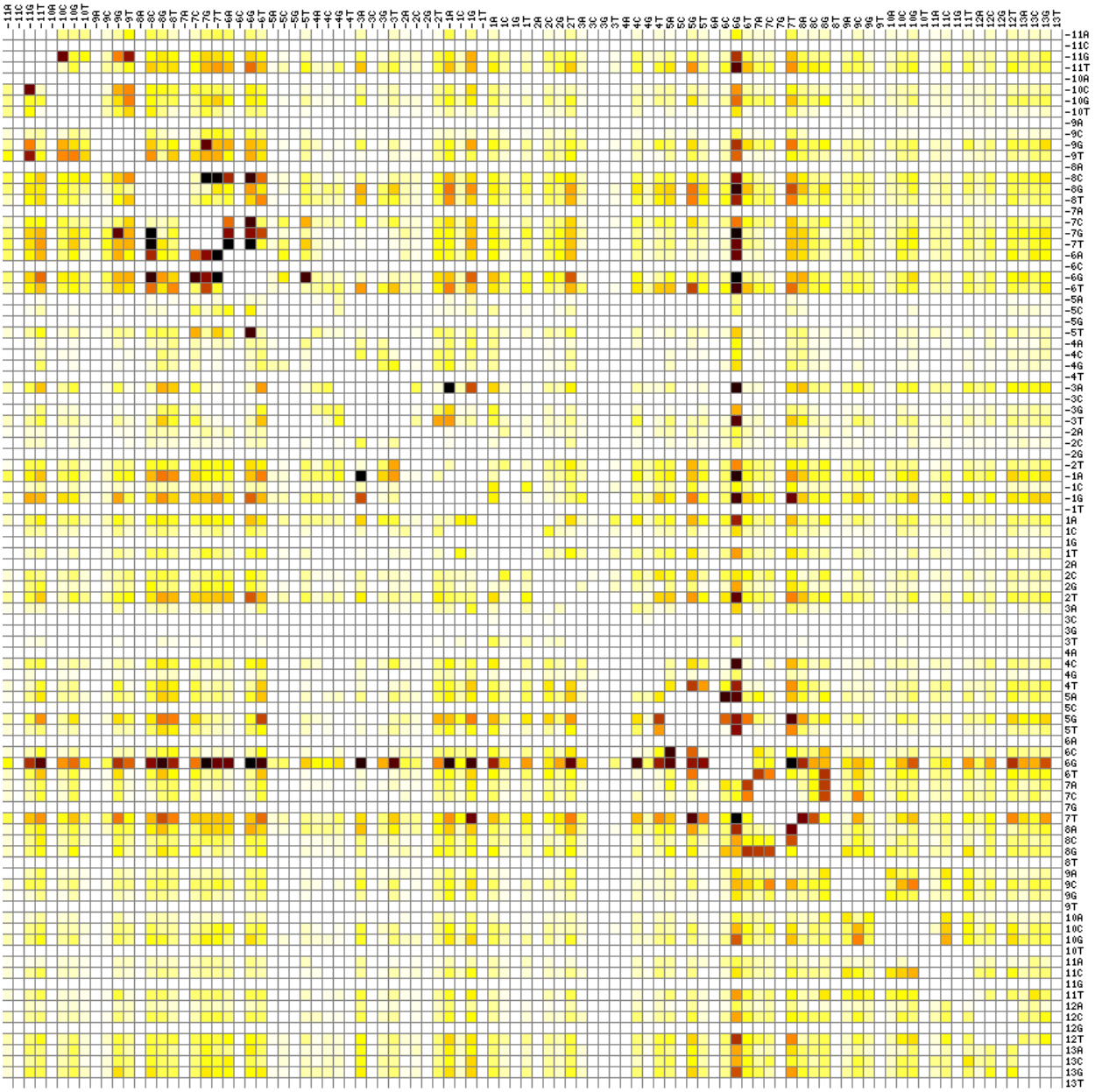


Figure 3.4: A heatmap of the most highly observed pairwise combinations from a set of 2900 cleavage sensitive DNA sequences that were cleavage sensitive based on experimental data but predicted to be cleavage resistant based off of single base pair degeneracy data. Rows and columns represent each of the possible base pair changes to the sequence.

One interpretation of these data is that positions which are the most permissive in this analysis, such as +6G positon, make the mCre enzyme more “relaxed” at this and a range of additional positions. Many of these permissive single base pair changes center on the ± 5 and 6 base pair positions in each I-CreI target half site.

Central 4 analysis:

The central 4 base pairs of mCre do not have direct DNA-amino acid contacts as shown in the derived crystal structure. In this case some changes to the canonical target site, particularly in a pairwise, triplet or higher combination fashion may be tolerated. Because there are no direct amino acid residue contacts in the central 4 base pairs of a target sequence, obvious enzyme engineering solutions are not directly available using modeling. To query all 256 possible central 4 bp combinations our randomized target site library data was again used to identify and rank target sites on the frequency with which each central 4 base pair sequence was cleaved. A rank order of all 256 central 4 base pair combinations was made on the basis of enrichment in the cleavage sensitive library (cleavage sensitive/cleavage resistant). Higher rank is an indicator that the central 4 bases are permissive to cleavage. This data can be used as an addition screen for sites that may have either a permissive or restrictive central 4 base pairs.

***Anopheles* target gene selection:**

Anopheles gene targets were developed by collaborators using the following criteria: DNA sequences in genes that excluding 5'UTRs, 3'UTRs, or intronic regions, up-weighting genes that have ovary-biased expression and up-weighting genes thought likely on the basis of biological information to give female sterility when targeted and inactivated in *Drosophila*. These criteria were scored via a "p(sterile)" value. A list of genes with the highest p(sterile) scores (approximately 300 genes) were then used to identify the subset for which a specific mCre variant could likely be generated.

Generation of *Anopheles* fertility gene-specific variants:

The HEG target sites for additional biological filtering were selected in the following manner: A search of all annotated *Anopheles* gene exonic regions was conducted to determine highest ranking candidate sites based on HEG engineerablity of the site as well as potential of causing sterility. The search was completed using a PSSM composed of single base pair degeneracy data, and single base pair protein design data. Both of these data sets were developed internally (Li et al., 2012; Ulge et al.,

2011). Each of the central 20 base pairs along the DNA/protein interface of a potential HEG target is scored. Low scores (lower scores are better) at each position are given to nucleotides which a) match the WT homing endonuclease recognition sequence, b) have high tolerance of substitution based on single bp degeneracy data, or c) have HEG protein designs for single base pair recognition changes. Higher scores are variable and given to sites which have low or no tolerance for degeneracy and no current protein design options.

The list of candidate sites was initially limited to a score of 120 and then sorted based on the p(sterile) value (probability of sterility). From this list, visual inspection of the remaining potential targets was performed. Sites were selected or eliminated based on the following criteria: a) minimal amount of contiguous target site base pair design changes (max of 1). b) minimal number of protein design changes adjacent to degenerate sites. c) "Central 4" base pair tolerance ranking determined from high through-put cleavage sensitivity and cleavage resistance data. d) Avoidance of more than one design change to the -3,-4,-5 or +3,+4,+5 target site sequence, based on previous design challenges encountered. e) Avoidance of targets with more than three contiguous non WT nucleotides either engineered or degenerate except in the case of the "central 4" bp region

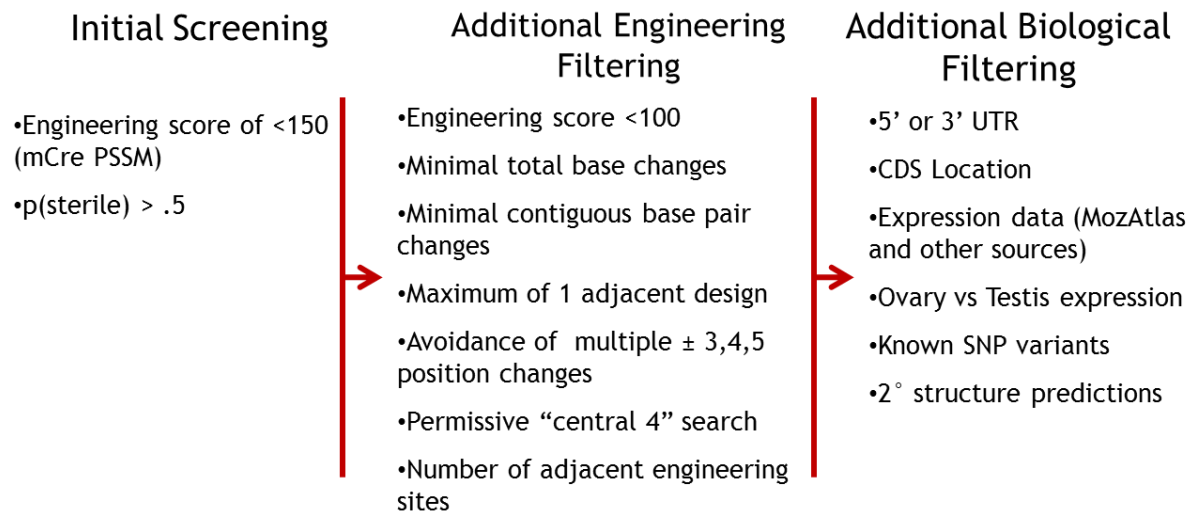


Figure 3.5: Gene candidate and targeting criteria used to identify genes that are both targetable and with potentially sterilizing effects.

Gene ID	p(sterile)	Target site	Central 4 (rank/Enrichment)
AGAP008417	0.51	AAcAgcTCGTtcGACAGaaT	11/1.8
AGAP009232	0.56	AAgtCGTCGTGcGtCAGaTg	3/2.56
AGAP011418	0.52	AAcACaaCcTGgGAgAGTaT	149/0.51
AGAP003087	0.56	AtAACcTCcaacGAtAGaTT	90/0.81
AGAP004667	0.59	AAAAtcTCaTtcGcCAtaTT	57/1.04
AGAP000642	0.72	AAcAgGTCGTagGcCAGcTg	51/1.07
AGAP004038v2	0.67	gAActGTCGcagGAgctgTT	48/1.09

Table 3.1: *Anopheles* gene targets identified from the gene candidate selection process outlined in figure 3.1. Uppercase black base pairs represent native sequences of the mCre binding site. Lowercase green bases represent predicted degeneracy base pairs well tolerated by mCre. Lowercase red base pairs are not well tolerated by mCre and require mCre engineering solutions.

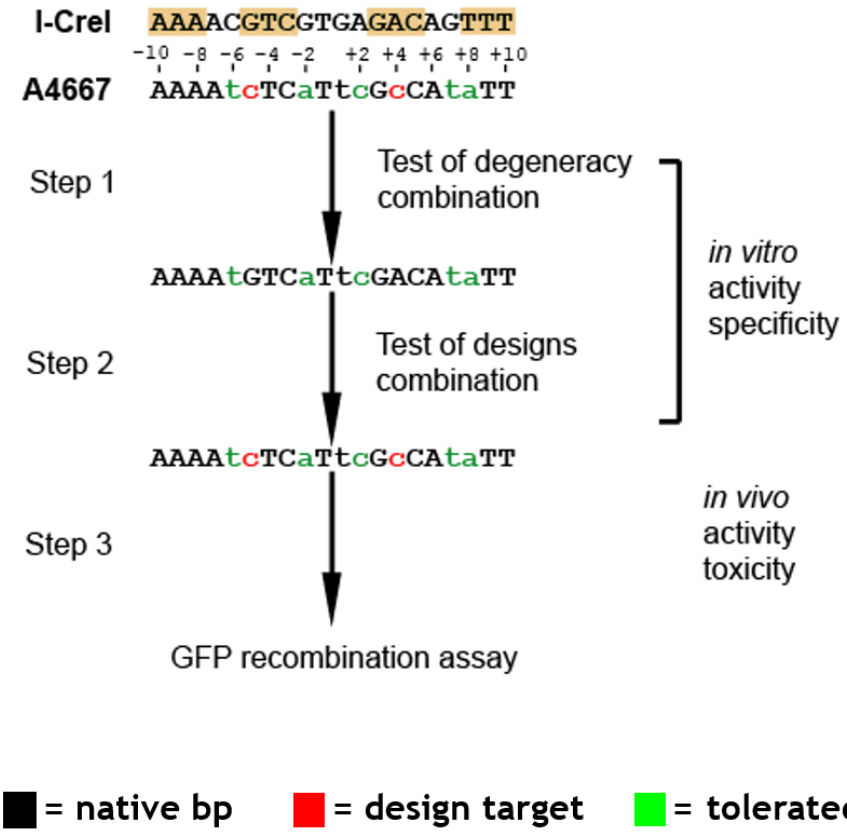


Figure 3.6: General methodology used to test potential mCre designs against *Anopheles gambiae* targets. Step 1 tests enzyme activity on DNA sequences where all degenerate base pairs are combined on a single target. If step 1 is passed with significant activity by the native mCre enzyme, additional engineering amino acid substitutions are made to the enzyme before expression. Black base pairs are

native to the mCre DNA recognition sequence. Red base pairs are bases that differ from the native mCre binding sequence, but have mCre engineering solutions that can be incorporated into the enzyme. Green base pairs also differ from the native mCre recognition sequence, but individual changes have been shown to be well tolerated by the mCre enzyme.

From the seven targets identified using this target selection methodology (Table 3.1), two targets were determined to be of sufficient activity. The methods for determining this activity are as follows (Summarized in Figure 3.6):

First, homing endonuclease gene targets are queried for tolerance of multiple degeneracy bases (shown in green). *In vitro* cleavage specificity profiling guides the generation of target site-specific mCre variants. The first step is to predict the cleavage sensitivity of target sites containing multiple base pair changes from single base pair cleavage sensitivity profiling data. These predictions can be experimentally verified in a second step, and then combined with mCre protein designs to generate a target site-specific mCre variant. A competitive cleavage assay is used to measure the effects of degeneracy or engineering changes for an HEG on a new target compared to the native target site (figure 3.7). Both native and novel target sites are cloned into a plasmid that is linearized prior to subjecting the DNA to a purified homing endonuclease protein. Cleaved products are visualized on an agarose gel to determine the relative cleavage efficiency of the native and test sites from relative band intensities. If this biochemical assay shows moderate activity and improved specificity toward the target site, we take the design to step three for *in vivo* characterization using the DR-GFP recombination assay in 293T cells.

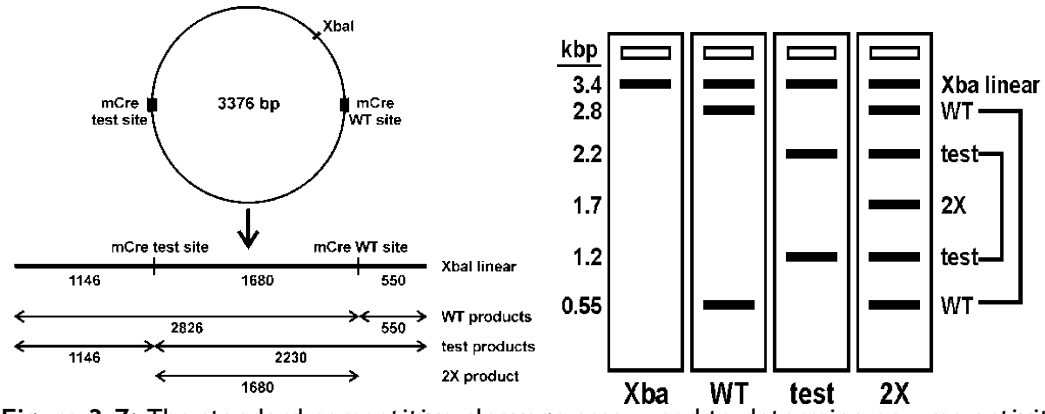


Figure 3.7: The standard competitive cleavage assay used to determine enzyme activity on a novel target sequence. Both the novel target site "test site" and the native mCre recognition sequence "WT site" are present in the test vector. Prior to testing, the vector is linearized using the XbaI restriction enzyme. Next purified mCre is added for a 1 hour digest. WT and novel recognition cleavage sites preference and specificity are tested in a single reaction.

Step 1: Test Combined Degeneracy:

Because more effort is required to incorporate design changes into a modified mCre enzyme, it is important to first determine the ability of the enzyme to tolerate the degeneracy changes in combination. If the combined degeneracy changes are not tolerated, additional engineering and selections would be required and can be suspended for later attempts. If all degeneracy sites are tolerated with comparable enzymatic activity to that of the wild type target sequence, then engineering solutions for the other non-degeneracy sites can be incorporated and tested against a full target site.

For the *Anopheles* gene targets shown in Figure 3.8, it is shown that the native mCre enzyme cleaves both native recognition sequences and the degenerate combined sequences with similar efficiency.

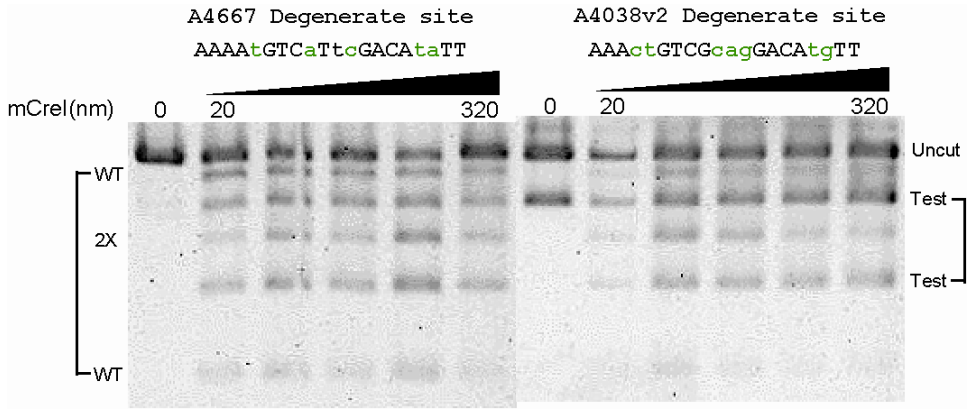


Figure 3.8: The result shows that when all predicted degeneracy sites are incorporated into a single target, the WT mCre enzyme can cleave the AGAP004667 and AGAP004038v2 target sites with comparable efficiency to the mCre WT target site.

Step 2: Test engineered mCre design combinations:

If combined degeneracy bases are tolerated by the native mCre enzyme, the next step is to incorporate engineered design changes into the mCre enzyme to see if combining all engineering changes is tolerated along with the degeneracy changes. The amino acid substitutions specific to each engineering solution are shown in Table 3.2

Gene ID	*Target site sequence*	*1st domain*	*2nd domain*
AGAP004667	**AAAAtcTCaTtcGcCAtaTT**	**-5C I24K/R68T**	**+4C Q44L/R70N/D75K**
AGAP004038v2	**gAActGTCGcagGAgctgTT**	**-10G N30R/S32R/Y33H**	**+5G I24K/R68T +6C Q26T/I77R**

key: ■ = native bp ■ = design target ■ = tolerated

Table 3.2: Combined mCre engineering solutions for two *Anopheles* gene target sequences. These now represent complete genomic target sites.

If the complete targets are tolerated, then additional *in vivo* testing is performed to see how the enzyme behaves in a cellular environment rather than an *in vitro* environment

The *in vitro* cleavage results (below) showed that WT mCre can cleave both A4667 and A4038v2 target sites, albeit with an efficiency about 20 fold lower than the mCre native target sites. However,

both A4667 and A4038v2 mCre variants exhibited a complete specificity shift towards their intended target sites with appreciable cleavage activity.

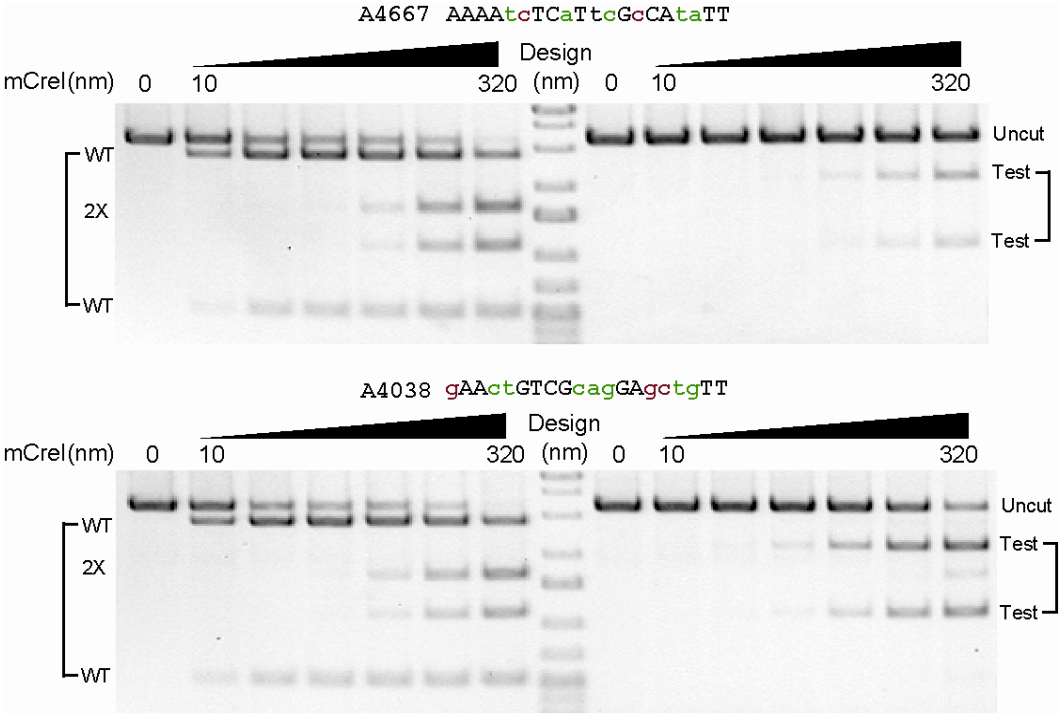


Figure 3.9: For each target site, both the WT mCre enzyme (left panels) and mCre enzymes with incorporated engineering design changes from table 3.3 (right panels) are tested against the native mCre binding sequence and the desired *Anopheles* genome target sequences.

Unfortunately, the target site cleavage activity was not as strong as expected. However, the enzyme designs did seem to be very specific and shift strongly away from the native mCre recognition sequence. Because the results were moderately promising, *in vivo* testing was carried out to determine enzyme activity in a cellular environment.

Step 3: Test *in vivo* activity (DR-GFP recombination assay)

The *in vivo* activity of A4667 and A4038v2 were determined using pDR-GFP recombination assay in 293T cells. The results indicated that A4667 and A4038v2 showed *in vivo* activity comparable to WT mCre, and about 10% *in vivo* activity of I-SceI (Table 1, column [d]).

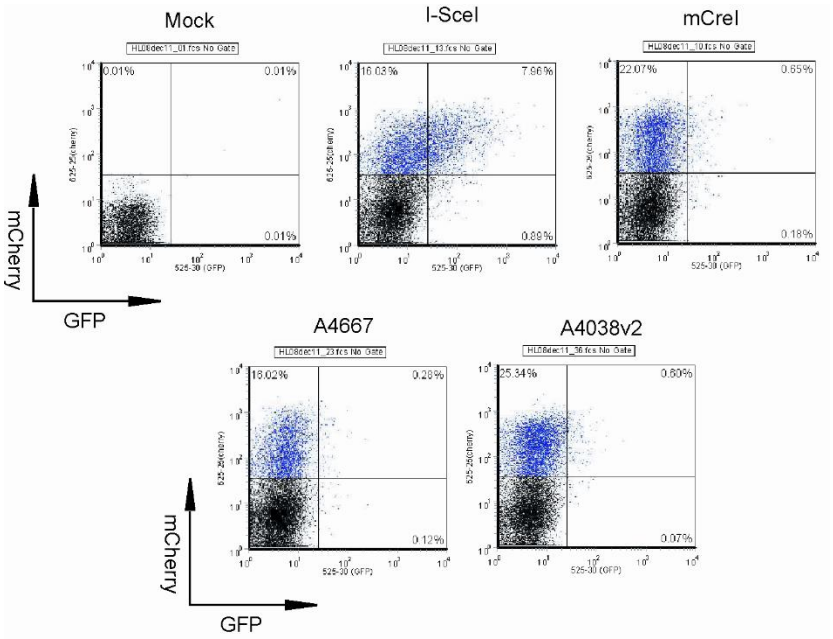


Figure 3.10: pDR-GFP recombination flow cytometric results

Enzyme	No gate			mCherry+
	Fold increase of GFP+ cells[a]	HEdependent GFP+ cells (%)[b]	GFP+mCherry+ (%)[c]	GFP+mCherry+ (%)[d]
I-SceI	3.01	74.53	0.46	2.76
mCre	8.38	89.29	7.24	32.44
A4667	7.40	81.70	0.48	2.16
A4038v2	6.87	86.53	0.53	4.07

[a] The ratio of UR (GFP+mCherry+) cells over LR (GFP+ only) cells.
[b] The percentage of HE activity dependent GFP+ cells.
[c] The ratio of UR (GFP+mCherry+) cells over whole population.
[d] The ratio of UR (GFP+mCherry+) cells over sum of UR and UL (mCherry+) cells, representing *in vivo* activity normalized against protein expression.
Table 3.3: Quantification of GFP recombination detected by each engineered mCre enzyme in comparison to I-SceI and mCre WT.

As with all genome engineering targets, there is a risk of genetic variation that could greatly reduce the enzymatic activity ability of a homing endonuclease. Because of this, we reviewed the presence of previously identified SNPs in *Anopheles* genomes. The AGAP004667 gene target does not have any known SNPs within the target site. AGAP004038v2 has a known SNP located at the -4 position on the mCre target site changing from a T(1) to A(0.31) at that position. It the frequency of this SNP is unknown.

Gene ID	p(sterile)	Target Site	Degeneracy activity	*In vitro* activity	Specificity Shift	*In vivo* activity
AGAP004667	0.59	**AAAAtcTCaTtcGcCAtaTT**	+++	Yes, 80nM	Yes	Yes
AGAP004038v2	0.67	**gAActGTCGcagGAgctgTT**	+++	Yes, 40nM	Yes	Yes

Table 3.4: Two potential *Anopheles* target mCre targets with results at each step of the testing pipeline.

Continuing directions and conclusions

The methodology described above represents an efficient way to identify engineerable target sites in any gene of interest. While the gene designs obtained for this particular project were successfully shifted away from the native mCre recognition site sequence, enzymatic activity of the enzyme leaves much to be desired as the protein designs had lower activity than native mCre enzyme. In this case, additional selections via random mutagenesis or yeast display methods (Baxter et al., 2013) would be ideal in an attempt to increase activity. In our observations, selection for increased activity often causes a general decrease in enzyme specificity. This outcome should be avoided though in order to reduce the occurrence of undesirable off target cleavage effects in most applications.

Materials and Methods

Anopheles target sites for I-CreI were determined using a Position-Specific Scoring Matrix (PSSM) which was derived from single base degeneracy data and single base pair protein design data. The PSSM was used in a search of the *Anopheles* gambiae genome AgamP3.5 (Mongin et al., 2004).

High scoring targets were synthesized by annealing complementary single stranded oligonucleotides (IDT, San Diego, CA). Target sites consisted of the 20 bp homing endonuclease target site with an additional 15 bp of the upstream and downstream *Anopheles* gambiae DNA sequence. The target site was integrated into a pCcdB plasmid using NheI and SacII restriction sites. The target site plasmid was transformed into XL-1 Blue competent cells using the manufactures protocol (Agilent

Technologies, Santa Clara, CA). Target sites were verified by Sanger sequencing (Genewiz, La Jolla, CA)

The bacterial protein expression plasmid used was pET15b and was obtained from Novagen (Gibbstown, NJ, USA). The *Escherichia coli* protein expression host strain C2566 was obtained from New England Biolabs (Ipswich, MA, USA). DNA oligonucleotides (50-nmol scale, salt-free) were synthesized by Operon (Huntsville, AL, USA). Qiaquick PCR purification kits and Ni-NTA HisSorb plates were obtained from Qiagen (Valencia, CA, USA). Other reagents, including restriction enzymes, Taq DNA polymerase and T4 DNA ligase were obtained from New England Biolabs (Ipswich, MA, USA) or Sigma-Aldrich (St Louis, MO, USA).

Protein expression and purification:

Homing endonuclease proteins were expressed overnight and purified as previously described by nickel affinity chromatography as described by (Ulge et. al., 2009). Briefly: Purification of mCre after 24 hour incubation at 18° C was accomplished over a 1 mL HisTrap Ni protein purification column (GE Healthcare) using an AKTAExpress protein purification machine (GE Healthcare), according to the manufacturer's instruction. The bacteria pelleted, resuspended in binding buffer (300 mM NaCl, 50 mM NaPO3 pH 8.0, 20 mM imidazole, 2 mM PMSF, 2.5 mM benzamidine), and lysed with lysozyme and sonicated. The supernatant was filtered and applied to the HisTrap column (GE Healthcare). The column was thoroughly washed with 20 mL wash buffer (300 mM NaCl, 50 mM NaPO3 pH 8.0, 20 mM imidazole) and then eluted using a 19 mL non-linear gradient of wash and elution buffer (elution buffer contained 300 mM NaCl, 50 mM NaPO3 pH 8.0, 500 mM imidazole): first, a 2 mL linear gradient ranged from 0% to 30% elution buffer, second, a 15 mL linear gradient ranged from 30% to 60% elution buffer, and finally a 2mL gradient ranged from 60% to 100% elution buffer. 1 mL fractions were collected throughout. Elution of mCre was confirmed by the presence of a 40 KD band after polyacrylamide gel electrophoresis of 10 mL of eluate. Multiple fractions containing mCre were pooled and concentrated using a 10 KD protein concentrator (Millipore) to ~100 mL. The process was repeated twice to dilute

out the imidazole in protein buffer (300 mM NaCl, 50 mM NaPO3 pH 8.0, 5% glycerol) and re-concentrated to ~100 mL. The volume of this sample was doubled in glycerol, and the final mCre sample containing 150 mM NaCl, 25 mM NaPO3 pH 8.0, 52% glycerol was stored at -20° C. The concentration of mCre in this sample was determined by standard Bradford assay (Bradford, 1976).

Proteins for *in vitro* binding assays were expressed from pET15b with an N-terminal 6 × His affinity purification/binding tag. Proteins used for *in vitro* cleavage assays were expressed and purified from pET24d without affinity tags.

HE proteins for *in vitro* cleavage assays were expressed in *E. coli* host strain C2566 from pET24d and purified as previously described in (Li et. al., 2011). *In vitro* cleavage assays were conducted in 20 mM Tris pH 8.0, 10 mM MgCl2 with 1:1 protein/DNA ratio under conditions where ~50% of the wild-type target site was cleaved; corresponding to 15 min at 37°C for I-CreI/mCre. Digests were stopped by adding loading buffer containing 0.1% SDS to samples, and the ladders of substrate and product fragments from each digest were separated by agarose gel electrophoresis.

***In vitro* competitive cleavage assay:**

The plasmid substrate for *in vitro* competitive cleavage assays were constructed by cloning both a native and a test target site into pCcdB (Doyon et al., 2006) at two different locations: the native target site into AflIII/BglII-cleaved plasmid DNA, and the test target site into NheI/SacII-cleaved plasmid DNA. In order to perform competitive cleavage assays, plasmid substrates were linearized by XbaI digestion, and 100 ng of linear plasmid substrate was incubated with LHEs in the presence of 20 mM Tris pH 8.0, 100 mM NaCl, 10 mM MgCl2 for 1 h at 37°C. Cleavage reactions were quenched by adding 10 mM EDTA and 1% SDS followed by heating for 10 min at 60°C. Plasmid substrate and cleavage products were separated by agarose gel electrophoresis, visualized by staining with ethidium bromide and photographed for quantitation.

Chapter 4: New safe harbors for targeted gene insertion in human cells

Background

Genomic 'safe harbor sites' (SHSs) are defined genomic locations where genes or other genetic elements can be safely inserted for purposes of human disease therapy, or for the study of gene structure, function or regulation. The most widely used human safe harbor sites have been identified by several strategies: serendipity (e.g., the AAVS1 adeno-associated virus insertion site on chromosome 19 (DeKelver et al., 2010; Mali et al., 2013); homology with known SHSs useful sites in other species (e.g., the human homolog of the murine Rosa26 locus (Irion et al., 2007); and most recently by recognition of the dispensability of a subset of human genes (e.g., the CCR5 chemokine receptor gene, that when deleted confers resistance to HIV infection)(Li et al., 2013; Lombardo et al., 2007; MacArthur et al., 2012).

In order to more systematically identify potentially useful new human SHSs, we located 20 bp target sites in the human genome predicted to be efficiently cleaved by the well-characterized I-CreI homing endonuclease protein or its single-chain form, monomerized mCre (Heath et al., 1997; Jurica et al., 1998; Li et al., 2009). We identified 35 potential new SHSs based on the presence of a high quality I-CreI/mCre-target site match, then further evaluated each of these SHSs on the basis of additional criteria. Three of these newly identified SHSs were subjected to additional experimental analyses including mCre homing endonuclease, SHS-specific TAL effector and CRISPR/Cas9 nuclease-mediated error-prone repair or the targeted insertion of a *loxP* recombinase site. These newly defined human chromosomal SHSs should be immediately useful, as they can be targeted individually or in multiplexed fashion by a common genome engineering nuclease (mCre) together with SHS-specific repair templates. Our strategy to identify and further validate human genomic SHSs should be useful for identifying and evaluating additional SHSs in the human genome or in other organisms that can be targeted by different site-specific genome engineering reagents including other homing/meganuclease, CRISPR/Cas9, TAL effector and zinc finger nucleases.

Materials and Methods

Identification of human genomic I-CreI cleavage sites:

In order to identify potential new human SHSs, we first searched the human genome for high quality matches to the target site of the well-characterized homing endonuclease protein I-CreI and its monomerized version mCre. We used detailed information on the cleavage specificity of I-CreI/mCre to construct a target site-specific position weight matrix (PWM) that quantified the contribution of each base pair across 20 target site base pair positions to target site cleavage activity (Figure 4.1). This PWM was used to identify 128 target site sequences predicted to be cleaved with ≥90% of the cleavage efficiency of the native I-CreI site (Li et al., 2012; Ulge et al., 2011). These 128 target site sequences were FASTA-formatted, uploaded into the NCBI BLAST search engine (http://blast.ncbi.nlm.nih.gov/), then searched against the human genome sequence (hg 19, build 37) using the following BLAST parameters: Optimize for 'Highly similar sequences (megablast)'; Max target seq = 50; Short queries: 'Adjust for short sequences'; Expect threshold = 1; Word size = 7; Match/mismatch: 4, −5; and Gap cost: Existence= 12/Extension= 8. All hits of ≥95% identity (19/20 or 20/20 bp matches) were subsequently evaluated as potential new safe harbor sites.

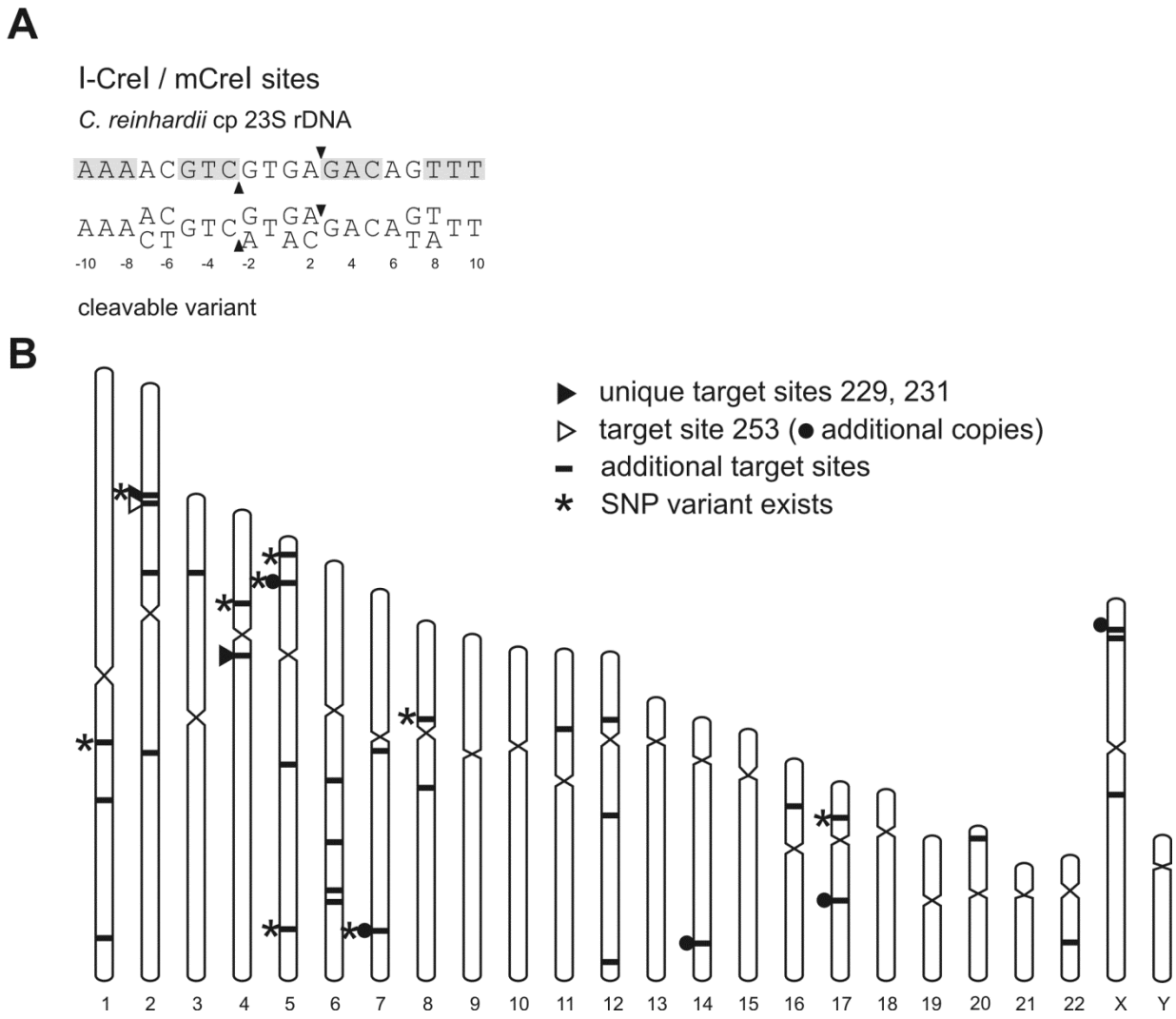


Figure 4.1: Identification of putative I-CreI safe harbor sites in the human genome. (A) The top row shows the wild type I-CreI/mCre 20 bp target site sequence. Palindromic base pair positions are shaded. The bottom sequence indicates possible base variants at specific locations that permit wildtype levels of cleavage, based on single base pair site scanning data summarized in the position weight matrix shown in Figure 3.2. A total of 128 target site variants incorporating these data was generated and used to BLAST search the human genome to identify 27 target sites with 19 or 20/20 bp identity to our target query library. (B) Target sites identified by BLAST search were present in 35 locations within the genome. Target site locations are marked to indicate their chromosome arm locations and positions. The three SHSs assayed in detail (SHS229, 231 and 253) are indicated by the triangles with the location of additional copies of the SHS253 target site by filled circles. Asterisks indicate SHSs in which SNP variants in the CreI target site sequence were identified in 1000 Genomes Project data (see text for additional detail).

Safe Harbor Site scoring:

Potential new human SHSs identified on the basis of a high quality I-CreI/mCre (hereinafter referred to as CreI) target site matches were further evaluated using 8 safety, functional and accessibility criteria in addition to site uniqueness (Table 4.1). These criteria included proximity to genes or miRNAs, especially those involved in cancer; genomic criteria from ENCODE and related data resources on chromatin state; and the mapping of regulatory or functional activity to within ≤300 kb of each potential SHS. SHS criteria were assessed by centering the CreI target site at each SHS chromosomal location, and using the target site to anchor the search of 300kb up- and downstream in UCSC genome browser track data to identify and assess target site proximity to genes, miRNAs, cancer- or miRNA-cancer relatedness; proximity to transcriptionally active regions, replication origins or ultra-conserved elements; location in open chromatin as assessed by nuclease sensitivity; and whether a given SHS was located in a region of copy number variation (CNV) (http://genome.ucsc.edu/). We also used the same criteria to assess the three most widely used human SHSs (AAVS1, hROSA26 and CCR5)(Mukherjee and Thrasher, 2013; Papapetrou et al., 2011).

We also determined whether there was base pair-level population genetic variation in the CreI target at each SHS. In order to search for SHS base pair level variation we used data from the 1000 Genomes Project (1KGP) data archive (http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/), which contains genome sequence data from 1092 individuals. For the subset of sites where we identified base pair variation, we determined the location and frequency of specific variants in 1KGP data within the CreI target site and used an I-CreI/mCre PWM developed from single base pair binding and cleavage scanning to predict to what extent a given base pair variant would alter cleavage efficiency.

	SHS criterion	UCSC browser track source
safety	1. > 300kb from any cancer-related gene on allOnco list	genes and gene predictions: UCSC Genes
	2. > 300kb from any miRNA/other functional small RNA	genes and gene predictions: sno/miRNA
	3. > 50kb from any 5' gene end	genes and gene predictions: RefSeq Genes
functional silence	4. > 50kb away from any replication origin	regulation: UW Repli-seq: Peaks
	5. > 50kb away from any ultraconserved element	regulation: VISTA Enhancers
	6. low transcriptional activity (no mRNA ± 25kb)	mRNA and EST: Human mRNAs
consistent/ accessible/ unique	7. not in copy number variable region	repeats: Segmental Dups
	8. in open chromatin (DHS signal ± 1kb)	regulation: ENC DNase/FAIRE: Uniform DNaseI HS
	unique (1 copy in human genome)	BLAST search output

Table 4.1: Criteria for Human Genomic Safe Harbor Sites. Table outlining the methodology of determining safety, function and accessibility of potential safe harbor sites. UCSC genome browser tracks or other sources used to assess a particular criterion are listed next to each criterion. Table 1 footnotes: SHS, safe harbor site; allOnco, a list of cancer-related genes compiled by the Bushman lab from 8 separate sources comprising a non-redundant list of 2,125 genes (see: http://www.bushmanlab.org/links/genelists); DHS, DNaseI hypersensitive site.

Experimental validation of new human SHS:

Potential new SHSs identified by BLAST search were experimentally validated by PCR amplification, DNA sequencing and mCre *in vitro* cleavage sensitivity analyses. Sites were PCR-amplified using primer pairs designed to amplify a ~300-400 bp region using the CLC Workbench Primer Design Tool [http://www.clcbio.com] (Supplementary Table 1). Genomic DNA was purified from human 293T cells obtained from the ATCC (ATCC line CRL-3216) using a Wizard® Genomic DNA Purification Kit (Promega), then amplified with SHS-specific primer pairs. SHS PCR amplification was performed in 25 µL reactions containing 150 ng of genomic DNA; 2.5 µL of 10X Thermo Pol buffer (New England Biolabs, Ipswich, MA), 200 µM of dNTPs, a 400 nM final concentration of each primer and 0.25 µL (1.25 unit) of Taq polymerase (New England Biolabs, Ipswich, MA). Amplification was performed using the following program: 30 cycles of 1 min @ 95°C; 30 sec @ 95°C; 30 sec @ 50°C and 30 sec @ 68°C followed by 5 min @ 68°C.

A subset of SHSs were sequenced by first amplifying the SHS region using flanking primer pairs, then using a 3rd primer located ~100 bp from the cleavage site as a sequencing primer to include the CreI target site in the sequenced region (Supplementary Table 1). Genomic DNA purified from human 293T-REX cells (InVitrogen) together with SHS-specific primer pairs were used in 25 µL reactions containing 50 ng of purified genomic DNA; 12.5 µL PrimeStar Max DNA polymerase premix (Takara, Mountain View, CA); and a 240 nM final concentration for each amplification primer. Amplifications were performed using the following program: 35 cycles of 10 sec @ 98°C; 15 sec @ 50°C and 3 min @ 72°C. The resulting PCR products were gel-purified using a QIAquick Gel Extraction Kit (Qiagen), quantified by spectrophotometer and then used for DNA sequencing and in vitro cleavage analyses.

Capillary sequencing was performed by Genewiz Inc, (South Plainfield, NJ), using the SHS-specific sequencing primers listed in Supplementary Table 1. Sequenced reads were aligned to expected reads using CLC Main Workbench (Boston, MA). PCR products were digested with purified mCre protein in 15 µL reactions containing 15 fmol DNA substrate and 0, 15 or 150 fmol of purified mCre protein (9) in 170 mM KCl, 10 mM MgCl2 and 20 mM Tris pH 9.0. Digestions were performed at 37°C for 1 hr, then stopped by adding 3 µL (1:6) of 6x stop buffer (60mM TrisHCl pH 7.4, 3% SDS, 30% glycerol, 150mM EDTA) prior to electrophoresis through a 1% agarose gel run in TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA).

Substrate and cleavage product bands in digestion experiments were identified following gel electrophoresis by ethidium bromide and digital image capture followed by band intensity quantification using ImageJ (http://imagej.nih.gov/ij/). A comparably-sized PCR product containing the native CreI target site (5'-AAAACGTCGTGAGACAGTTT-3') was included in experiments as a positive control.

***In vivo* cleavage sensitivity of SHS:**

The *in vivo* cleavage sensitivity of individual SHSs was analyzed by co-expressing mCre with the TREX2 3' to 5' repair exonuclease in human 293T cells, followed by the recovery and mCre digestion of target sites to identify the fraction that were mutant and cleavage-resistant (Certo et al., 2012). We focused on *in vivo* cleavage and targeted modification experiments on three representative SHSs: SHS231, a Chr4 site with the highest SHS score (Table 4.2); SHS229, a Chr2 SHS with perfect nucleotide sequence identity to a member of our 20 bp site query library; and SHS253, the Chr2-specific member of the small family of 6 identical target sites represented once each on 6 different chromosomes (Chr2, 5, 7, 14, 17 and X; Figure 4.1B, Table 4.2).

	genomic location	sequence	match	site criterion 1	2	3	4	5	6	7	8	site score	site ID
	current human SHSs												
	chr19:55,625,241-55,629,351			−	+	−	+	+	−	+	+	5	*AAVS1*
	chr3:46,414,443-46,414,942			−	+	−	+	+	−	+	+	5	*CCR5*
	chr3:9,415,082-9,414,043			−	+	−	−	+	−	+	−	3	*hROSA26*
	new human SHSs												
	canonical I-CreI/mCreI site	AAAACGTCGTGAGACAGTTT											
	chr1:152,360,840-152,360,859	AAAATGTCAgGAGACATTTT	19	+	+	−	−	+	−	+	−	4	323
	chr8:68,720,172-68,720,191		19	+	+	+	+	+	+	+	−	7	325
	chr1:175,942,362-175,942,381	AAACTGTCATGAGACATTTg	19	−	−	−	−	+	−	+	−	2	289
	chr1:231,999,396-231,999,415	AAACTGTCATGgGACAGATT	19	+	+	−	+	+	−	+	−	5	227
*	chr2:45,708,354-45,708,373	AAAATGTCATGCGACATTTT	20	+	+	−	+	+	−	+	−	5	229
*	chr2:48,830,185-48,830,204	AAACTGaCATAAGACAGATT	19	−	+	−	+	+	−	+	−	4	253
	chr5:19,069,307-19,069,326		19	+	+	−	+	+	−	+	−	5	255
	chr7:138,809,594-138,809,613		19	−	+	−	−	+	−	+	+	4	257
	chr14:92,099,558-92,099,577		19	+	+	−	+	+	−	+	−	5	259
	chr17:48,573,577-48,573,596		19	−	+	−	+	+	−	+	−	4	261
	chrX:12,590,812-12,590,831		19	+	+	−	+	+	−	+	−	5	263
	chr2:77,263,930-77,263,949	AAAATGTgGTGAGACATTTT	19	+	+	−	+	+	−	+	+	6	317
	chr2:150,500,675-150,500,694	AAACTGTCATAAGACAGATc	19	+	+	+	+	+	+	+	−	7	303
	chr3:31,670,871-31,670,890	AAAATGTCATACtACAGATT	19	+	+	−	+	+	−	+	−	5	331
	chr4:37,769,238-37,769,257	AAACCGTCGTGAtACATTTT	19	+	+	−	+	+	−	+	+	6	283
*	chr4:58,976,613-58,976,632	AAACTGTCATAtGACAGATT	19	+	+	+	+	+	+	+	−	7	231
	chr5:7,577,728-7,577,747	AAAATGTCATGAGACAGTcT	19	+	+	−	+	+	−	+	−	5	315
	chr5:93,159,222-93,159,241	AAAATGTCAaGAGACATTTT	19	−	−	−	+	+	−	+	−	3	327
	chr5:159,922,029-159,922,048	AAACTGTCAaAAGACAGATT	19	−	−	−	+	+	−	+	−	3	305
	chr16:19,323,777-19,323,796		19	+	+	−	+	+	−	+	−	5	307
	chr20:5,055,245-5,055,264		19	−	+	−	−	+	−	+	+	4	309
	chr6:89,574,320-89,574,339	AAACTGTCcTAAGACAGTTT	19	+	+	−	+	+	−	+	−	5	285
	chr6:114,713,905-114,713,924	AAAATtTCATGAGACATTTT	19	+	+	+	+	+	+	+	−	7	233
	chr6:134,385,946-134,385,965	AAAATGTCATGAGgCAGTTT	19	+	+	−	+	+	−	+	+	6	311
	chr6:138,972,461-138,972,480	AAACTGTCATACcACAGTTT	19	+	−	−	+	+	−	+	−	4	299
	chr7:113,327,685-113,327,704	AAACTGTCATACaACAGTTT	19	+	−	+	+	+	+	+	−	6	301
	chr8:40,727,927-40,727,946	AAACTGaCGTAAGACAGATT	19	+	+	−	+	+	−	+	−	6	293
	chr11:32,680,546-32,680,565	AAAATGTCcTGAGACAGATT	19	−	+	−	+	+	−	+	+	5	319
	chr12:27,543,737-27,543,756	AAAAaGTCATGAGACATTTT	19	−	+	−	+	+	−	+	−	4	333
	chr12:66,516,386-66,516,405	AAACTGTaGTAAGACAGATT	19	−	+	−	+	+	−	+	−	4	295
	chr12:126,152,581-126,152,600	AAAATGTCATGAGAtATTTT	19	+	+	−	+	+	−	+	−	5	329
	chr17:14,810,285-14,810,304	AAACaGTCATAAGACAGATT	19	+	−	−	+	+	−	+	−	4	297
	chr22:35,770,121-35,770,140	AAACTGaCATGAGACAGATT	19	−	+	−	+	+	−	+	−	4	291
	chrX:16,059,732-16,059,751	AAAATGTCATGAGAaAGTTT	19	−	+	+	+	+	+	+	−	6	313
	chrX:79,674,328-79,674,347	AAAATGTCATAAGgCAGTTT	19	−	+	−	−	+	−	+	−	3	321

Table 4.2: Safe Harbor Site Scoring: Table summarizing SHS scoring for all 9 criteria presented in Table 1. Each site is listed with its hg19 build 37 genomic coordinates, target site sequence and match to the 20bp query library of 128 potential, highly cleavable I-CreI/mCre target sites. The 20 bp sequence of the site is listed (lower case letters denoting a variance from the canonical or expected I-CreI/mCre binding site) and the number of which each site matches the canonical or expected I-CreI/mCre binding site. All sites are then scored for site criteria listed in Table 4.1, and given a composite 'site score' representing the sum of the site criteria the site meets. Site ID is a unique identifying number of each newly identified human genomic SHS. The three sites with asterisks beside them are those that were used to integrate a *loxP* site at that location.

Table 4.2 footnotes: SHS, safe harbor site; *AAVS1*, adeno-associated virus integration site 1; *CCR5*, C-C chemokine receptor type 5; *hROSA26*, human homologue of the ROSAßgeo26 locus.

A modified calcium phosphate (CaPO4) transfection protocol (Chen and Okayama, 1987) was used to introduce a pRRL-based lentiviral expression vector encoding mCre, TREX2, and mCherry proteins into human 293T cells. The three coding sequences were contained within a single transcript and open reading frame separated by T2A sequences. Cells (2-4 x 105/well) were plated in a 6-well plate 24 hrs prior to transfection, and were ~70% confluent at the time of transfection. Expression vector plasmid DNA (1.5 µg in 10 µL H2O) was mixed with 40 µL of freshly prepared 0.25M CaCl2 and 40 µL of 2x BBS buffer (50mM BES pH6.95 (NaOH), 280mM NaCl, 1.5mM Na2HPO4; Boston BioProducts), then incubated at room temperature for 15 min before being added dropwise to wells. Plates were incubated overnight in 3% CO2 at 37°C. The following day medium was replaced and cells were grown for an additional 24 hrs at 5% CO2, 37°C. Transfection efficiency was checked by determining the fraction of mCherry-positive cells by flow cytometry. In brief, cells were trypsinized, counted, then fixed with formaldehyde (1% final concentration, 10 min at room temperature followed by the addition of 1/20 volume of 2.5M glycine) prior to flow cytometric analysis of ~20,000 cells/transfection on a BD FACSCanto II flow cytometer (BD Biosciences, San Jose, CA) . Genomic DNA prepared from co-transfected and control cells was used for PCR amplification and *in vitro* mCre cleavage analysis of SHSs as described above.

Construction of SHS-specific repair templates and TALEN/CRISPR-Cas9 nucleases:

In order to determine whether SHS cleavage *in vivo* could catalyze homology-directed repair, we co-transfected human 293T cells with a SHS-specific repair template and an expression vector for mCre or for a SHS-specific TAL effector nuclease pair, CRISPR/Cas9 cleavase or nickase (Figure 4.2).

Repair templates consisted of 500 bp homology arms that were SHS-specific and flanking a 48 bp insert containing a canonical *loxP* recombinase site and PvuI and SacII restriction endonuclease cleavage sites. The repair template design for SHS homology-dependent repair is shown in Figure 4.2. Repair templates were made by overlap extension PCR using oligonucleotide primers to generate PCR products that, when re-amplified, incorporated the 48 bp *loxP* insert at the center of the repair

template (Figure 4.2). The overlap extension PCR primers for SHS231 repair template construction are listed in Supplementary Table 1 as 231: Repair Template Construction: 'Right Reverse' and 'Left Forward'. The corresponding primers used to re-amplify the SHS231 overlap extension PCR products are listed as 231: Repair Template Construction: 'Right Forward' and 'Left Reverse'.

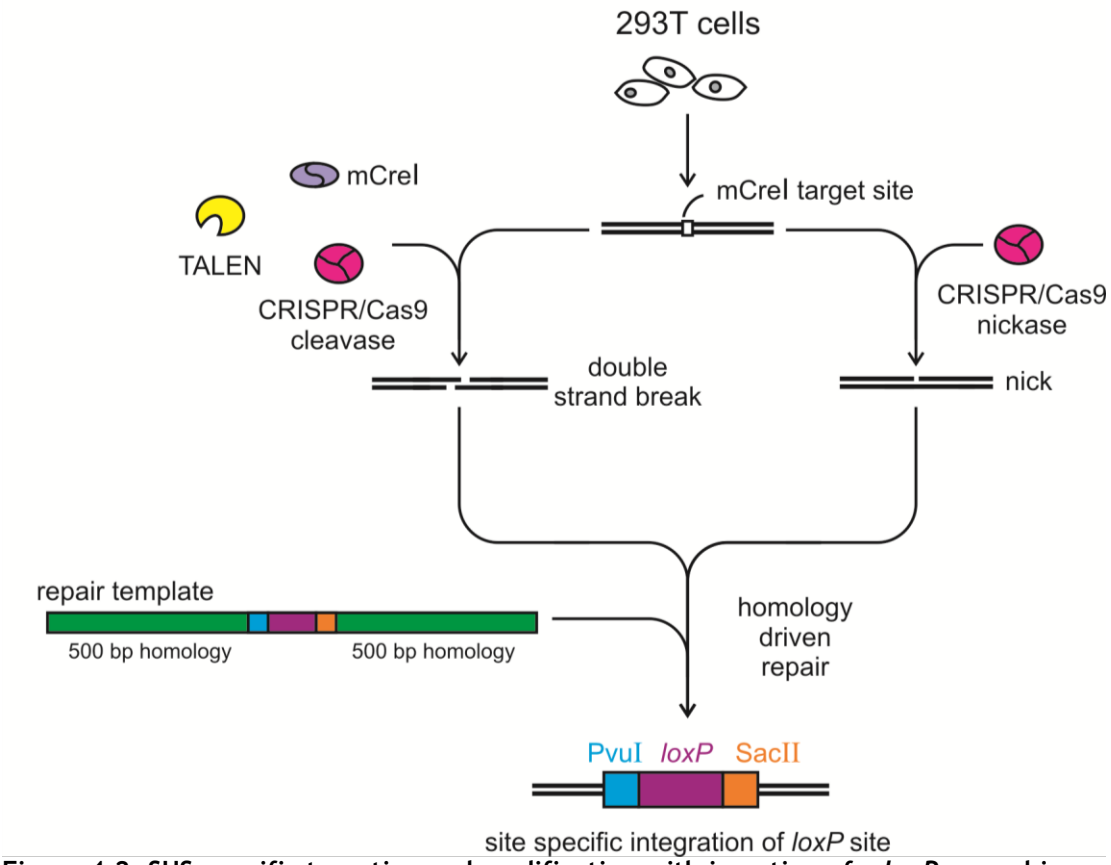


Figure 4.2: SHS-specific targeting and modification with insertion of a *loxP* recombinase site. Three different genome engineering nucleases were used to cut selected SHSs in human 293T cells to promote homology-dependent repair of the cleaved site with incorporation of a new *loxP* recombinase site flanked by two new restriction cleavage sites at the SHS. Three genome engineering nucleases were used to generate DNA double strand breaks or, in the case of the SRISPR/Cas9 nickase a DNA single strand break, at the SHS to promote homology-dependent repair off the co-transfected repair template.

A SHS231-specific TALEN protein pair to cleave the SHS231 I-CreI/mCre cleavage site was designed using the TALEN Targeter 2.0 web design engine (https://tale-nt.cac.cornell.edu/node/add/talen) (Cermak et al., 2011). Forward and reverse strand, 20 bp-specific TALEN sequences were inserted into the expression vector RKSXX-pCVL-UCOE.7-SFFV-BFP-2A-HA-NLS2.0-TruncTAL, and each TALEN open reading frame was generated by assembling the following repeat variable di-residues

(RVDs): left TALEN: NG NG NN NN HD NG NI NH NN NH HD NG NI NI NN NN NI NG NG NI, corresponding to the nucleotide sequence TTGGCTAGGGCTAAGGATTA (Chr4: 58,976,594-58,976,613); and right TALEN: NG NN NG NI NG NH HD NG NG NG HD HD NG HD NG NG NN NG NG NI, corresponding to the nucleotide sequence TGTATGCTTTCCTCTTGTTA (Chr4:58,976,613-58,976,632).

A corresponding SHS231-specific CRISPR/Cas9 cleavase expression vector was constructed in pX260 (Cong et al., 2013; Hsu et al., 2013) that contained expression cassettes for the *S. pyogenes* Cas9 nuclease, the CRISPR RNA array, and the tracrRNA. The Cas9 target site, located 110 bp downstream of the mCre/TALEN cleavage site, was identified using the CRISPR Design Tools Resource developed by Zhang and colleagues [http://crispr.mit.edu/]. A corresponding SHS231-specific Cas9 nickase expression vector was constructed in pX334 that encoded Cas9 with a D10A mutation to confer nickase, as opposed to cleavase, activity. A common guide sequence was inserted into both expression vectors: 5'-CTAATCTGGACAAAACATTTATATACTGCG-3' followed by a TGG proto-spacer adjacent motif (PAM).

Targeted modification of SHSs in human cells:

Nuclease expression and repair template pairs were transfected into 293T cells as described above: cells (2 – 4 x 10^5 cells/well) were plated in 24-well plates 24 hrs prior to transfection, then transfected with 0.3 pmol DNA of both nuclease and repair template vectors. Vector DNAs in 10 µL H2O were mixed successively with 40 µL of freshly prepared 0.25M CaCl2 and 40 µL of 2x BBS buffer, then incubated 15 min at room temperature prior to being added to cells for incubation in 3% CO2, 37°C overnight. Media was replaced the following day, and cells were grown for an additional 48 hrs prior to genomic DNA preparation as described above. SHSs were amplified from purified genomic DNA as described above, then gel-purified and cloned into a pGEM®-T Easy plasmid vector (Promega) for sequencing. White colonies were picked following transformation into α-Select Chemically Competent Gold Efficiency Cells (Bioline). We initially identified no targeted insertions in TAL SHS-targeted cells among 21 randomly picked, insert-containing colonies. Thus we picked 60 additional colonies (10 pools

of 6 colonies/pool) and used DNA prepared from these pooled colonies for restriction digestion to identify PvuI-sensitive pools on a 1% agarose gel. The pool that was positive was de-convoluted to identify a *loxP*-positive clone for SHS DNA sequencing. Plasmid DNAs were sequenced using a T7 promoter sequencing primer to generate sequence data that covered the SHS target site region (Figure 4.4C). Sequencing results were aligned with the repair template sequence using the CLC Main Workbench (Qiagen) software.

Results:

Identification of potential new human Safe Harbor Sites:

In order to identify potential new human SHSs we compiled a list of 128 DNA target site variants for the I-CreI/monomerized mCre homing endonuclease protein that were predicted to be cleaved with high efficiency (≥90% of the native site cleavage efficiency; Figure 4.1A). We then used BLAST to search for the closest matches to these 128 CreI sites in the human genome (hg19). The rationale for using the cleavage specificity of CreI to drive SHS searching was the long (20-22 bp) target site and high site specificity of cleavage of the I-CreI/mCre homing endonuclease proteins. We reasoned that CreI target sites would be correspondingly infrequent in the human genome, and that any CreI SHSs in the human genome would be candidates for immediate targeting by the compact, easily vectorized mCre protein. We also reasoned that any SHS defined by a CreI target site also had a high likelihood of being targeted by a SHS-specific TALEN pair or by or CRISPR/Cas9 nucleases as both classes of nuclease have less stringent target site selection criteria than does CreI.

Among the 27 CreI target sites and potential SHSs we identified by BLAST search, one as a perfect match (20/20 bp) to one of our predicted highly cleavage 128 target sites, while the others contained single base pair differences (19/20 bp identity). A majority of the target sites were found only once in the human genome (24/27 or 89%), while the remaining 3 were represented 2, 3 or 6 times in the human genome at different genomic locations (Table 4.2). The 35 different genomic SHS

locations identified by CreI target site searches were on 17 of the 23 human chromosome pairs including the X chromosome, and 24 of 48 chromosome arms (Figure 4.1B, Table 4.2).

Safe Harbor Site scoring:

All 35 potential new SHSs together with three other widely cited SHSs (AAVS1, CCR5 and hROSA26) were next evaluated using 8 safety, functional and accessibility criteria in addition to site uniqueness. Among our 35 SHSs, 25 (or 71%) fulfilled more than half (≥5/9) of our SHS criteria as did the previously defined AAVS1 and CCR5 SHSs (Table 4.2). When we examined SHS safety criteria alone (SHS criteria 1-6 in Table 1), 21/35 (60%) of our SHSs met ≥4 of 6 criteria. Three SHSs matched all 6 safety criteria (SHS231, 233 and 303). In contrast, the AAVS1, CCR5, hROSA26 SHSs each matched only 3 of these 6 safety criteria (Table 4.2).

None of our potential SHSs was located in a copy number-variable (CNV) region of the human genome. However, we identified base pair variation in 10 of our 35 SHS in the CreI target site region using whole genome sequencing data from the 1000 Genomes Project. Base pair variation was restricted to single nucleotide polymorphic variants (SNPs or SNVs); no indels were identified in any of our 35 SHSs. Ten of our 27 CreI target sites, representing 11 different SHSs, contained SNP variants. The three most frequent SNPs, with frequencies ranging from 7.5% to 50% in 1000 Genomes samples, were predicted to reduce CreI cleavage efficiency by ≥70%. Among the remaining CreI SHS target sequences, only one low frequency SNP (0.37%) was predicted to suppress CreI cleavage by >80%. Of note, all 35 sites were present in non-mutated form in 80% of individuals in 1000 Genomes data, and all 35 of our SHS were predicted to be fully I-CreI/mCre cleavage-sensitive in 94% of individuals included in 1000 Genomes data (additional results not shown).

SHS experimental validation: In order to experimentally validate the above *in silico* analyses, we amplified 28 of our 35 SHSs, then cloned and sequenced 10 SHSs. We also used an independently generated PCR product from specific SHSs and purified mCre protein to determine in vitro SHS cleavage

sensitivity. The rationale for performing in vitro cleavage analyses was that the I-CreI/mCre PWM that drove our BLAST search was based on single base pair variant data, not base pair variant combinations. Thus we wanted to determine the cleavage sensitivity of SHSs found by BLAST search but were not perfect 20/20 bp matches or had not previously been tested for mCre cleavage sensitivity. Four of the 6 target site sequences we analyzed, representing 10 SHS (unique SHS227, 229, 231, 233, 251 and SHS253, 255, 257, 259, 263 at different chromosomal locations) were cleaved by mCre as efficiently as a control native CreI target site (data not shown).

SHS cleavage with targeted modification in human cells:

The *in vivo* cleavage sensitivity of SHSs was analyzed by co-expressing mCre with the TREX2 3'-to-5' repair exonuclease in human 293T cells, followed by PCR amplification and mCre digestion of target sites. This experiment was designed to identify a cleavage-resistant target site fraction in nuclease-expressing cells that would provide a minimum estimate of the proportion of SHS molecules that were cleaved with error-prone rejoining in vivo (Certo et al., 2012). Five of the 6 SHS we assayed in this way (SHS227, 229, 231, 253, 257, 263) displayed an increase in the fraction of mCre-resistant target sites of from 3.8% to 31.3% over a mock-transfected control.

In order to determine whether SHS cleavage *in vivo* could catalyze high fidelity homology-dependent repair, we co-transfected human 293T cells with an expression vector for mCre, a TAL effector nuclease pair or a CRISPR/Cas9 cleavase/nickase expression vector together with a SHS-specific repair template containing a *loxP* site flanked by two restriction sites (Figure 4.2). SHS229, 231 and 253 were analyzed following mCre expression, as were SHS229 and 231 after CRISPR/Cas9 nuclease expression and SHS231 after TAL effector nuclease expression. We also performed parallel targeted integration assays at the other two SHS we subjected to more comprehensive analysis, SHS229 and 253.

We also sequenced cloned SHS231 PCR amplicons from cells transfected with each of our 3 genome engineering nucleases in order to determine the fidelity of cleavage-dependent integration.

We reasoned that this would be an independent way to estimate the frequency of SHS-targeted *loxP* cassette integration. SHS231 PCR amplicons generated from cells expressing mCre, a SHS-specific TAL effector nuclease pair or a CRISPR/Cas9 nuclease were cloned into pGEM®-T Easy plasmid DNA, and insert-containing colonies were sequenced across the SHS cleavage region (Figure 4.3). The frequency of targeted integration at SHS231 among otherwise unselected insert-positive colonies was 4.8% for mCre/TREX2 (3/63 clones), 6.1% (2/33) for CRISPR/Cas9 nuclease and 16.1% (5/31) for the CRISPR/Cas9 nickase. We identified no targeted integration event among the initial 21 pGEM colonies isolated from TAL-effector nuclease expressing cells. When we assayed 10 additional pools of 6 clones/pool we identified 1 SHS-targeted integration event (1/81 or 1.23%). Sequence data for all three nucleases indicated accurate, targeted integration of the *loxP* site together with low frequency single base substitutions that most likely reflect PCR errors introduced by Taq DNA polymerase

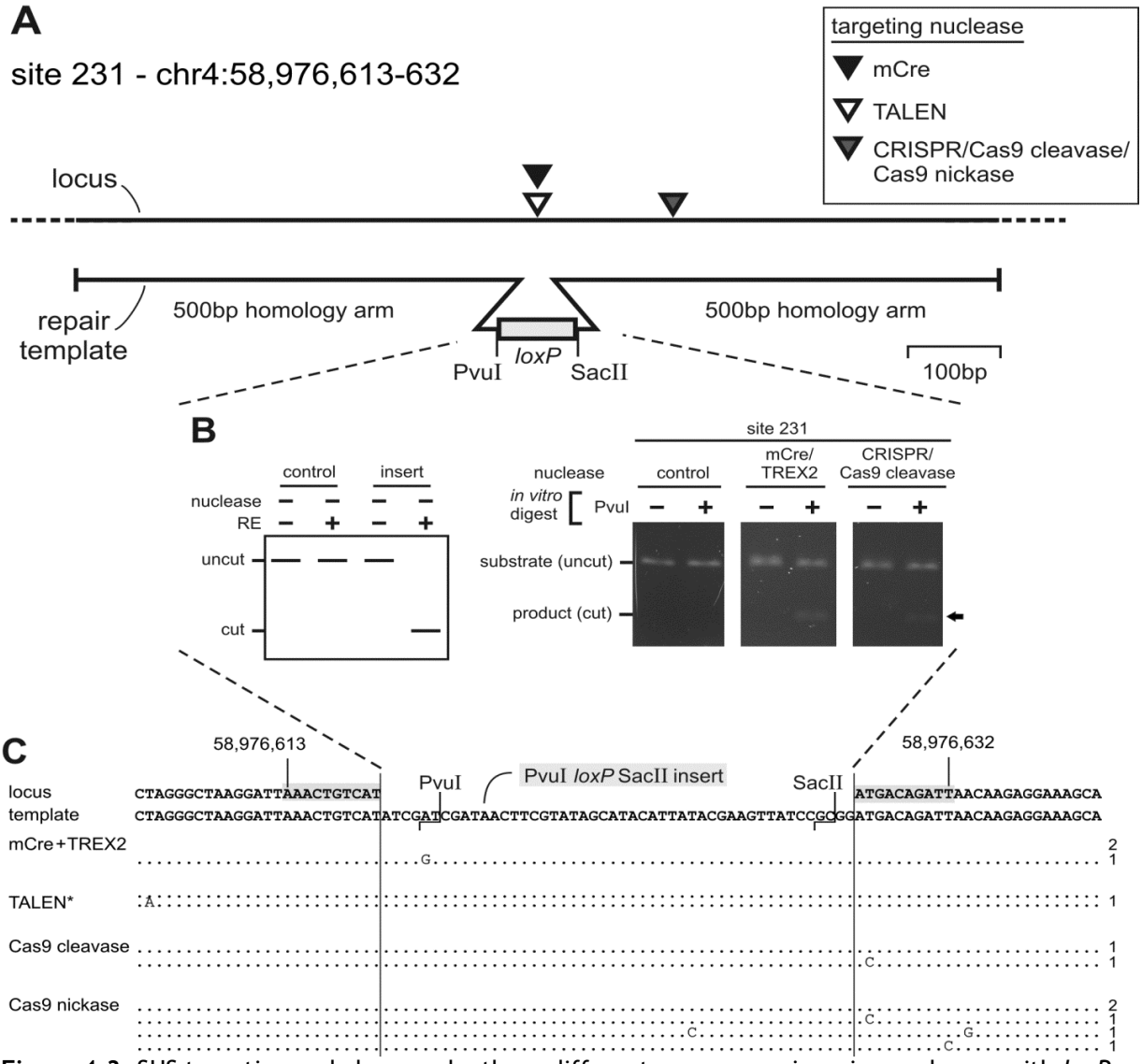


Figure 4.3: SHS targeting and cleavage by three different genome engineering nucleases with *loxP* recombinase site insertion. (A): The chromosome 4 SHS231 is shown as a horizontal line with the position of genome engineering nuclease cleavage sites indicated by triangles. The mCre homing endonuclease (black triangle) and a SHS231 site-specific TAL effector nuclease pair (white triangle) cleave exactly at the mCre target site, whereas the corresponding CRISPR/Cas9 nuclease (or nickase) cleaves 147bp downstream of the mCre target site. The SHS231-specific repair template is depicted below the SHS as a horizontal line with a novel insert of 49 bp that included a *loxP* recombinase site flanked by two new restriction cleavage sites for PvuI and SacII. (B): Targeting and repair of SHS231 was detected by a PCR-based assay in which transfected cells were amplified with primers that flanked the chromosomal SHS231 and were outside the region covered by the repair template. A PCR product of ~1kb was used for cleavage assays with PvuI as shown schematically in the left panel and as data in the right panel. (C) A second 400bp PCR product was generated from each transfection experiment and cloned into a p-GEM T-Easy vector to facilitate sequencing of the target site region. The locus and repair template sequences are shown at the top, aligned with the sequence of targeted clones to indicate identity (dots) or base substitutions with the substitution indicated by the variant base identified. The frequency of targeted clones among PCR products analyzed by this approach provides an estimate of targeting frequency in vivo: 3/63 (4.76%) for mCre/TREX2; 1/81 (1.23%) for TALEN; 2/33 (6.06%) for Cas9 cleavase and 5/31 (16.1%) for Cas9 nickase. See text for additional detail.

Discussion

Despite their utility only a small number of SHSs have been well-characterized or are in wide use in human cells. We used a systematic approach to identify and evaluate 35 potential new human SHSs. All 35 new SHSs contained a 20 bp I-CreI target site at its center, and are potentially targetable by a single genome engineering nuclease, the small, easily vectorized and active mCre homing endonuclease protein (Heath et al., 1997; Jurica et al., 1998; Li et al., 2009). All of these sites should also be candidates for targeting by CRISPR/Cas9, TAL effector and zinc finger nucleases. We demonstrated the potential utility of a subset of these SHSs by using mCre, SHS-specific CRISPR/Cas9 cleavase/nickase proteins and a SHS-specific TAL effector nuclease pair to promote mutagenic end joining with site disruption, or homology-dependent insertion of a *loxP* recombinase site. The experimental workflow is summarized in Figure 4.4.

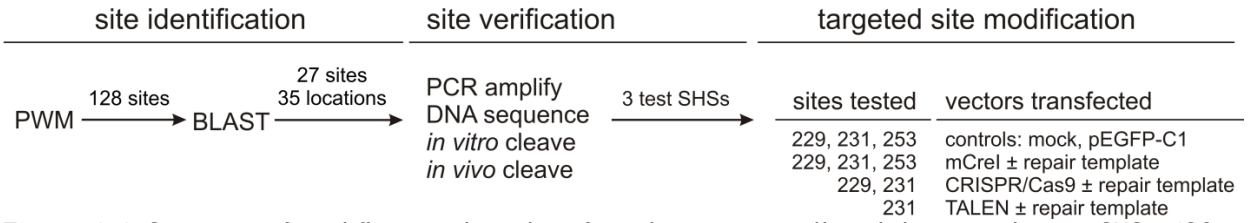


Figure 4.4: Summary of workflow used to identify and experimentally validate new human SHSs. 128 potential cleavable sites were generated using the target site degeneracy data of I-CreI/mCre-I. A BLAST search of these 128 sites identified 27 target sites at 35 genomic locations. These 35 locations were then verified via PCR amplification, sequencing, *in vitro* cleavage, and *in vivo* cleavage. Three representative SHSs were then used in targeted repair experiments in living cells to demonstrate site cleavage and repair with the incorporation of a novel *loxP* site.

Our newly identified SHSs cover 71% of the human chromosomal complement (22 autosomes and the X and Y chromosomes) and half of all chromosome arms. These sites were scored on the basis of 9 site criteria driven largely by safety concerns, and made extensive use of the rapidly growing collection of human genome-scale structural, genetic and regulatory data (ENCODE Project Consortium, 2012). Over half of our SHS (20/35, or 57%) met 4 of our 6 safety criteria (Tables 4.1 and 4.2). In contrast, the existing human SHSs AAVS1, CCR5 and hROSA26 met only 2 or 3 of these 6 safety criteria (Table 4.2). Among our 35 SHSs, none was in a region of copy number (CNV) variation, though 10 sites did contain CreI target site basepair/SNP variation. We used 1000 Genomes Project data to identify 3 of these sites, SHS255 on Chr5 (variant frequency = 0.5041), SHS301 on Chr7 (variant frequency =

0.2231) and SHS297 on Chr17 (variant frequency = 0.0753) where target site SNP variation was predicted to substantially reduce mCre cleavage efficiency: the predicted reductions in cleavage efficiency ranged from 72% to 84% when compared with the native target site. This analysis indicated that a majority of our new SHS should be cleavage-sensitive, and thus potentially accessible, in any individual.

Several aspects of SHS use remain to be explored and/or optimized. Chief among these are safety and efficiency, both particular concerns for therapeutic applications. While constitutive high level expression of I-CreI or mCre is toxic (Monnat et al., 1999), transient expression has little appreciable toxicity in DNA double strand break/homology-dependent repair-competent cells. This likely reflects the small number of CreI cleavage-sensitive sites in the human genome, although CreI off-target cleavage sites have not as yet been rigorously quantified. Cleavage-dependent toxicity may be further attenuated by using nickase, as opposed to cleavase, variants, and by transfecting an mRNA encoding the desired SHS-specific nuclease together with a SHS-specific repair template.

Repair efficiency at our newly identified SHS also remains to be systematically explored and optimized. Important variables here beyond cell type-specific gene transfer efficiency include the type of repair template (single vs double-stranded), and the length and degree of nucleotide sequence identity between the repair template and target site flanking sequences. The highest efficiency of repair is likely to occur when there is substantial (>200bp), perfect DNA sequence identity in each repair template donor arm. Shorter, less perfectly base-paired or single-stranded repair templates should also work, though appear to be less efficient repair templates. It should be possible to further optimize all of these variables to ensure high efficiency gene editing for specific SHS, cell types and recipients. Long term, stabile gene expression with the maintenance of chromatin state may depend critically on the location of the SHS, and the nature of the genetic modification targeted to a specific SHS.

The new human SHSs we identified can be used to facilitate a wide range of genome engineering applications. Site modifications can also be multiplexed directly when the mCre homing endonuclease is used together with 2 or more site-specific repair templates. SHS multiplexing with simultaneous or sequential modification of different SHSs on the same chromosome or chromosome arm would help ensure the stable association and mitotic inheritance of introduced genes. This approach could be used to facilitate e.g., genomic interaction analyses or the transfer of groups of genes or whole pathways for consistent expression. The new human SHSs we have identified should thus be a useful starting point for many different types of genome engineering to facilitate biology as well as therapeutic goals. The approach we used to identify and evaluate SHSs should be useful as a guide for identifying additional potential human SHSs. One useful starting point might be the human genomic regions being identified as dispensable in at least a subset of individuals (MacArthur et al., 2012).

Conclusions

Safe Harbor Sites (SHSs) are defined genomic locations where new genes or genetic elements can be introduced without disrupting the expression or regulation of known genes. We identified 35 potential new human SHSs, and determined their suitability as safe harbors on the basis of genomic and experimental criteria. New SHSs were identified initially on the basis of a high quality match to target sites that are cleaved efficiently by the native I-CreI/mCre homing endonuclease protein. We assessed all 35 potential SHSs, together with the existing AAVS1, CCR5 and hROSA26 SHSs, on the basis of 9 different safety, functional and accessibility criteria. We then used a combination of *in vitro* and *in vivo* assays to further characterize three of our newly identified SHSs in greater detail. *In vivo* SHS validation assays included the use of the mCre homing endonuclease and SHS-specific CRISPR/Cas9 and TAL nucleases to catalyze either SHS error-prone repair or the SHS-targeted insertion of a *loxP* recombinase site. A unique feature of using SHSs defined on the basis of I-CreI/mCre cleavage sensitivity is the potential for multiplexing: native mCre together with SHS-specific repair templates can be used for the simultaneous or sequential targeting of different SHSs. Our results identify

potentially useful new human SHSs, and provide site- and sequence-specific tools to enable the use of these sites for basic as well as clinical applications.

Chapter 5: Targeted correction of human disease-causing mutations

Background

Shwachman-Bodian-Diamond Syndrome (SBDS) is an autosomal recessive disorder which is characterized by bone marrow dysfunction and pancreatic insufficiency (Burroughs et al., 2009). This disorder is caused by a mutation in a gene on chromosome 7 (*SBDS*) which was named after the disease. SBDS protein is nucleocytoplasmic with a particular concentration in the nucleolus, consistent with its role in ribosome biogenesis (Ganapathi et al., 2007). Studies in yeast and in human cells support a role for the SBDS protein in ribosome biogenesis: SBDS loss results in diminished ratios of the 60S ribosomal subunit relative to the 40S subunit in yeast (Menne et al., 2007). Two mutations, which are found in over 95% of known cases of the disease in humans, are located in exon 2 of the *SBDS* gene (Boocock et al., 2003). Because the phenotype caused by this disease is localized to mutations on a single gene, SBDS becomes an ideal target for gene correction as a therapeutic approach. Additionally, because this syndrome is found to cause bone marrow phenotypes it is an ideal candidate for hematopoietic or induced pluripotent stem cell modification for re-implanting as an end goal for therapy. This is particularly useful for this disease, as over expression of SBDS is potentially toxic and may disrupt developmental roles for SBDS in the heme lineage. Therefore, it is hypothesized that an endogenous range of regulated gene expression is required for normal cell function. In general, hematopoietic stem cell transplantation can cure associated diseases such as aplastic anemia or leukemia. However, stem cell transplant is limited by: donor availability, the sensitivity of patients with certain IBMFS, such as Fanconi anemia, Shwachman-Diamond syndrome, along with transplant-related toxicities, and abnormalities in non-hematologic organ systems. These properties make SDBS an ideal candidate for patient derived iPS cell targeted gene repair as a gene therapy approach. Repairing the gene in its native locus is ideal due to the fact that required gene expression will be controlled by the endogenous locus.

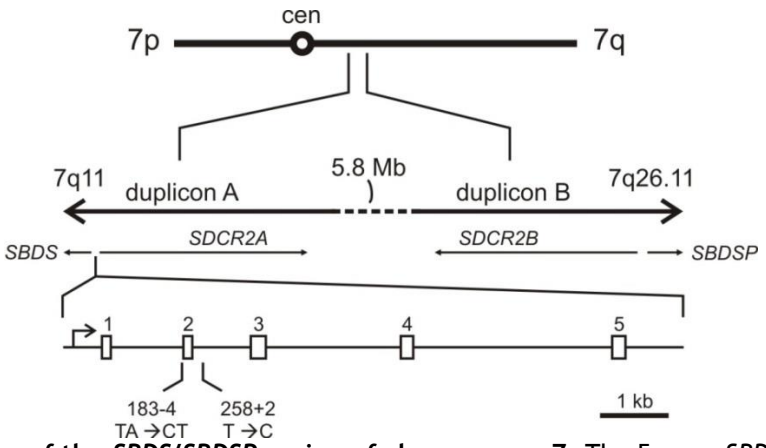


Figure 5.1: Structure of the *SBDS*/*SBDSP* region of chromosome 7. The 5 exon *SBDS* gene is located in the centromere-proximal end of the 305 kb duplicon A at 7p11. A corresponding duplicon B at 7p26.11 is oriented in the opposite orientation, and has at its telomere-proximal end the *SBDSP* pseudogene that serves as a gene conversion template and the source of the two most common SDS patient mutations, 258+2 T>C and 183-184 TA>CT.

The first approach in attempting to modify this gene was with the homing endonuclease platform. In this case, a very fortuitous site was identified using the PSSM based search as described in Chapter 2. Using the PSSM search, a target sequence in the ideal location, between the two most common disease causing mutations loci on chromosome 7 (Figure 5.2) was identified.

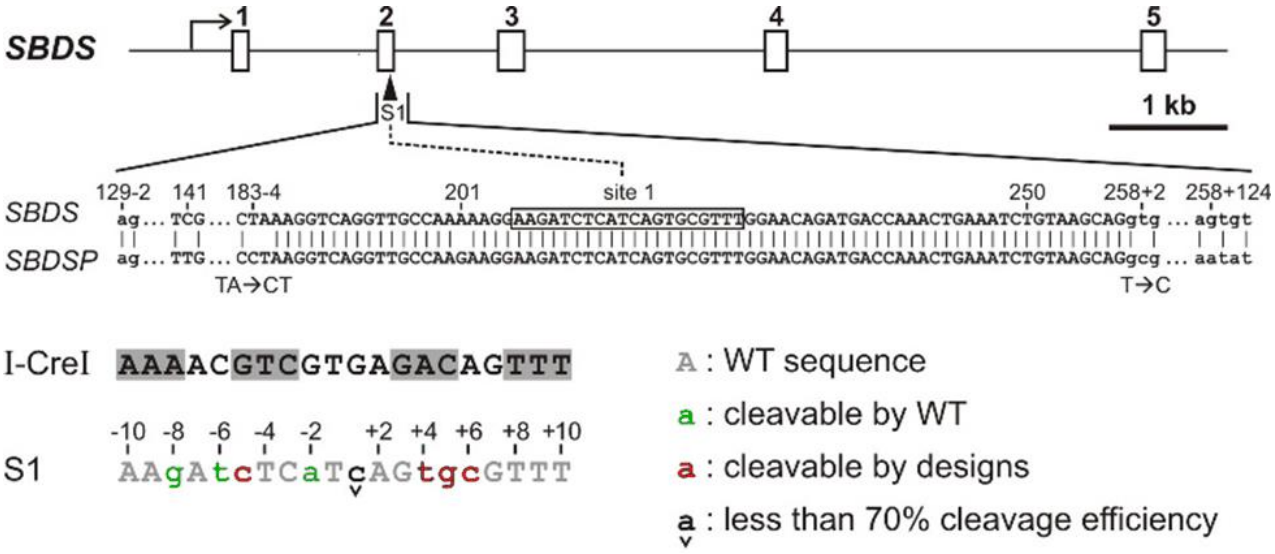


Figure 5.2: A high quality, fortuitous site for genome engineering via mCre homing endonuclease. *SBDS* contains 5 exons, 7.9kb SBDS locus is shown on the top line, with mCre target sites S1 indicated by a box. The sequences of SBDS and the SBDSP pseudogene are shown below the gene model together with sequence differences which likely cause common mutations seen in disease phenotypes. The bottom panel show sites the site details and summarizes in the letter/color code pre-existing cleavage degeneracy and design data that will guide the generation of site-specific mCre variants.

To test the ability of mCre to cleave this target, we utilized the same methodology as detailed in Chapter 3. The first step is to test for the ability of the homing endonuclease to cleave a target test site containing only the combined degeneracy bases using the in vitro competitive cleavage assay (Figure 3.2). For this test, a target sequences containing the following position changes from wild type were tested: -8G, -6T, and -2A. In a separate test, the +1C base change was also added to the other three degeneracy positions. This position was tested in a separate reaction due to the fact that the predicted cleavage efficiency would be significantly affected by the +1C mutation which in single base pair degeneracy testing has 0% of mCre WT activity due to that mutation (Figure 3.2). However because there are multiple changes in the central 4 base pairs (-2G as well) it was important to verify this result. Surprisingly, the effect of both cleavage analysis tests gave similar results with and without the presence of the +1C base pair change (Figure 5.3) Results show a loss of approximately 10 fold activity due to the combined degeneracy bases, regardless of the +1C base change. While the results were not ideal, work should continue on this site on the basis of its ideal location. Additionally, because of the nature of SDBS, there is a likely growth advantage of *SBDS* corrected cells, so even a low percentage of cells successfully modified are estimate to be able to grow at an advantage for selection. Another point to consider is additional selection methods being applied to an enzyme to better select for mutants with higher specificity of activity via randomized mutagenesis and selection and screening.

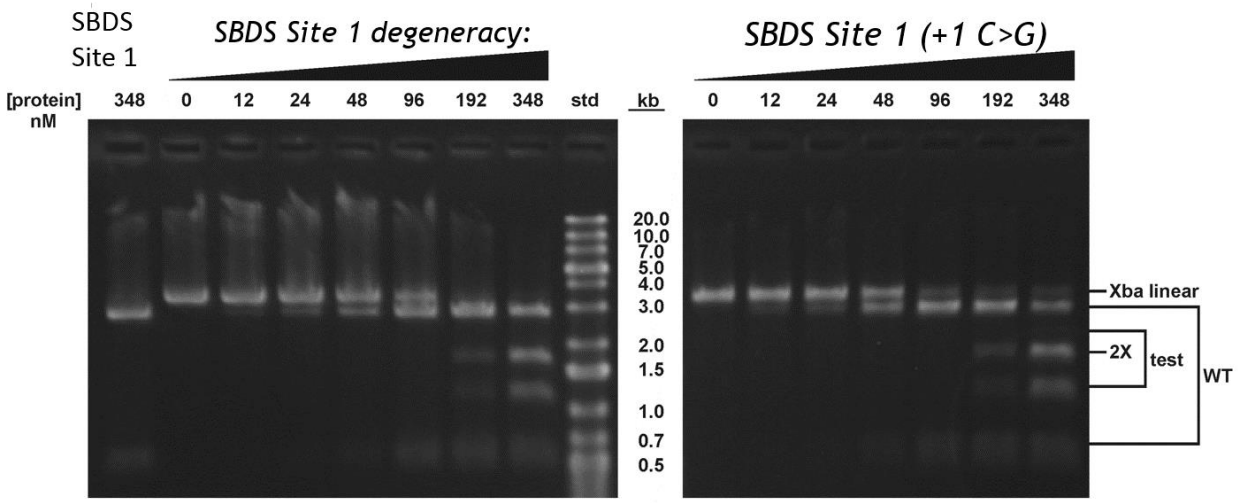


Figure 5.3: *In vitro* competitive cleavage assay to test combined degeneracy of the *SBDS* genomic target sequence. The panel on the left represents the cleavage assay on the native mCre DNA recognition sequence with the following changes: -8G, -6T, and -2A. The panel on the right tests a sequence with identical base changes as the left panel with the addition of the +1C base pair change. In both tests enzymatic activity is reduced on by the degeneracy combination sequence approximately 10 fold.

The next step of enzyme testing and verification included adding engineering design changes to the -5C: 24K, 68T; +4T: 44T, 70N, 75Q; +5G: 24K, 68T; +6C 26T, 77R. First the -5C mutations alone were tested with the combined degeneracy sites (Figure 5.4). In this test, a surprisingly strong specificity shift toward the modified mCre enzyme's ability to cleave the target site occurred.

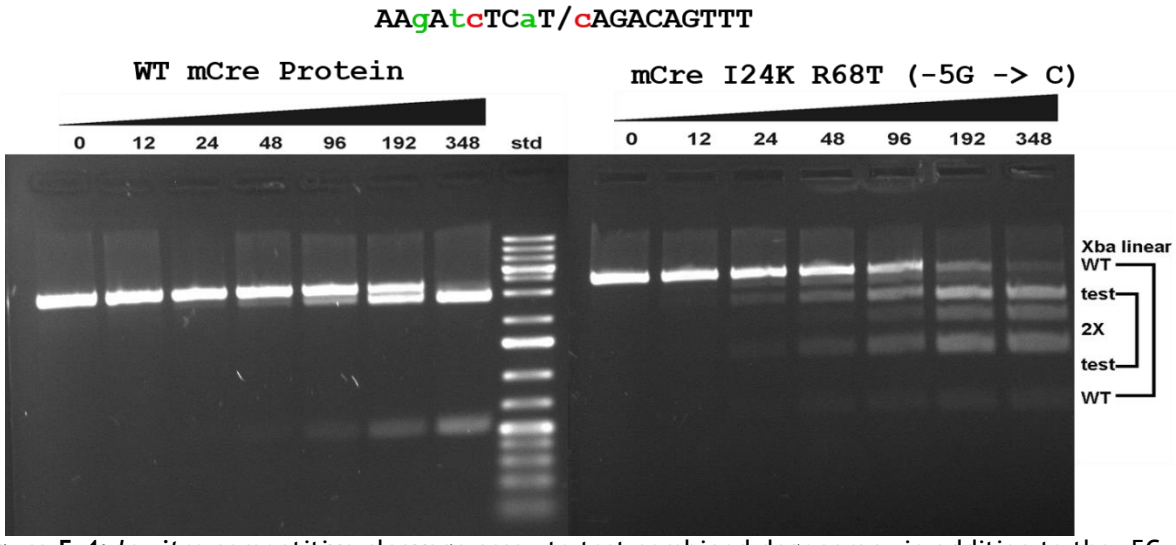


Figure 5.4: *In vitro* competitive cleavage assay to test combined degeneracy in addition to the -5C permitting enzyme mutation (24 I>K, 68R>T) on the target sequence shown at the top of the figure. The left panel shows activity of native mCre protein on the normal target site only and the target sequence remains uncleaved. The right panel shows the modified mCre (24 I>K, 68R>T) activity on both the native binding site as well as the desired target site.

Unfortunately, when the other engineering mutations on the right side of the molecule were incorporated into the enzyme (+4T: 44T, 70N, 75Q; +5G: 24K, 68T; +6C 26T, 77R), all activity was lost (data not shown). We hypothesize that is likely caused by having too many enzyme residue changes within a small region of the DNA binding domain interfere with each other, making an overall incompatible binding pocket and/or destabilizing the mCre protein.

For future experiments, we determined that adjacent base pair changes requiring engineering designs should be avoided for this reason. The engineering problem lies within the +4,5,6 region of the mCre enzyme. Alternative HE solutions for this target site could possibly be developed by library selection or additional computational predictions, however these approaches have not been thoroughly attempted at this time.

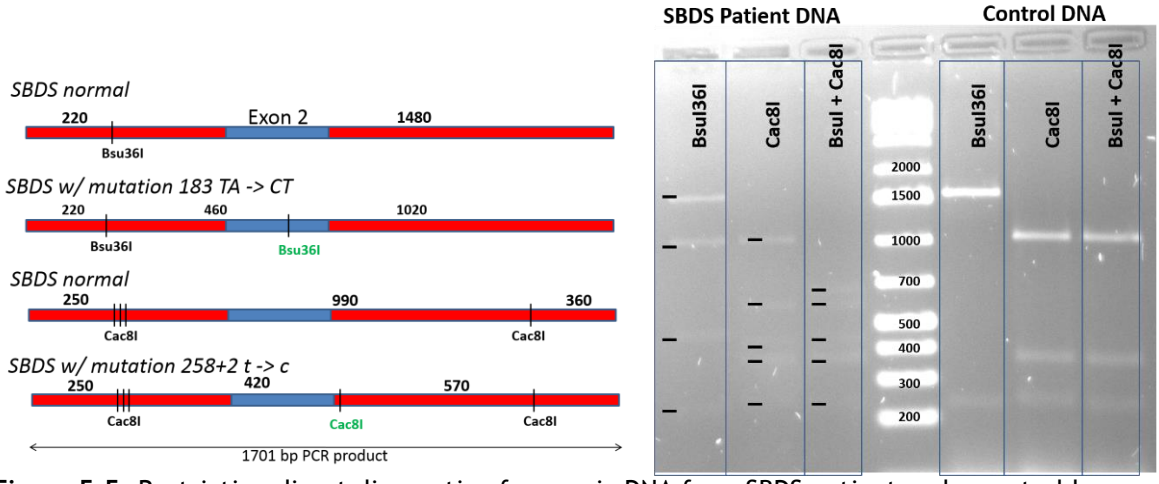


Figure 5.5: Restriction digest diagnostic of genomic DNA from SBDS patient and a control human DNA. The patient genome represents a compound heterozygote, in which one patient allele contains the 258+2 T>C mutation and the other allele contains the 183-184 TA>CT mutation. Heterozygosity of *SBDS* has no detectible phenotype. This diagnostic can also be used to query genome repair in an edited population.

Rather, the alternative approach, which is now underway, uses the TALEN platform. In this approach, we have identified the TALEN pairs that successfully cleave a target sequence in exon 2 of SBDS (Figure 5.6). Two versions of the TALEN pairs were made. While both pairs of TALENs bind the same sequences, they differ in the amino acid sequence of the FokI domain dimer interface which is either wildtype FokI or modified FokI which forms an obligate heterodimer pair (ELD/KKR pair)(Doyon 2010).

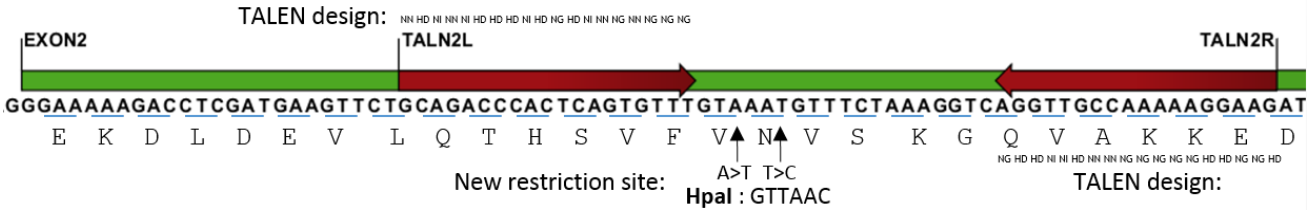


Figure 5.6: TALEN based targeting of *SBDS* exon 2.

Repair template construction:

The repair template plasmids provided with TALEN transfection plasmids include genomic sequence identical to the patient derived iPS cells 500 bp upstream of the left TALEN binding site and

500 bp downstream of the right TALEN except for the desired 258+2 T>C mutation is corrected back to 258+2 C>T in the repair template sequence. The repair template sequence within and including the TALEN binding sites includes additional silent mutations designed to make a TALEN-resistant DNA sequence when repair takes place. The TALEN resistant designs were assigned based on the following observation that substitutions that are likely most disruptive to repeat variable diresidues (RVD) would be to switch a cytosine being recognized by an HD RVD, or an adenine or guanine being recognized by an NN RVD, into a thymine. This substitution is likely to cause a steric clash with the aspartate or asparagine side chain to inhibit TALEN binding and therefore prevent the repaired gene from being re-cut (Barry Stoddard, personal communication). Additionally, a silent mutation was added to the repair template to provide a novel HpaI restriction site. This addition will allow fast and efficient screening of iPS clones by PCR amplification and subsequent enzymatic cleavage to identify cells containing successfully repaired SBDS allele(s).

Transfection optimization of iPS cells:

Transfection has been optimized in the Shimamura lab (FHCRC) using the Continuum transfection reagent from Gemini Biosciences. These transfection conditions were chosen to optimize introduction of TALEN nuclease and repair template plasmids into the SBDS-iPSCs without the use of viral gene transfer. In trial experiments using a GFP construct, achieved transfection efficiency measured upwards of 97% with high cell viability.

Results and next steps

At the time of writing, two targeted correction experiments have been performed using the TALEN reagents described. In both, experiments, the expression of blue fluorescent protein upstream of the TALEN construct was not detected, nor have single cell derived colonies with a corrected mutation been found.

A second correction strategy is also being attempted in parallel. In this strategy we are embedding a promoter and *SBDS* cDNA between two 500bp homology arms specific to one of the newly

defined human chromosomal safe harbor sites described in Chapter 4. We have shown these loci can be cleaved by the mCre homing endonuclease, as well as by TALEN and CRISPR/Cas9 platforms. The safe harbor approach also requires us to identify a promoter which provides an appropriate amount of transcription, as overexpression is possibly toxic.

Both approaches will be aided by improving transgene expression in the patient derived iPS cells. An alternative approach would be to use direct mRNA transfection to avoid the expression problems we are currently having with the TALEN expression.

Additional Methods

iPS cell transfection: 150,000 iPS cells on matrigel were transfected with 4 µg of DNA containing SBDS-specific TALEN pair A or B coding plasmids (1.5µg/each plasmid) or, in the case of mCreI-driven safe harbor insertion, an mCreI expression vector + 1 µg of the repair template using the Continuum transfection reagent discussed above. At 48 hrs after transfection, cells were reseeded at single cell density on matrigel in 60 mm dishes using 3 dilutions (1:50, 1:10 and remainder of population) to generate single cell-derived colonies for targeted mutation correction or transgene insertion.

A total of 25 colonies from TALEN pair A and 27 colonies from TALEN pair B were isolated, PCR-amplified to generate a 1.7kb fragment covering the endogenous SBDS locus, the cleaved with HpaI to reveal targeted repair events. None of the 52 colonies was positive.

A positive control standard to determine the sensitivity of detection of repair events should enable the rapid screening of pools of transfectants as opposed to individual colonies. This will speed the rapid screening of much larger numbers of transfected cells to focus on the subset of pools that contain putative repair events. Successful mutation-repaired iPS cells or iPS cells containing an *SBDS* transgene inserted into a safe harbor site will be recovered from positive pools by dilution cloning as

outlined above, then verified by DNA sequencing prior to biochemical characterization and differentiation assays.

Chapter 6: Conclusions and Continuing directions

Genome engineering as a whole has grown faster than imaginable over the course of the last five years. When I started work in the field, my focus was on a development of homing endonucleases as a genome engineering tool. The methods and strategies I employed when using homing endonucleases are generally applicable to the newer genome engineering tools as well. Additionally, improvements in genomics provided an increasing set of organisms for genome engineering applications. The diversity of potential genome engineering targets will continue to require multiple nucleases as demonstrated by the several different approaches described in this dissertation.

There is a clear set of next steps that can be followed to progress each project described in this thesis. First, regarding the development of the algal genomics project, *Chrysochromulina tobin* now has a draft genome that was completed due to my efforts. With this genome, I have identified new genes and metabolic pathways of interest that are of considerable interest for genome engineering applications. For example, overexpression of TAG synthesis enzymes, or perturbation of one or more lipase genes could provide a method for increasing high value lipid production or storage in this alga. However, in order to perform the desired genome engineering, a transformation system is still required (additional details in Appendix 2). Transformation in *C. tobin* is almost realized in my opinion, though I suggest utilizing a new selection plasmid for detection of drug resistance, and/or a fluorescent signal. The plasmid should be built using an endogenous promoter from the *C. tobin* genome now that this information is available. Electroporation methods for *C. tobin* are the furthest advanced at this point and should be pursued. At this point, increasing throughput on electroporation experiments using the conditions I developed (Appendix 2) would likely yield transformants.

Regarding the *SBDS* gene repair project and *Anopheles* projects, we seem to have pushed to the limits of engineering homing endonuclease enzymes using the degeneracy and single design

methods described in both chapters 3 and 4. As described, engineering changes and degeneracy are tolerated at the single base level as well and combinatorial, however, multiple engineering changes, especially in adjacent positions, cause a significant loss of activity. Two solutions should help improve the outcome of homing endonuclease specific genome engineering projects. 1) Discovery and characterization of more homing endonucleases; which will increase the starting library of materials for genome targets of interest and potentially allow for more heterodimer fusions. 2) Improved methods for homing endonuclease selections. Yeast selection systems have previously been described, but are only viable for some HEs. A phage assisted continuous evolution system as described by Esvelt et al 2010, may be a better method for improving homing endonuclease targeting for a novel target choice.

I have also described novel safe harbor sites that can be utilized immediately using existing genome engineering reagents. Safe harbor insertion provides an alternative approach for gene repair problems, as highlighted by the Shwachman-Diamond syndrome project. To utilize this approach however, proper promoter selection is required which is a practical next step to follow up with.

Finally, genome engineering tools can be repurposed for functions other than just DNA cleavage. Appendix 3 describes work I did on creating epigenetic modification tools. Such tools utilize the targeting function of existing genome engineering tools and link targeting to other biochemical functions such as DNA methylation or histone modification. These novel tools will expand our ability to go beyond engineering DNA base pair content alone.

References:

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinforma. Oxf. Engl. *21*, 2104–2105.

Adams, C.C., and Stern, D.B. (1990). Control of mRNA stability in chloroplasts by 3' inverted repeats: effects of stem and loop mutations on degradation of psbA mRNA in vitro. Nucleic Acids Res. *18*, 6003–6010.

Aldrich, J., Cherney, B., Merlin, E., Williams, C., and Mets, L. (1985). Recombination within the inverted repeat sequences of the Chlamydomonas reinhardii chloroplast genome produces two orientation isomers. Curr. Genet. *9*, 233–238.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Arnould, S., Delenda, C., Grizot, S., Desseaux, C., Pâques, F., Silva, G.H., and Smith, J. (2011). The I-CreI meganuclease and its engineered derivatives: applications from cell modification to gene therapy. Protein Eng. Des. Sel. PEDS *24*, 27–31.

Aron, Z.D., Dorrestein, P.C., Blackhall, J.R., Kelleher, N.L., and Walsh, C.T. (2005). Characterization of a new tailoring domain in polyketide biogenesis: the amine transferase domain of MycA in the mycosubtilin gene cluster. J. Am. Chem. Soc. *127*, 14986–14987.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science *289*, 905–920.

Baurain, D., Brinkmann, H., Petersen, J., Rodríguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A.J., Burger, G., Lang, B.F., and Philippe, H. (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol. Biol. Evol. *27*, 1698–1709.

Baxter, S.K., Lambert, A.R., Scharenberg, A.M., and Jarjour, J. (2013). Flow cytometric assays for interrogating LAGLIDADG homing endonuclease DNA-binding and cleavage properties. Methods Mol. Biol. Clifton NJ *978*, 45–61.

Belfort, M., and Perlman, P.S. (1995). Mechanisms of intron mobility. J. Biol. Chem. *270*, 30237–30240.

Bennett, S. (2004). Solexa Ltd. Pharmacogenomics *5*, 433–438.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bigelow, N.W., Hardin, W.R., Barker, J.P., Ryken, S.A., Macrae, A.C., and Cattolico, R.A. (2011). A Comprehensive GC-MS Sub-Microscale Assay for Fatty Acids and its Applications. J. Am. Oil Chem. Soc. *88*, 1329–1338.

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E.S., Santini, S., Ogata, H., Probert, I., Edvardsen, B., and de Vargas, C. (2013). Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol. Ecol. *22*, 87–101.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., and Galaxy Team (2010). Manipulation of FASTQ data with Galaxy. Bioinforma. Oxf. Engl. *26*, 1783–1785.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science *326*, 1509–1512.

Boissel, S., Jarjour, J., Astrakhan, A., Adey, A., Gouble, A., Duchateau, P., Shendure, J., Stoddard, B.L., Certo, M.T., Baker, D., et al. (2014). megaTALs: a rare-cleaving nuclease architecture for therapeutic genome engineering. Nucleic Acids Res. *42*, 2591–2601.

Boocock, G.R.B., Morrison, J.A., Popovic, M., Richards, N., Ellis, L., Durie, P.R., and Rommens, J.M. (2003). Mutations in SBDS are associated with Shwachman-Diamond syndrome. Nat. Genet. *33*, 97–101.

Boysen, R.I., and Hearn, M.T. (2001). The metal binding properties of the CCCH motif of the 50S ribosomal protein L36 from Thermus thermophilus. J. Pept. Res. Off. J. Am. Pept. Soc. *57*, 19–28.

Brugmans, L., Kanaar, R., and Essers, J. (2007). Analysis of DNA double-strand break repair pathways in mice. Mutat. Res. *614*, 95–108.

Burki, F., Shalchian-Tabrizi, K., and Pawlowski, J. (2008). Phylogenomics reveals a new "megagroup" including most photosynthetic eukaryotes. Biol. Lett. *4*, 366–369.

Burroughs, L., Woolfrey, A., and Shimamura, A. (2009). Shwachman-Diamond syndrome: a review of the clinical presentation, molecular pathogenesis, diagnosis, and treatment. Hematol. Oncol. Clin. North Am. *23*, 233–248.

Cai, Z., Guisinger, M., Kim, H.-G., Ruck, E., Blazier, J.C., McMurtry, V., Kuehl, J.V., Boore, J., and Jansen, R.K. (2008). Extensive Reorganization of the Plastid Genome of Trifolium subterraneum (Fabaceae) Is Associated with Numerous Repeated Sequences and Novel DNA Insertions. J. Mol. Evol. *67*, 696–704.

Cattolico, R.A., Jacobs, M.A., Zhou, Y., Chang, J., Duplessis, M., Lybrand, T., McKay, J., Ong, H.C., Sims, E., and Rocap, G. (2008). Chloroplast genome sequencing analysis of Heterosigma akashiwo CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. BMC Genomics *9*, 211.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res. *39*, e82.

Certo, M.T., Gwiazda, K.S., Kuhar, R., Sather, B., Curinga, G., Mandt, T., Brault, M., Lambert, A.R., Baxter, S.K., Jacoby, K., et al. (2012). Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. Nat. Methods *9*, 973–975.

Chen, C., and Okayama, H. (1987). High-efficiency transformation of mammalian cells by plasmid DNA. Mol. Cell. Biol. *7*, 2745–2752.

Cheng, L., Blazar, B., High, K., and Porteus, M. (2011). Zinc fingers hit off target. Nat. Med. *17*, 1192–1193.

Chevalier, B.S., Kortemme, T., Chadsey, M.S., Baker, D., Monnat, R.J., and Stoddard, B.L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. Mol. Cell *10*, 895–905.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinforma. Oxf. Engl. *21*, 3674–3676.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messié, M., Tringe, S.G., Woyke, T., Welsh, R.M., Ishoey, T., Lee, J.-H., et al. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc. Natl. Acad. Sci. U. S. A. *107*, 14679–14684.

Darii, M.V., Cherepanova, N.A., Subach, O.M., Kirsanova, O.V., Raskó, T., Ślaska-Kiss, K., Kiss, A., Deville-Bonne, D., Reboud-Ravaux, M., and Gromova, E.S. (2009). Mutational analysis of the CG recognizing DNA methyltransferase SssI: Insight into enzyme–DNA interactions. Biochim. Biophys. Acta BBA - Proteins Proteomics *1794*, 1654–1662.

DeKelver, R.C., Choi, V.M., Moehle, E.A., Paschon, D.E., Hockemeyer, D., Meijsing, S.H., Sancak, Y., Cui, X., Steine, E.J., Miller, J.C., et al. (2010). Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. Genome Res. *20*, 1133–1142.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636–4641.

Delihas, N. (2011). Impact of Small Repeat Sequences on Bacterial Genome Evolution. Genome Biol. Evol. *3*, 959–973.

Doyon, J.B., Pattanayak, V., Meyer, C.B., and Liu, D.R. (2006). Directed evolution and substrate specificity profile of homing endonuclease I-SceI. J. Am. Chem. Soc. *128*, 2477–2484.

Drager, R.G., Zeidler, M., Simpson, C.L., and Stern, D.B. (1996). A chloroplast transcript lacking the 3' inverted repeat is degraded by 3'-->5' exoribonuclease activity. RNA *2*, 652–663.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edvardsen, B., Eikrem, W., Green, J.C., Andersen, R., Moon-van der Staay, S.Y., and Medlin, L. (2000). Phylogenetic reconstructions of the Haptophyta inferred from 18S ribosomal DNA sequences and available morphological data. Phycologia *39*, 19–35.

Emmanuel D Ladoukakis, A.E.-W. (2008). The excess of small inverted repeats in prokaryotes. J. Mol. Evol. *67*, 291–300.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Feng, S., Xue, L., Liu, H., and Lu, P. (2009). Improvement of efficiency of genetic transformation for Dunaliella salina by glass beads method. Mol. Biol. Rep. *36*, 1433–1439.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science *281*, 237–240.

Fisch, K.M. (2013). Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. RSC Adv. *3*, 18228–18247.

Ganapathi, K.A., Austin, K.M., Lee, C.-S., Dias, A., Malsch, M.M., Reed, R., and Shimamura, A. (2007). The human Shwachman-Diamond syndrome protein, SBDS, associates with ribosomal RNA. Blood *110*, 1458–1465.

Gelderblom, W.C., Thiel, P.G., Jaskiewicz, K., and Marasas, W.F. (1986). Investigations on the carcinogenicity of fusarin C--a mutagenic metabolite of Fusarium moniliforme. Carcinogenesis *7*, 1899–1901.

Green, B.R. (2011). After the primary endosymbiosis: an update on the chromalveolate hypothesis and the origins of algae with Chl c. Photosynth. Res. *107*, 103–115.

Grishin, A., Fonfara, I., Alexeevski, A., Spirin, S., Zanegina, O., Karyagina, A., Alexeyevsky, D., and Wende, W. (2010). Identification of conserved features of LAGLIDADG homing endonucleases. J. Bioinform. Comput. Biol. *8*, 453–469.

Haberle, R.C., Fourcade, H.M., Boore, J.L., and Jansen, R.K. (2008). Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. J. Mol. Evol. *66*, 350–361.

Han, C.S., and Chain, P. Finishing Repetitive Regions Automatically with.

Heath, P.J., Stephens, K.M., Monnat, R.J., and Stoddard, B.L. (1997). The structure of I-CreI, a group I intron-encoded homing endonuclease. Nat. Struct. Biol. *4*, 468–476.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

Hu, Q., Sommerfeld, M., Jarvis, E., Ghirardi, M., Posewitz, M., Seibert, M., and Darzins, A. (2008). Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J. Cell Mol. Biol. *54*, 621–639.

Irion, S., Luche, H., Gadue, P., Fehling, H.J., Kennedy, M., and Keller, G. (2007). Identification and targeting of the ROSA26 locus in human embryonic stem cells. Nat. Biotechnol. *25*, 1477–1482.

Jacobs, M.A., Connell, L., and Cattolico, R.A. (1999). A conserved His-Asp signal response regulator-like gene in Heterosigma akashiwo chloroplasts. Plant Mol. Biol. *41*, 645–655.

Jacoby, K., Metzger, M., Shen, B.W., Certo, M.T., Jarjour, J., Stoddard, B.L., and Scharenberg, A.M. (2012). Expanding LAGLIDADG endonuclease scaffold diversity by rapidly surveying evolutionary sequence space. Nucleic Acids Res. *40*, 4954–4964.

John, U., Beszteri, S., Glöckner, G., Singh, R., Medlin, L., and Cembella, A.D. (2010). Genomic characterisation of the ichthyotoxic prymnesiophyte Chrysochromulina polylepis, and the expression of polyketide synthase genes in synchronized cultures. Eur. J. Phycol. *45*, 215–229.

Jurica, M.S., Monnat, R.J., and Stoddard, B.L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. Mol. Cell *2*, 469–476.

Karol, K., Jacobs, M.A., Zhou, Y., Sims, E., Gillett, W., and Cattolico, R.A. Comparative analysis of complete mitochondrial genome sequences from two geographically distinct Heterosigma akashiwo (Raphidophyceae) strains. Proc. Seventh Int. Chrysophyte Symp. 261–282.

Keeling, P.J. (2004). Diversity and evolutionary history of plastids and their hosts. Am. J. Bot. *91*, 1481–1493.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005). The tree of eukaryotes. Trends Ecol. Evol. *20*, 670–676.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, L. and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics *18*, 1427–1431.

Li, H., Pellenz, S., Ulge, U., Stoddard, B.L., and Monnat, R.J. (2009). Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. Nucleic Acids Res. *37*, 1650–1662.

Li, H., Ulge, U.Y., Hovde, B.T., Doyle, L.A., and Monnat, R.J. (2012). Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. Nucleic Acids Res. *40*, 2587–2598.

Li, L., Krymskaya, L., Wang, J., Henley, J., Rao, A., Cao, L.-F., Tran, C.-A., Torres-Coronado, M., Gardner, A., Gonzalez, N., et al. (2013). Genomic editing of the HIV-1 coreceptor CCR5 in adult hematopoietic stem and progenitor cells using zinc finger nucleases. Mol. Ther. J. Am. Soc. Gene Ther. *21*, 1259–1269.

Liu, H., Probert, I., Uitz, J., Claustre, H., Aris-Brosou, S., Frada, M., Not, F., and de Vargas, C. (2009). Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proc. Natl. Acad. Sci. U. S. A. *106*, 12803–12808.

Lombardo, A., Genovese, P., Beausejour, C.M., Colleoni, S., Lee, Y.-L., Kim, K.A., Ando, D., Urnov, F.D., Galli, C., Gregory, P.D., et al. (2007). Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. Nat. Biotechnol. *25*, 1298–1306.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Maiya, S., Grundmann, A., Li, X., Li, S.-M., and Turner, G. (2007). Identification of a hybrid PKS/NRPS required for pseurotin A biosynthesis in the human pathogen Aspergillus fumigatus. Chembiochem Eur. J. Chem. Biol. *8*, 1736–1743.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature *437*, 376–380.

Mascher, T., Helmann, J.D., and Unden, G. (2006). Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases. Microbiol. Mol. Biol. Rev. *70*, 910–938.

Matthews, J.M., and Sunde, M. (2002). Zinc Fingers–Folds for Many Occasions. IUBMB Life *54*, 351–355.

Mausz, M.A., and Pohnert, G. (2014). Phenotypic diversity of diploid and haploid Emiliania huxleyi cells and of cells in different growth phases revealed by comparative metabolomics. J. Plant Physiol.

McDonald, S.., Sarno, D., Scanlan, D.., and Zingone, A. (2007). Genetic diversity of eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle. Aquat. Microb. Ecol. *50*, 75–89.

Medlin, L.K., Kooistra, W.H.C.F., Potter, D., Saunders, G.W., and Andersen, R.A. (1997). Phylogenetic relationships of the "golden algae" (haptophytes, heterokont chromophytes) and their plastids. In Origins of Algae and Their Plastids, D.D. Bhattacharya, ed. (Springer Vienna), pp. 187–219.

Medlin, L.K., Sáez, A.G., and Young, J.R. (2008). A molecular clock for coccolithophores and implications for selectivity of phytoplankton extinctions across the K/T boundary. Mar. Micropaleontol. *67*, 69–86.

Menne, T.F., Goyenechea, B., Sánchez-Puig, N., Wong, C.C., Tonkin, L.M., Ancliff, P.J., Brost, R.L., Costanzo, M., Boone, C., and Warren, A.J. (2007). The Shwachman-Bodian-Diamond syndrome protein mediates translational activation of ribosomes in yeast. Nat. Genet. *39*, 486–495.

Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y., Lee, Y.-L., Rupniewski, I., Beausejour, C.M., Waite, A.J., Wang, N.S., Kim, K.A., et al. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. Nat. Biotechnol. *25*, 778–785.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011). A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. *29*, 143–148.

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., and Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. *14*, 193–202.

Mongin, E., Louis, C., Holt, R.A., Birney, E., and Collins, F.H. (2004). The Anopheles gambiae genome: an update. Trends Parasitol. *20*, 49–52.

Monnat, R.J., Hackmann, A.F., and Cantrell, M.A. (1999). Generation of highly site-specific DNA double-strand breaks in human cells by the homing endonucleases I-PpoI and I-CreI. Biochem. Biophys. Res. Commun. *255*, 88–93.

Moret, B., and Warnow, T. Cyberinfrastructure for Phylogenetic Research (CIPRES) project.

Mukherjee, S., and Thrasher, A.J. (2013). Gene therapy for PIDs: progress, pitfalls and prospects. Gene *525*, 174–181.

Newman, S.M., Boynton, J.E., Gillham, N.W., Randolph-Anderson, B.L., Johnson, A.M., and Harris, E.H. (1990). Transformation of Chloroplast Ribosomal RNA Genes in Chlamydomonas: Molecular and Genetic Characterization of Integration Events. Genetics *126*, 875–888.

Oudot-Le Secq, M.-P., and Green, B.R. (2011). Complex repeat structures and novel features in the mitochondrial genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana. Gene *476*, 20–26.

Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E.V., Bowler, C., and Green, B.R. (2007). Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. Mol. Genet. Genomics MGG *277*, 427–439.

Palmer, J.D., and Thompson, W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell *29*, 537–550.

Panina, E.M., Mironov, A.A., and Gelfand, M.S. (2003). Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proc. Natl. Acad. Sci. U. S. A. *100*, 9912–9917.

Papapetrou, E.P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L.M.S., Kadota, K., Roth, S.L., Giardina, P., Viale, A., et al. (2011). Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. Nat. Biotechnol. *29*, 73–78.

Radakovits, R., Jinkerson, R.E., Darzins, A., and Posewitz, M.C. (2010). Genetic engineering of algae for enhanced biofuel production. Eukaryot. Cell *9*, 486–501.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

R Core Team (2014). R: A language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing).

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature *499*, 209–213.

Rice, D.W., and Palmer, J.D. (2006). An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. BMC Biol. *4*, 31.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. TIG *16*, 276–277.

Rocap, G., and McKay, C. The Stramenopile Chloroplast Genomics Project.

Rohde, C., Zhang, Y., Reinhardt, R., and Jeltsch, A. (2010). BISMA - Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. BMC Bioinformatics *11*, 230.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Rott, R., Liveanu, V., Drager, R.G., Stern, D.B., and Schuster, G. (1998). The sequence and structure of the 3′-untranslated regions of chloroplast transcripts are important determinants of mRNA accumulation and stability. Plant Mol. Biol. *36*, 307–314.

Sánchez Puerta, M.V., Bachvaroff, T.R., and Delwiche, C.F. (2004). The complete mitochondrial genome sequence of the haptophyte Emiliania huxleyi and its relation to heterokonts. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *11*, 1–10.

Sánchez Puerta, M.V., Bachvaroff, T.R., and Delwiche, C.F. (2005). The complete plastid genome sequence of the haptophyte Emiliania huxleyi: a comparison to other plastid genomes. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *12*, 151–156.

Shalchian-Tabrizi, K., Reier-Røberg, K., Ree, D.K., Klaveness, D., and Bråte, J. (2011). Marine-freshwater colonizations of haptophytes inferred from phylogeny of environmental 18S rDNA sequences. J. Eukaryot. Microbiol. *58*, 315–318.

Simon, M., López-García, P., Moreira, D., and Jardillier, L. (2013). New haptophyte lineages and multiple independent colonizations of freshwater ecosystems. Environ. Microbiol. Rep. *5*, 322–332.

Singh, A., Nigam, P.S., and Murphy, J.D. (2011). Renewable fuels from algae: an answer to debatable land based fuels. Bioresour. Technol. *102*, 10–16.

Smit, A., Hubley, R, and Green, P (2010). RepeatMasker Open-3.0.

Smith, D.R., and Keeling, P.J. (2012). Twenty-fold difference in evolutionary rates between the mitochondrial and plastid genomes of species with secondary red plastids. J. Eukaryot. Microbiol. *59*, 181–184.

Smith, D.R., Arrigo, K.R., Alderkamp, A.-C., and Allen, A.E. (2014). Massive difference in synonymous substitution rates among mitochondrial, plastid, and nuclear genes of Phaeocystis algae. Mol. Phylogenet. Evol. *71*, 36–40.

Sondergaard, T.E., Hansen, F.T., Purup, S., Nielsen, A.K., Bonefeld-Jørgensen, E.C., Giese, H., and Sørensen, J.L. (2011). Fusarin C acts like an estrogenic agonist and stimulates breast cancer cells in vitro. Toxicol. Lett. *205*, 116–121.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.

Starkenburg, S.R., Kwon, K.J., Jha, R.K., McKay, C., Jacobs, M., Chertkov, O., Twary, S., Rocap, G., and Cattolico, R.A. (2014). A pangenomic analysis of the Nannochloropsis organellar genomes reveals novel genetic variations in key metabolic genes. BMC Genomics *15*, 212.

Staunton, J., and Weissman, K.J. (2001). Polyketide biosynthesis: a millennium review. Nat. Prod. Rep. *18*, 380–416.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-Component Signal Transduction. Annu. Rev. Biochem. *69*, 183–215.

Stoddard, B.L. (2005). Homing endonuclease structure and function. Q. Rev. Biophys. *38*, 49–95.

Taylor, G.K., Petrucci, L.H., Lambert, A.R., Baxter, S.K., Jarjour, J., and Stoddard, B.L. (2012). LAHEDES: the LAGLIDADG homing endonuclease database and engineering server. Nucleic Acids Res. *40*, W110–W116.

Tesler, G. (2002). GRIMM: genome rearrangements web server. Bioinforma. Oxf. Engl. *18*, 492–493.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trentacoste, E.M., Shrestha, R.P., Smith, S.R., Glé, C., Hartmann, A.C., Hildebrand, M., and Gerwick, W.H. (2013). Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. Proc. Natl. Acad. Sci. U. S. A. *110*, 19748–19753.

Ulge, U.Y., Baker, D.A., and Monnat, R.J. (2011). Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. Nucleic Acids Res. *39*, 4330–4339.

Wang, L., Mao, Y., Kong, F., Li, G., Ma, F., Zhang, B., Sun, P., Bi, G., Zhang, F., Xue, H., et al. (2013). Complete Sequence and Analysis of Plastid Genomes of Two Economically Important Red Algae: Pyropia haitanensis and Pyropia yezoensis. PLoS ONE *8*.

Warner, J.R., and McIntosh, K.B. (2009). How common are extraribosomal functions of ribosomal proteins? Mol. Cell *34*, 3–11.

Windbichler, N., Papathanos, P.A., Catteruccia, F., Ranson, H., Burt, A., and Crisanti, A. (2007). Homing endonuclease mediated gene targeting in Anopheles gambiae cells and embryos. Nucleic Acids Res. *35*, 5922–5933.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhao, L., and Stoddard, B.L. (2014). Rapid determination of homing endonuclease DNA binding specificity profile. Methods Mol. Biol. Clifton NJ *1123*, 127–134.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Appendices

Appendix 1: Additional files from genomes

Supplementary Tables

Table A1.1: Inventory of algal inverted repeat sequences in chloroplast genomes

Organism	SNPs or indels found in 23S ribosomal subunit	SNPs or indels found in 16S ribosomal subunit	Inverted Repeat length†	Accession number
Haptophytes				
Chrysochromulina tobin	5	6	4656	This work
Emiliania huxleyi CCMP373	0	0	4674	NC_007288
Emiliania huxleyi CCMP1516	0	0	4868	JN022705
Phaeocystis antarctica	0	0	4674	JN117275
Phaeocystis globosa	8	0	4611	NC_021637
Pavlova lutheri	No Inverted Repeat			NC_020371
Stramenopiles				
Apedinella radians	2	2	4732	Unpublished*
Aureococcus anophagefferens	No Inverted Repeat			NC_012898
Aureoumbra lagunensis	No Inverted Repeat			NC_012903
Botrydium cystosum	0	0	4924	Unpublished*
Ectocarpus siliculosus	0	0	8616	NC_013498
Fistulifera sp. JPCC DA0580	0	0	12031	NC_015403
Fucus vesiculosus	0	0	5242	NC_016735
Heterosigma akashiwo	0	0	21665	NC_010772
Nannochloropsis gaditana	0	0	5109	NC_020014
Nannochloropsis oculata	4	5	7541	Unpublished*
Nannochloropsis salina	0	0	5131	Unpublished*
Nereocystis luetkeana	0	0	5416	Unpublished*
Odontella sinensis	0	0	7725	NC_001713
Pelagomonas calceolata	No Inverted Repeat			Unpublished*
Phaeodactylum tricornutum	0	0	6916	NC_008588
Pinguiococcus pyrenoidosus	0	0	5070	Unpublished*
Saccharina japonica	5	1	5312	NC_018523
Synedra acus	1	0	6795	NC_016731
Synura petersenii	0	0	22465	Unpublished*
Thalassiosira oceanica CCMP1005	0	0	23698	NC_014808
Thalassiosira pseudonana	0	0	18345	NC_008589
Tribonema aequale	0	3	5749	Unpublished*
Vaucheria litorea	0	0	4935	NC_011600
Cryptophytes				
Cryptomonas paramecium	No Inverted Repeat			NC_013703
Guillardia theta	3	2	4922	NC_000926
Rhodomonas salina	6	2	4959	NC_009573

Chlorophytes				
Acutodesmus obliquus	0	0	12023	NC_008101
Bryopsis hypnoides	No Inverted Repeat			NC_013359
Chlamydomonas renhardii	0	0	22211	NC_005353
Chlorella variabilis	No Inverted Repeat			NC_015359
Chlorella vulgaris	No Inverted Repeat			NC_001865
Coccomyxa subellipsoidea C-169	No Inverted Repeat			NC_015084
Dunaliella salina	0	0	14409	NC_016732
Floydiella terrestris	No Inverted Repeat			NC_014346
Gonium pectorale	0	0	14750	NC_020438
Leptosira terrestris	No Inverted Repeat			NC_009681
Micromonas sp. RCC299	0	0	7307	NC_012575
Monomastix sp. OKE-1	No Inverted Repeat			NC_012101
Nephroselmis olivacea	0	0	46137	NC_000927
Oedogonium cardiacum	0	0	35492	NC_011031
Oltmannsiellopsis viridis	0	0	18510	NC_008099
Ostreococcus tauri	1	0	6825	NC_008289
Parachlorella kessleri	0	0	10913	NC_012978
Pedinomonas minor	0	0	10639	NC_016733
Picochlorum sp.	No Inverted Repeat			Unpublished*
Pleodorina starrii	0	0	16608	NC_021109
Pseudendoclonium akinetum	0	0	6110	NC_008114
Pycnococcus provasolii	No Inverted Repeat			NC_012097
Pyramimonas parkeae	0	0	12865	NC_012099
Schizomeris leibleinii	No Inverted Repeat			NC_015645
Stigeoclonium helveticum	No Inverted Repeat			NC_008372
Trebouxiophyceae sp. MX-AZ01	No Inverted Repeat			NC_018569
Rhodophytes				
Calliarthron tuberculosum	No Inverted Repeat			NC_021075
Chondrus crispus	No Inverted Repeat			NC_020795
Cyanidioschyzon merolae strain 10D	No Inverted Repeat			NC_004799
Cyanidium caldarium	No Inverted Repeat			NC_001840
Gracilaria tenuistipitata var. liui	No Inverted Repeat			NC_006137
Grateloupia taiwanensis	No Inverted Repeat			NC_021618
Porphyra purpurea	11	20	4827	NC_000925
Pyropia haitanensis	1	7	4828	NC_021189
Pyropia yezoensis	9	6	4827	NC_007932
Euglenoids				
Euglena gracilis	0	0	6127	NC_001603
Euglena viridis	No Inverted Repeat			NC_020460
Eutreptiella gymnastica	No Inverted Repeat			NC_017754
Monomorphina aenigmatica	No Inverted Repeat			NC_020018
Streptophytes				
Chaetosphaeridium globosum	0	0	12431	NC_004115
Chara vulgaris	0	0	10919	NC_008097
Chlorokybus atmophyticus	0	0	7640	NC_008822
Mesostigma viride	0	0	6056	NC_002186

Staurastrum punctulatum	No Inverted Repeat	NC_008116
Zygnema circumcarinatum	No Inverted Repeat	NC_008117

†Large repeats containing the ribosomal operon in chloroplast genomes were queried for size and sequence homology between the 16S and 23S ribosomal subunits. The three organisms *Cyanidioschyzon merolae (*Rhodophyta), *Bryopsis hypnoides* (Chlorophyta) *and Monomorphina aenigmatica* (Chlorophyta) were omitted from the small repeat statistical analyses because of the presence of a greatly expanded tandem repeat structure or large (>200 bp) repeat structures, which causes an over representation the repeat structure due to the counting nature of the analyses presented.
*Cattolico RA, Jacobs M, Rocap G. unpublished genomic data

Appendix 2: Transformation methods for *Chrysochromulina tobin*

Introduction:

In order to genetically modify an organism a viable method for nucleic acid transfer to the organism is required. While characterizing the genome of *Chrysochromulina tobin*, I was also attempting to develop a transformation method for *C. tobin*. The following transformation methods were developed and tested after determining a variety of drug tolerances and sensitivities, a semi-solid plating method, and constructing multiple transformation vectors for *C. tobin*. Transformation methods attempted were: Electroporation, particle bombardment, glass bead and bacterial feeding. These protocols were based on published methods for other algal species, though no haptophyte has yet been transformed. No successful *C. tobin* transformants has been detected at the time of this writing. Many of these methods were developed between myself and Ryan Sinit, an undergraduate research assistant.

Drug selections:

Drug sensitivity of *C. tobin* was useful for a variety of reasons including monoculture development and transformant selection. The following drugs were tested for growth inhibition on *C. tobin*:

	Antibiotic	Zeocin	Puromycin	Chloram-phenicol	G418	Strepto-mycin	Spectin-omycin	Blasticidin	Hygro-mycin
Drug Concentration (µg/mL)									
1000		X	X	X	X	X	-	+	+
500		X	X	X	X	+	-	+	+
250		X	X	-	-	+	-	+	+
125		X	X	+	-	+	-	+	+
62.5		+	-	+	+	+	-	+	+
31.25		+	+	+	+	-	+	+	+
15.625		+	+	+	+	-	+	+	+
0		+	+	+	+	+	+	+	+

Table A2.1: The antibiotic treatment effects on *C. tobin* culture growth. X = total kill, - = growth inhibition, + = no observed effect.

Cells were cultured in a 100mL to a cell density ~2 x 10^6 cells/mL in RAC1 media. 100,000 cells were plated in each well in a total of 175 µL of RAC1, the plate was a sterile 96 well round bottom plate. Upon visual inspection after 2 days, the cells had settled to the bottom of the wells and I transferred 150 µL of cells to a sterile flat bottom plate for better separation. 2 hours after transfer, 50 µl of the listed antibiotics were added to each well to make the defined final concentration. Cells were visualized under light microscopy to determine the overall health of each individual well.

Follow up testing was performed with larger cultures (100 mL) and growth curve determination over multiple days. Zeocin, Puromycin, G418, Hygromycin, Streptomycin and Chloramphenicol were all tested to determine if antibiotic concentrations estimated to be tolerated in the small scale experiment would allow cell proliferation. In the end, the two antibiotics that allowed for the least impeded growth were Hygromycin and Streptomycin, each which were tolerated by *C. tobin* at a concentration of 400 µg/mL in RAC-1.

Transformation vectors:

A dual purpose transformation vector was developed for testing various transformation methods. The vector was derived from an EGFP vector that initially contained a 35s promoter for the eGFP and SV40 promoter for the Kanamycin resistance which also acts as a G418 resistance cassette. A variety of alternative promoters were tested to drive the eGFP cassette including SV40, CaMV (cauliflower mosaic virus) and a endogenous ubiquitin promoter from *Chrysochromulina* sp. Using this vector, flow cytometry could be used to screen large numbers of cells for GFP signal, or G418 could be used as a selection marker.

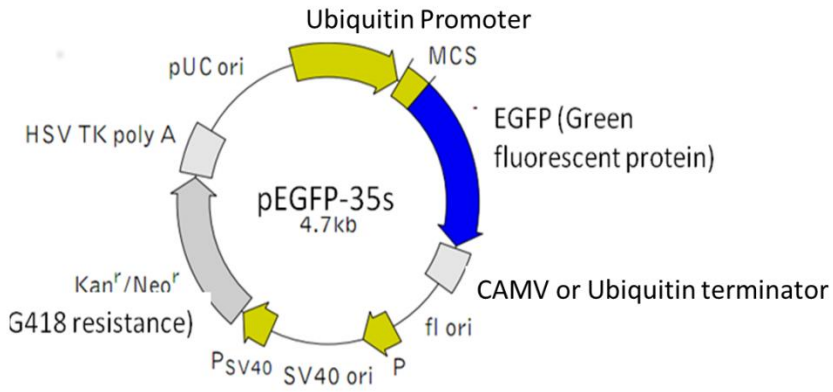


Figure A2.1: transformation plaid used in *C. tobin* experimentation

Chrysochromulina Plating Protocol:

A semi-solid plating method for *Chrysochromulina* sp. was developed in order to allow for drug selection based screens of transformation.

1) Making 1% Agarose
   I) Calculate the amount of 1% agarose necessary to get desired concentration of agarose per plate
      - We have observed that the best brand for this is SeaKem LE Agarose (Cat No 50004), but the difference is minimal.
   II) In an appropriately sized media container, add the media and agarose and autoclave on the liquid cycle for 30 minutes
   III) After taking the agarose out of the autoclave, keep it in a 55° water bath to keep the agarose in a liquid form.
      - For future use: It is safe to make a 100 mL stock of agarose and keep it for another experiment. However, it should only be used once more after your initial experiment (as the gel will get too dry and not be quite 1%). The best way to melt the agarose is by using a microwave for 45 -90 seconds (depending on volume).
2) Preparing Cell/Media Solution
   I) Use the flow cytometer to get an estimate of the cell concentration
   II) Calculate the volume of cells needed to be added to media from stock culture to obtain target cell concentration
      - Use $C_1V_1 = C_2V_2$ to get the appropriate volume necessary. Ex) With a reading of 3,000,000 cells per 33 µL and a target volume of 600 mL (for 30 plates), only about 0.197 mL of the cells from the stock solution is necessary.
      - With the volume of 20 mL per plate in mind, calculate the volume of cell/media solution you need to make. Ex) For 10 plates at 20 mL per plate you'll need 200 mL of solution
   III) Add the appropriate amount of cells from the stock solution to the measured out media.
      - Keep at RT
   IV) Aliquot into 125 mL flasks for easier pouring
      - Measure out the appropriate volume at this point. Ex) For a 0.08% concentration, add 92 mL of the cell/media solution
3) Pouring Plates
   I) Lay out 5 x 100 mm bacterial plates (20 mL per plate)

II) Add in the appropriate volume of agarose from the 1% solution (8 mL for 0.08%)
III) Swirl the flask
IV) Pour out each 100 mL aliquot into the 5 plates (estimating volume) using sterile technique
V) Cover the plates and let sit for 1 – 2 hours
VI) Parafilm the sides
VII) Move plates to incubation (20 C, 100 microeinsteins)

Transformation of *Chrysochromulina tobin*

Electroporation:

Electroporation seems like the most viable method at this point and continued experimentation using this method is warranted based on experiments using Cy3 labeled DNA oligonucleotides. The electroporation buffer that I optimized was a minimal RAC1 buffer which contains the following:

Minimal RAC1: 7.72 mM NaCL, .7mM CaCl2, .4mM KCl, 3 mM Tris-base pH 8.5.

This media does not include some of the reagents found in the regular RAC1 media in order to avoid increasing the amount of metal ions in the solution which affects electroporation.

Electroporation cell viability testing:

A variety of electroporation conditions have been tested and the best settings are found using approximately 2750 V/cm field strength for the electroporation. This setting gives a cell viability that is sufficiently high for cell growth, but high enough that approximately 30-60% cell death is achieved which is suggested as an ideal range for electroporation.

I used the Bio-Rad electroporator for all transformations except for the "long time constant" experiment, in which I used the Hitchcock Bio-rad electroporator II in the biology department. For this experiment, cells I used were 4 day post transfer P4 *Chrysochromulina* sp. at a concentration of 3.7 million cells/ml. I electroporated cells in half RAC1 and half Minimal RAC1 media after 5 minute centrifugation and resuspension (final concentration of cells ~6 million cells/mL).

DNA was added to a final concentration of 20 ng/µL. Cy3 labeled oligos were provided by Weiliang Tang of and had been previously used for RNAi transfection efficiency measurements. I used 2mm electroporation cuvettes. After electroporation, cells were allowed to recover for 10 minutes prior to transferring cells to 8 mL of RAC1 media

Cy3 has an excitation maximum at 550 nm but I am using a 488 nm laser for excitation which results in excitation ~20% of the maxima. The emission maximum for Cy3 is 570 nm, therefore, majority detection is via the FL2 Channel (585/40 nm) and also smaller detection level via FL1 (533/30 nm).

Loeb Lab (Bio-rad)	(Used 25 ohms resistance and a 1 mF capacitor)	
Voltage	Field Strength (V/cm)	Time constant (use)
Cy3 Oligos		
300	1500	4.9
500	2500	2.9
600	3000	2.5
700	3500	2.5
35s eGFP plasmid		
600	3000	2.4
No DNA control		
600	3000	2.4
Cy3 + no electroporation control		
0	0	-

Table A2.1: Electroporation conditions for the Cy3 dsDNA electroporation experiment.

Cells were detected using flow cytometry (Accuri C6) after electroporation's. Gating was done to isolate full size cells and then each set of cells were counted by raw value or Cy3 detection. Cy3 can be detected mainly by FL2 but also by FL1 at a lower level on the Accuri C6 flow cytometer. Cell survival was measured 1 hour after electroporation and 24 hours after electroporation to account for cells that were intact, but dead. In addition, the amount of Cy3 association with cells was also measured using the FL1 channel on the flow cytometer (Figure A2.2). It was determined that between 2500 and 3000 V/cm field strength yielded an ideal proportion of Cy3 associated cells and appropriate cell survival.

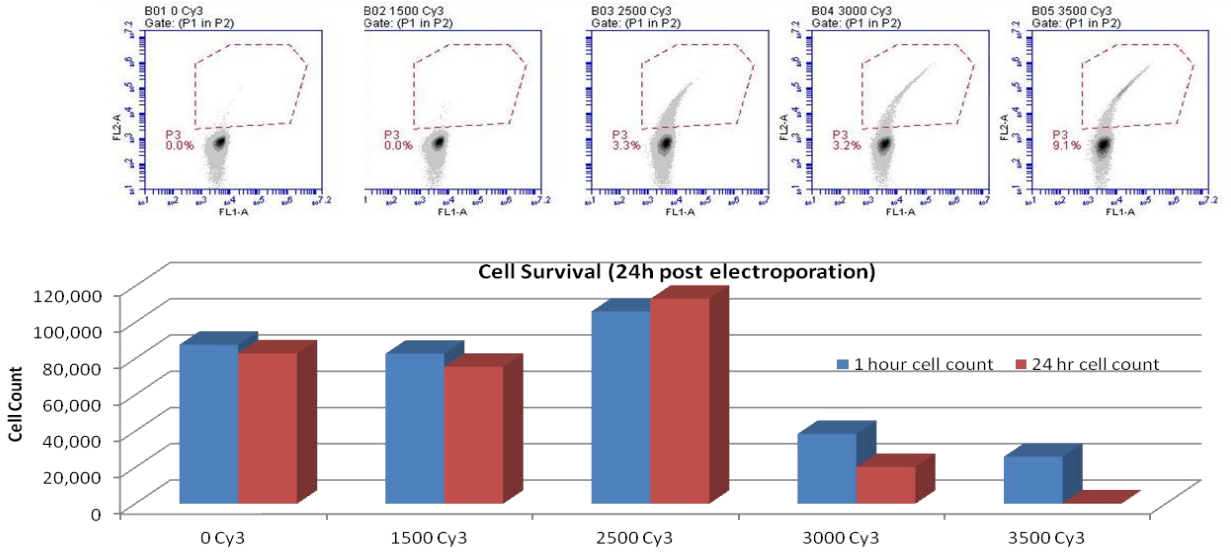


Figure A2.2: Cell associated Cy3 detection (top of figure) and cell survival at 1 hour and 24 hours post electroporation.

	All Cells		Reads of correct cell size		Cy3 Gated cells	
	Count	Volume (µL)	Count	% of This Plot	Count	% of This Plot
B01 0 Cy3	132,400	653	87,657	66.21%	34	0.04%
B02 1500 Cy3	129,135	653	82,716	64.05%	28	0.03%
B03 2500 Cy3	185,053	653	106,038	57.30%	3,525	3.33%
B04 3000 Cy3	111,226	653	38,503	34.62%	1,235	3.21%
B05 3500 Cy3	159,128	653	25,849	16.24%	2,347	9.09%
B06 Long TC Cy3	84,115	653	548	0.65%	543	99.82%
C02 3000 -DNA	71,075	326	27,000	37.99%	2	0.01%
C03 3000 EGFP	52,614	326	16,255	30.89%	0	0.00%

Table A2.2: Cell counts obtained from electroporation of *Chrysochromulina tobin* with Cy3 ssDNA oligos

Other methods:

Glass Bead Transformation Protocol: Adapted from (Feng et al., 2009)

-Centrifuge cells and resuspend in RAC-1 to a concentration of 1 x 10^5 cells/ml

-Aliquot 800 µL of cells into a 1.5 mL tube

-Add 300 mg of .5 mm glass beads, 100 µL of 20% PEG (Polyethylene glycol), and 50 µg of plasmid DNA

- Invert tube a few times to mix

-Vortex for 15 seconds @ 2000 rpm

-Allow beads to settle, and gently transfer cells to 25mL RAC1.

-Culture for 24 hrs before testing protein expression or drug resistance.

Biolistics (Particle bombardment):

1. Prepare Particles
    i. Weight out 60 mg of the microcarrier into a 1.7 Eppendorf
    ii. Add 1 mL of 70% EtOH
    iii. Vortex vigorously for 3-5 minutes on a platform vortexer
    iv. Allow the particles to soak in 70% EtOH for 15 minutes
    v. Pellet microparticles by spinning in a centrifuge for 5 seconds
    vi. Remove the supernatant
    vii. Wash three times:
        a. 1 mL of sterile water
        b. Vortex for 1 minute
        c. Allow the particles to settle for 1 minute
        d. Pellet the microparticles by briefly spinning in centrifuge
        e. Remove the supernatant
    viii. Resuspend in 1000 µL of 50% glycerol for a final concentration of 60 mg/mL. Store at -20°C until needed
    ix. Add 5 µL DNA suspension (1 µg/µL of TE buffer) and 50 µL of tungsten particle suspension (0.05 mg/mL of 50% glycerol) in a 1.7 Eppendorf tube
    x. Add 50 µL of 2.5 M $CaCl_2$ and 20 µL of 0.1 M spermidine free base to the Eppendorf tube
    xi. Incubate for 10 minutes at RT
    xii. Centrifuge for 30 seconds at 10,000 x g
    xiii. Remove supernatant
    xiv. Add 250 µL of 100% EtOH
    xv. Centrifuge for 30 seconds at 10,000 x g
    xvi. Remove 250 µL of supernatant
    xvii. Resuspend using 120 µL 100% EtOH
2. Prepare Samples (Use 4-5 day old cultures)
    i. Count cultures (for reference)
    ii. Preparation of the Liquid Samples
        a. Pour a thin layer of the liquid culture into 100 mm petri dish
        b. Let the plates sit for 1 hour
        c. Incubate at 20°C until ready to bombard
    iii. Preparation of the Semi-solid State Media Samples
        a. Autoclave a 1% agarose RAC1 mixture
        b. Keep the agarose mixture in a 55°C water bath after autoclaving
        c. Pour out 110.4 mL of the liquid culture into 125 mL flask
        d. Pipette 9.6 mL of the 1% agarose mixture into the 125 mL flask
        e. Swirl
        f. Pour into 6 plates (try to keep volume among all plates consistent)
        g. Incubate at 20°C until ready to bombard

iv. Preparation of the Biolistic System (http://www.bio-rad.com/LifeScience/pdf/Bulletin_9075.pdf)

3. Preparation of the Macrocarriers (directly prior to bombardment)
    i. Place the macrocarrier onto its holder
    ii. Pipette 10 µL of the suspended particle solution onto the middle of the macrocarrier
    iii. Place the macrocarrier assembly into a large petri dish containing desiccant and wait 5 minutes or until the EtOH has evaporated
4. Bombardment
    i. Bombardment of the Liquid Samples
        a. Place the macrocarrier assembly from before into its chamber
        b. Place the liquid culture on the sample holder
        c. Bombard
        d. Reset macrocarrier assembly with new rupture disk and macrocarrier
        e. Bombard a second time
        f. Incubate at 20°C
    ii. Bombarding of the Semi-solid State Media Samples
        a. Using a Pasteur pipette vacuum, remove the excess liquid on the sides of the plate
        b. Place the macrocarrier assembly into its chamber
        c. Place the semi-solid state sample on the sample holder
        d. Bombard
        e. Pour 10 mL of 20°C RAC1 media into the plate
        f. Incubate at 20°C
5. Selection
    i. pEGFP-C1 Selection
        a. No selection will be done with the pEGFP-C1 plasmid
    ii. pTG 1889 Selection
        a. Mix together X-Gluc reagents
        b. If a semi-solid plate, use a Pasteur pipette vacuum to remove some of the excess media solution
        c. Add the staining buffer to the sample
        d. Incubate at 20°C for 24 hours
        e. Remove the staining buffer
        f. Wash several times with 50% EtOH
6. Counting
    i. pEGFP-C1
        a. For semi-solid state samples
            I. Funnel the contents of the plate into a 50 mL Falcon tube
            II. Vortex the sample for 30 seconds to break up the agar
            III. Let the sample incubate for a day at 20°C
            IV. Decant using a cheese cloth filter into a test tube
            V. Use flow cytometry to determine if EGFP is present
        b. For liquid samples
            I. Use flow cytometry to determine if EGFP is present
    ii. pTG 1889
        a. For semi-solid state samples
            I. Count blue colonies after 3-4 days

b. For liquid samples
    I. Use a microscope to determine if there are any cells that are stained blue

Bacterial feeding:

Because *C. tobin* is known to feed on bacterial prey we postulated that feeding the alga that had the desired transformation vector in the bacteria to be phagocytized may be used as a transformation method, similar to what is done with RNA interference experiments performed on the nematode C. elegans.

This was attempted as a proof of concept using a GFP expressing E. coli (to detect bacterial ingestion) as well as a eukaryotic GFP vector containing E. coil in separate experiments. Under microscope examination however, it was observed that the E. coli cell size is actually too large for *C. tobin* ingestion. Therefore continued experiments using this method would require a smaller bacterial host to transform in order to subsequently feed them to *C. tobin*.

Other methods that were not attempted but might yield success:

LiAc Transformation, which is a method commonly used in yeast transformation. Because *C. tobin* is lacking a cell wall this poration might work.

Other heat shocking methods as Chrysochromulina cells are observed to swell when the temperature of the environment is significantly increased 5-10 degrees.

Appendix 3: Targeted epigenetic modification in human cells

Background:

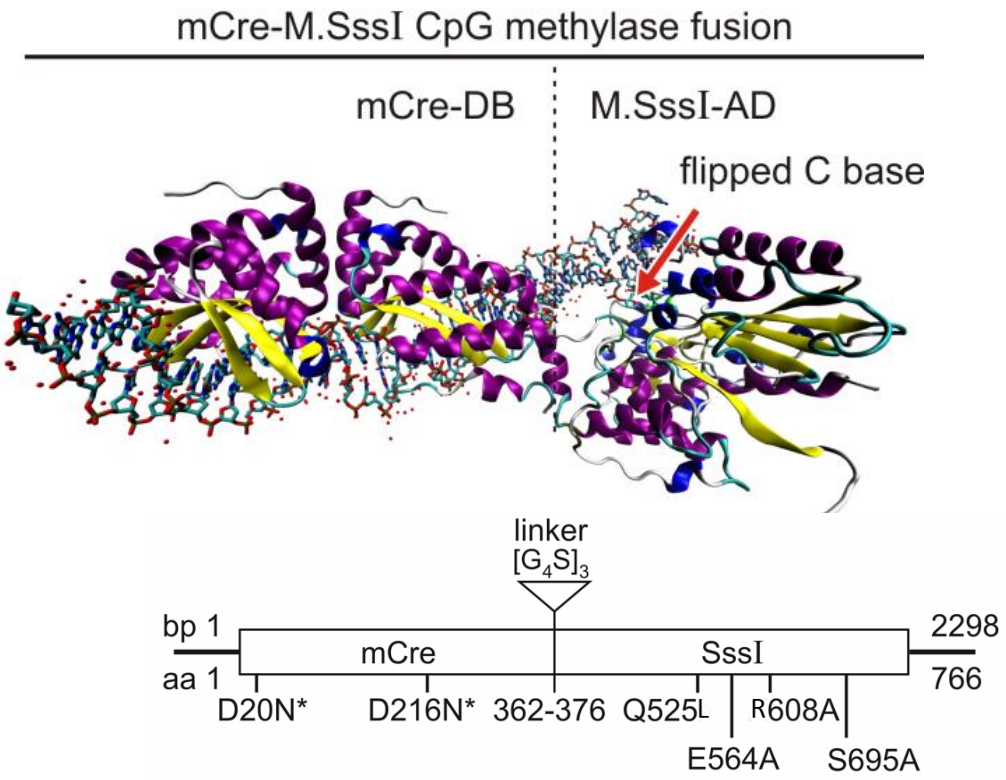
The epigenetic state of a gene or chromosomal region is determined by covalent DNA base and histone tail post-translational modifications that collectively regulate gene or locus-specific structure and function. This project addresses the need for new experimental tools to write, modify or erase specific epigenetic marks with a high degree of gene/locus-specificity. We are developing new tools that are general (i.e., not organism-specific), but gene-targeted. These tools can be built to target and modify a wide range of epigenetic modifications in many organisms, and have the potential to support epigenetic research in numerous fields of biology and medicine.

These new epigenomic tools consist of novel chimeric proteins, called epigenetic Mark-modifying Fusion proteins (or EMFs). EMFs are built from highly site-specific, programmable DNA binding (DB) modules with which we work, and epigenetic chemical activity domains (AD) derived from DNA-or histone methyltransferase/demethylase proteins. They are designed to write, modify or erase specific epigenetic chemical modifications, at pre-determined genomic target sites, with high activity and specificity. EMFs are general—and generalizable—reagents: thus they should catalyze new gene-targeted epigenetic research in a broad range of organisms.

We have demonstrated the feasibility of this general approach by building an EMF protein that writes site targeted DNA 5meC epigenetic marks in bacteria and in human cells using a fusion protein containing two domains. The first domain is that of mCre, which has been discussed in detail in the main text of this document. There is a specific change to the

Fusion protein design:

The fusion protein was constructed using a flexible linker sequence of GGGGS, repeated 3 times, to bind the C terminus of mCre to the N terminus of the M SssI protein.



	Methylation activity (nM/min)	DNA binding affinity (nM)
WT	2.62	42.4
"dead": R608A	none	51.3
"low" Q525L	0.199	532.4
"half": S695A	1.42	132.7
"2x low" Q525L+S695A	?	?

Figure A3.1: mCre-SssI fusion construct with SssI activity mutations. The top figure represents a theoretical structure of the mCre-SssI fusion protein. Below, mutations in the SssI portion of the fusion construct are derived from Darii et. al. 2009 (Darii et al., 2009), who characterized mutations in SssI that reduced the activity of the methyltransferase via methylations activity and DNA binding affinity. The mutations that we are using are shown on the left and naming corresponds to the activity level of the SssI portion of the fusion protein.

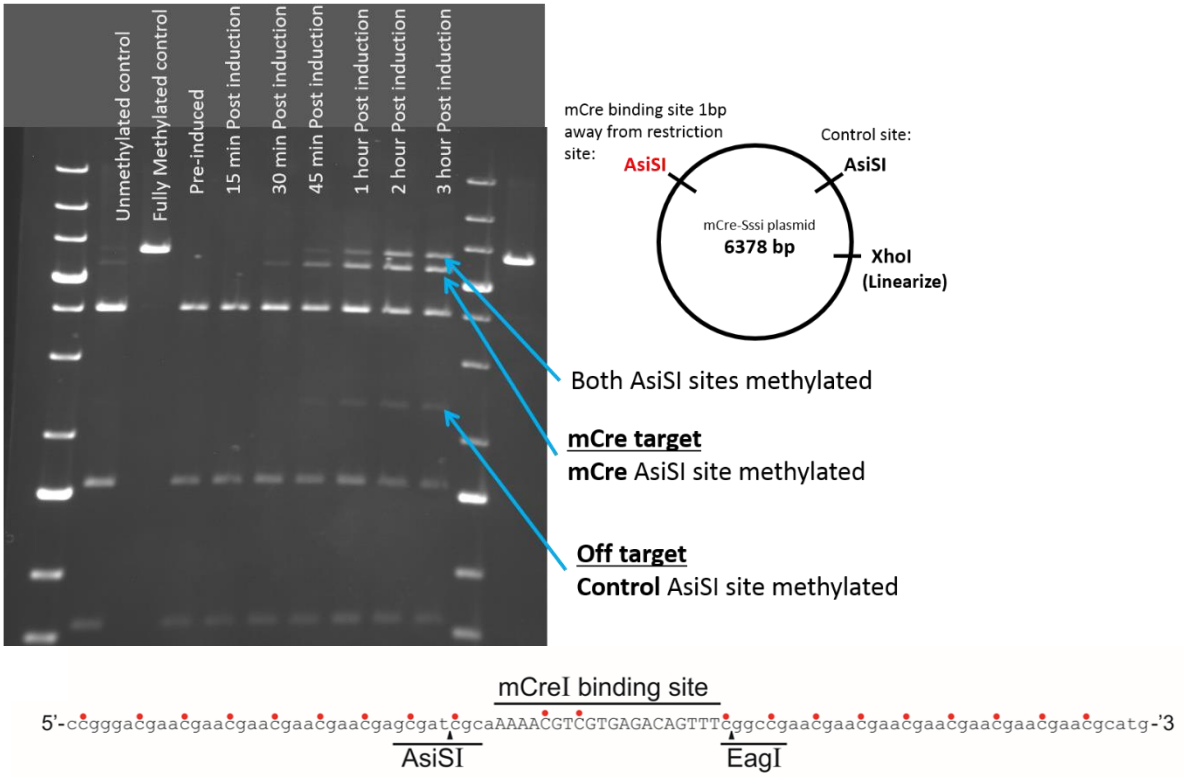


Figure A3.2: The mCre binding site (bottom) was inserted into the test plasmid resulting in an AsiSI restriction site located immediately adjacent to an mCre DNA binding site. AsiSI is methylation sensitive, so a methylated CpG in the AsiSI DNA recognition site prevents AsiSI cleavage, therefore different banding patterns occur in in-vitro testing depending on which CpGs are methylated in the test plasmid.

Minutes post induction (low activity Mutant SssI)	CpGs methylated near mCre target Site:	CpGs methylated near off target Site:
0	0.88%	0.58%
15	0.85%	0.00%
30	**4.55%**	0.35%
60	**13.70%**	**2.63%**
180	**20.37%**	**10.00%**
Minutes post induction (Double Mutant SssI)		
0	0.00%	**3.45%**
60	**5.92%**	**1.45%**
180	**5.28%**	**2.73%**

Table A3.1: In vitro targeting of methylation in the bacterial system.

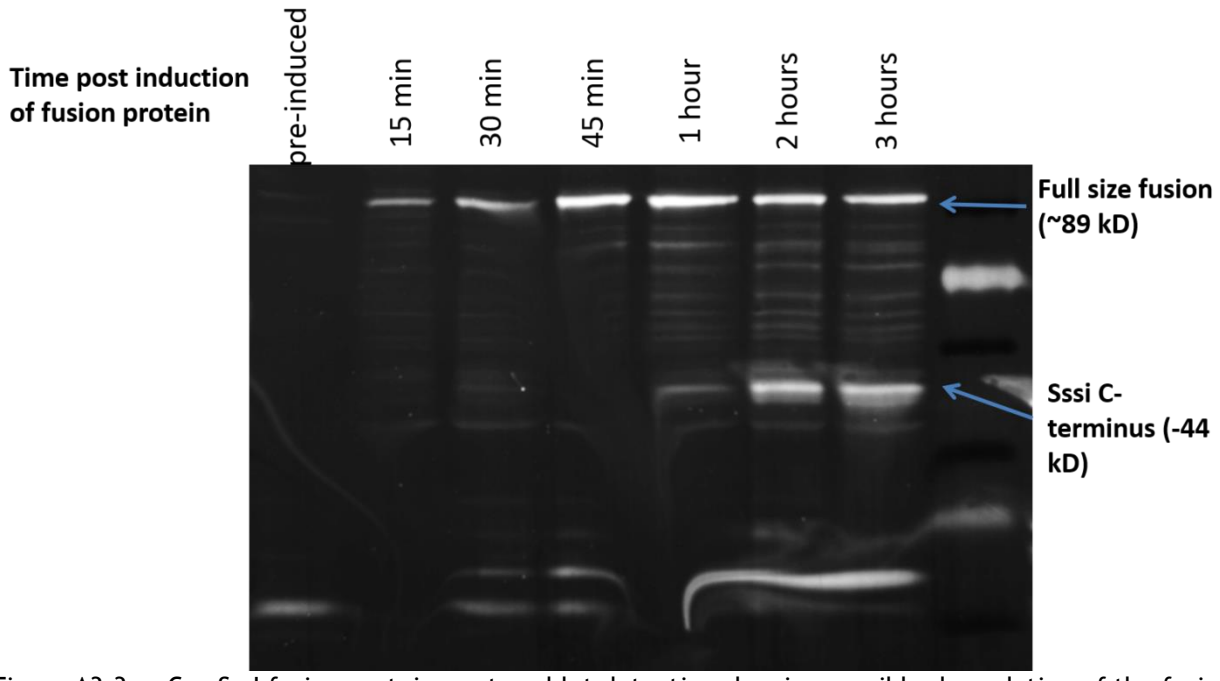


Figure A3.3: mCre-SssI fusion protein western blot detection showing possible degradation of the fusion protein product.

All the mutant M.SssI derivative fusion proteins showed a low level of expression at the expected molecular weight (~89kD) at all-time points of induction on SDS-PAGE gels stained by protein stains (GelCode Blue Stain reagent, ThermoScientific; SYPRO Ruby protein gel stain, Invitrogen), more sensitive than standard coomasie gel stains. However, because the methylation status of the plasmids purified from arabinose-induced cultures of two mCreI-mutant M.SssI derivatives (Q525L M.SssI and Q525L/S695A M.SssI) showed methylation specific for the mCreI DNA target site, it was still necessary to determine if the mutant fusion proteins being expressed in *E. coli* were intact, and not degraded by proteases. Western blot analysis of the Q525L M.SssI and Q525L/ S695A M.SssI mutant proteins using the mouse monoclonal anti-myc primary antibody showed no signal, whereas the positive control lysate supernatant pGlue-WRN with myc epitope tag at its N-terminus showed the expressed full-length product (~187kD) and endogenous WRN (~62kD). This result suggests that, even though the plasmid DNA sequences were verified to contain myc epitope tag sequence, the myc epitope tag at the C-terminus of the fusion proteins was degraded by proteases, and therefore the primary antibody did not react with the target protein. Western blot analysis of the Q525L M.SssI and Q525L/S695A M.SssI mutant fusion proteins using the rabbit anti-I-CreI whole protein primary antibody showed multiple bands in

the non-induced sample, suggesting that there is still background methylation even though glucose was added to growth. At earlier time points of induction (15, 30, 45, 60 minutes) of the Q525L M.SssI mutant, multiple bands were also observed, but with the target protein showing a pronounced full-length product (~89kD). A full-length product band is likely to be present at the later time points of induction (2h, 3h), but the Western blot transfer was not efficient. At later time points of induction (1h, 2h, 3h), in addition to the full-length product, smaller products (~65kD, 44kD) were observed, corresponding to either mCre ORF or M.SssI ORF. The positive control mCre purified protein showed an expected product (~40kD). At all-time points of induction of the Q525L/ S695A M.SssI mutant, multiple bands were also observed, but the target protein showed a pronounced full-length product (~89kD) and smaller products (~44kD) corresponding to either mCre ORF or M.SssI ORF.

This result suggests that, even though these mCre- mutant M.SssI fusion proteins have little methylation activity compared to mCre- wildtype M.SssI protein, and should therefore be less toxic to the cell, they are still not stable in E. coli and are being degraded.

***In vitro* bacterial methylation:**

Methylation status of plasmids isolated from *E. coli* cells producing the fusion products mCre-wild-type and mutant CpG methyltransferases (M.SssI) Q525L and Q525L/S695A analyzed on a 1% agarose gel in 1x TAE buffer. All samples were digested with *Asi*SI-*Xho*I restriction endonucleases Methylation of CG sites by M.SssI blocks *Asi*SI cleavage. Negative control is the plasmid encoding the mCre- wild-type M.SssI that is not methylated by CpG methyltransferase (M.SssI; NEB). Positive control is the plasmid encoding the mCre- wild-type M.SssI that is methylated by CpG methyltransferase (M.SssI; NEB).

Bisulfite conversion and DNA sequencing are useful in the analysis of 5-methylcytosine and cytosine in DNA. This method is based on the selective deamination of cytosine residues but not 5-methylcytosines by treatment with sodium bisulfite, followed by PCR and sequencing of single clones,

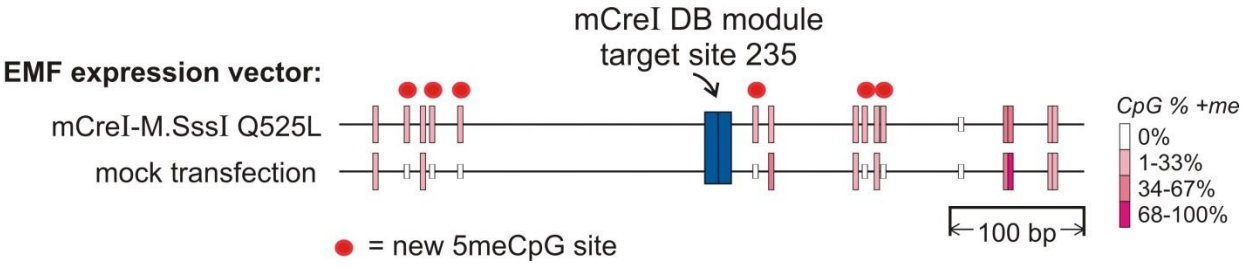
thereby accurately detecting the presence of 5-methylcytosine in the region of interest such as the target site and off-target site. It will determine the methylation specificity of the mCre-M.SssI mutant fusion proteins at the target site and off-target site. Reaction between cytosine and sodium bisulfite results in cytosine being converted to uracil, which is then replaced with thymine during PCR amplification, while methylated cytosine is retained and is thus maintained as C during amplification. The EZ DNA Methylation-Gold kit (Zymo Research) was used which consolidates DNA denaturation and bisulfite conversion in 1 step with >99% of non-methylated cytosine residues converted to uracil and >99% protection of methylated cytosines. Recovered bisulfite-converted DNA was then used for PCR amplification using primers specific for the modified DNA and "hot start" DNA polymerase (GemTaq, MGQuest) which modifies the PCR products by adding a 3'-terminal deoxyadenosine ("A-tailing") for ligation into the pGEM-T Easy vector (Promega). MethPrimer (Li and Dahiya, 2002) was the program used for designing bisulfite-conversion-based methylation PCR primers at the target site and off-target site. These primers have all non-CpG 'C's converted to 'T', contain no CpG 'C' sites in their sequences but span a region with the maximum number of CpG sites at the target site and off-target site. The amplified PCR product was then ligated into the linearized pGEM-T Easy vector with a single 3'-terminal thymidine at both ends in the *lacZ* gene, and clones were selected using blue/white screening on plates. Insertion of the PCR product in the *lacZ* gene of the pGEM-T Easy vector will disrupt active β-galactosidase fomation, resulting in white colonies. Sequencing of randomly selected white and light-blue colonies will give the methylation pattern for every CpG site in a long stretch of sequence (83 CpGs in 876bp PCR product at the target site; 28 CpGs in 552bp PCR product at the off-target site). Plasmid samples for bisulfite conversion, PCR amplification and sequencing included the unmethylated DNA (negative control), DNA methylated by M.SssI CpG methyltransferase (NEB) (positive control), plasmids purified from non-induced cultures and arabinose-induced cultures at 30 minutes and 3 hours from each mCreI-Q525L M.SssI mutant and mCre-Q525L/S695A M.SssI double mutants. BISMA (Bisulfite Sequencing DNA Methylation Analysis) (Rohde et al., 2010) program was used for analysis of bisulfite sequencing methylation data. The methylation percentage of each CpG site can be calculated according to the number of methylated and unmethylated cytosines in different clones. So far, clones

from the non-induced mCreI-Q525L M.SssI and mCreI-Q525L/S695A M.SssI mutants showed no methylation at the mCre binding site within the target region and off-target site. However, some clones have background methylation in the 65 CpGs of the 650bp PCR product at the target region and 29 CpGs of the 552 bp PCR product at the off-target region. Data past the 65th CpG site in the target region was not readable. At the 3-hour time point, clones from the arabinose-induced mCreI-Q525L M.SssI mutant showed 20% of the 65 CpG sites in the 650bp PCR product in the target region was methylated, while clones from the arabinose-induced mCre-Q525L/ S695A M.SssI mutant showed 4-fold less methylated CpG sites (5.3%). Clones from the arabinose-induced mCreI-Q525L M.SssI mutant showed 10% of the 29 CpG sites in the 552bp PCR product in the off-target region was methylated, while clones from the arabinose-induced mCre-Q525L/S695A M.SssI mutant also showed 4-fold less methylated CpG sites (2.7%). The consistency in the result suggests that the mCre-Q525L/S695A M.SssI double mutant has lower activity than the mCreI-Q525L M.SssI mutant. Also, the 2-fold more CpG sites being methylated at the target region than at the off-target region for both mCre-M.SssI mutants suggests that methylation occurred after induction of fusion protein by arabinose, and this methylation is directed by the low activity mutant plasmids, with its mCre homing endonuclease delivering the M.SssI to bind and methylate at the target region. Bisulfite-sequencing/PCR of the same low activity mCre-M.SssI variants at early time points of induction will be performed to determine the methylation specificity of these fusion proteins at the target site region compared to the off-target region.

To determine full-length protein expression and site-specific methylation activity in human cells, the mCre-M.SssI ORFs from 5 plasmids (mCre-wild type M.SssI, mCre-M.SssI mutants such as R608A, S695A, Q525L and Q525L/S695A) are being cloned into inducible vectors such as pFTSH-WRN and pLKO, both of which the protein expression is under control of a human CMV promoter. The pLKO vector contains a puromycin selection marker for selection of pools and clones. The 293 T-Rex Flp-in system will be used for transfection of the pFTSH/mCre-M.SssI vectors, allowing their integration at a specific site and selection of pools and clones using a hygromycin selection marker. Stable inducible

clones for immunoblotting can then be generated by induction of mCre-M.SssI ORF expression using tetracycline/ doxycycline.

Primers for PCR and cloning of the mCre-M.SssI ORF into the pFTSH-WRN vector at the BsrGI sites are oligonucleotide primer #1 (5'-
GGGCGATGTACAAAAAAGCAGGCTTCATGGACACCAAGTATAACAAGGAG
TTCCTGC -3') and oligonucleotide primer #2 (5'-
CGCATGTGTACACTAGGTACGCGTAGAATCGAGACCGAGGAGAGGGTTAG
GGATAGGCTTACCGGCGCGCCCACCTCCAATTTTATCTATAATCGCTTC -3'). The reverse oligonucleotide primer #2 contains a V5 epitope tag at the C-terminus for detection of fusion protein by immunoblotting. An additional epitope tag (hemagglutinin, HA) in the N-terminus of the pFTSH vector can also be used for detection of fusion protein by immunoblotting. The pFTSH-WRN vector will be digested with *BsrG*I, and the band of interest that does not contain the WRN ORF will be gel-purified, dephosphorylated and ligated to the PCR-amplified, *Bsr*GI-digested mCre-M.SssI products. The pLKO vector carries mCre-mCherry ORFs bounded by unique sites *Nhe*I and *Bsr*GI, which will be digested, and the band of interest that does not contain the mCre-mCherry ORFs will be ligated to the PCR-amplified, *Nhe*I-*Bsr*GI digested mCre-M.SssI products. Primers for PCR and cloning of the mCre-M.SssI ORF into the pLKO vector are oligonucleotide primer #1 (5'- CTAAgctagcATGGACACCAAGTATAACAAGGAGTTCCTGC -3') and oligonucleotide primer #2 (5'- CTCGCCTGTACAttaACCTCCAATTTTATCTATAATCGCTTC -3'). The hemagglutinin (HA) epitope tag in the N-terminus of the pLKO vector will be used for detection of fusion protein by immunoblotting.



	methyltransferases	*demethylases*
cytosine (C)	M.SssI	TET2
H3K27	KMT6/EZH2	KDM6A/UTX KDM6B/JMJD3

Table A3.2: Future constructs of interest to pursue that allow for histone modification and DNA de-methylation

Materials and Methods:

Polymerase chain reactions were carried out using the reaction conditions for *LA Taq* DNA polymerase (Takara). The plasmid pAIT2 was a generous gift from Bill Jack (New England Biolabs). Restriction enzymes and ligase were obtained from New England Biolabs and from Fermentas.

Construction of mCre SssI expression plasmids:

Complementary oligonucleotide primers and independent PCR reactions were used to generate two DNA fragments having overlapping ends that could be fused by combining them in a subsequent PCR reaction. In the first PCR, amplification of a DNA fragment from a plasmid template monomeric homing endonuclease CreI (or mCre) used the 5' external PCR primer (5'-cgccgccatggacaccaagtataacaaggagttcc -3') made up of a *Nco*I restriction site for cloning and a portion of the mCre sequence, and a 3' internal PCR primer (5'-cactaccgccaccgccagaaccgccgccaccgccggctggagagctcttcttcttctcc -3') that includes a portion of the flexible linker sequence $(GGGGS)_n$ (n = 3) which overlaps with that of the 5' internal PCR primer in the second PCR, a *Ngo*MIV restriction site and a portion of the mCre sequence. In the second PCR, the linker and M. SssI methyltransferase DNA fragment was amplified using the plasmid pAIT2 (from NEB)

which contains the M. SssI methyltransferase gene as a template, a 5' internal PCR primer (5' - gttctggcggtggcggtagtggtggtggcggctcgggcgccatgagcaaa gtagaaaataaaacaaaaaaac -3') that includes a portion of the flexible linker sequence $(GGGGS)_n$ (n = 3) which overlaps with that of the 3' external PCR primer in the first PCR, a *N*arI restriction site and a portion of the SssI methyltransferase sequence, and a 3' external PCR primer (5' – ggaagatctacctccaattttatctataatcgcttc -3') made up of a *Bgl*II restriction site for cloning and a portion of the SssI methyltransferase sequence. By using two internal primers that overlap, the two DNA fragments generated in the first and second PCRs were mixed by denaturing and annealing them in a subsequent PCR reaction using the external oligo primers, which allows the fusion product between mCre and methyltransferase to be formed. After amplification, the product (~2308 bp) was cleaved with *Nco*I and *Bgl*II. This DNA insert was ligated into *Nco*I and *Bgl*II-digested pBadmycHisC vector, transformed into competent DH5α *E. coli* cells, and selected on LB/carbenicillin plates. Individual transformants were picked, screened by restriction digestion for the presence of the insert, and sequenced to verify that the plasmid contained the correct sequence. The fusion plasmid encodes a catalytically inactive mCre homing endonuclease fused to a wild-type M.SssI and was cloned into the pBAD/myc-HisC expression vector (Invitrogen) that carries a C-terminal *myc* epitope tag and C-terminal polyhistidine (His_6) region designed for expression and purification of protein from the *E. coli* araBAD promoter, regulated by AraC gene, and can be induced by adding L-arabinose to the medium. Protein resulting from the expression can be purified in one step, by affinity chromatography on a Ni^{2+}- or Co^{2+}-containing resin.

Mutant plasmids encoding a Q525L, E564A, R608, S695A and Q525L/S695A substitutions in the M.SssI gene were prepared from the fusion plasmid as template using the QuikChange site-directed mutagenesis (Agilent). The oligonucleotide primers create point mutations on the fusion plasmid mCre-M.SssI that reverts the A residue of the glutamine codon (C<u>A</u>G) at amino acid position 525 to a T residue to produce the leucine residue (C<u>T</u>G). Primers used for making the Q525L mutant are as follows: oligonucleotide primer#1 (5'- TCAAGACTTATCTCAAC<u>T</u>GGGTATTCAAAAGGG -3') and

oligonucleotide primer#2 (5'- CCCTTTTGAATACCCAGTTGAGATAAGTCTTGA -3'). Primers used for making the E564A mutant are as follows: oligonucleotide primer #1 (5'- caaaatacttgttaatggCGaatgtaggagctcttc -3') and oligonucleotide primer #2 (5'- gaagagctcctacattCGccattaacaagtattttg -3'). Primers for making the R608A mutant are as follows: oligonucleotide primer#1 (5'- cggttcctcacaagcaGCGagaagagtttttatg -3') and oligonucleotide primer #2 (5'- cataaaaactcttctCGCtgcttgtgaggaaccg -3'). Primers used for making the S695A mutant are as follows: oligonucleotide primer #1 (5'- AGGACCAACCTTAACTGCAGCGGGTGCAAATTCAAG -3') and oligonucleotide primer #2 (5'- CTTGAATTTGCACCCGCTGCAGTTAAGGTTGGTCCT -3'). The plasmid encoding a Q525L substitution in the M.SssI gene was used as a template and the primers used for making the Q525L/S695A mutant were the same primers used for making the S695A mutant. Sequencing confirmed that the only mutation introduced was the desired one.

Protein expression:

The mCre-M.SssI ORFs from the constructs cloned into pBAD/myc-HisC were transformed into the chemically competent *E. coli* TOP10 cells (Invitrogen) for the expression of mCre-wildtype M.SssI fusion protein and for plasmid DNA preparation. This strain is deleted for *mrr* and *mcr*BC SssI methyltransferase restriction genes, and so DNA restriction at $G^{m}C$ is blocked. Culture of single colonies was grown at 37 °C, inoculated into fresh LB medium containing carbenicillin (100 μg/ml) and then grown to OD_{600} = 0.5-0.7. A 1-ml sample (noninduced control) was taken immediately, cell was pelleted and resuspended in $1/10^{th}$ volume of 20mM Tris-HCl pH 8.0 buffer, and sample was frozen at -20°C until needed for SDS-PAGE. This clone was also stored as a glycerol stock. To reduce transcription from the pBAD promoter ("leaky" promoter), in the absence of L-arabinose, the uninduced level is repressed by growth in the presence of glucose. Therefore, culture using this glycerol stock was grown overnight to saturation in LB medium containing carbenicillin (100 μg/ml) and 2% (v/v) glucose, inoculated into fresh LB medium containing carbenicillin (100 μg/ml) and 2% (v/v) glucose and grown to OD_{600} = 0.5-0.7. Glucose from the cells was washed out by centrifuging the remaining cell culture and the resulting cell pellet was resuspended in fresh LB medium containing carbenicillin (100 μg/ml) prior to inducing

protein expression by the addition of 0.1% (v/v) L-arabinose followed by growth for 1 h, 2h, 3h, 4h, overnight at 37°C and for overnight at 16°C to determine the condition for optimal expression of the fusion protein. The fusion protein should yield an ~89kD product. A pBAD/myc-His/lacZ plasmid was used as a positive control to confirm that growth and induction was done properly. The positive control should yield a 120kD protein. A second sample (induced) was collected, cell was pelleted and resuspended in 20mM Tris-HCl pH 8.0 buffer. The remaining cell cultures were centrifuged and cell pellets were frozen at -80°C. Expression of the protein from the induced cells was analyzed by SDS-PAGE using NuPAGE 4-12% Bis-Tris gel (Invitrogen) run at 1x NuPAGE MOPS SDS running buffer (Invitrogen).

The *E. coli* TOP10 was also used for expression of the mutant derivatives. After transformation, culture of single colonies was grown overnight to saturation in LB medium containing carbenicillin (100 μg/ml) and 2% (v/v) glucose, inoculated into fresh LB medium containing carbenicillin (100 μg/ml) and 2% (v/v) glucose and grown to OD_{600} = 0.5-0.7. A 1-ml sample (noninduced control) was taken immediately, and the cell suspension was frozen at -20°C. For two of the mutant derivatives, additional samples of mCreI-Q525L M.SssI and mCreI-Q525L/S695A M.SssI mutants (equivalent of 1x 10^7 cells and 5x 10^7 cells) were also taken, cells were pelleted and frozen at -80 °C until needed for immunoblotting. Glucose from the cells was washed out by centrifuging the remaining cell culture and the resulting cell pellet was resuspended in fresh LB medium containing carbenicillin (100 μg/ml) prior to inducing protein expression by the addition of 0.1% (v/v) L-arabinose followed by growth for 1 h, 2h, 3h, 4h, overnight at 37°C and for overnight at 16°C to determine if they result in better expression than the mCre-wildtype M.SssI fusion protein. For both mCreI-Q525L M.SssI and mCreI-Q525L/S695A M.SssI mutants, a shorter time course of expression was also performed for 15, 30, 45 minutes, 1h, 2h and 3h at 37°C. A second sample (induced) was collected for SDS-PAGE, and additional samples from the same two mutant derivatives were also taken for immunoblotting.

For Western blot analysis, cell pellets (1x 10^7 cells; 5x 10^7 cells) were resuspended in 1x SDS-PAGE sample buffer, samples were heated, separated by SDS-PAGE electrophoresis using NuPAGE 4-12% Bis-Tris gel (Invitrogen) and 1x NuPAGE MOPS SDS running buffer (Invitrogen) and then electroblotted onto PVDF membrane (BIO-RAD) at 20V at 4 °C overnight in transfer buffer (25mM Tris, 192mM glycine, 20% (v/v) methanol). Non-specific antibody binding was blocked by incubation in TBS-T buffer (150mM NaCl, 10mM Tris-HCl pH7.5, 0.05% (v/v) Tween-20 and 5% (w/v) nonfat dry milk). The mCre-M.SssI mutant fusion proteins were detected with a mouse monoclonal anti-myc primary antibody (clone 9E-10), which recognizes the peptide myc epitope (EQKLISEEDL), in a 1:1000 dilution and an increased concentration (1:200 dilution). A clone of pGlue-WRN which carries a myc epitope tag at its N-terminus (70 µg clear lysate supernatant; 187kD) was used as a positive control. The mCre- M.SssI fusion proteins were also detected with a rabbit anti-I-CreI whole protein primary antibody (800912nd $(NH_4)_2SO_4$ ppt.) in a 1:2000 dilution. A purified mCre protein (0.1 µg; 40kD) was used as a positive control. Bound antibodies were detected with donkey anti-mouse IgG secondary antibody (Alexa Fluor 647; Invitrogen; 1:1000 dilution) and donkey anti-rabbit secondary antibody (Alexa Fluor 546; Invitrogen; 1:2000 dilution) and chemiluminescence detection (FluorChemQ Alpha Innotech).

Methylation detection of plasmid DNA:

Overnight culture of *E.coli* TOP10 harboring mCre-wildtype M.SssI and mCre-M.SssI mutant derivatives using the glycerol stock and single colonies, respectively, was diluted into fresh LB medium containing carbenicillin (100 µg/ml) and 2% (v/v) glucose and grown to OD_{600} ~0.6 prior to harvesting 5ml sample from pre-induced cells for plasmid preparation. To obtain plasmid DNA from induced cells, cells were washed out from glucose prior to adding 0.1% (v/v) L-arabinose before harvesting 5ml sample for plasmid preparation at each time course of expression. Methylation of CpG sites was assayed by digesting the plasmids with excess of *Eag*I/*Xho*I, *Asi*SI/*Xho*I and *Hpa*II/*Xho*I restriction enzymes. *Eag*I, *AsiS*I and *Hpa*II are methylation sensitive restriction enzymes while *Xho*I restriction enzyme is methylation insensitive. Restriction products were analyzed on a 1% agarose gel run in TAE buffer.