Life at the front: Defining cyanobacterial traits for plasticity in changeable oceans

By

Pramita Ranjit

A thesis submitted to Macquarie University

for the degree of Master of Research

Department of Molecular Sciences

20 April 2019

MACQUARIE
University
SYDNEY·AUSTRALIA

Declaration

The material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university or institution. The contribution and assistance of others have been appropriately acknowledged.

Pramita Ranjit

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr Deepa Varkey for her guidance, support, encouragement and patience throughout the duration of this project. The amount of work and effort you put forth for me is truly appreciated and I feel extremely lucky to have had you as my supervisor. My sincerest thanks to my co-supervisor Dr Martin Ostrowski for his incredible guidance, motivation and insightful feedback throughout the research journey. Completion of this project would not have been possible without both of your expertise and guidance.

My sincere thanks to my co-supervisor Professor Ian Paulsen for giving me an opportunity to work as a part of his research team and for his continued support and guidance. I would also like to thank Dr. Lisa Moore for dedicating a lot of her time in providing me her insightful feedback that helped me improve my thesis. Also, special thanks to both Dr. Lisa Moore and Dr. Bhumika Shah for their continuous support and providing me guidance in various aspects of the project. It was great to be a part of the "cyano-team" with them. I am also thankful to all the members of the Paulsen group for their continuous enthusiasm and support.

A huge thanks to my parents, parent's in-law and rest of my family members for supporting and encouraging me to pursue this research degree. And finally, I want to specially thank my husband for always being there for me and believing in me no matter the situation. Without you none of this would have been possible.

Table of Contents

Abbreviations

NRS	National Reference Stations
NSI	North Stradbroke Island
PHB	Port Hacking
MAI	Maria Island
EAC	East Australian Current
GTI	Genotype Thermal Index
HMW	High Molecular Weight
LB	Luria Bertani
SOC	Super Optimal Broth with catabolite repression
IPTG	Isopropyl β-D-1-thiogalactopyranoside
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
PAM	Protospacer Adjacent Motif
gRNA	Guide RNA
HRT5	Homologous Repair Template upstream of target gene
HRT3	Homologous Repair Template downstream of target gene
bp	Base pair
DDH	DNA- DNA hybridisation
PS	Photo System
CTAB	Cetyl Trimethyl Ammonium Bromide
ANI	Average Nucleotide Identity
petB	cytochrome b6 subunit of cytochrome b6f complex
PCR	Polymerase Chain Reaction
ASW	Artificial Seawater
HGT	Horizontal Gene Transfer
Km	Kanamycin
MUSCLE	Multiple sequence comparison by log-expectation

Abstract

Microbial prokaryotes and eukaryotes thrive in the oceans where they represent the majority of biomass. They have adapted to and exploit almost every available niche by employing a wide diversity of functional traits. It is well known that distinct genetic lineages of marine picocyanobacteria occupy large geographical areas, such as the vast subtropical ocean gyres, or defined latitudinal slices encompassing the temperate, mesotrophic regions. However, a big proportion of the ocean corresponds to regions that are dynamic, such as the neritic coastal zone, or environments influenced by boundary currents. The aim of this thesis is to identify whether there are lineages of picocyanobacteria that are specifically adapted to variable ocean regions using comparative genomics and ecology. Secondly, to identify what specific molecular mechanisms they utilise to adapt. And finally, build and refine an efficient molecular toolbox to genetically manipulate the lineages of picocyanobacteria that could, for example, experimentally define the role of putative genes involved in adaptations to dynamic ocean regions. In the light of rapidly changing environmental conditions, a molecular-ecological understanding of how these abundant model cyanobacteria adapt will provide important insights into the future productivity and the health of the ocean in general.

Chapter 1: Introduction

1.1 Marine picocyanobacteria and their global significance

Marine picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus* are unicellular photosynthetic organisms belonging to a diverse group of photoautotrophic prokaryotes, cyanobacteria, which were the first organisms to evolve the ability to carry out oxygenic photosynthesis[1]. This metabolic innovation is known to have played an important role in the formation of an aerobic biosphere through their contribution towards the great oxidation event that occurred ~2.3 Ga years ago[2]. Hence, cyanobacteria paved the way for other complex life forms to evolve by generating an oxygen-rich environment for their survival.

Marine picocyanobacteria play an indispensable role in ecosystem functioning[3] and occupies a key position at the base of marine food web as primary producers. These photosynthetic microbes numerically dominate the phytoplankton in the world's ocean with the estimated global annual mean abundance for *Synechococcus* being 7.0 ± 0.3 x 10^{26} cells and 2.9 ± 0.1 x 10^{27} cells for *Prochlorococcus*[4] and contribute significantly to biomass production, accounting for up to 50% of the global primary production in some regions[5]. Therefore, they are one of the major sources of energy and organic matter that help sustain other members of the marine ecosystem. Due to their photosynthetic ability to fix inorganic carbon such as CO_2 and HCO_3^-, they are fundamental components of the carbon cycle. The also play integral roles in the biogeochemical cycling of phosphorus and nitrogen[5].

1.2 Ecological partitioning and global distribution of marine picocyanobacteria

Despite their abundance, marine picocyanobacteria were discovered only around 40 years ago. Marine *Synechococcus* (henceforth referred to as *Synechococcus*, unless specified otherwise) was first reported in 1979[6], whereas *Prochlorococcus* was discovered only a decade later in 1986[7]. These genera share a common ancestor and often co-occur but have morphological (e.g. cell size), physiological (e.g. light-harvesting pigments) and genetic differences.

Synechococcus genome size ranges from ~2.2 – 2.86 Mbp and comprises of a single chromosome with no plasmids[3]. The major light harvesting complex in *Synechococcus* are phycobilisomes that comprises different combinations of phycobiliproteins each binding to one or more chromophores such as phycocyanin (A_{max} 620nm), phycoerythrin (A_{max} 560nm) and phycourobilin (A_{max} 495 nm) which determines the light absorption spectra[8]. Phycoerythrin is the main light harvesting pigment and the orange fluorescence emitted by it is a key distinguishing feature for detection and identification of *Synechococcus.* In comparison, the genome size of *Prochlorococcus* ranges from

1.64 to 2.7 Mb[3]. Its major photosynthetic pigments are divinyl chlorophyll a and b, that allows the absorption of blue light of 460nm- 480nm at lower depths[7]. These differences are key factors in their ecological partitioning in the marine environment.

Synechococcus are ubiquitous and have a widespread distribution across the oceanic environment from the equator to the polar circle (Figure 1.1). They are more abundant close to the surface waters and occur with cell densities of up to 10^{5}-10^{6} cells ml^{-1} in mesotrophic water whilst in oligotrophic waters, they are less abundant (~ 10^{3} cells ml^{-1}) [9]. In contrast, *Prochlorococcus* occurs at high cell densities up to 10^{6} cells ml^{-1} but are mostly restricted to the warm oligotrophic waters. Furthermore, they have the capacity to colonize both the surface as well as the water column down to the depth of 150-200 m with sub-lineages adapted to specific light quality and intensity [10].

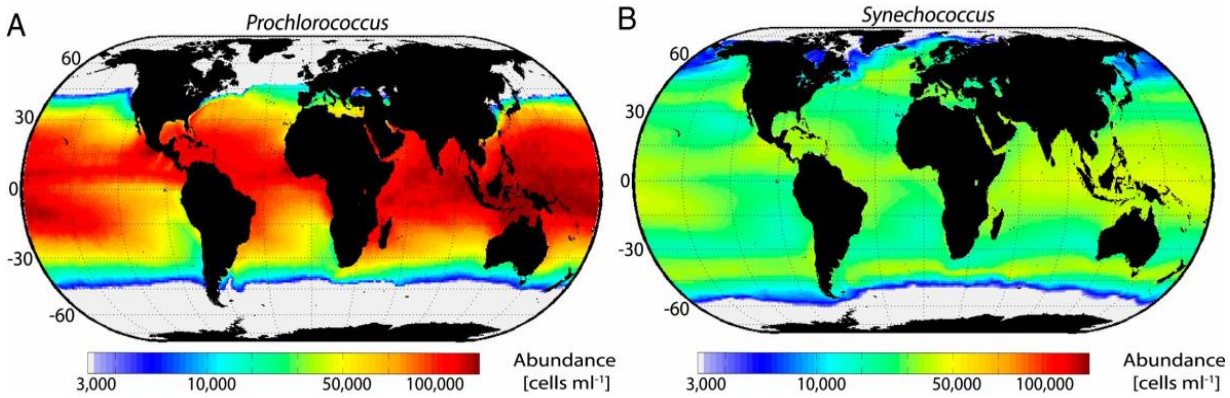


Figure 1.1 Global distribution of marine picocyanobacteria. A) *Prochlorococcus* B) *Synechococcus* Figure adapted from Flombaum *et al.,* 2013[4]

1.2.1 Phylogenetic diversity in marine picocyanobacteria

Marine picocyanobacteria are a phylogenetically diverse group comprising several sub-lineages (or clades or ecotypes) of *Prochlorococcus* and *Synechococcus* that occupy different niches in the oceanic environment.

The evolutionary diversification of *Prochlorococcus* has been closely associated with environmental factors[9]. The two main sub-lineages (or ecotypes) of *Prochlorococcus* are the high-light (HL) and the low-light adapted (LL) clades [11,12]. The HL ecotypes are most abundant in the uppermost region of the euphotic zone [13] and can be sub-divided into the high iron-adapted (HLI- low temperature and HLII- high temperature) clades and low-iron adapted (HLIII/HLIV) clades[13]. The LL ecotypes dominates the lower region of the euphotic zone [1314] with LLI found near the nutricline followed by LLII/III and LLIV at the bottom. Other less abundant sub-lineages include NC1, HNLC, HLIII - HLVI and LLV – LLVI [3]. Hence, *Prochlorococcus* ecotypes display clear connection between the phylogeny and environmental niches they occupy [15].

In contrast, *Synechococcus* are phylogenetically more diverse. It is subdivided into sub clusters 5.1, 5.2 and 5.3. The major sub cluster 5.1 has at least 20 recognizable lineages with clades I, II, III and IV being the most abundant lineages [3] and depicts four distinct groups of frequently co-occurring clades[16]. Clades I and IV are predominantly found along the coastal boundary zone in colder nutrient-rich water and generally co-occur in higher latitude regions above ca. 30°N and below 30°S whereas clade II is more abundant in the warm, subtropical/tropical regions, mostly confined within 30°N and 30°S [9]. Clade III is prevalent in oligotrophic open ocean environments and sometimes found to co-occur with clade II [9,17] Other *Synechococcus* lineages are either present at low abundance or are predominant in certain restricted regions only. For example, clades CRD1, CRD2 and envB/C are found to be successful in low iron upwelling regions[16,18] and clades XV and XVI occur at the transitional sites near the junctions of other biomes. Other lineages of sub-clusters 5.2 and 5.3 are mostly present in estuarine environments[17,19]. Thus, *Synechococcus* sub-lineages occupy distinct niches along a horizontal gradient of mainly light quality, temperature and nutrient availability.

1.2.2 Distribution of Synechococcus population- along the east Australian coast

Given that marine picocyanobacteria display vast phylogenetic and functional diversity and numerically dominate the phytoplankton in the world's ocean, their combined metabolic activities control the biogeochemical cycles that drive critical ecosystem functioning. To understand how they contribute to the global flow of nutrients and their niche adaptation, it is of key importance to characterise the biogeography and ecology of diverse subpopulations of marine picocyanobacteria. However, many clades of *Synechococcus* are still not well characterised as compared to *Prochlorococcus*.

As previously described, the global distribution pattern of *Synechococcus* depicts four distinct groups of frequently co-occurring clades. Among these distinct groups of clades, the preferred habitat for clade XV and XVI and their overall contribution to the *Synechococcus* population is largely unexplored. In a study conducted by Sohm *et.al.,*2016 clades XV and XVI were detected near junctions or transition of other three biomes including in waters between Australia and New Zealand[16] possibly including the Tasman Sea. Direct comparison of 16S rRNA and *petB* gene sequences indicate that clade XV is equivalent to clade IIh *sensu* Mazard *et al.,* 2012 and hereafter will be referred to as clade IIh.

Transition zones or oceanic fronts are narrow zones that exist between different water masses, creating enhanced horizontal gradients of water properties such as temperature, salinity, nutrients etc[20]. For example, when cold mesotrophic water converges with warm oligotrophic water, frontal zones with sharp gradients and filaments of sea surface temperature and chlorophyll often forms at

the junction[21]. The biophysicochemical factors such as temperature and chlorophyll rapidly change at the oceanic fronts which may result in fertilisation and horizontal advection[22] of nutrients leading to growth and accumulation of plankton at the fronts. Sharp changes in microbial structure, composition and their activity occur at the oceanic fronts leading to enhanced bacterial carbon production and increased primary production.

The Tasman Sea, has been identified as a 'sea hotspot' for rapid oceanographic changes[23,24]. It is situated between Southeast Australia and New Zealand and has shown surface warming of + 2.28°C[24] per century in contrast to global surface-ocean warming of 1°C per century. The intensification and southerly extension of the East Australian Current (EAC), a major poleward flowing Western Boundary Current, has been attributed to increased temperature rise in this area[25]. The EAC originates from the South Pacific Ocean Gyre and extends southward along the Australian east coast between 18°S and -33°S where it separates into two stream layers. The upper layer of the current flows eastwards across Tasman Sea whereas the deep layer flows southward as far as the Tasmanian east coast[25]. EAC transports warm, oligotrophic waters from the tropics to temperate latitudes[26] and is strongest during summer[27]. EAC is very variable and show seasonal, interannual and decadal trends, that are known to change the local oceanography such as by enhancing upwelling and local biology including fisheries[28]. The changes in intensity and timing of the southerly flow have a wide impact on both water temperatures and species ranges in south east Australia influencing fisheries and endangering native populations. Hence, south-eastern Australia is an area experiencing one of the strongest shifts and/or intensification of the dominant surface current and is a hotspot of high biological and commercial significance[29].

However, we have limited knowledge concerning the movements, adaption or range expansion of marine picocyanobacteria and their consequences on the food webs and biogeochemical cycles in the Tasman Sea and it would be interesting to know what lineages of picocyanobacterial populations, particularly clades belonging to *Synechococcus,* adapt to the dynamic conditions at the ocean fronts and furthermore, what features of these lineages would contribute to their adaptation. Therefore, addressing this knowledge gap about the changes in the picocyanobacterial community caused by EAC is a critical issue, particularly because EAC is forecasted to both warm and strengthen significantly resulting in various changes effecting the weather patterns to shifts in marine species distribution [30].

1.3 Factors defining *Synechococcus* distribution

Because of the *Synechococcus* microdiversity and their combined effect in influencing the ecosystem functioning, it is important to understand what factors contribute to their ecological partitioning.

1.3.1 Light

Being photosynthetic organisms, *Synechococcus* are strongly influenced by the quality and quantity of available light, important for maintenance of photosynthesis and prevention of photoinhibition and cellular death[31]. *Synechococcus* is present in high number in surface waters with their abundance decreasing down the water column[12,13]. However, in well mixed waters, *Synechococcus* cells have been observed even around the depth of ~200m, albeit in low numbers. The ubiquitous nature of *Synechococcus* is partly due to its use of the wide variety of photosynthetic light harvesting pigments[32]. Their light-harvesting complex, phycobilisomes, comprise different phycobiliproteins that bind to different ratios of chromophores, phycourobilin (PUB) and phycoerythrobilin (PEB*)*. These distinct pigment types differ in the wavelength of incident light absorbed optimally. Therefore, based on the light niches, strains have different pigment types. For example, strains prevalent in oligotrophic waters have high PUB to PEB ratio whereas strains predominant in the mesotrophic environments have lower PUB to PEB ratio[2].

1.3.2 Temperature

Temperature plays a fundamental role on the rate of photosynthesis, enzyme activities and nutrient uptake rate [33,34]. Habitat ranges may be defined by adaptability to various temperature conditions as is evident in the distribution of major *Synechococcus* clades. Clades I and IV, are mostly prevalent across the temperate region and occur at lower temperature range from 10°C to 20°C whilst clades II and III are abundant across the tropical region at temperatures 20°C to 30°C [9,16].

1.3.3 Nutrients

The limiting factors that affect *Synechococcus* growth includes macronutrients such as nitrogen and phosphorus that differ in their concentrations and bioavailability across the oceanic regimes. In addition to this, trace metals such as iron, zinc, copper, nickel and manganese also influence their growth. *Synechococcus* utilise different strategies to adapt to different and often fluctuating nutrient availability, including reduced nutrient requirements, altered elemental stoichiometry, changes in regulatory systems and the number of nutrient transporters and hence have specific nutrient niche preferences[35,36]. Clades I and IV are mostly restricted to the coastal regions where they are exposed to a wide variety of nutrient concentrations whilst Clade III dominates oligotrophic regions and therefore, prevail in conditions with limited range of macronutrient concentrations [3,9,16].

1.3.4 Biotic factors

Biotic factors such as presence of viruses in oceans, grazing and competition for survival play an important role in the abundance and distribution of *Synechococcus.*

Cyanophages i.e. viruses that infect cyanobacteria are abundant and ubiquitous in marine environment and viral lysis potentially plays a role in structuring *Synechococcus* population[37,38]. *Synechococcus* is also subject to grazing by heterotrophic nanoflagellates (HNFs), dinoflagellates and ciliates and some of these grazers display strain specificity which could be due to the difference in their cell surface properties and grazer evasion strategies[39,40]. Furthermore, competition for survival, even within the same genera i.e. presence of one strain hindering the growth of other also influences niche selection. It was seen that presence of *Synechococcus sp*. CC9605 impaired the growth of CC9311 and WH8102 when cultured together[41]. Additionally, *Synechococcus* inhibits the growth of its competitors by producing secondary metabolites which also functions as a defence mechanism[41].

1.4 Molecular adaptations contributing to marine *Synechococcus* niche adaptation and microdiversity

Specific molecular adaptions to biotic and abiotic factors likely shape niche selection in *Synechococcus* populations. Unravelling these adaptations and physiological traits which contribute to the fitness of lineages in specific environmental regimes are key to understanding the current and future distributions of *Synechococcus* lineages.

1.4.1 CRP-Type Regulators

Presence of regulatory systems could provide an advantage to certain clades in adaption to a certain niche. For example, genomic data for coastal environment adapted clade I *Synechococcus* strain CC9311, has revealed that it has high capacity to respond and sense changes in its coastal environment with features such as higher number of response regulators and histidine kinase sensors as compared to the open ocean *Synechococcus* sp WH8102, along with larger capacity for transport, uptake, use and store metals specially iron and copper[9,42]. However, whether this increased capacity of transport and storage capacity of Fe is due the greater Fe requirements or changing Fe supply in the coastal environment is still uncertain[3]. Therefore, it is important to understand how these regulatory systems function and how they assist in adaptation to a certain niche.

CRP (catabolite repressor protein) regulators function as transcriptional activators that control the expression of the target genes by interacting with the particular DNA sequences upstream of it [43]. In bacteria such as *E.coli,* CRP alters its conformation when it binds with 3'-5' cyclic AMP(cAMP) and interacts with DNA sequences which results in positive or negative regulation of gene expression

depending on its position of interaction with respect to RNA polymerase binding site[44]. CRP found in marine picocyanobacteria show similarity to these family of bacterial regulatory proteins[43].

Marine picocyanobacteria possesses four main orthologous clusters of CRP regulators including cluster 468 (NtcA), cluster 1606 (PtrA), cluster 2049 and cluster 1390 (Crp1390).

Two CRP regulators NtcA and PtrA have been studied in detail. NtcA is a highly conserved global nitrogen regulator and member of catabolite activator protein family that upregulates the transcript level of its own gene in response to ammonium deprivation[3,45]. PtrA is only found in some *Synechococcus* clades and has a role in regulating the response to severe phosphorus limitation[3,36]. The functions of the remaining two regulators needs further exploration and experimental verification. For example, *crp* 1390 genes occur in close association in the genome with iron acquisition and storage genes and hence has been suggested to play a role in iron acquisition[3]. Iron has been considered as one of the most important trace metals for marine picocyanobacteria with its potential role in controlling the extent of primary production in marine environment including functions such as electron transport from water to $NADP^+$ that requires cytochrome b_6f (5 Fe), PSI (12 Fe), PSII (2 or 3 Fe)., and ferrodoxin (2Fe)[3]. However, very little is known about the molecular mechanisms required for regulating Fe uptake and acquisition. Similarly, cluster 2490 often co-occur with the genes known to encode for adenylate cyclases or guanylate cyclases and display similarity with CRP of *E. coli* suggesting it may play an analogous role[3].

1.4.2 Membrane lipid composition

Oceanic surveys have established distinct temperature preferences for the most abundant lineages, clades I, II, III and IV. Clades I and IV are most prevalent at 10°C – 20°C whilst clades II and III are abundant at higher temperatures of 20°C – 30°C. Laboratory-based studies have confirmed these distinct optimal growth temperatures amongst the representative lineages suggesting that marine *Synechococcus* may have developed efficient adaptive strategies to cope with temperature variations. A study conducted by Varkey *et. al.,*2016 compared the growth of different *Synechococcus* strains and determined that membrane lipid composition was a potential key adaptive factor that influenced growth at different temperature conditions[46] (Figure 1.2). The level of unsaturation in the clade IV strain, BL107 was higher than in the strains of clades II (WH8109) and III (WH8102), with a further increase in unsaturated lipids at lower temperatures. This suggests that changes in membrane lipid composition may be an important adaptation at lower temperatures. It was also found that clades adapted to temperate environments, i.e. lower temperature niches, had an additional fatty acid desaturase ortholog which might likely have a role in regulating membrane fluidity at low temperature and thus might play a role in helping them colonize temperate niches [46].

Membrane fluidity is a one of the key factors for determining the fitness of an organism in a specific thermal niche and is largely regulated by the ambient temperature, the degree of unsaturation and the length of the constituent fatty acid (i.e. membrane thickness)[47]. Membranes are sensitive to temperature change and can affect their fluidity and therefore the activity of the membrane proteins[48]. For example, the activity of most enzymes and bio-membranes directly depend upon temperature and hence may impact major metabolic processes as well as constrain growth. Furthermore, changes in membrane (thylakoid) fluidity due to temperature variation may result in photo-system II inactivation and particularly impacts processes such as photosynthesis[4950]. However, the adaptive physiological processes providing ability to different *Synechococcus* temperature ecotypes to thrive in their respective thermal niches remain under explored[33].

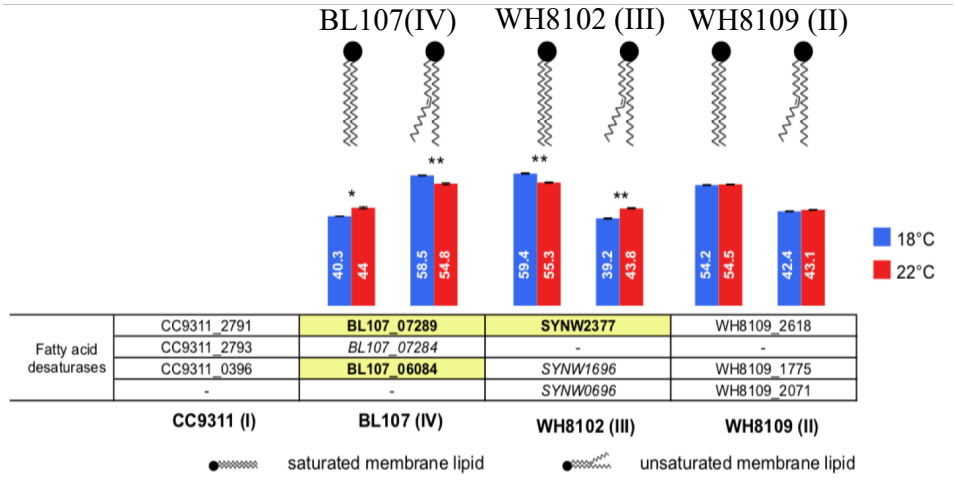


Figure 1.2: Lipid composition of *Synechococcus* isolates BL107 (clade IV), WH8102 (clade III) and WH8109 (clade II) when grown at 18 °C (blue) and 22 °C (red). Included in the figure are genes encoding fatty acid desaturases in representative isolates of clades I, II, III and IV. Gene IDs in bold represent proteins that are detected but not differentially expressed and those in italics represent proteins that are not detected. Figure adapted from Varkey *et al.,* 2016[46]

1.5 Tools for exploring marine picocyanobacterial ecology

Because marine *Synechococcus* exhibit extensive microdiversity, methods that provide higher taxonomic resolution is necessary to determine the genetically distinct or identical populations that may be adapted to specific environmental niche.

Over the past half century, the overall similarity of the two-genome sequence had been measured using DNA- DNA hybridisation (DDH) experimental methods[51,52] with 70% DDH being widely accepted as the standard cut-off for bacterial species delineation[53]. However, due to the error prone and tedious nature of the DDH method, 16S rRNA gene is now used widely used as measure for gene sequence similarity with its threshold of 97% roughly corresponding to 70% DDH[54]. However, in several instances, species sharing high level of 16S rRNA sequence similarity (>99%) has made it difficult to distinguish two distinct species using 16S rRNA gene sequences alone[55] even though

DDH analysis showed a distinct separation. Another method, multi locus sequence analysis (MLSA) that analyses the genealogical relationship based on catenating several housekeeping genes has also been evaluated as an alternative to DDH and has been successful for taxonomic resolution in certain groups[56] but has a major drawback due to its putative bias in gene selection and availability of amplification primers[57]. Additionally, average nucleotide identity (ANI), based on the computational comparisons of two genome sequences seem to be good alternative for gene sequence similarity measurement[58] since ANI values are an overall descriptor of genome relatedness that is simple and useful.

Picocyanobacterial phylogenetic diversity and biogeography has been studied utilising various conserved phylogenetic markers such as 16S rRNA, 16S-23S internal transcribed spacer region (ITS), cytochrome b_6 subunit of cytochrome b_6f complex (*petB*), RNA polymerase subunit (*rpoC1*), RuBisCO large subunit (*rbcL*), photosystem II (*psbA*), phycoerythin (*cpeB*), nitrate reductase (*narB*), nitrogen regulator (*ntcA*) genes, shedding light into their distinct niche partitioning[19,59 45 60].

1.6 Molecular toolbox to unravel marine picocyanobacterial adaptation factors

Synechococcus is a highly diverse group with its lineages occupying distinct niches and possess molecular adaptations that shape their fitness. Population diversity in *Synechococcus* is also mediated by horizontal gene transfer (HGT) via phages or plasmids or gene duplication[61]. Acquiring or losing genes via HGT may contribute to forming a complete set of unique genes further contributing to complexity of their diversity and distribution. Discovering the function of these unique genes would improve our understanding of how these genes contribute to the fitness of different lineages in distinct niches. To identify the role of these unique genes as well as to unravel the molecular adaptations, construction of gene knockouts and comparison of the phenotypic variations with the mutant is an important aspect to better understand how *Synechococcus* are adapting to different marine settings. However, lack of precise, modern genetic tools for engineering of marine *Synechococcus* is one of the major challenges[62].

Most of the generated mutants are from freshwater *Synechococcus,* which are very different from the ones found in the marine environment. The current strategies for genetic manipulation are very time consuming to generate the desired mutant. One of the techniques for genomic alteration includes creating a deletion mutant by double homologous recombination between a suicide vector and host chromosome where a selective marker replaces the gene of interest[63]. Further genetic changes are made by integrating other antibiotic resistant markers which impacts the pathway engineering since there are only limited number of antibiotic cassettes available.

Recently, CRISPR (clustered regularly interspaced palindromic repeats)/ Cas9 (CRISPR associated protein 9) genome editing technology has revolutionised the field of biotechnology by efficient and

precise manipulation of gene sequences in a single step in variety of organisms. CRISPR/Cas9 system is a natural adaptive immunity system in bacteria that protects bacteria from invading viruses and plasmids by degrading the exogenous DNA[64]. CRISPR system has been efficiently used in targeted genome editing by changing the spacer sequences in the CRISPR array to be complementary to the gene target and providing the repair template with the desired change sequence[65], hence making mutations or deletions at the enzyme - cut site.

The application of CRISPR/Cas9 technology has not been used widely in the *Synechococcus* due to the toxicity of *Cas9* enzyme in these organisms[66], which is due to the enzyme's off-target effects. Recently, another nuclease Cpf1, a RNA directed double stranded dsDNA nuclease, has been used to overcome this problem which has been shown to be nontoxic to freshwater *Synechococcus*[62]. Cpf1 enzyme makes a staggered cut and allows for more precise selection of cleavage target, hence reducing off target effects[62] and has been successfully used to generate mutants in freshwater cyanobacteria. Therefore, given the effectiveness and preciseness of the CRISPR/Cpf1 on freshwater cyanobacteria and the lack of precise genetic modification tools for marine *Synechococcus* the work presented here aims to test if CRISPR based gene editing technique works as effectively in marine *Synechococcus* as in freshwater *Synechococcus* and thereby ease the process of genetic manipulation including simplifying and accelerating markerless modifications that was tedious and cumbersome with previously existing technologies.

1.7 Objective of the project

Marine picocyanobacteria are abundant and ubiquitous, photosynthetic microorganisms that contribute significantly to global primary production. Therefore, they form the basis of marine food web and play a key role in the cycling of nutrients[3].

Given the diverse nature of marine *Synechococcus* and many clades discovered but understudied, it is of key importance to characterize the biogeography and ecology of novel clades such as clade IIh in order to understand their overall contribution to *Synechococcus* populations, their possible niche adaptation as well as how they contribute to global flow of nutrients.

The widespread distribution of *Synechococcus* can be attributed to their genetic and genomic diversity which underpins a wide array of adaptation strategies. Distinct lineages of oceanic *Synechococcus* partition into environmental niches that are defined by different light, temperature and nutrient regimes [9,67]. Some adaptations have been well characterised, for example, the molecular mechanisms of phosphate and nitrogen accumulation [36] and chromatic adaptation[68] have been described in multiple strains. However, the mechanisms that are responsible for their adaptation to grow in waters spanning 30˚C of temperature (i.e. from the equatorial to polar regions) are poorly understood.

Temperature plays a fundamental role on the rate of photosynthesis, enzyme activities and nutrient uptake rates [33].

Similarly, the function of the CRP regulators such as cluster 1390 which are thought to play a role in iron storage and acquisition are not well studied and requires further experimental verification[3]. Iron has been considered as one of the most important trace metals for marine picocyanobacteria with its potential role in controlling the extent of primary production in marine environment[3]. Therefore, it is expected that molecular adaptations tuning the cell to different temperature conditions as well as the capacity to regulate the expression of genes via transcriptional regulators will have a significant effect on fitness.

Molecular adaptations are likely, a significant contributor optimizing the fitness of strains. Unravelling what these adaptations are, may help define physiological traits that influence the future distribution of *Synechococcus* lineages in light of the climate change-associated environmental perturbations. The south-eastern coast of Australia is a highly productive region but is also one of the fastest warming regions in the world, with numbers in *Synechococcus* populations predicted to change[4]. Being at the base of the marine food web, such changes in the abundance and distribution of *Synechococcus* could have serious implications for the future productivity in this region and the health of the oceans in general. Therefore, developing a precise molecular toolbox to genetically manipulate the marine *Synechococcus* is an important aspect to unravel these molecular adaptations.

The specific aims of this project were to

1. Investigate the distribution of *Synechococcus* clade IIh along east coast of Australia to verify its hypothesised designation as a transition zone clade.
2. Provide whole genome sequences of isolates from Australian marine habitat, compare genome relatedness of new sequenced isolates and identify unique traits contributing to the niche adaptation of clade IIh
3. Build and refine a molecular toolbox to genetically manipulate marine *Synechococcus* to confirm the role of genes contributing to molecular adaptations using targeted gene deletions.

Chapter 2. Experimental Design, Materials and Method

2.1 Biogeography of marine *Synechococcus* from Australian marine habitats

This study capitalised on the sustained microbial observations, in the form of 16S rRNA amplicons and shotgun metagenomes, undertaken at National Reference Stations (NRS) around Australia since 2012[69] providing an opportunity to capture oceanographic and ecosystem responses to seasonal changes.

Processed amplicon sequencing data of the 16S rRNA gene (amplified with bacterial 27f-519r primers) was obtained from the Australian Microbiome data portal of Bioplatforms Australia < https://data.bioplatforms.com/bpa/otu/> which provides data generated using standardised methods. Briefly, the amplicon data was generated from pelagic seawater samples collected monthly from 2012-2016 from multiple depths at seven NRS, as detailed in Brown *et al.,* 2018[69] and those from multiple voyages around Australia including Seymour et al., 2010 and Moore *et al.,* 2018 voyage. DNA extracted from all samples was amplified using bacterial 16S rRNA gene primers 27F – 519R and sequenced at the Ramaciotti Centre for Genomics (UNSW, Australia) using the duel indexed paired end approach on the Illumina MiSeq platform. All sequenced data were processed following a standardised method as outlined in Brown *et al.,* 2018[69] to produce the single nucleotide variant (zOTU) abundance table. The downloaded data was verified by an additional chimera check, taxonomic assignment was carried out using the non-redundant Silva database (v132) following the DADA2 workflow. All zOTUs sequences classified as *Synechococcus* were extracted and high-resolution taxonomy was assigned using assignSpecies in DADA2 with a curated in-house database refined by manual sequence alignment and phylogenetic placement in ARB[70].

This study focused on capitalising the metagenomic dataset from three NRS namely, North Stradbroke Island (NSI), Port Hacking (PHB), Maria Island (MAI) on east coast of Australia which are situated within the context of East Australian Current (EAC). The abundance of major *Synechococcus* clades I, II, III and IV along with clade IIh were analysed using package tidyverse and visualised using package ggplot2 in R version 1.1.453 thereafter focusing on defining the genome, traits and ecology of a novel clade IIh which was found to occur at the transition of the other major *Synechococcus* biomes (i.e. temperate and tropical lineages)[16].

Abundance versus temperature relationships were calculated for the most abundant clades including clade IIh. To control for sampling bias due to uneven sampling of temperature, the relative abundances for each clade, or sub-clade, were collected in 1˚C temperature bins and mean abundance for each bin was calculated. Kernel density estimation weighted by mean abundance was used to define the ecological temperature optima or Genotype thermal index (GTI).

Spearman rank correlations were computed in R Version 1.1.453 using package Hmisc 4.2-0 to investigate non- linear relationships between Clade IIh abundances and environmental variables. Correlations were visualised using package corrplot. Abundance information of clade IIh was scaled using square root transformation and the environmental variables were scaled using scale function in base-R. We further analysed whether clade IIh was correlated to other *Synechococcus* clades using Linear Pearson's correlation and whether those clades further correlated with the same environmental variables with p-value cut off <0.01.

2.2 Genome sequences of local isolates

A selection of putative *Synechococcus* strains isolated from eastern Australian marine habitats were being propagated in the Paulsen lab but had not been characterised (Table 2.1).

Table 2.1: List of *Synechococcus* strains studied in this project isolated from east coast of Australia from surface water (Depth <25m)

Strain *(Synechococcus sp.)*	Isolation site	Isolation date
HB1133	Sydney Freshwater Beach	14/09/2011
HBA1120	Sydney Freshwater Beach	14/09/2011
PH41509	Port Hacking National Reference Station	09/2015
MU1611	Tasman Sea	03/09/2016
MU1625	Tasman Sea	11/09/2016
MU1642	Tasman Sea	16/09/2016
MU1643	Tasman Sea	17/09/2016
MU1644	Tasman Sea	17/09/2016
MU1648	Tasman Sea	18/09/2016
MU1651	Tasman Sea	20/09/2016
MU1652	Tasman Sea	20/09/2016
MU1655	Tasman Sea	20/09/2016

2.2.1 High molecular weight (HMW) DNA extraction for whole genome sequencing

500 ml of all *Synechococcus* cultures (Table 2.1) were grown in artificial sea water medium[71] (ASW) at 22°C, 100 rpm, 40 µmol photons m^{-2} s^{-1}. The cultures were regularly subcultured (i.e. before it reaches the exponential phase) to ensure a high population of *Synechococcus* cells as compared to heterotrophic bacteria. Heterotrophic bacteria: *Synechococcus* ratio was determined by staining the

cells using SYBR® Green I nucleic acid gel stain 10,000 × in DMSO (Invitrogen) and visualising under fluorescence microscope (Zeiss Axioskop 2). Once high number of *Synechococcus* was ensured, the cells were harvested at mid-log phase 7 x 10^7 - 4 x 10^8 cells/ml by centrifugation (8000 g, 15 mins), washed twice with 1X Tris-EDTA buffer, then stored at -80°C until further use. DNA extraction was done using MagAttract HMW DNA kit (Qiagen). DNA concentration was checked using Qubit 3.0 Fluorometer (Invitrogen) and its purity was assessed using a NanoDrop® ND-1000 spectrophotometer and ND-1000 v3.5.2 software (Thermo Fisher Scientific).

To verify the *Synechococcus* clades, prior to genome sequencing, the extracted DNA were PCR amplified using reaction mixture : petB_F primer (10mM, 0.5 µl), petB_R primer (10mM, 0.5 µl), dNTPs (10mM, 0.5 µl), $MgCl_2$ (25mM, 1µl), buffer (10X, 2.5 µl), Taq polymerase (5units/µl, 0.2 µl) and nuclease free water to make up a total volume of 24 µl. All the primers are listed in Appendix II. The PCR amplified products were cleaned using PCR Clean-Up System (catalogue number: A9281, Promega) and sent for sequencing to Macrogen, Korea. The sequencing results were analysed using Unipro Ugene v1.30.0[72] and aligned with other *petB* sequences using MUSCLE (Multiple sequence comparison by log-expectation)[73]. The alignment results were then visualised on a phylogenetic tree using ARB[70] allowing identification of the closest related species for verification of clades before sending for whole genome sequencing.

2.2.2 Genome Assembly

DNA extracted from 12 samples of the local isolates were sent to Ramaciotti Centre for Genomics (UNSW, Australia) for sequencing on the Illumine MiSeq platform. Genome assemblies were done using data from two individual Illumina MiSeq runs (Nextera XT DNA library prep; 250 bp paired-end sequences). Prior to assembly the paired-end fastq files for each strain were filtered using Trimmomatic-0.38[74] with the following options: ILLUMINACLIP:/usr/local/Trimmomatic-0.38/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. Genome assemblies were carried out using SPAdes[75] version 3.12.0 using –careful option. Each cyanobacterial was unialgal but not axenic, i.e. it harboured heterotrophic co-isolates which also contributed to the DNA. The assemblies were initially screened and sorted by differential coverage binning either manually or using MetaBat2[76] with filtered pre-assembled sequences mapped to the assembly using bamM (http://bamm-project.org/documentation.html). The resulting bins were assigned to a taxonomic source using phylogenetic placement checkM[77] and the cyanobacterial bin was used as the basis for assembly refinement.

2.2.3 Assembly refinement: De novo improvements of gaps (16S rRNA operon assemblies)

Ribosomal RNA operons assemble poorly from NGS sequencing of mixed strains[78] (e.g. metagenomes). In order to retrieve the correct sequence, an rRNA operon from another similar strain (e.g. *Synechococcus* sp. PROS-U-1), was used as a bait sequence to pull out the fastq sequences from each of the paired-end fastq files. The initial inputs were a set of paired end fastq files and the bait sequence. The bait sequence came from the GenBank file sourced from Cyanorak database [http://application.sb-roscoff.fr/cyanorak/]. The index for bait was built using Bowtie2[79].The ids of the sequences that matched the bait sequence were extracted using samtools [80] which were then used to extract the matching fastq sequences to a new file for *de novo* assembly. The extracted forward and reverse fastq files were imported as lists into Geneious® 11.1.5 and assembled *denovo* using the geneious assembler. The assembled sequences (which is 16S rRNA sequence of IIh) were then used to blast against the bait sequence to confirm the alignment with bait sequence. This directed assembly of the 16S-23S rRNA operons was carried out for three of the twelve assembled genomes.

2.2.4 Phylogenetic placement for assembled genomes

A cyanobacterial genome tree was compiled with 21,388 genomes (120 concatenated conserved core genome protein sequences), including ~220 complete or near complete cyanobacteria genomes available in GenBank, 25 pre-release marine genomes (courtesy of the University of Warwick and Station Biologique de Roscoff, twelve new draft genomes of local isolates and 22 'clean' metagenome assembled genome from local marine cyanobacteria (assembled using data from the Marine Microbes Project) The Genome Tree and assembly was carried out using option de_novo_wf from GTDB-TK[81] software toolkit (https://github.com/Ecogenomics/GtdbTk) that uses user supplied as well as reference genomes from Genome Taxonomy Database (http://gtdb.ecogenomic.org).

2.2.5 Verify genome relatedness of assembled genomes

To verify the genome relatedness of the assembled genomes, average nucleotide identity (ANI), 16S rRNA gene identity and *petB* gene identity, of all sequenced cyanobacterial genomes were computed and compared. ANI were calculated for the assembled genomes using FastANI[82]. A single contig pseudomolecule for each genome assembly was generated by joining the contigs in a random order with 20 N's. FastANI fragments the query genome into non-overlapping fragments which are then mapped onto the reference genome using Mashmap[58]. 16S rRNA sequences were extracted from assemblies using Barrnap (BAsic Rapid Ribosomal RNA Predictor) using for i in *fa; do barrnap $i

–-outseq $i.rrna; done. *petB* sequences were extracted using the Perl script: perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV' petb_genomes.id > petb_genomes.fasta.

Alignment for 16S rRNA and *petB* sequences were done using MUSCLE in Geneious® 11.1.5 and the percentage identity matrix was analysed in R Version 1.1.453. The plot for ANI versus 16S rRNA and *petB* sequences were visualised using package ggplot2 in R Version 1.1.453. For further verification of genome distinctiveness, the number of base pairs that were not identical during the alignment of 16S rRNA and *petB* sequences were computed. 16S rRNA values were plotted against ANI to investigate the percentage difference in genome.

2.3.6 Identification of unique genes for clade IIh

Curated tables summarising of the presence, absence and abundance of gene orthologs and paralogs in 104 *Synechococcus* and *Prochlorococcus* genomes were made available by the MaCUMBa consortium who curate the Cyanorak database http://application.sb-roscoff.fr/cyanorak/[83]. The relative abundance of each orthologous cluster was determined by mapping the filtered forward reads against a database of all predicted gene sequences in Cyanorak using Blastn with default parameters (min 75% identity, 60% alignment length for the mapped read to the subject). Blastn output was filtered to report a one or zero hits for each read. In this study the unique genes for clade IIh were selected on the basis that they were found in at least one of the three isolates including *Synechococcus* sp. HB1133, HBA1120, PH41509 and Pros-U-1 (a previously sequenced clade IIh genome). The functions for each of the unique genes were derived from the Cyanorak database as well as conducting Blast search in NCBI. To examine the genomic context of genes unique to clade IIh, genomic islands for *Synechococcus* sp. HB1133 draft genome were first identified using Alien_hunter - 1.7[84] and aligned against the whole genome using Blast to determine the genetic loci of unique genes.

2.3 Genetic manipulation of marine *Synechococcus*

To develop an efficient molecular toolbox for the construction of gene-knockout mutants in marine *Synechococcus* species, we tested the utility of CRISPR/Cpf1 system using two different transformation techniques, electroporation and conjugation. This required optimisation of several parameters including the number of cells used for plating, the molar ratio of plasmid to *Synechococcus* cells, and the concentration of kanamycin for the selection of transformants. These variables were systematically tested in two different *Synechococcus* strains WH8109 and CC9311 for knocking out of gene Syncc8109_1050 (hereafter denoted as 1050) in WH8109 that encodes for a

CRP regulator that possibly plays a role in iron acquisition and sync_2793 (hereafter denoted as 2793) in CC9311 that encodes for fatty acid desaturase gene.

2.3.1 Consideration of cloning vector

Vector pSL2680 (Addgene plasmid # 85581), a derivative of broad host range plasmid RSF1010, originally isolated from *E.coli* incompatibility group Q[85], was used to facilitate rapid cloning of the editing plasmid into marine *Synechococcus* strains CC9311 and WH8109. The vector has been shown to be stably maintained in other *Synechococcus* sp. strains WH7803, WH8102, and WH8103 [86] and encodes for proteins essential for replication of vector in its host [62]. It has also been found to replicate well in most Gram- negative bacteria and allows the expression of the Cpf1 enzyme for genome editing without integrating the *cpf1* gene into the cyanobacterial chromosome[62]. Furthermore, the vector incorporates a kanamycin resistance gene which has been used in this study to verify plasmid uptake by *Synechococcus* and also as a marker for selection of transformants.

2.3.2 Genomic DNA extraction

Genomic DNA was used as template DNA for the primers to amplify the homologous repair template (outlined in section 2.3.3.4). The *Synechococcus* spp. strains (CC9311, WH8109) used for this study were cultivated in ASW[71] media at 22°C, 100 rpm, 40 µmol photons m^{-2} s^{-1}. Genomic DNA was extracted from the *Synechococcus* strains using the CTAB method previously described [87]. Briefly, *Synechococcus* culture (50 ml) was harvested by centrifugation (7,000 × g, 15 mins) and the cell pellet resuspended in nuclease-free TE buffer (567 µl). Sodium dodecyl sulphate (10% w/v, 30 µl) and proteinase K (20 mg/ml ,3 µl) was added onto the cell suspension and incubated in water bath (60ºC, 4 h). After incubation, NaCl (5M, 100 µl) and 10% cetyl trimethylammonium bromide (CTAB) in NaCl (0.7% w/v, 80 µl) was added. The mixture was then incubated again (65°C, 10 min). One volume of chloroform: isoamyl alcohol (24: 1 v/v) was added to the mixture and centrifuged (12,000 x g, 5min). The supernatant was collected and 1 volume of phenol: chloroform: isoamyl alcohol (25: 24: 1 v/v/v) was added and centrifuged (12000 g, 5 min). The previous three steps were repeated, and the supernatant was collected. DNA was precipitated by adding 0.6 volumes of isopropanol. The solution was mixed by gentle inversion and centrifuged (15000 g, 10 min, 4°C). The supernatant was discarded, and the pellet was washed with 70% (v/v) ice cold ethanol and centrifuged (15000 x g, 10 min, 4°C). The supernatant was discarded leaving the pellet to air dry before resuspending in nuclease-free water (30 µl).

2.3.3 Construction of editing plasmids for CRISPR-Cpf1 gene editing

Construction of editing plasmids for CRISPR/Cpf1 gene editing involves a series of processes including guide RNA (gRNA) design and ligation, homologous repair template (HRT) synthesis and NEB assembly of plasmids which are described in detail in sections below.

2.3.3.1 guide RNA Design (gRNA)

Nucleotide sequences namely, gRNA_2793 and gRNA_1050 (Appendix II), 20bp in length within the gene of interest preceded by a PAM (Protospacer adjacent motif) sequence, were chosen as a genomic target for Cpf1 enzyme. Primers, complimentary to the target sequences namely, gRNA_9311_2793F, gRNA_9311_2793R for gene 2793 in *Synechococcus* sp CC9311 and gRNA_8109_1050F, gRNA_8109_1050R for gene 1050 in *Synechococcus* sp WH8109 (Appendix II) were determined and ordered from Integrated DNA technologies.

2.3.3.2 Ligation of gRNA into plasmid via golden gate assembly

Plasmid pSL2680 (Appendix I) was grown overnight at 37°C in 10 ml of liquid Luria Bertani (LB) media containing tryptone (10 g/L), yeast extract (5 g/L), NaCl (5 g/L) supplemented with kanamycin (50 μg/ml). The plasmid was extracted the following day using SVminipreps DNA purification system (catalogue number: A1460, Promega). The quality and concentration of the plasmid was determined using a NanoDrop® ND-1000 spectrophotometer and ND-1000 v3.5.2 software (Thermo Fisher Scientific).

Goldengate assembly was used to ligate following gRNA oligos into the AarI site of pSL2680: gRNA_8109_1050F/ gRNA_8109_1050R (Appendix II) to yield plasmid pSL1050 and gRNA_9311_2793F/gRNA_9311_2793R (Appendix II) to yield plasmid pSL2793. The oligos were phosphorylated using the reaction mixture: gRNA_8109_1050F and gRNA_8109_1050R for 1050 gene (100 μM, 3 μl each) along with T4 polynucleotide kinase buffer (10X, 3 μl), T4 Polynucleotide Kinase enzyme (10,000 units/ml, 2 μl) and water (19 μl) and placed in a thermocycler at 37°C for 1.5 hours. NaCl (0.5 M, 4 μl) was then added to the reaction mixture. Similar reaction mixture except 100 μM of primer pairs gRNA_9311_2793F and gRNA_9311_2793R were used for 2793 gene. Following this, the oligos were annealed by cooling the reaction mixture to 12°C, starting at 95°C for 10 minutes and then decreasing the temperature by 10°C at each step at the rate of 1°C/second initially which is then followed by 0.3°C /second until it reaches 12°C. The oligos were then diluted at the ratio of 1:200 and ligated in to pSL2680 (100 ng) using golden gate reaction mixture of diluted annealed oligos (2 μl), AaRI oligo (50X, 0.4 μl), AarI restriction enzyme (2U/μl, 1 μl), T4 DNA

ligase (40,000U/ml, 1 µl), T4 DNA ligase buffer (10X, 2 µl), and nuclease free water making up a total volume of 20 µl and placed into a thermocycler at following conditions 37º C for 5 mins, 16º C for 10 mins, repeat the cycle for 10 cycles; 50º C for 5 mins, 80º C for 5 mins and 12º C hold.

2.3.3.3 Transformation of gRNA ligated plasmids

The ligation reaction (5µl) was then transformed into *E. coli* Stellar^{TM} competent cells PT5055-2 using heat shock treatment as described in section 2.3.3.5. Transformants were plated onto LB agar supplemented with kanamycin (50 µg/ml), 1M IPTG, 20% X-GAL for blue/white screening and incubated overnight at 37˚C. White transformant colonies were picked and patched onto new LB agar plates supplemented with kanamycin (50 µg/ml) and verified using colony PCR. All primers are listed in Appendix II. Colony PCR was done in 25 µl volumes using the reagents: GoTaq Buffer (5X, 5 µl), dNTPs (10mM, 0.5 µl), gRNA_PCR_seqF (5mM, 1 µl), gRNA_8109_1050R (5mM, 1 µl) for WH8109, GoTaq polymerase (5U/µl, 0.12 µl), nuclease free water (15.38 µl), Magnesium Chloride (25mM, 2 µl) in a thermocycler at following conditions: 95ºC for 5 min, 95ºC for 30s, 55ºC for 30s, 72ºC for 1min; repeat the three cycles for 30 cycles, 72ºC for 5 min. PCR products were visualized on a 1.2% agarose gel. Colony PCR for CC9311 were done with same concentration of the reaction mixture but different primer pairs which were: (5mM) gRNA_PCR_seqF and gRNA_9311_2793R. Colonies positively identified with gRNA inserts were then cultured overnight (37°C, 200 rpm, 12-16 h) in 10ml LB supplemented with kanamycin (50 µg/ml). Glycerol stocks (in 25% glycerol final concentration) were made with 800 µl of overnight culture and stored at -80°C. Plasmids pSL2793 and pSL1050 were extracted using the rest of the overnight culture using Wizard® *Plus* SV Minipreps DNA Purification System (catalogue number: A1460, Promega). The purified plasmid products were sequenced at Macrogen (Korea) to verify the gRNA insert was cloned without errors.

2.3.3.4 Homologous repair template (HRT) synthesis

Homologous repair templates (HRT) are used to repair the DNA breakage caused by the Cpf1enzyme in the target genome. 1000bp sequences upstream (HRT5) and downstream (HRT3) of the target gene were used to synthesize the repair templates. HRT's were amplified using high fidelity PCR and cloned into the KpnI site on the plasmid. The reaction mixture comprised Platinum SuperfiMix (2X, 12.5µl, Invitrogen), genomic DNA (4140.5ng/µl, 0.5µl) and nuclease free water (9.5µl) along with the primer pairs listed in Appendix III for markerless and marked editing plasmids for 1050 and marked editing plasmid for 2793 to synthesize the homology regions from genomic DNA. The spectinomycin resistance gene was used as a selection marker. The amplified homologous regions

were then separated on 1.5% agarose gel and cleaned with Wizard SV Gel and PCR Clean-Up System (catalogue number: A9281, Promega).

2.3.3.5 NEB assembly of editing plasmids

Plasmids pSL2793 and pSL1050 were digested with restriction enzyme *Kpn*I using the following reaction: pSL1050 (70.6 µl, 132.8 ng/µl) was mixed with FD buffer (10X, 15 µl), *Kpn*I (10U/µl, 3 µl), FastAP (1U/µl, 4 µl) and nuclease free water to make up a total volume of 150 µl and incubated (37°C, 4 h). The amount of vector pSL2793 used was (104.2 ng/µl, 75 µl). The digested vectors were visualized following electrophoresis on 0.8% agarose and the observed high molecular weight band was extracted under blue light transilluminator (Safe Imager^{TM}2.0). Vector DNA in the extracted gel slice was cleaned using Wizard SV Gel and PCR Clean-Up System (catalogue number: A9281, Promega) and concentrated to desired concentration using concentrator plus (Eppendorf). The final vector was assembled using NEB assembly as follows: all fragments were assembled in 1:1:1 stoichiometry. The concentration of each of the homology region required were calculated using NEB calculator. The insert DNA length was set as 1000 bp (which is the length of our homology region) or 1145 bp for the spectinomycin resistance gene, vector DNA length set as 12384 bp (which is the length of the plasmid) and the vector DNA mass as 227.5 ng/µl for pSL1050 and 170 ng/µl for pSL2793. Each of the fragments were mixed with NEBuilder HiFi DNA Assembly Master Mix (2X, 10 µl) and nuclease free water to make up a total volume of 20µl and incubated (50°C, 60 minutes) in a thermocycler. The assembled vector was either stored at -20˚C until further use or diluted (4-fold) and transformed immediately into *E. coli* Stellar^{TM} competent cells PT5055-2.

For transformation, four-fold diluted assembled vector (2 µl) was added to Stellar^{TM} competent cells (50 µl), mixed gently by pipetting and placed on ice for 30 minutes. The cells were heat shocked (42°C, 45 s) in a waterbath, then returned to ice (2 min). Warmed super optimal broth with catabolite repression (SOC) media (37°C, 400µl) was added to the cells and incubated (37°C, 200 rpm, 1 h). The cells (50 µl and 100 µl) were then plated onto LB agar plates supplemented with kanamycin (50 µg/ml) plates for selection and incubated overnight (37°C). Randomly selected colonies (20) were picked and patched onto a new LB agar plates supplemented with kanamycin (50 µg/ml) plate and HRT inserts verified using colony PCR. Individually picked colonies were mixed with Go Taq Green buffer (5X, 5 µl), dNTPs (10 mM, 0.5 µl), Go Taq polymerase (5 U/µl, 0.12 µl), $MgCl_2$ (25 mM, 2 µl), water along with primers HRT_PCR_F and 1050_HRT5_R_hrt3OH for 1050 and HRT_PCR_F and 2793_HRT5_R_hrt3OH for 2793 (10mM, 1 µl each; Appendix II and III) to a total volume of 25 µl reaction mix. Colonies positively identified with HRT inserts were then cultured overnight (37°C,

200 rpm, 12-16 h) in 10 ml LB supplemented with kanamycin (50 μg/ml). Glycerol stocks (in 25% glycerol) were made with 800 μl of overnight culture and stored at -80°C. The resulting editing plasmids with HRT inserts were epSL2793 for 2793 gene in CC9311, marked epSL1050 and unmarked un_epSL1050 for 1050 gene in WH8109. All editing plasmids were extracted using the rest of the overnight culture using Wizard® *Plus* SV Minipreps DNA Purification System (catalogue number: A1460, Promega). The purified plasmid products were sequenced at Macrogen (Korea) to verify the sequence of the homologous region inserts.

To facilitate the uptake of editing plasmids epSL2793, epSL1050 and un_epSL1050 into *Synechococcus* spp. CC9311 and WH8109 respectively, two different transformation techniques electroporation and conjugation were tried.

2.3.4 Electroporation

Electroporation involves the process of exposing cells to high intensity electric field pulses to induce temporary pores in the cell membrane therefore, allowing cells to take up the external DNA[88].

Synechococcus spp. strains WH8109 and CC9311 were grown in 200 ml batch cultures in ASW[71] media at 22°C, 100 rpm, 40 μmol photons m^{-2} s^{-1}. The cells were harvested at mid log phase 7 x 10^7 - 4 x 10^8 cells/ml by centrifugation (7,800 rpm, 10 mins) and resuspended in sorbitol (1M, 25 ml). This process was repeated twice before resuspending into 100 μl of 1M sorbitol to final concentration of 10^8-10^9 cells/ml. Different ratio of the editing plasmid epSL2793 or epSL1050 or un_epSL1050 (50, 100, 250 and 1000 ng) with different volumes of *Synechococcus* cells ranging from 10^6, 10^7 , 10^8 and 10^9 cells/ml were mixed in a 1.5 ml microfuge tube and transferred to prechilled 2 mm electroporation cuvettes and incubated on ice for 5 min. Cell suspension without plasmid was used as a negative control. Cells were electroporated using a Bio-Rad Gene Pulser at field strength of 2.5 kV, resistance 200 Ω and capacitance 25 μF making sure no blue arc appears. Immediately after electroporation 1ml of ASW was added onto the cells directly into the cuvettes and incubated at room temperature. The electroporated cells were then transferred to 4 ml ASW into a culture flask. After 48 h of recovery under 5-10 μmol photons m^{-2} s^{-1}, the cells were mixed with 0.4% LMP agarose supplemented with different concentrations of kanamycin (5, 10, 15, 25, 30μg/ml) and pour plated [86]. The concentration of *Synechococcus* cells plated were 10^3,10^4,10^5,10^6 and 10^7 cells/ml. The plates were incubated at low light i.e. <10μmol photons m^{-2} s^{-1} for 24 hours and then transferred to 10-20 μmol photons m^{-2} s^{-1} of cool white light.

2.3.5 Conjugation

Conjugation is another method facilitating the uptake of genetic material via direct contact of cells. *Synechococcus* spp. strains WH8109 and CC9311 were grown in 200 ml PCRS11 Red Sea medium[89] at 22°C, 100 rpm, 15µmol photons m^{-2} s^{-1}. The cells were harvested at mid log phase 7 x 10^7 - 4 x 10^8 cells/ml by centrifugation (7800 rpm, 20 mins) and resuspended in 200µl of PCRS11 Red Sea medium. Conjugative *E. coli* strain MC1061 comprised of 3 plasmids pRK25 (conferring ampicillin resistance), pRL528 (conferring chloramphenicol resistance) and one of the editing plasmids (epSL2793 or epSL1050 or un_epSL1050) was grown overnight (37 °C, 200rpm) in 20 ml of liquid LB media supplemented with kanamycin (50 µg/ml), chloramphenicol (30 µg/ml) and ampicillin (50 µg/ml). *E. coli* cells were harvested by centrifugation (7830 rpm, 10 mins) and resuspended in 70 µl of LB. For biparental mating, 50 µl of *Synechococcus* cells was mixed with 10 µl, 20 µl and 40 µl of conjugative *E. coli* MC1061 and spotted onto 0.2 µm filter placed on a 0.6% agar plate made with PCRS11 Red Sea media and incubated at room temperature for 48 h at light intensity of 10 µmol photons m^{-2} s^{-1}. 50 µl *Synechococcus* cells without *E. coli* MC1061 was spotted onto the filter as a control. The cells on the filters were resuspended into 1 ml of PCRS 11 media. 10µl and 100µl of each of the resuspended cells were mixed with 0.3% PCRS11 (20ml) agar supplemented with kanamycin (25 µg/ml and 50 µg/ml). The temperature of PCRS11 agar should be less than 25°C. The mixture was poured into petri dishes and solidified. The plates were then incubated at room temperature at 5 µmol photons m^{-2} s^{-1} for 3 days and then transferred to 15 µmol photons m^{-2} s^{-1}. Individual colonies appeared within 2-3 weeks.

2.3.5.1 Separation of transformants and mutants

Plasmid uptake by *Synechococcus* spp. strains were confirmed with PCR using the reaction mixture comprising primers kan_F (10 mM, 1µl), kan_R (10 mM, 1µl) (Appendix II), dNTPs (10 mM, 1µl), buffer (10X, 2.5 µl), Taq polymerase (5U/µl, 0.2 µl) and nuclease free water to make up a total volume of 24 µl. For separation of mutants from transformants a series of passaging of transformants and regular PCR checks were done. Briefly, liquid culture for colonies that gave positive PCR results for plasmid uptake were started. This was done by aspirating colonies with a pipette tip and dispersing it into PCRS11 media supplemented with kanamycin (25 and 50 µg/ml). They were left to grow for a week and three PCR checks were done. One to make sure the editing plasmid has not been lost with primers kan_F and kan_R. Second, to check if the gene has been knocked out with gene specific primers 2793_detect_F and 2793_detect_R for 2793 and 1050_detect_F and 1050_detect_R for 1050

and third, to ensure gene has been knocked out and DNA repair has been done with homologous regions with primers 2793_sc_genhrt5spec_F and 2793_sc_spechrt3gen_R for 2793 and 1050_sc_genhrt5spec_F and 1050_HRT_screen2_R for 1050. All the primer sequences are listed in Appendix II. After a few passages assuming that the edit may have been done, we attempted to cure the plasmid out of the conjugated *Synechococcus* strains by allowing them to grow for few days in non-selection media followed by pour- plating in non-selection media, allowing spontaneous loss of the plasmid. After 12 regular passages and PCR check with the primer pairs to verify that the edit has been made, bands of approximately right size were extracted from the gel under blue light transilluminator (Safe Imager^{TM}2.0) and cleaned using PCR Clean-Up System (catalogue number: A9281, Promega) and sent for sequencing to (Macrogen, Korea) to verify the sequences.

Chapter 3: Results

3.1 Distribution of *Synechococcus* along the eastern coast of Australia

This study utilised the microbial observations in form of 16S rRNA amplicons and shotgun metagenomics that have been sustained since 2012 across the three NRS namely NSI, PHB and MAI (Figure 3.1) to provide the first in-depth look at the distribution and community structure of a key phytoplankton, marine *Synechococcus* in the eastern coast of Australia. The composition of *Synechococcus* community showed distinct latitudinal and seasonal variation (Figure 3.2).

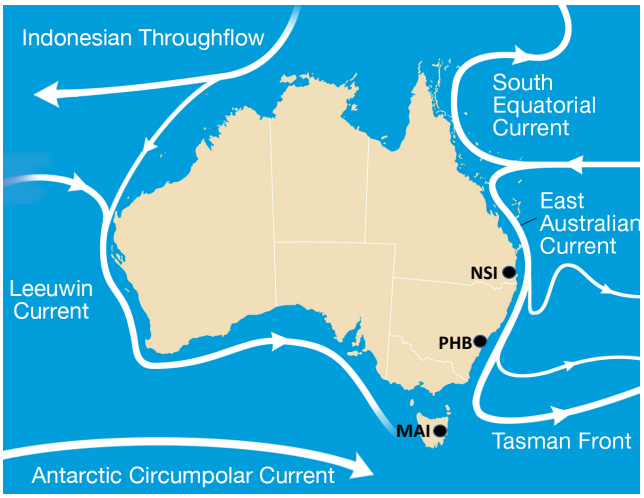


Figure 3.1: Three National Reference Stations of interest including North Stradbroke Island (NSI), Port Hacking (PHB) and Maria Island (MAI) around east coast of Australia within the context of East Australian current. Figure adapted from Australian Government Bureau of Meteorology

At NSI, the most abundant *Synechococcus* was clade IIabc (43.99%). *Synechococcus* subclades IIa, IIb and IIc share identical 16S rRNA gene sequence in the V1-V3 region and cannot be further resolved using this marker, therefore, they are referred to as clade IIabc. Similarly, the most abundant *Synechococcus* clade at PHB was also clade IIabc (47.84%). However, the community here encompassed large relative proportions of other *Synechococcus* genotypes including clade IIh (6.51%), clade IIe (9.66%), clade I (1.73 %) and clade IV (1.00%) which varied seasonally. Other picocyanobacterial community at PHB comprised mainly *Prochlorococcus* HLII (18.85%) and combination of all other less defined *Synechococcus* and *Prochlorococcus* clades each of which were present in very small numbers. The most abundant picocyanobacteria at MAI was *Synechococcus* genotypes clade IV (54.30%) and clade Ia (12.88%). The relative abundance of clade IIabc at MAI was very low (0.72%) whereas clade IIh (4.79%) and clade IIe (5.22%) occurred at higher relative abundance at certain time points. *Synechococcus* clades observed at MAI (i.e. temperate strains) is distinct from the ones observed at NSI and PHB (i.e. tropical strains).

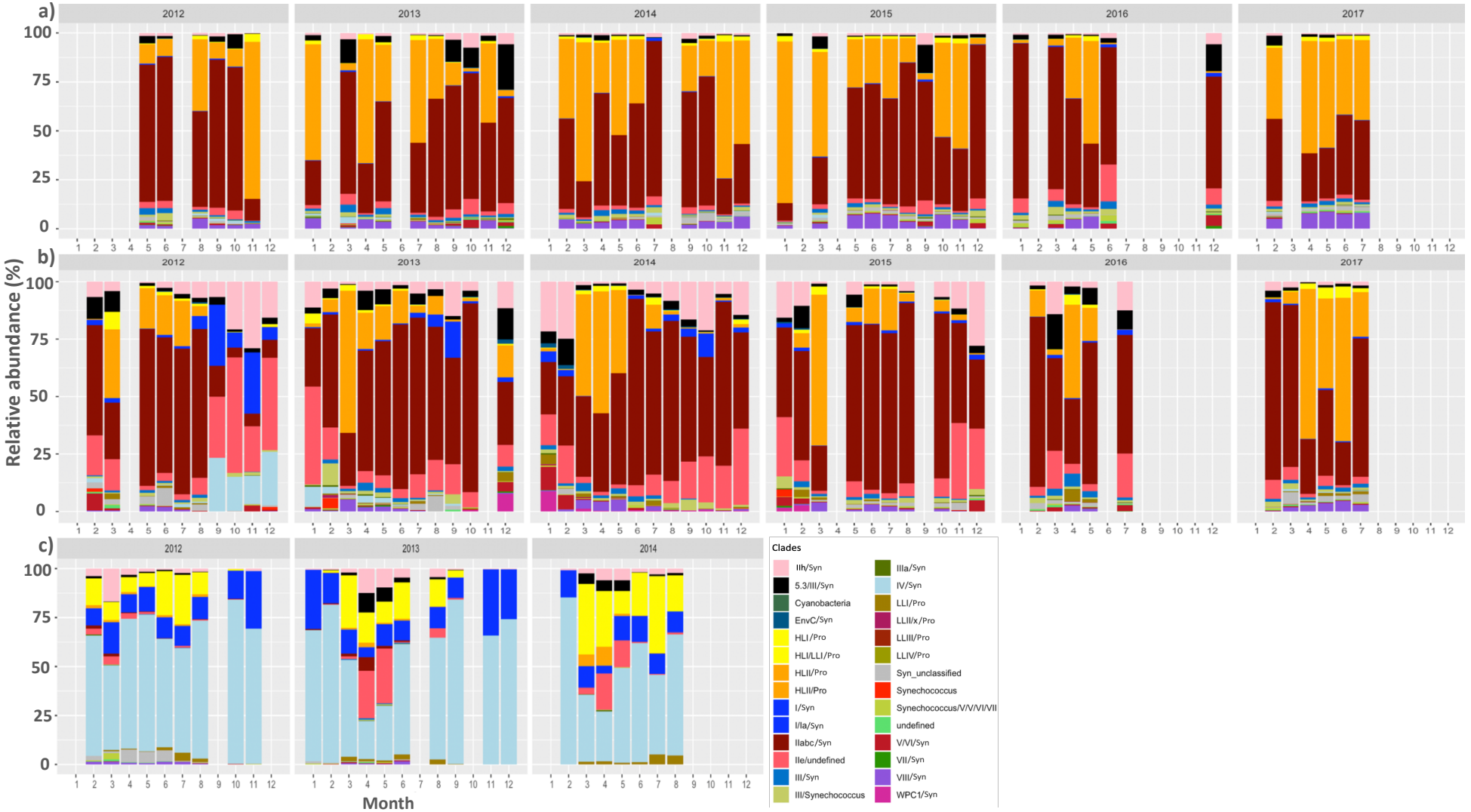


Figure 3.2: Bar chart representation of marine cyanobacteria composition in oceanic regions along the east Australian coast across three National Reference Stations: a) North Stradbroke Island (NSI), b) Port Hacking (PHB) and c) Maria Island (MAI). The community composition was derived from 16S rRNA gene. Each letter represents clade belonging to *Synechococcus* (Syn) or *Prochlorococcus* (Pro) and corresponding colour represents composition in the figure

Synechococcus clade IIh that occurred at higher abundances at PHB displayed seasonal variation and the changes in their abundances were distinct. They were present in highest number during the austral summer months i.e. December, January, February (42.53%) and dropped dramatically during the austral winter months i.e. June, July, August (27.25%). More interestingly, they co-occur with clade IIabc (tropical) and clade Ia (temperate clade) potentially under intermediate conditions (i.e. in-between the two 'stable' biomes) suggesting clade IIh could possibly be a 'transitional' clade. Additionally, relative abundance of clade IIh positively correlated with that of tropical clade IIabc and temperate clade Ia (Figure 3.3) further indicating that clade IIh is possibly a 'transitional clade' that occur in intermediary conditions between the two major biomes. Another noticeable thing is the distribution pattern of clade IIe which is very similar to clade IIh i.e. it varies seasonally and occurs in higher abundance during summer (43.82%) and co-occurs with both clade IIabc and Ia. Furthermore, clade IIh is positively correlated with clade IIe suggesting that they may co-occur under similar environmental conditions.

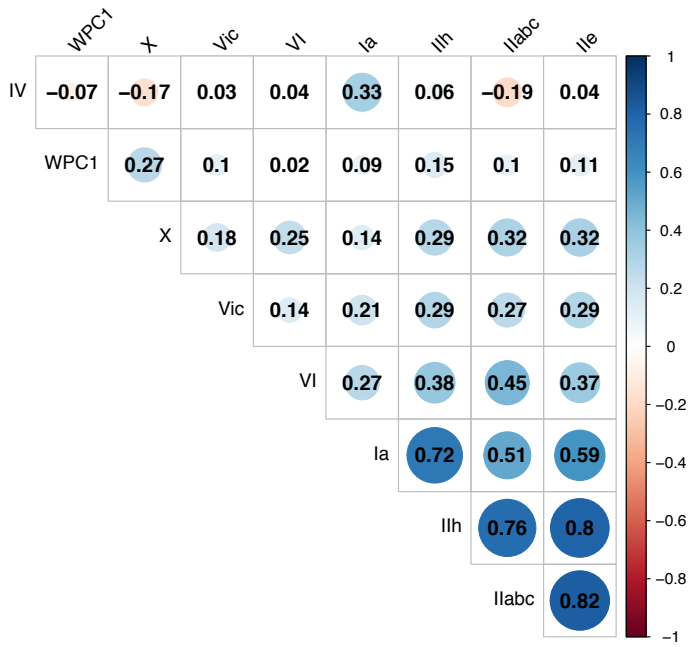


Figure 3.3: Pearson's linear correlation coefficient of clade IIh with clade IIabc, clade IIe and clade Ia ($p<0.01$). Blue indicates positive correlation and red indicates negative correlation. Numbers inside the circle represent correlation coefficient (r). An 'r' of 1 indicates perfect linear relationship; An 'r' of -1 indicates perfect negative relationship and 'r' of 0 indicates no linear relationship.

3.1.1 Sea surface temperature profiles at each national reference station

The average annual temperature at NSI was 23.6°C with a summer average of 25.5°C and winter average of 21.6°C. The highest temperature was observed during March (27.5°C) and the lowest temperature was observed during September (19.9°C). The average annual temperature at PHB was 19.9°C with a summer average of 20.5°C and winter average of 18.3°C. The highest temperature was observed during March (24.5°C) and the lowest temperature was observed during September

(16.3°C). As for MAI, the average annual temperature was 14.6°C with a summer average of 16.5°C and winter average of 13.4°C. The highest temperature was observed during February (18.1°C) and the lowest temperature during October (11.91°C) (Figure 3.4). Overall, the temperature profiles depict NSI as a tropical site, PHB (where clade IIh was most abundant) intermediate between tropical and temperate region and MAI as a temperate site.

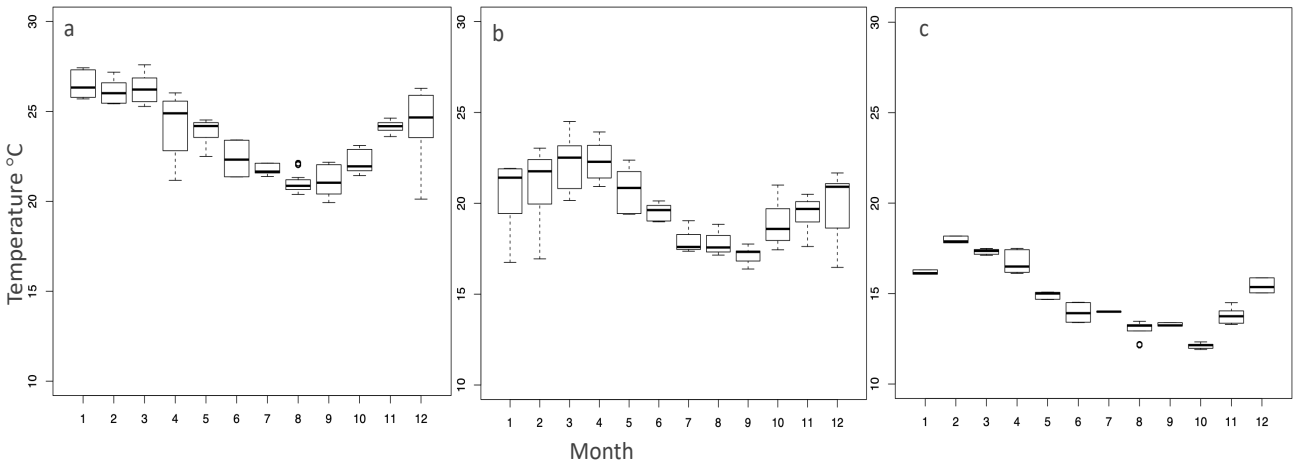


Figure 3.4: Boxplots showing temperature profiles at three different National Reference Stations (depth < 25m) a) North Stradbroke Island (Years: 2012-2017) b) Port Hacking (Years: 2012-2017) c) Maria Island (Years: 2012-2014)

3.1.2 Environmental parameters influencing clade IIh

Spearman rank correlations were computed to check the relationship of clade IIh with environmental parameters that may influence their distribution (Figure 3.5). Clade IIh showed moderate negative correlations to macronutrients (silicate, nitrate, phosphate) and depth, and positive correlations with temperature and salinity. For *Synechococcus* clades that appeared to co-occur with Clade IIh, there were some differences in their relationships with environmental parameters. Clade IIabc was highly positively correlated to temperature while moderately correlated positively to salinity and negatively to depth, nitrate and phosphate. Clade IIe showed a moderate positive correlation to salinity and temperature. In contrast, Clade Ia showed moderate negative correlation to temperature and its negative correlation to depth and macronutrients were weak. Similarly, clade IV has weak negative correlation with temperature. Temperature was found to be a key variable influencing clade IIh and other co-occurring clades i.e. while tropical clade (IIabc) was positively correlated to temperature, negatively correlation was observed for temperate clades (clade Ia and IV).

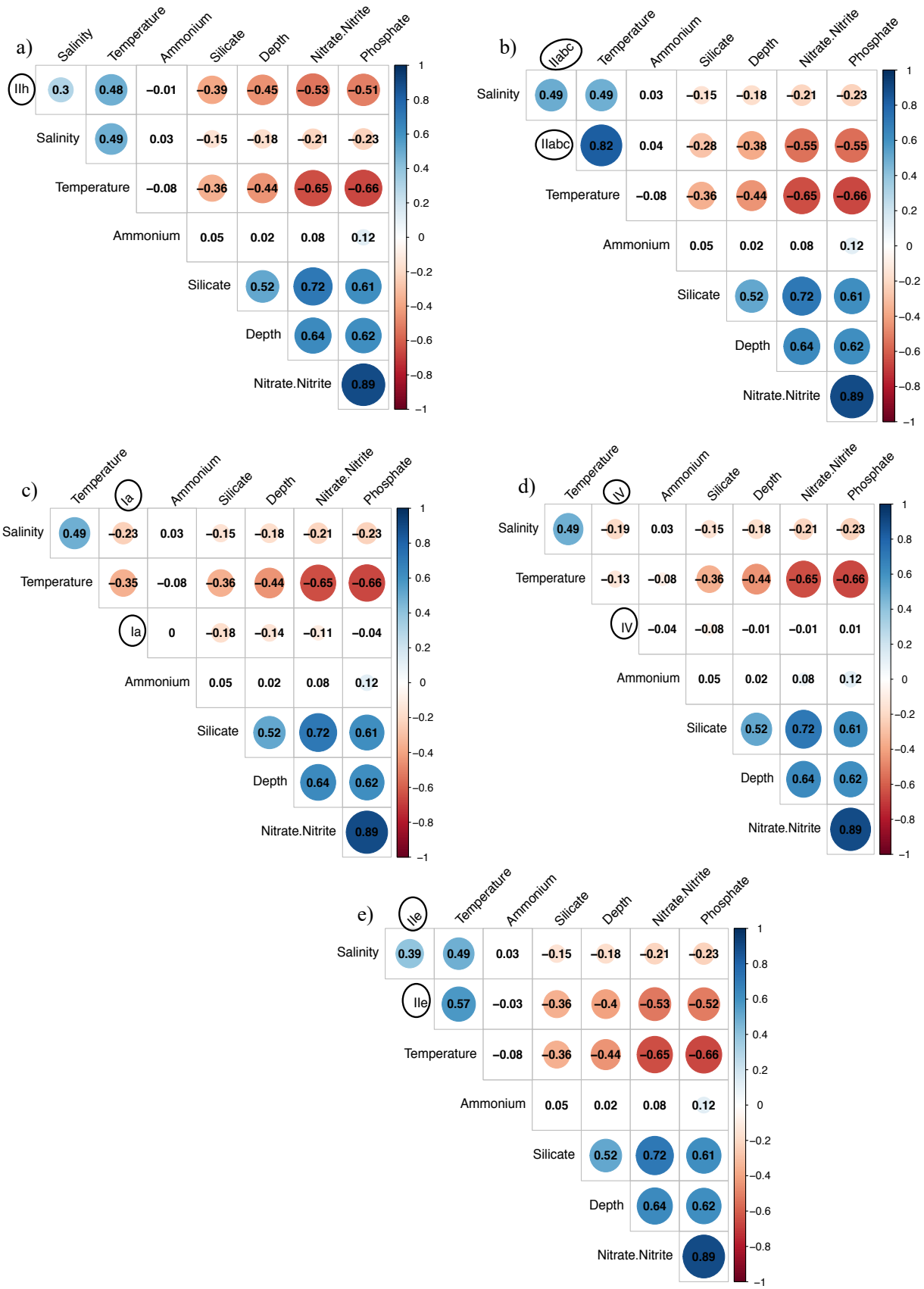


Figure 3.5: Spearman's correlation coefficient of a) clade IIh b) clade IIabc c) clade Ia d) clade IV and e) clade IIe with environmental parameters ($p < 0.01$). The clades are shown inside black circles. Blue indicates positive correlation and red indicates negative correlation. Numbers inside the coloured circles represent correlation coefficient (r). An 'r' of 1 indicates perfect linear relationship; An 'r' of -1 indicates perfect negative relationship and 'r' of 0 indicates no linear relationship.

Abundance versus temperature relationships for clade IIh along with other major clades were determined by collecting abundance information of each clade into 1°C temperature bins and calculating the mean abundance for each bin. Genotype Thermal Index (GTI) or ecological temperature optima was defined using Kernel density estimation weighted by mean abundance. GTI for clade IIh was found to be 15-25°C, which is intermediate between the temperature optima observed for temperate clades (I and IV) i.e. 10-20°C, and tropical clade (IIabc) i.e. 20-30°C (Figure 3.6).

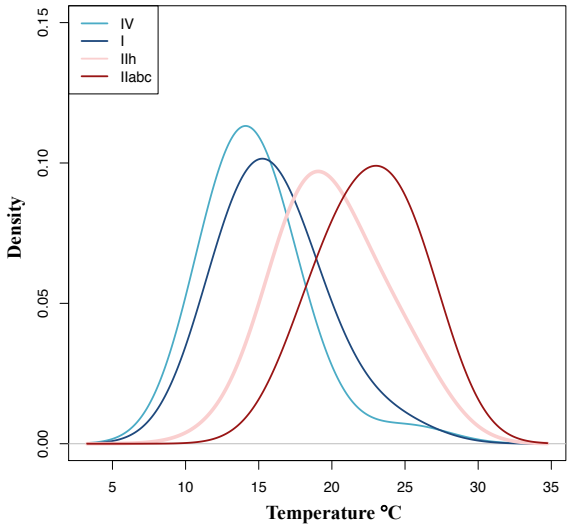


Figure 3.6: Genotype thermal index showing intermediate ecological temperature optima for clade IIh (pink) in comparison to other major clades IIabc (maroon), I (dark blue) and IV (light blue)

3.2 *Synechococcus* genomes from the east Australian coast

Of the currently available 51 *Synechococcus* genomes (Cyanorak Database [http://application.sb-roscoff.fr/cyanorak/]), there is only one for clade IIh thus limiting detailed study of the adaptations unique to this potential transition zone clade. Furthermore, there were no genomes for any picocyanobacterial isolates from the Australian marine habitat Therefore, this study fills that knowledge gap by providing 12 sequenced genomes of *Synechococcus* strains, isolated from the coastal and oceanic waters in the east Australian marine environment, including the NRS stations PHB and MAI. Prior to sequencing, clades of these new isolates verified using *petB* phylogenetic marker (Table 3.1) determined that 7 of the 12 isolates belonged to clade IIh, which were isolated from Sydney Freshwater Beach, Port Hacking and the Tasman Sea.

3.2.1 Genome assembly and relatedness

All of the genomes assembled in this project has been deposited at GenBank (Bioproject accession number: PRJNA516148) but is not publicly available yet. The majority of the assembled genomes

were near completion (i.e. >90%) with less than <3% contamination verified using CheckM[77] and comprised <25 contigs. Interestingly, most of the assembled genomes also provided genomes for heterotrophic bacteria belonging to family Rhodobacteraceae (UID3360), often found associated with phytoplankton. Within the scope of this thesis, genome assemblies were refined iteratively, with manual extraction and de novo assembly of the ribosomal RNA operons and isolated de-novo re-assembly of the regions corresponding to gaps. However, further work (e.g. primer walking and repeat resolution) in future is needed to complete and close the genome sequences.

Table 3.1: Summary of genome assemblies of 12 isolates along with clades verified using *petB* phylogenetic marker

Strain *Synechococcus sp.*	*petB* Classification	Genome completion *Synechococcus* (%)	Number of contigs	Length	Genome Completion *Rhodobacteraceae*
. HB1133	II h	99.46	10	2373219	-
HBA1120	II h	99.46	10	2363388	-
PH41509	II h	90.19	28	2161481	-
MU1611	II h	99.73	11	2338323	80.45
MU1625	II h	99.18	13	2334076	-
MU1642	II g	99.7	22	2116729	89.94
MU1643	II e	100	16	2111352	61.25
MU1644	II g	99.73	22	2199727	26.78
MU1648	II h	96.47	21	2282239	99.47
MU1651	II a	71.01	20	2128818	-
MU1652	II h	94.29	24	2158180	0
MU1655	I a	75.69	277	1889680	-

The phylogenetic relationship of the new sequenced genomes was visualised using a genome tree (Figure 3.7) The genome tree resolved the major *Synechococcus* lineages (I, II, III and IV) into distinct clades with 100% boot- strap support for neighbour-joining (NJ). To identify the genome relatedness and resolve the taxonomic structure of the new sequenced genomes ANI, 16S rRNA and *petB* gene identity were computed. (Table 3.2; Appendix VI).

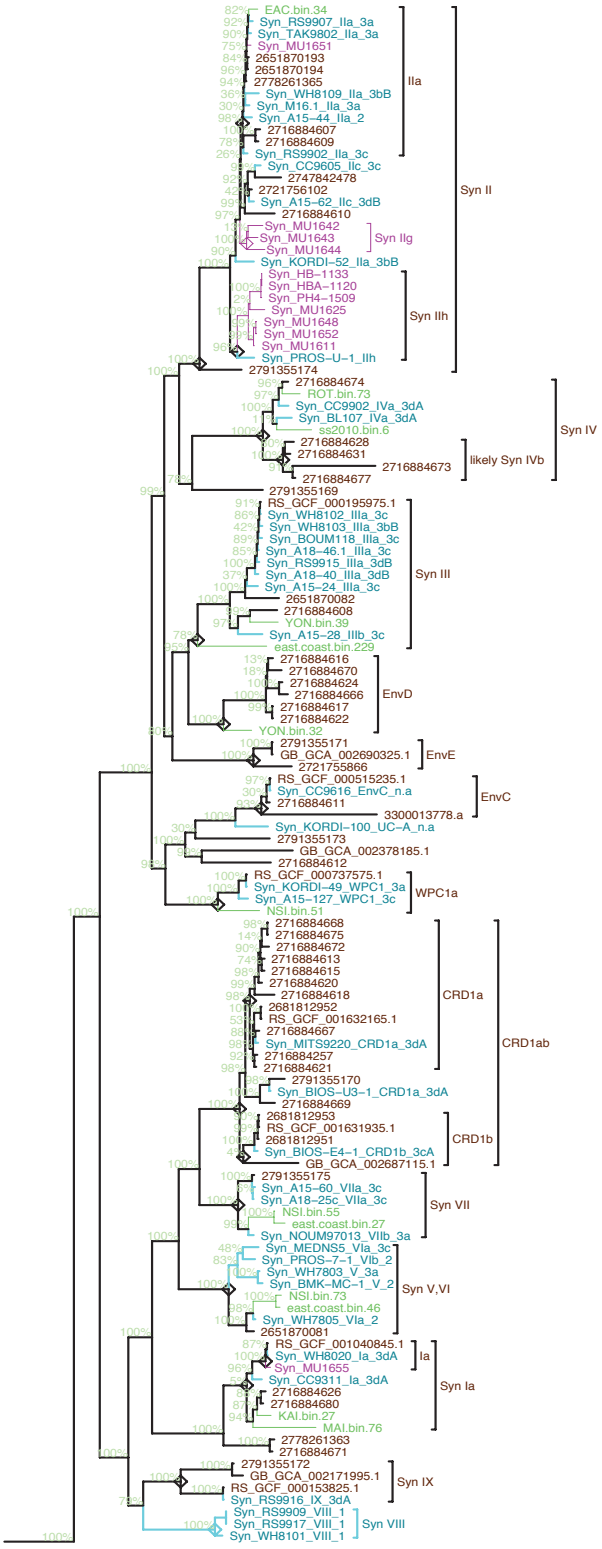


Figure 3.7: Subset of picocyanobacterial genome tree compiled with 21,388 genomes (120 concatenated conserved core genome protein sequences), including ~220 complete or near complete cyanobacteria genomes available in GenBank, 25 pre-release marine genomes (courtesy of the University of Warwick and Station Biologique de Roscoff), twelve new draft genomes of local isolates (shown in pink) and 22 'clean' metagenome assembled genome from local marine cyanobacteria (using data from the Marine Microbes Project) showing phylogenetic placement of the new sequenced genomes (in pink).

Table 3.2: A subset matrix showing genome relatedness of *Synechococcus* sp. HB1133 (clade IIh) with other isolates. Individual clades are shown in brackets.

Synechococcus sp.	**HB1133 ANI (%)**	**HB1133 16S rRNA (%)**	**HB1133 Number of bp difference 16S rRNA**	**HB1133 *petB* identity (%)**	**HB1133 Number of bp difference *petB***
HBA 1120 (IIh)	99.98	100	0	100	0
MU1611 (IIh)	95.53	99.932	1	91.172	58
MU1625 (IIh)	95.1	99.932	1	91.324	57
MU1651 (IIa)	86.03	99.932	9	92.084	59
MU1643 (IIe)	85.88	99.46	8	91.02	58
MU1644 (IIg)	85.85	99.46	8	91.174	63
Pros-U-1 (IIh)	89.34	99.932	1	90.411	52
CC9311 (Ia)	77.37	99.055	14	84.17	104
WH8109 (II)	85.62	99.325	10	90.411	63
WH8102 (III)	80.0726	99.19	12	86.758	87
BL107 (IV)	78.2173	99.595	6	84.323	103

Clear species delineation was observed for the new sequences using ANI values, with the current accepted threshold for species demarcation of <94% (i.e. isolates with >94% ANI values belong to same clades)[90] while 16S rRNA gene identity for the same were >99% providing no clear delineation (Table 3.2; Appendix VI). An exception was observed for *Synechococcus* sp. Pros-U-1 (a previously sequenced clade IIh strain, isolated from Moroccan upwelling region of Atlantic Ocean) which though phylogenetically close to Australian clade IIh genomes had ANI values ranging from 89.11-89.53% suggesting PROS-U-1 and Australian clade IIh isolates should belong to different clades. Therefore, ANI threshold for delineating 'species' in marine *Synechococcus* could potentially lie between 87 and 91 % in agreement to previously proposed threshold by Dufresne *et.al.,* 2008[83]. As for *petB* gene identity, isolates belonging to same clade had values >90% providing clear species

demarcation. Genome relationship between ANI vs 16S rRNA and *petB* percentage identity (Figure 3.8) show most of the picocyanobacteria genomes with 95-100% 16S rRNA identity and 85-90% of *petB* gene identity is equivalent to ANI values less than 85% further suggesting ANI threshold for marine *Synechococcus* to be less than standard 94%.

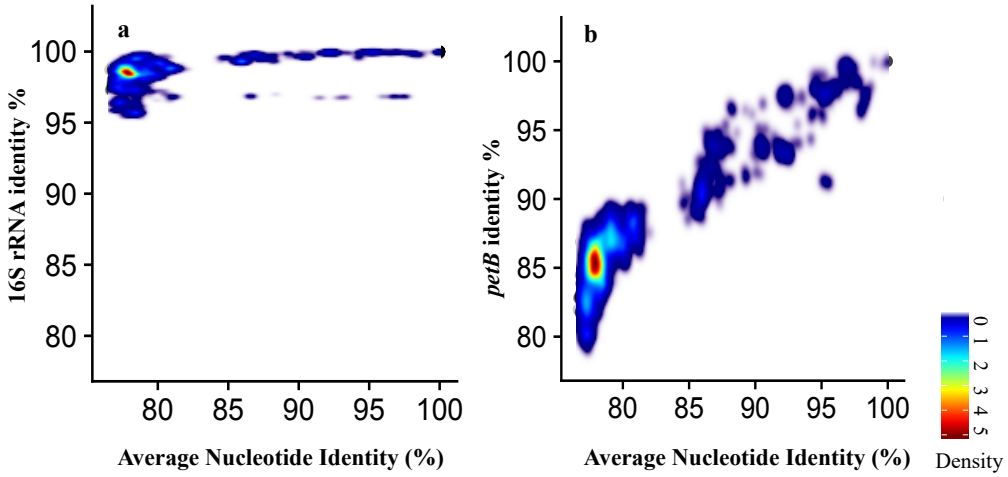


Figure 3.8: Genome relationship a) ANI versus percentage of 16S rRNA identity b) ANI versus percentage of *petB* identity

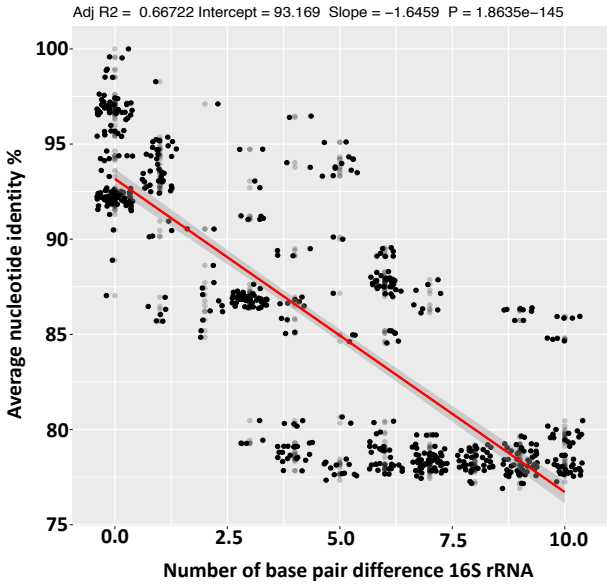


Figure 3.9: Plot of genome wide ANI versus number of base pair difference in 16S rRNA for ~220 complete or near complete cyanobacteria genomes available in GenBank, 25 pre-release marine genomes, twelve new draft genomes of local isolates and 22 'clean' metagenome assembled genome from local marine cyanobacteria (using data from the Marine Microbes Project). Regression line (shown in red)

To verify if isolates within the same clades are genetically distinct or identical, change in number of base pairs were computed (Table 3.2; Appendix VI; Figure 3.9). It was found that a single nucleotide base pair change in 16S rRNA correlates with a >4.47% change in the entire genome suggesting isolates having 16S rRNA identity >99% may still be genetically distinct. *petB* sequences had higher

number of changes in base pair sequences giving a clearer indication of genetic distinctiveness between representatives belonging to same clade (Table 3.2). On other hand ANI values >99.98% possibly could indicate towards two isolates being genetically identical such as *Synechococcus* sp. HB1133 and HB1120 with no base pair change seen between the two genomes supported by both *petB* and 16S rRNA sequences.

3.2.2 Genes unique to clade IIh

Clade IIh is mostly seen to be abundant in the possible transitional conditions between the temperate and tropical provinces during austral summer (section 3.1). Determining genes that are specific to clade IIh may provide an insight into specific mechanisms that allow their adaptability to specific marine habitats. Genome comparisons of four clade IIh isolates including three local isolates (HB1133, HBA1120, PH41509) and PROS-U-1 (a previously sequenced clade IIh isolate) identified 42 unique genes to clade IIh. While most of the unique genes were hypothetical, unique genes also encoded an osmosensory histidine kinase, RsbU serine phosphatase, Che-Y like response regulators and putative multidrug efflux ABC transporter (Appendix VIII). Many of these unique genes (10/42) occurred in one of the 16 genomic islands in *Synechococcus* sp. HB1133 (Appendix VII).
Higher numbers of response regulators have been associated with picocyanobacteria that occur in more variable environments[42]. In addition to five unique possible response regulators, *Synechococcus* sp. HB1133 was found to have another 16 genes encoding response regulators (i.e. 21 in total). In comparison, WH8020 (clade Ia) had 22 response regulators and CC9311 (clade Ia) is known to have 28[42] and open ocean strains *Synechococcus* sp. WH8109 (clade II) had 15 and *Synechococcus* sp. WH8102 (clade III) is known to have 14 response regulators[42].

3.3 Molecular toolbox for genetic manipulation

Deciphering molecular adaptations that contribute to specific ecological niche selection, such as dynamic oceanic regions/transition zones, requires the verification of the functional role of genes likely involved. Construction of gene knockout mutants is commonly used to verify gene function, however an efficient method for genetic manipulation of marine *Synechococcus* has not been established. Therefore, we aimed to optimise a molecular toolbox for marine *Synechococcus* using model representatives, WH8109 of clade IIa and CC9311 of clade Ia to test the utility of gene editing technique based on CRISPR/Cpf1. The number of cells required to obtain isolated colonies, concentration of kanamycin (Km) required to select for transformants, and two different transformation techniques electroporation and conjugation were tested.

3.3.1 Optimising growth conditions

A range of cell concentrations (10^3 to 10^6 cells ml^{-1}) of *Synechococcus* culture was pour plated to test the optimal concentration required to obtain isolated colonies. We determined that 10^3 cells ml^{-1} of *Synechococcus* culture was ideal to obtain sufficient isolated colonies for screening (Table 3.3).

Table 3.3: Determining number for cells for isolated colonies

Concentration of sample (cells ml^{-1})	Number of colonies observed
10^3	Isolated; >200
10^4	Isolated; >200
10^5	Lawn of cells; non-isolated
10^6	Lawn of cells; non-isolated

To determine the concentration of kanamycin (Km) required for selection of transformants from wildtype cells, we used minimum inhibitory concentration assays.

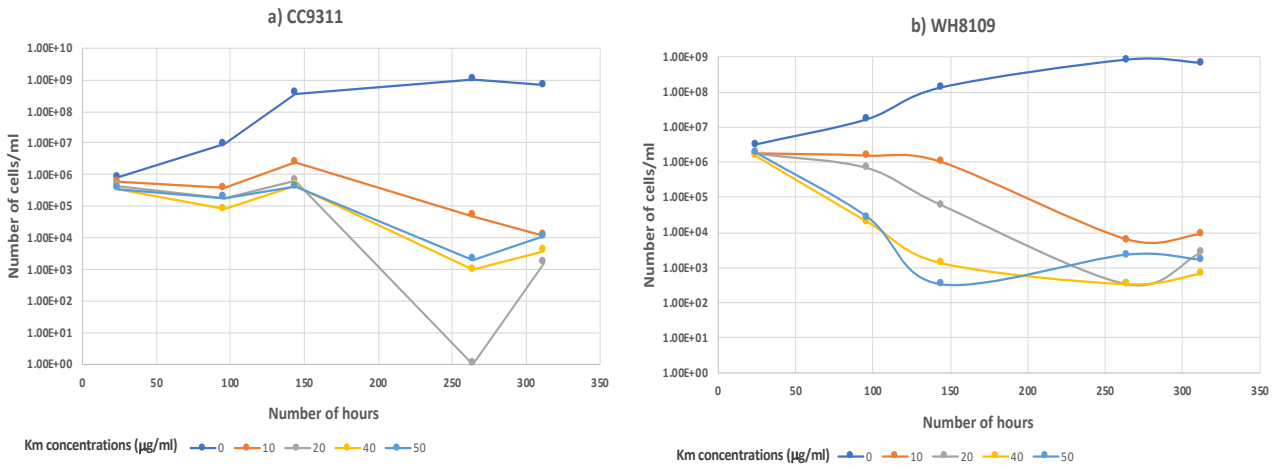


Figure 3.10: Growth curves of *Synechococcus* sp. a) CC9311 b) WH8019 at different concentrations of kanamycin 0 µg/ml (dark blue), 10 µg/ml (orange), 20 µg/ml (grey), 40 µg/ml (yellow), 50 µg/ml (light blue)

For both *Synechococcus* strains, WH8109 and CC9311, a sharp decline (90-100%) in cell numbers were observed after 150 hours for all tested Km concentrations (Figure 3.10). Highest decrease in cell numbers were observed for Km 20µg/ml for both strains wherein cell concentration dropped from 6.47 x 10^5 cells ml^{-1} to below the detection limit for CC9311 cultures (Figure 3.10a) while for WH8109 cultures cell concentrations decreased from 6 x 10^4 cells ml^{-1} to 3 x 10^2 cells ml^{-1} (Figure 3.10b). From these observations 20 µg/ml Km seems to be the concentration after which cell growth declines rapidly. Therefore, two different concentrations, higher than 20 µg/ml of Km (i.e. 25 and

50µg/ml) were chosen for selection of transformants and to exclude cells that may have ability for potential spontaneous mutation. However, lower concentrations i.e. Km 5 and 10 µg/ml were used following electroporation since *Synechococcus* cultures are highly stressed after electroporation and tend to die at higher concentrations of Km (25 and 50 µg/ml).

3.3.2 Electroporation as a method of transformation

Electroporation of two *Synechococcus* strains with editing plasmids (Appendix IV), resulted in mixed results. For strain CC9311, colonies were visible only on control plates (i.e. for cells that had not been electroporated). Electroporation of CC9311 was further tested using series of concentrations for number of cells and plasmids (section 2.5). However, no colonies were visible on any Km selection plates. For strain WH8109 many colonies (~40 per plate) were obtained on both control and Km (5-10 µg/ml) plates but no colonies were visible on plates with higher kanamycin concentrations.

3.3.3 Transformation using conjugation

For conjugation via biparental mating, 50µl (~10^8 cells/ml) of *Synechococcus* cells was mixed with 10µl, 20µl and 40µl of conjugative *E. coli* MC1061 following which 10µl and 100µl of each resuspended cells were plated on PCRS11 media supplemented with Km (25µg/ml and 50µg/ml).
For CC9311, ~40 transconjugant colonies were obtained on a single plate that had 100 µl volume of resuspended culture from a biparental mating of 40 µl of *E. coli* and 50 µl (10^8 cells/ml) of *Synechococcus* spp. CC9311 in PCRS11 agar plate supplemented with Km 25µg/ml.
As for WH8109, conjugation with both marked (i.e. epSL1050) and unmarked (i.e. un_epSL1050) plasmids were done. More than 100 colonies in total were obtained on Km (25 and 50µg/ml) selection plates with both marked and unmarked plasmids.
Therefore, higher numbers of transformant colonies were obtained with conjugation as a method for transformation compared to electroporation, for both strains WH8109 and CC9311.
Plating with two different concentration of Km (50µg/ml and 25 µg/ml) also showed different resistant capacities of these strains. *Synechococcus* sp. WH8109 seemed to be more resistant to kanamycin selection when compared to CC9311. For WH8109, PCRS11 agar plates supplemented with kanamycin 25µg/ml had almost double the number of colonies as compared to agar plates supplemented with 50µg/ml Km but for CC9311 plates supplemented with 50µg/ml Km had no colonies at all. Therefore, it seemed that CC9311 is more sensitive to the presence of kanamycin, and hence concentrations lower than 25µg/ml should be used for CC9311 but 50µg/ml Km can be used for WH8109.

3.3.4 Verification of possible completion of editing process

Plasmid uptake via conjugation by both *Synechococcus* strains were verified (gel image not shown) using primers designed to target the kanamycin resistant gene present in the backbone of the editing plasmids (Appendix I). Positive colonies were then allowed to grow for 3-4 days in PCRS11 media with Km selection (50µg/ml) following which they were subcultured or passaged (~1000 times dilution). Passaging of liquid transformant cultures were repeated for 5-6 times, allowing them to grow for 3-4 days after each passage, after which PCR checks were done. No amplicons (gel image not shown) were obtained when primers specific to the gene of interest were used, which is the expected result when the target gene has been knocked out, i.e. a successful gene knockout should not yield any amplicon. When primers homologous to sequences flanking the edited region were used, amplicons of multiple sizes were obtained (Figure 3.11). This included an amplicon of ~1289 bp, which is the expected size if the edit was completed, and another at ~1838bp, indicating that the original intact gene is still present. These results suggested that the gene may have been knocked out in some cells but not all. Furthermore, it also suggested ploidy nature (i.e. multiple chromosomes) of *Synechococcus* and that the editing process may have been completed in some chromosomes but not all.

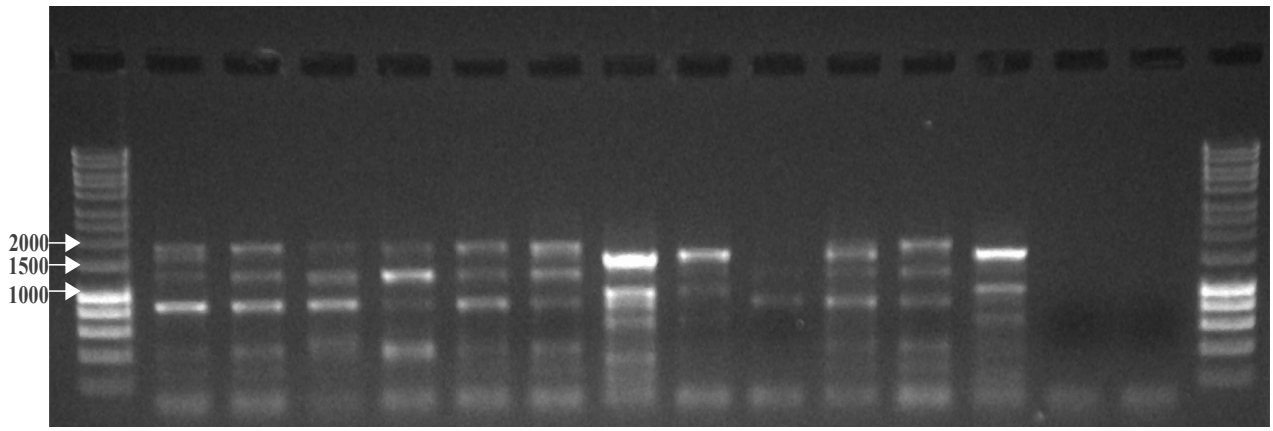


Figure 3.11: Gel image showing multiple bands after six passages with bands showing possibilities of gene knockout (~1289 bp) and no knockout (~1838 bp) in the same lane. First and last lane: 1kbp hyperladder

Therefore, two different conditions were tried with these liquid cultures. One, assuming that the gene may have been knocked out in at least some of the population, the cultures were grown in non-selection liquid medium for a week to cure the plasmid out, and then plated (without selection) to obtain individual colonies of possible mutants. However, screening of colonies with gene specific primers, yielded an amplicon indicating that the gene had not been knocked out. This suggested that gene knockout may have been lost while growing them non selection media and that the growth conditions may have selected for wildtype over possible mutants.

Second, the liquid cultures were passaged further (6 times) in liquid media with Km selection to allow more cells in the population to be edited. During this process, PCR checks after each passage with primers for amplification of edited region could not verify the completion of editing process as amplicons of multiple sizes were obtained. Furthermore, sequencing results for the right sized amplicon still showed the presence of original intact gene.

Chapter 4. Discussion

4.1 Clade IIh as a 'transitional' clade

This study provides the first in-depth look at the distribution and community structure of a key phytoplankton, marine *Synechococcus* in the eastern coast of Australia. We capitalised on the sustained microbial observations, in the form of 16S rRNA amplicons and shotgun metagenomes, undertaken at NRS around Australia since 2012 which provides an opportunity to capture oceanographic and ecosystem responses to seasonal changes.

Analyses of 16S rRNA amplicons revealed latitudinal preference and seasonal patterns of the major *Synechococcus* clades (I, II, and IV) between the three stations that lie within the context of the dynamic EAC system i.e. NSI lies near the source latitudes of EAC, PHB occurs below the separation zone where the EAC turns away from the coast and spawns an eddy field and MAI is affected by the southerly extension of EAC[29,69]. The abundances of these three lineages observed across the NRS in this study agree well with predictions based on a synthesis of previous global studies[9,16,59,91], with clade II abundant in the stations with lower-nutrient subtropical waters (NSI and PHB) and clades I and IV prevalent in the higher latitude station with cooler mesotrophic waters (MAI).

Synechococcus clade IIabc are the most abundant lineages at NSI consistent with it being known to predominate subtropical and tropical oceanic latitudes at temperatures between 20°C to 28°C. The temperature range at NSI (i.e. 20-26°C) as well as at PHB (i.e. 18-24°C) (Figure 3.4) fits within the preferred thermal range of clade IIabc, thus potentially supporting their abundance. In contrast, clades Ia and IV which are predominant in cold temperate waters[9] at temperatures between 10-20°C, were the most abundant clades at MAI which is consistent with the temperature range 12-18°C observed for this region.

While distinct *Synechococcus* clades are observed at the NSI and MAI, the environmental conditions at PHB seems to be supporting the growth of both temperate and tropical zone clades. For example, *Synechococcus* clades IIabc along with (Ia and IV), which are usually found in the tropical and temperate regions respectively, appear sporadically at PHB. This indicates towards a major transition between sub-tropical and temperate biomes occurring between PHB and MAI. Analyses of picocyanobacterial genotypes using the time series dataset from NRS provide new insights regarding their fine scale biogeography, particularly the occurrence of clade IIh which appeared episodically at PHB, and less often at MAI. The higher abundance of clade IIh at these potential transition sites were interesting since it has previously been seen to be associated with ecotones[16].

Clade IIh is found to be the most abundant in PHB during austral summer months and found in co-occurrence with clade IIabc (tropical) and clade Ia (temperate) potentially under intermediate

conditions between two 'stable' biomes with possible influence from the EAC. The EAC is known to be strongest during summer. It separates from the coast at around 31- 32°S. By the time the EAC reaches 33°S it starts re-circulating back northward and the majority of EAC that does not recirculate moves eastward[24,28,92]. PHB is located at 34°S[29,69] which is close to where the EAC separates and starts recirculating and therefore could be a potential site where the effect of EAC recirculation (i.e. the eddies) as well as the effect of uncirculated water of EAC is possibly highest. The water conditions at PHB is highly stratified during the austral summer months but upwelling usually occurs via the eddies or the wind driven slope water intrusions[69] causing vertical mixing of the water and thereby potentially bringing nutrient from the deep thermocline layer to the surface. Therefore, since PHB is situated just downstream of the EAC separation zone these variable or changing conditions due to EAC may highly affect the water conditions at PHB including mixing of nutrients and change in temperature resulting in moderate conditions at PHB. The temperature at PHB is subtropical in nature (Figure 3.5). The average annual temperature observed at PHB is 19.9°C which is intermediate between the average temperature observed at NSI (23.62°C) and MAI (14.68°C). Higher abundance of clade IIh at these variable conditions support the hypothesis that it potentially is a 'transitional clade' that occurs in a dynamic mixed condition between the two major 'stable' biomes. Temperature appears to be a key variable influencing clade IIh and other co-occurring clades even though other variables such as nutrients, light etc may have an important role, investigation of which was beyond the scope of this thesis.

Based on these observations we predicted that it would display an ecological temperature optima (or Genotype Thermal Index (GTI)) in between that of clade II(abc) (tropical clade) and clades I and IV (temperate clade), and that the physiology and genome of this strain would display adaptations that reflect the changeable conditions found between microbial provinces. As predicted, the GTI (Figure 3.5) shows that its ecological temperature optima for clade IIh is between 15-25°C, in between the optima for the major temperate and sub-tropical clades further supporting the hypothesis that clade IIh potentially is a 'transitional clade'.

The hypothesis that clade IIh is potentially a 'transitional clade' and the evidence of its co-occurrence with both tropical and temperate clades is supported by linear Pearson's correlation that show strong positive correlation of clade IIh with clade IIabc (tropical strain) and clade Ia (temperate strain) (figure 3.3). Furthermore, its correlation with the environmental parameters agree with the general conditions expected for Port Hacking that support its high abundance at this region. For example, Port Hacking is located on the continental shelf region south of the EAC[69] separation zone and is impacted by the dynamics of flow of EAC and its eddy field (Appendix V) that may result in intermediate conditions and hence support a transitional clade. Moderate positive correlation of clade IIh with temperature and salinity also supports its occurrence in high number at PHB than at other

NRS stations which are more of tropical (NSI) and temperate (MAI) nature. Therefore, these correlations provide us some interesting clues of its adaptation to the areas between 'stable' subtropical and temperate provinces.

4.2 Genomes of local isolates

Compared to the major *Synechococcus* clades, very little is known about clade IIh with only one completed genome of a Clade IIh isolate from the Moroccan upwelling site of Atlantic Ocean [http://application.sb-roscoff.fr/cyanorak/]. Therefore, this study aimed to contribute to that knowledge gap by providing sequences for seven new isolates belonging to *Synechococcus* clade IIh which was used to explore the adaptations unique to this potential transitional clade

4.2.1 Optimisation of culture conditions for maximum recovery of Synechococcus DNA

Nearly all laboratory strains of marine cyanobacteria harbour one or more heterotrophic co-isolates and it is important to note that very few, if any, axenic strains are in culture for marine cyanobacteria (M. Ostrowski *pers comm*). Common heterotrophic species encountered in *Synechococcus* cell lines include *Flavobacteria sp.* and members of the marine *Roseobacter* clade[93]. In order to maximise the recovery of *Synechococcus* DNA from non-axenic cultures for efficient whole- genome assembly, it was important to optimise the culture conditions. The strategy to increase the ratio of *Synechococcus* to heterotrophs (10:1) involved maintaining the balanced growth i.e. monitoring the growth of culture and subculturing before it reaches exponential phase (i.e. ~ 9 x 10^6 cells/ml) and confirmation of decreasing number of heterotrophic (non-pigmented bacteria) by SYBR green staining and fluorescence microscopy. This strategy contributed to higher cellular concentration of *Synechococcus* DNA and hence high near to completion genomes during assembly. Therefore, despite the presence of non-pigmented cells in the cultures used to prepare DNA, the *de novo* assembly of paired-end Illumina MiSeq resulted in almost near complete genomes for six strains.

On other hand, it was also interesting that most of the assembly results (Table 3.2) also resulted in genome assembly of heterotrophic bacteria from the Rhodobacteraceae (UID3360). It is surprising that the heterotrophs seemed to be growing and contributing to the overall DNA even though there was no organic carbon supplemented in artificial seawater media used to grow *Synechococcus* cultures. This could be possibly because of the mutual interaction between the heterotrophs and the phototrophs [94]. *Synechococcus* has seen to be leaky organisms per se, either by release of numerous extracellular vesicles or by cell lysis leading to release of cytoplasmic polypeptide[95] that supports the growth of the heterotrophic microorganisms and in turn the heterotrophs hydrolyse the organic matter

released by the phototrophs including *Synechococcus* before other higher trophic organisms can acquire it.

Therefore, having obtained near complete genomes for Rhodobacteraceae, this study provides a foundation for future studies to further analyse the interaction between the two communities in depth at a genomic level.

4.2.2 Genome-resolved classification with reference to 16S rRNA and petB

The availability of complete genomes has improved our ability to resolve different strains and species, based on nucleotide polymorphisms and shared DNA content. Marine *Synechococcus* exhibits broad genomic diversity with distinct species adapted to specific ecological niche. Therefore, it is crucial to understand what taxonomic resolution is required to determine the genetically distinct populations that may be adapted to specific environmental niche. With this in mind, to verify the genetically distinct lineages, the project compared genome relatedness of the new assembled genomes using ANI, 16S rRNA and *petB* gene percentage identity.

As compared to 16S rRNA gene identity, clear species delineation was observed from the ANI values. For example, *Synechococcus sp.* MU1651(clade IIa) and MU1643 (IIe) that are found in the different branches of the genome tree (Figure 3.7) had 16S rRNA gene identity of >99% (Appendix VI) pointing towards the possibility of them being genetically identical however the ANI values for the pair was 90.543% which is below the currently accepted species threshold of approximately 94%[90]. Similar results were also observed for *Synechococcus* MU1651 (clade IIa) and MU1655 (clade Ia) where their 16S rRNA gene identity was 98.65% while ANI was 77.0635%. Therefore, ANI values, whose calculations are based on computational comparisons of two genomes and involve a large number of widely distributed (typically >1000) lineage specific genes appears to be a better estimator than 16S rRNA gene identity, for assigning strains to different clades/species thereby increasing the strength and resolution of the extracted phylogenetic signals[90]. For all the new sequenced genomes, while a continuum of values ranging from 98.5-100% with no clear limit is observed for 16S rRNA sequence identity for assigning clades ANI values clearly lie below the threshold of 94%. However, neither 16S rRNA nor ANI values completely resolve identification of genetically identical representatives within a certain clade or sub-clades and requires further refinement to determine percentage identity. For example, a previously sequenced isolate belonging to clade IIh, PROS-U had ANI values 89.11-89.53% against other IIh isolates pointing towards them belonging to different clades. Therefore, the threshold for delineating 'species' in the marine *Synechococcus* could possibly lie somewhere between 87 and 91% which is consistent with values suggested previously[83]. The

decrease in threshold is particularly important as one base pair change in the 16S rRNA may be consistent with a ~4.4% change in the entire genome.

As compared to 16S rRNA gene identity values, *petB*, a core genome marker[59] for marine *Synechococcus*, provided better taxonomic resolution at fine scale. *petB* values are closer to ANI values i.e. *petB* values 85-90% are equivalent to less than 85% ANI values (Figure 3.8). Furthermore, higher base pair differences in *petB* is found between isolates belonging to same clades (Table 3.2) thereby providing more resolution to identify identical or distinct lineages within the same clade. Therefore, *petB* gene identity was a better estimator for picocyanobacterial species delineation in our study than both 16S rRNA and ANI for determining genetically distinct isolates belonging to same clade.

4.2.3 Coastal/opportunist lifestyle of clade IIh

Two major lifestyle strategies, 'open ocean/specialists' found in warm-oligotrophic or temperate-mesotrophic waters and 'coastal/opportunists' that often occupy coastal habitats or present in low number in a broad range of ecosystem or occasionally occur in high number in upwelling areas following environmental perturbations have been suggested for marine *Synechococcus* which is supported by the number of genes encoding two-component system histidine kinases and response regulators[83]. *Synechococcus* clades I, V, VI, VIII, IX and sub-cluster 5.2 exhibit high numbers of regulatory systems allowing them to adapt to variable environmental conditions whereas clades II, III, IV, prevalent in open-ocean waters, has low numbers of regulatory systems[42,83].

The positive correlations of clade IIh observed with the temperate and tropical strains (Figure 3.3) along with the hypothesis that clade IIh may have the ability to adapt to the dynamic conditions influenced by the EAC is further supported by the presence of higher number of response regulators when compared to open ocean clades such as II(abc) and almost equal number of response regulators (including histidine kinases) as the coastal environmental clade CC9311 (clade I). Like coastal waters, that are typified as being highly dynamic and more productive than the open oceans[42], transition zones are defined as being highly dynamic and highly productive due to rapid changes in biophysicochemical factors such as temperature and nutrients, resulting in fertilisation and increased primary production[21,96]. This suggests that clade IIh shares similar features as the 'coastal/opportunists' group of marine *Synechococcus* that has higher regulatory capacity to cope with a more variable environment as compared to the 'open ocean/specialists' that have low number of regulatory systems[83]. Therefore, clade IIh having features similar to the 'coastal/opportunists' reflects their capability to adapt to the changeable conditions at oceanic fronts.

4.2.4 Role of unique genes of clade IIh

We predicted that Clade IIh may have several adaptation features that allow them to respond to the wide spectrum of environmental changes similar to the opportunist/coastal group of marine *Synechococcus* because of the presence of high number of regulatory systems. The analysis of multiple representative genome sequences from the oligotrophic clade II (*n=6*), and the temperate clades I (*n* =8) and IV (*n* =2), allowed us to identify the unique genes only found in clade IIh genomes. Comparative genomics identified 42 unique clade IIh genes. It has been shown that accessory genes (i.e. unique to individual lineages) in *Synechococcus* have been suggested to provide a benefit under certain stress conditions, thereby aiding in occupying certain niche and adaptation[36].

The genes that had assigned functions were mostly related to two-component signal transduction systems (TCS) (Appendix VIII). TCS represent the dominant sense-response mechanisms for the regulation of a wide range of biological functions, such as cell–cell signalling, nutrient assimilation, chemotaxis, osmolarity, sporulation and virulence[97,98]. A typical TCS is composed of a transmembrane sensor histidine kinase and a cytoplasmic response regulator. The sensor kinase is often localized to the cytoplasmic membrane, where it senses external stimuli and transduces this information to the response regulator. The response regulator, is normally a transcriptional regulatory protein that controls expression of a set of related genes, mediates the proper cellular response to the stimuli[99]. Sensor histidine kinases are key component for bacterial adaptation to the rapidly changing environment[100].

The gene 01697 of *Synechococcus sp.* HB1133 (a representative of clade IIh), encodes for possible EnvZ-OmpR two-component system which is a well characterised sensor kinase in *E. coli* and responds to an increase in external osmolarity. Together, EnvZ and OmpR enable cells to sense the external osmolarity and respond to it by regulating the transcription of two porin genes: *ompF* and *ompC*. Such responses are essential for acclimation to a new osmotic environment[100].

Gene 02391 encodes for membrane-anchored PPM-type phosphatase i.e. serine phosphatase[101] RsbU. RsbU is a positive regulator of the activity of σ^B, which is the general stress-response σ factor of Gram-positive microorganisms. The phosphatase activity of RsbU is stimulated greatly during the response to stress by associating with a kinase, RsbT[102].

Gene 02505 encodes for CheY, which is also a member of the response regulator family in bacterial two-component signalling systems. It has been found to play a role in general stress response of the Gram-negative bacteria[99].

Therefore, the presence unique genes of clade IIh that encode for response regulators further supports our hypothesis that clade IIh possess the ability to respond to rapidly changing conditions of transition

zones. Potential transition zones at the east coast of Australia have highly dynamic and changing conditions possibly due to the effect of EAC that brings high temperature, less nutrient and highly saline water from the tropics to the temperate latitudes and the cold nutrient rich water existing on the surface of the coast. Therefore, presence of the response regulators such as EnvZ-OmpR, PPM-type phosphatase and CheY would potentially allow clade IIh to sense and respond to these variable environmental stresses including changing temperature and high salinity and provide an additional capacity to cope via dynamic transcriptional control.

4.2.5 Role of genomic islands

Genomic islands have been found to play a significant role in *Synechococcus* diversity by acting as a repository for novel genes[83]. A key mechanism for gaining novel phenotypes and ecological functions involves horizontal transfer of novel genes within genomic islands. The genes in the islands appears to be unique to each species suggesting that these genes may serve an important function in adaptation to local environmental conditions[83]. Sixteen genomic islands were found in *Synechococcus* sp. HB1133 and 10 out of 42 unique genes were found to be present in the genomic island. The presence of the few unique genes encoding for response regulators in the islands (Appendix VIII) suggests that clade IIh may have acquired the regulatory capacity via horizontal gene transfer that may assist this group of *Synechococcus* to acclimate to the dynamic conditions of the oceanic fronts by sensing change and responding to stress conditions.

To summarise, based on the distribution pattern of marine picocyanobacteria around east Australian coast we hypothesized that clade IIh may be adapted to the conditions found 'in between' the major biomes and that it is a 'transitional' clade. The hypothesis was supported by a series of observations: optimal genotype thermal index between 15-25°C which is between the two major biomes, co-occurrence with both temperate and tropical strain depicted by the correlation patterns with other clades as well as with abiotic factors, presence of high number of response regulators in comparison to open ocean clades and the unique genes having functions as two-component signal transduction systems which could be one of the key components for adaptation to the rapidly changing environment. This study is a 'beginning' to understand the characteristics and habitat of clade IIh. Therefore, for future studies high throughput techniques such as quantitative label-free shotgun proteomics and RNA sequencing can used to determine the gene expression profiles of clade IIh in response to the variable environmental conditions at the potential transition zones. This would provide us an insight into the acclimation responses and the mechanisms employed by them to cope with prevalent conditions.

4.3 Molecular toolbox for genetic manipulation

Gene knockout experiments are important, to better understand the molecular adaptations as well as to verify the role of unique genes contributing to the fitness of marine *Synechococcus* to adapt to certain environmental niches. The current strategies available for generating mutants in marine *Synechococcus*, involving processes such as double homologous recombination and integrating other antibiotic resistant markers are very tedious and cumbersome[63]. Therefore, given the effectiveness and preciseness of the CRISPR/ Cpf1 system in freshwater cyanobacteria[62], this technique was tried in marine *Synechococcus* to address the need of precise, modern genetic tools for engineering marine *Synechococcus* and build an optimal toolbox for genomic manipulation. With the aim of capitalising on the gene editing precision and efficiency of the CRISPR/Cpf1 system, we tried knockout experiments on gene 1050, encoding a CRP regulator, in *Synechococcus* spp. WH8109 and gene 2793, encoding a fatty acid desaturase, in *Synechococcus* spp. CC9311. Two different transformation methods, electroporation and conjugation were tested to facilitate the uptake of editing plasmids by *Synechococcus* sp. WH8109 and CC9311. We successfully constructed three gene-specific editing plasmids (Appendix IV) containing CRISPR/Cpf1 and transformed both strains via conjugation.

Genetic manipulation via electroporation facilitates direct transfer of genes that may be toxic in alternate hosts such as *E.coli,* circumvents the alteration of plasmids occurring during conjugation[103] as well as removes the need for the counter selection of *E. coli* after conjugation. Therefore, given the less tedious process and benefits of electroporation, the project aimed to check if electroporation would be successful in facilitating plasmid transfer to marine *Synechococcus* cells. It was found that no transformants were observed for CC9311 and whereas ~40 transformants were obtained for WH8109. Hence, it can be summarised that electroporation may be used as a transformation method for WH8109, but it seemed to be a very poor method of transformation for CC9311. The reasons for electroporation not being successful in transforming DNA into CC9311 could be because of the electric shock to *Synechococcus* cells that is being introduced as a stress factor that could have possibly decreased the cell survival rate. Furthermore, several other experiments involving electroporation using circular DNA have had no transformants and enhanced transformation efficiency with linearized plasmids[104,105]. Therefore, future studies for genetic manipulation using electroporation should consider alternatives such as direct electroporation of PCR-amplified fragments containing an antibiotic resistant marker. The marker as well as the homology regions can be amplified using the same method used in this project such as high-fidelity PCR and assembling them using Goldengate assembly, purifying and directly electroporating the assembled fragments into the cells.

Higher number of transformants colonies were obtained using conjugation as method of transformation possibly because conjugation excludes stress factors such as electric stock to cells as well as multiple centrifugation steps as compared to electroporation.

From our experiments using CRISPR/Cpf1, multiple PCR checks for gene knockout with gene specific primers as well as primers specific for targeted edit suggested mixed results i.e. the edit may have been completed in some chromosomes but not in others (section 3.3.4). The lack of completion of the editing process could be potentially due to the ploidy (i.e. multiple chromosomes) nature of cyanobacteria that imparts many evolutionary advantages to them such as resistance to double stranded breaks in chromosome, gene redundancy allowing them to retain wildtype information in many chromosomes while mutation occurs in only few other chromosomes under unfavourable conditions and ability of gene conversion from heterozygous to homozygous cells in absence of selection[106]. If the tested strains are in fact polyploidy in nature, then they could also be potentially using information from other copies of chromosome to complete the repair process rather than using the homologous repair template provided and thereby continue retaining the gene preventing the gene knockout.

One of the major challenges and time-consuming process in genomic modification of cyanobacteria is ensuring that all the chromosome copies carry identical sequences of modified DNA which require multiple round of subculture/passages in selection media and as many as 12 passages in liquid media (PCRS11) supplemented with antibiotic (kanamycin 25µg/ml for CC9311 and 50µg/ml for WH8109) were done to allow the plasmid to complete the editing process along with PCR checks after each passage to verify the completion of edit. However, the genes of interest were still retained by both *Synechococcus* spp. as confirmed by sequencing results. Therefore, it seemed that 12 passages were not enough to allow the editing process to go to completion. Therefore, for future work knowing the genome copies of the strains for genetic manipulation would be advantageous to get an estimation of how many passages would be needed before the gene is knocked out. Previous study for genetic manipulation with CRISPR/Cpf1 on freshwater cyanobacteria showed that additional three passages were enough to allow the editing process to go to completion[62] which is strikingly different in marine *Synechococcus* as shown by this study in which as many as 12 passages did not seem to be enough to allow the editing process to undergo completion.

Additionally, a previous study has shown that most of the marine *Synechococcus* including CC9311 and WH8109 lack CRISPR/cas system based on absence of the *cas1* and *cas2* genes as diagnostic markers[107] which may suggest that CRISPR system may not even work on these species.

Another reason for retention of gene could potentially be lack of expression of Cpf1enzyme which could be preventing a desired cut to be made in the genome sequence. Even with the assumption that Cpf1is being expressed, it could be under expressed or could even be toxic to marine *Synechococcus.*

Enzyme Cas9 was found to be toxic to freshwater *Synechococcus* and therefore an alternative enzyme Cpf1 was proposed but the toxicity of Cpf1 has not been tested in marine *Synechococcus*. Therefore, more investigation is needed to determine the expression level and cytotoxicity of Cpf1 enzyme in marine cyanobacteria for application of CRISPR system in marine *Synechococcus*

4.4 Conclusions and future directions

Marine picocyanobacteria contribute significantly to the global primary production as well as play a key role in biogeochemical cycling. Given the diverse nature of marine *Synechococcus* they are ubiquitous and well known to adapt to various conditions of the oceans ranging from subtropical to temperate regions. However, a major portion of the ocean corresponds to dynamic regions between the boundary currents. Distinct *Synechococcus* lineages (such as clade IIh) that adapt to these conditions are not well-studied.

This project explored the ecology of marine *Synechococcus* community across the east Australian coast that is significantly affected by the EAC. The EAC is predicted to both strengthen and warm significantly causing rapid oceanographic changes and shifts in marine species distribution. Along the east coast of Australia, a novel *Synechococcus* clade IIh was found to be most abundant in between the major biomes i.e. at potential transition zones formed due to union of the cold mesotrophic water at the coast and the warm water brought by the EAC. However, for in-depth study of this novel clade, only one isolate belonging to clade IIh had been sequenced prior to this study. Therefore, this project contributed to the study of clade IIh by making whole genome sequences of other isolates belonging to clade IIh accessible and investigated various adaptation factors using comparative genomics to identify its unique genes and functions, its coastal lifestyle and relationship with environmental parameters that may have allowed them to acclimate to the potential 'transition' zones. The project also successfully constructed three editing plasmids for its utilisation in testing the utility of gene editing technique, CRISPR/Cpf1 in marine *Synechococcus*. Furthermore, optimisation of several parameters including testing two different transformation techniques, number of cells and kanamycin selection required for selection of transformants for two strains CC9311 and WH8109 were done as a step to build an optimal genetic manipulation system for marine *Synechococcus*.

In the future, the draft genomes can be completed via primer walking and repeat resolution and utilise it for further understanding of evolutionary processes as well as unravelling of molecular adaptations that contribute to ubiquitous distribution of this genus and niche specificity. Furthermore, gene knockout experiments for the unique genes of clade IIh such as sensor kinases can be conducted to test the sensitivity of clade IIh to osmotic stresses. Similarly, experiments such as temperature and salt concentrations dependent growth measurement of clade IIh can be conducted to determine the

optimal growth rate as well as its growth response to osmotic stress as compared to other clades that occur in more 'stable' conditions of ocean. In addition to this, high throughput techniques such as RNA-Sequencing and label-free shotgun proteomics can be utilised to investigate the regulation and expression of genes to identify role of genes that may be key in defining it as a 'transitional' clade. It would also be very interesting to further investigate the global distribution and adaptation factors of clade IIh using other metagenomic datasets and extrapolate our findings and in overall contribute to better understanding of microdiversity of marine *Synechococcus*. Furthermore, determining the ploidy level of marine *Synechococcus* and fine tuning of CRISPR enzymes in order to understand the molecular adaptations via genetic manipulation as well as exploit its potential as bio factories is an important aspect for application of the revolutionary gene editing technique such as CRISPR/Cpf1 system.

In broader prospective, the information provided by this project may help us determine how marine *Synechococcus* will adapt to predicted environmental changes and warming oceans and its effect on the marine food web, health of the oceans and ecosystem in general.

References:

1. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506,** 307–315 (2014).
2. Scanlan, D. J. *Ecology of cyanobacteria II: Their diversity in space and time*. (Springer Science + Business Media, 2012).
3. Scanlan, D. J. *et al.* Ecological Genomics of Marine Picocyanobacteria. *Microbiol. Mol. Biol. Rev.* (2009). doi:10.1128/MMBR.00035-08
4. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria Prochlrococcus and Synechococcus. *Pnas* **110,** 9824–9829 (2013).
5. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* (2005). doi:10.1038/nature04159
6. Waterbury, J. B., Watson, S. W., Guillard, R. R. L. & Brand, L. E. Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium [10]. *Nature* (1979). doi:10.1038/277293a0
7. Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334,** 340 (1988).
8. Partensky, F., Blanchot, J. & Vaulot, D. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters : a review. *Bull. l'Institut océanographique* **19,** 457–475 (1999).
9. Zwirglmaier, K. *et al.* Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* (2008). doi:10.1111/j.1462-2920.2007.01440.x
10. Partensky, F., Hess, W. R. & Vaulot, D. Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol.Mol Biol.Rev.* (1999). doi:doi:1092-2172/99/$04.00
11. Rocap, G. *et al.* Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* (2003). doi:10.1038/nature01947
12. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature* (1998). doi:10.1038/30965
13. Bouman, H. A. *et al.* Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. *Science (80-.).* (2006). doi:10.1126/science.1122692
14. Zinser, E. R. *et al.* Prochlorococcus ecotype abundances in the North Atlantic ocean as revealed by an improved quantitative PCR method. *Appl. Environ. Microbiol.* (2006). doi:10.1128/AEM.72.1.723-732.2006
15. Kent, A. G. *et al.* Parallel phylogeography of Prochlorococcus and Synechococcus. *Isme J* (2018). doi:10.1038/s41396-018-0287-6
16. Sohm, J. A. *et al.* Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J.* **10,** 333–345 (2016).
17. Huang, S. *et al.* Novel lineages of prochlorococcus and synechococcus in the global oceans. *ISME J.* (2012). doi:10.1038/ismej.2011.106
18. Gutiérrez-Rodríguez, A. *et al.* Fine spatial structure of genetically distinct picocyanobacterial populations across environmental gradients in the Costa Rica Dome. *Limnol. Oceanogr.* (2014). doi:10.4319/lo.2014.59.3.0705
19. Ahlgren, N. A. & Rocap, G. Culture isolation and culture-independent clone libraries reveal new marine Synechococcus ecotypes with distinctive light and N physiologies. *Appl. Environ. Microbiol.* (2006). doi:10.1128/AEM.00358-06
20. Belkin, I. & Cornillon, P. Fronts in the world ocean's large marine ecosystems. *Prog. Oceanogr.* **81,** 1–33 (2009).
21. Baltar, F., Currie, K., Stuck, E., Roosa, S. & Morales, S. E. Oceanic fronts: Transition zones for bacterioplankton community composition. *Environ. Microbiol. Rep.* **8,** 132–138 (2016).
22. Samo, T. J. *et al.* Microbial distribution and activity across a water mass frontal zone in the California Current Ecosystem. *J. Plankton Res.* **34,** 802–814 (2012).

23. Hobday, A. J. & Pecl, G. T. Identification of global marine hotspots: Sentinels for change and vanguards for adaptation action. *Rev. Fish Biol. Fish.* (2014). doi:10.1007/s11160-013-9326-6
24. Ridgway, K. R. Long-term trend and decadal variability of the southward penetration of the East Australian Current. *Geophys. Res. Lett.* **34,** 1–5 (2007).
25. Ridgway, K. R. & Dunn, J. R. Mesoscale structure of the mean East Australian Current System and its relationship with topography. *Progress in Oceanography* (2003). doi:10.1016/S0079-6611(03)00004-1
26. Baird, M. E. *et al.* Biological properties across the Tasman Front off southeast Australia. *Deep. Res. Part I Oceanogr. Res. Pap.* (2008). doi:10.1016/j.dsr.2008.06.011
27. Wijeratne, S., Pattiaratchi, C. & Proctor, R. Estimates of Surface and Subsurface Boundary Current Transport Around Australia.pdf. (2018).
28. Ridgway, K. & Hill, K. The East Australian Current. *A Mar. Clim. Chang. Impacts Adapt. Rep. Card Aust. 2009* 1–16 (2009). doi:10.1029/2003JC001833
29. Kelly, P., Clementson, L., Davies, C., Corney, S. & Swadling, K. Zooplankton responses to increasing sea surface temperatures in the southeastern Australia global marine hotspot. *Estuar. Coast. Shelf Sci.* **180,** 242–257 (2016).
30. Suthers, I. M. *et al.* The strengthening East Australian Current, its eddies and biological effects - an introduction and overview. *Deep. Res. Part II Top. Stud. Oceanogr.* **58,** 538–546 (2011).
31. Six, C., Finkel, Z. V., Irwin, A. J. & Campbell, D. A. Light variability illuminates niche-partitioning among marine picocyanobacteria. *PLoS One* **2,** (2007).
32. Shukla, A. *et al.* Phycoerythrin-specific bilin lyase-isomerase controls blue-green chromatic acclimation in marine Synechococcus. *Proc. Natl. Acad. Sci.* (2012). doi:10.1073/pnas.1211777109
33. Pittera, J. *et al.* Connecting thermal physiology and latitudinal niche partitioning in marine Synechococcus. *ISME J.* (2014). doi:10.1038/ismej.2013.228
34. Mackey, K. R. M. *et al.* Effect of Temperature on Photosynthesis and Growth in Marine Synechococcus spp. *PLANT Physiol.* (2013). doi:10.1104/pp.113.221937
35. Palenik, B. Molecular Mechanisms by Which Marine Phytoplankton Respond to Their Dynamic Chemical Environment. *Ann. Rev. Mar. Sci.* (2015). doi:10.1146/annurev-marine-010814-015639
36. Ostrowski, M. *et al.* PtrA is required for coordinate regulation of gene expression during phosphate stress in a marine Synechococcus. *ISME J.* (2010). doi:10.1038/ismej.2010.24
37. Huang, S., Zhang, S., Jiao, N. & Chen, F. Marine cyanophages demonstrate biogeographic patterns throughout the global ocean. *Appl. Environ. Microbiol.* **81,** 441–452 (2015).
38. M??hling, M. *et al.* Genetic diversity of marine Synechococcus and co-occurring cyanophage communities: Evidence for viral control of phytoplankton. *Environ. Microbiol.* (2005). doi:10.1111/j.1462-2920.2005.00713.x
39. Strom, S. L., Brahamsha, B., Fredrickson, K. A., Apple, J. K. & Rodríguez, A. G. A giant cell surface protein in Synechococcus WH8102 inhibits feeding by a dinoflagellate predator. *Environ. Microbiol.* (2012). doi:10.1111/j.1462-2920.2011.02640.x
40. Zwirglmaier, K., Spence, E., Zubkov, M. V., Scanlan, D. J. & Mann, N. H. Differential grazing of two heterotrophic nanoflagellates on marine Synechococcus strains. *Environ. Microbiol.* **11,** 1767–1776 (2009).
41. Paz-Yepes, J., Brahamsha, B. & Palenik, B. Role of a Microcin-C-like biosynthetic gene cluster in allelopathic interactions in marine Synechococcus. *Proc. Natl. Acad. Sci.* **110,** 12030–12035 (2013).
42. Palenik, B. *et al.* Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment. *Proc. Natl. Acad. Sci.* (2006). doi:10.1073/pnas.0602963103
43. Scanlan, D. J., Bourne, J. A. & Mann, N. H. A putative transcriptional activator of the Crp/Fnr family from the marine cyanobacterium Synechococcus sp. WH 7803. *J. Appl. Phycol.* **8,** 565–567 (1997).

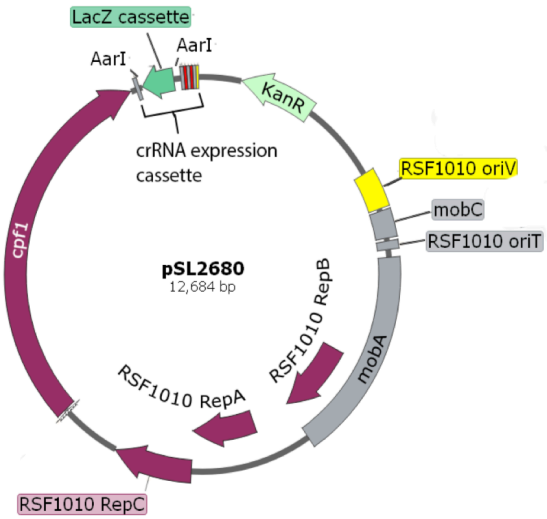
44. Soberón-chávez, G. *et al.* The Transcriptional Regulators of the CRP Family Regulate Different Essential Bacterial Functions and Can Be Inherited Vertically and Horizontally. **8,** 1–8 (2017).
45. Penno, S., Lindell, D. & Post, A. F. Diversity of Synechococcus and Prochlorococcus populations determined from DNA sequences of the N-regulatory gene ntcA. *Environ. Microbiol.* (2006). doi:10.1111/j.1462-2920.2006.01010.x
46. Varkey, D. *et al.* Effects of low temperature on tropical and temperate isolates of marine Synechococcus. *ISME J.* (2016). doi:10.1038/ismej.2015.179
47. Pittera, J. *et al.* Thermoacclimation and genome adaptation of the membrane lipidome in marine Synechococcus. *Environ. Microbiol.* (2018). doi:10.1111/1462-2920.13985
48. Mikami, K. & Murata, N. Membrane fluidity and the perception of environmental signals in cyanobacteria and plants. *Prog. Lipid Res.* **42,** 527–543 (2003).
49. Murata, N. & Los, D. A. Membrane Fluidity and Temperature Perception. *Plant Physiol.* (2016). doi:10.1104/pp.115.3.875
50. Takahashi, S., Whitney, S. M. & Badger, M. R. Different thermal sensitivity of the repair of photodamaged photosynthetic machinery in cultured Symbiodinium species. *Proc. Natl. Acad. Sci.* (2009). doi:10.1073/pnas.0808363106
51. Schildkraut, C. L., Marmur, J. & Doty, P. The formation of hybrid DNA molecules and their use in studies of DNA homologies. *J. Mol. Biol.* (1961). doi:10.1016/S0022-2836(61)80024-7
52. McCarthy, B. J. & Bolton, E. T. AN APPROACH TO THE MEASUREMENT OF GENETIC RELATEDNESS AMONG ORGANISMS. *Proc. Natl. Acad. Sci.* (2006). doi:10.1073/pnas.50.1.156
53. Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W. & Kämpfer, P. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.* (2010). doi:10.1099/ijs.0.016949-0
54. STACKEBRANDT, E. & GOEBEL, B. M. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int. J. Syst. Evol. Microbiol.* (2008). doi:10.1099/00207713-44-4-846
55. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. in *FEMS Microbiology Reviews* (2001). doi:10.1016/S0168-6445(00)00040-1
56. Gevers, D. *et al.* Re-evaluating prokaryotic species. *Nature Reviews Microbiology* (2005). doi:10.1038/nrmicro1236
57. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* **106,** 19126–19131 (2009).
58. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. in *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty597
59. Mazard, S., Ostrowski, M., Partensky, F. & Scanlan, D. J. Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.* **14,** 372–386 (2012).
60. Toledo, G. & Palenik, B. Synechococcus diversity in the California Current as seen by RNA polymerase (rpoC1) gene sequences of isolated strains. *Appl. Environ. Microbiol.* (1997).
61. Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* (2005). doi:10.1186/gb-2005-6-2-r14
62. Ungerer, J. & Pakrasi, H. B. Cpf1 Is A Versatile Tool for CRISPR Genome Editing Across Diverse Species of Cyanobacteria. *Sci. Rep.* (2016). doi:10.1038/srep39681
63. Golden, S. S., Brusslan, J. & Haselkorn, R. Genetic Engineering of the Cyanobacterial Chromosome. *Methods Enzymol.* (1987). doi:10.1016/0076-6879(87)53055-5
64. Horvath, P., Barrangou, R., Horvath1, P. & Barrangou2, R. CRISPR/Cas, the Immune System of Bacteria and Archaea WÊÊÊÊ^^^BÊ^^ÊÊÊ^^^ÊÊÍ^nÊSSMBM CRISPR/Cas, the Immune System of Bacteria and Archaea. *Source Sci. New Ser.* **327,** 167–170 (2010).

65. Li, B., Zeng, C. & Dong, Y. Design and assessment of engineered CRISPR-Cpf1 and its use for genome editing. *Nat. Protoc.* (2018). doi:10.1038/nprot.2018.004
66. Wendt, K. E., Ungerer, J., Cobb, R. E., Zhao, H. & Pakrasi, H. B. CRISPR/Cas9 mediated targeted mutagenesis of the fast growing cyanobacterium Synechococcus elongatus UTEX 2973. *Microb. Cell Fact.* **15,** (2016).
67. Humily, F. *et al.* Development of a targeted metagenomic approach to study a genomic region involved in light harvesting in marine Synechococcus. *FEMS Microbiol. Ecol.* **88,** 231–249 (2014).
68. Sanfilippo, J. E. *et al.* Self-regulating genomic island encoding tandem regulators confers chromatic acclimation to marine *Synechococcus*. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1600625113
69. Brown, M. V. Systematic, continental scale temporal monitoring of marine pelagic microbiota by the Australian Marine Microbial Biodiversity Initiative. *Sci. Data* 1–10 (2018).
70. Ludwig, W. *et al.* ARB: A software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
71. Wilson, W. H., Carr, N. G. & Mann, N. H. THE EFFECT OF PHOSPHATE STATUS ON THE KINETICS OF CYANOPHAGE INFECTION IN THE OCEANIC CYANOBACTERIUM SYNECHOCOCCUS SP. WH7803. **23,** 506–516 (1994).
72. Rose, R., Golosova, O., Sukhomlinov, D., Tiunov, A. & Prosperi, M. Flexible design of multiple metagenomics classification pipelines with UGENE. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty901
73. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh340
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu170
75. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* (2012). doi:10.1089/cmb.2012.0021
76. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* (2015). doi:10.7717/peerj.1165
77. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from. *Genome Res.* **25,** 1043–1055 (2015).
78. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics* (2012). doi:10.1038/nrg3117
79. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* (2012). doi:10.1038/nmeth.1923
80. Wysoker, A. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
81. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36,** (2018).
82. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9,** 1–8 (2018).
83. Dufresne, A. *et al.* Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.* (2008). doi:10.1186/gb-2008-9-5-r90
84. Vernikos, G. S. & Parkhill, J. Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the Salmonella pathogenicity islands. *Bioinformatics* **22,** 2196–2203 (2006).
85. Scholz, P. *et al.* Complete nucleotide sequence and gene organization of the broad-host-range plasmid RSF1010. *Gene* (1989). doi:10.1016/0378-1119(89)90273-4

86. Brahamsha, B. *A Genetic Manipulation System for Oceanic Cyanobacteria of the Genus Synechococcus*. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY* **62,** (1996).
87. Murray, M. G. & Thompson, W. F. Nucleic Acids Research Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8,** 4321–4326 (1980).
88. Carter, M. & Shieh, J. C. *Guide to Research Techniques in Neuroscience*. *Guide to Research Techniques in Neuroscience* (2010). doi:10.1016/C2009-0-01891-1
89. Hess, W. *et al.* Prochlorococcus marinus Chisholm et al. 1992 subsp. pastoris subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (Oxyphotobacteria). *Int. J. Syst. Evol. Microbiol.* **50,** 1833–1847 (2015).
90. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryots. *Proc. Natl. Acad. Sci.* **102,** 2567–2572 (2005).
91. Farrant, G. K. *et al.* Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc. Natl. Acad. Sci.* **113,** E3365–E3374 (2016).
92. Suthers, I. M. *et al.* The strengthening East Australian Current, its eddies and biological effects - an introduction and overview. *Deep-Sea Research Part II: Topical Studies in Oceanography* (2011). doi:10.1016/j.dsr2.2010.09.029
93. Zheng, Q. *et al.* Dynamics of heterotrophic bacterial assemblages within Synechococcus cultures. *Appl. Environ. Microbiol.* (2018). doi:10.1128/AEM.01517-17
94. Christie-oleza, J. A., Scanlan, D. J. & Armengaud, J. " You produce while I clean up " , a strategy revealed by exoproteomics during Synechococcus – Roseobacter interactions. 3454–3462 (2015). doi:10.1002/pmic.201400562
95. Biller, S. J. *et al.* Bacterial Vesicles in Marine Ecosystems. **998,** 183–187 (2014).
96. Belkin, I. M., Cornillon, P. C. & Sherman, K. Fronts in Large Marine Ecosystems. *Prog. Oceanogr.* **81,** 223–236 (2009).
97. Stock, A. M., Robinson, V. L. & Goudreau, P. N. Two-Component Signal Transduction. *Annu. Rev. Biochem.* (2002). doi:10.1146/annurev.biochem.69.1.183
98. Qian, W., Han, Z.-J. & He, C. Two-Component Signal Transduction Systems of Xanthomonas spp.: A Lesson from Genomics . *Mol. Plant-Microbe Interact.* (2008). doi:10.1094/mpmi-21-2-0151
99. Liu, Y. F. *et al.* GsmR, a response regulator with an HD-related output domain in Xanthomonas campestris, is positively controlled by Clp and is involved in the expression of genes responsible for flagellum synthesis. *FEBS J.* **280,** 199–213 (2013).
100. Hsing, W., Russo, F. D., Bernd, K. K. & Silhavy, T. J. Mutations that alter the kinase and phosphatase activities of the two- component sensor EnvZ. *J. Bacteriol.* (1998).
101. Zhang, W. & Shi, L. Evolution of the PPM-family protein phosphatases in Streptomyces: Duplication of catalytic domain and lateral recruitment of additional sensory domains. *Microbiology* (2004). doi:10.1099/mic.0.27480-0
102. Delumeau, O. *et al.* Functional and structural characterization of RsbU, a stress signaling protein phosphatase 2C. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M405464200
103. Puri, A. W. *et al.* Genetic tools for the industrially promising methanotroph Methylomicrobium buryatense. *Appl. Environ. Microbiol.* (2015). doi:10.1128/AEM.03795-14
104. Khatiwada, B., Kautto, L., Sunna, A., Sun, A. & Nevalainen, H. Nuclear transformation of the versatile microalga Euglena gracilis. *Algal Res.* **37,** 178–185 (2019).
105. Zhang, C. & Hu, H. High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. *Mar. Genomics* (2014). doi:10.1016/j.margen.2013.10.003
106. Griese, M., Lange, C. & Soppa, J. Ploidy in cyanobacteria. *FEMS Microbiol. Lett.* (2011). doi:10.1111/j.1574-6968.2011.02368.x
107. Cai, F., Axen, S. D. & Kerfeld, C. A. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. 687–693 (2013).

Appendix

Appendix I: Backbone plasmid pSL2680 (Addgene plasmid # 85581), replicating base vector for constructing cpf1/CRISPR editing plasmids. Image adapted from Addgene [https://http://n2t.net/addgene:85581; RRID: Addgene_85581]



Appendix II: PCR amplification primers

Abbreviations: F, forward; R, Reverse

Primer Name	Sequence	Tm °C	GC %
petB_F	TACGACTGGTTCCAGGAACG	56.6	55
petB_R	GAAGTGCATGAGCATGAA	50.4	44.44
gRNA_2793	GCTGGAACGCTTTTTCGCAA		
gRNA_9311_2793F	tctactgttgtagatGCTGGAACGCTTTTTCGCAA	63.7	42.9
gRNA_9311_2793R	ttaaagttcttagacTTGCGAAAAAGCGTTCCAGC	61.9	40
gRNA_PCR_SEQ_F	GTATCAAAGCAGCTATTTGCG	52.3	42.9
HRT_PCR_seq_Spec_F	GATTTGCTGGTTACGGTGAC	54	50
2793_detect_F	TCCGTCACGATTGGATTCAT	54	45
2793_detect_R	CGGATGTGTTCCCAGGTG	55.9	61.1
2793_sc_genhrt5spec_F	GGGGTTAGCTCTGTATTGGTACG	56.9	52.2
2793_sc_spechrt3gen_R	AACGTGCTGGTGAGGATTTC	55.5	50
kan_F	GGTGATTTTGAACTTTTGCTTTG	51.8	34.8
kan_R	TTATGCCTCTTCCGACCATC	54.3	50
HRT_PCR_seq_vector_F	GAGCCTTTTGTATTAGTAGC	48.2	40
gRNA_1050	TGAGGAAACCGAGGGAATGA		
gRNA_8109_1050F	AGA TTG AGG AAA CCG AGG GAA TGA	57.9	45.8

gRNA_8109_1050R	AGA CTC ATT CCC TCG GTT TCC TCA	59.5	50
1050_detect_F	CTTCAAACCGTTCTGCTGGA	55.4	50
1050_detect_R	TCAAAATTGGAGTGGTGAAGAG	53	40.9
1050_screen_genhrt5spec_F	GAGCTGCTGTCCCAGGTTG	58.6	63.2
1050 HRT_screen2_R	TATTTCTGCAACGAGGCGGT	57	50

Appendix III: Primers used for HRT amplification for markerless and marked editing plasmid for 1050 and 2793 genes. Overhang regions for annealing is denoted in lower case.

Primers used for HRT amplification for markerless editing plasmid for 1050		
1050_HRT5_F_plasOH	GGTCATTTTTTTGTCTAGCTTTAATGCGGTAGTTGGTAC Cttttgcgtgttcaagtcacc	HRT5
1050_HRT5_R_hrt3OH	GATCGCCAAGAACACCAAAGGAATTCACGATCTTCAA AATttggcatcgacaaaggtgcc	
1050_HRT3_F_hrt5OH	TTCTTGCCGGATGATCCAGCGGCACCTTTGTCGATGCC AAattttgaagatcgtgaattc	HRT3
1050_HRT3_R_plasOH	GCGCTGCCCGGATTACAGATCCTCTAGAGTCGACGGTA CCgttgtcaacatccttcgctt	

Primers used for HRT amplification for marked editing plasmid for 1050		
1050_HRT5_F_plasOH	GGTCATTTTTTTGTCTAGCTTTAATGCGGTAGTTGGTAC Cttttgcgtgttcaagtcacc	HRT5
1050_HRT5_R _specOH	TTGGGTAGCAGCGAAGTCGAGGCATTTCTGTCCTGGCT GGttggcatcgacaaaggtgcc	
1050_Spec_F_hrt5OH	TTCTTGCCGGATGATCCAGCGGCACCTTTGTCGATGCC AAccagccaggacagaaatgcc	Spec
1050_Spec_R_hrt3OH	GATCGCCAAGAACACCAAAGGAATTCACGATCTTCAA AATttatttgccgactaccttgg	
1050_HRT3_F_specO H	CTACGTGAAAGGCGAGATCACCAAGGTAGTCGGCAAA TAAattttgaagatcgtgaattc	HRT3
1050_HRT3_R_plasOH	GCGCTGCCCGGATTACAGATCCTCTAGAGTCGACGGTA CCgttgtcaacatccttcgctt	

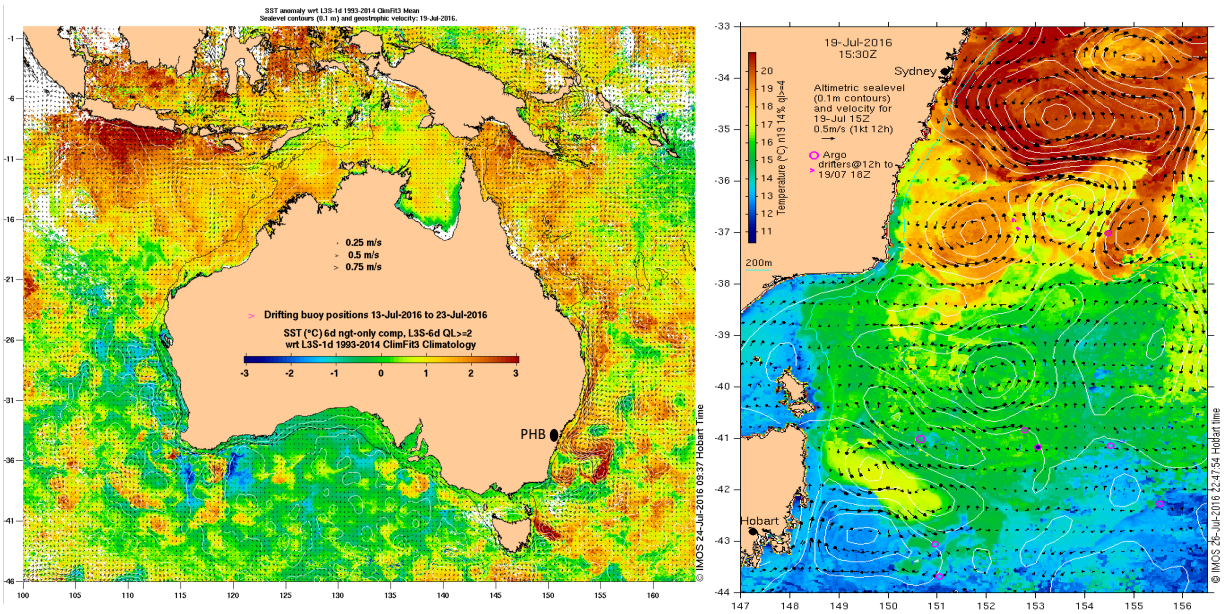
Primers used for HRT amplification for marked editing plasmid for 2793		
2793_HRT5_F_plasOH	GGTCATTTTTTTGTCTAGCTTTAATGCGGTAGTTGGTAC Ccagtcctgatctttcacgta	

2793_HRT5_R_hrt3OH	TTGGGTAGCAGCGAAGTCGAGGCATTTCTGTCCTGGCT GGttggcatcgacaaaggtgcc	HRT5
2793_Spec_F_hrt5OH	GGGTGTAAATTATGGACATCAAGCGGATTAGCCATGG TTTccagccaggacagaaatgcc	Spec
2793_Spec_R_hrt3OH	AGAAGAGGTTGAAGATGAGAAGCGATTGCGAAGTTTA CGAttatttgccgactaccttgg	
2793_HRT3_F_specO H	CTACGTGAAAGGCGAGATCACCAAGGTAGTCGGCAAA TAAtcgtaaacttcgcaatcgct	HRT3
2793_HRT3_R_plasOH	GCGCTGCCCGGATTACAGATCCTCTAGAGTCGACGGTA CCcgcaggaacttatccatcac	

Appendix IV: All plasmids used in this study

Plasmids	Source
pSL2680	Addgene
pSL2793	pSL2680 with gRNA for 2793 (This Study)
pSL1050	pSL2680 with gRNA for 1050 (This Study)
epSL2793	pSL2793 with homologous repair template and spectinomycin resistance gene inserts (This Study)
epSL1050	pSL1050 with homologous repair template and spectinomycin resistance gene inserts (This Study)
un_ epSL1050	pSL1050 with homologous repair template inserts (This Study)
pRK25	Plasmid present in conjugative *E. coli* strain MC1061
pRL528	Plasmid present in conjugative *E. coli* strain MC1061

Appendix V: Sea Surface Temperature Anomaly (a) and productivity (b) showing higher temperature and high productivity at east coast of Australia for the time period when clade IIh was seen to be most abundant



Appendix VI: ANI, 16S rRNA and *petB* percentage identity for genomes sequenced in this project

strain 1	strain 2	ANI	petb identity %	number of bp difference petb	16S rRNA % identity	number of bp difference 16S rRNA
syn_MU1611	syn_MU1611	100	NA			
syn_MU1611	Syn_HB1133	95.4904	91.172	58	99.932	
syn_MU1611	Syn_HBA1120	95.5283	91.172	58	99.932	1
syn_MU1611	syn_MU1625	96.8987	99.696	2	100	0
syn_MU1611	syn_MU1642	85.5084	88.28	77		
syn_MU1611	syn_MU1643	85.7674	90.715	61	99.527	7
syn_MU1611	syn_MU1644	85.9715	89.65	68	99.527	7
syn_MU1611	syn_MU1648	99.0931	99.848	1		
syn_MU1611	syn_MU1651	86.093	90.259	64	99.325	10
syn_MU1611	syn_MU1652	99.0148	99.848	1		
syn_MU1611	syn_MU1655	77.1766	84.699	28	99.055	14
syn_MU1611	Syn_PH4-1509	95.4863	91.172	58		
syn_MU1611	Syn_PROS-U-1	89.4472	91.476	56	99.865	2
syn_MU1625	syn_MU1625	100	NA	NA	NA	
syn_MU1625	syn_MU1611	96.9054	99.696	2	100	0
syn_MU1625	Syn_HB1133	95.1499	91.324	57		
syn_MU1625	Syn_HBA1120	95.191	91.324	57		
syn_MU1625	syn_MU1642	85.6339	88.128	78		
syn_MU1625	syn_MU1643	85.665	90.868	60	99.527	7
syn_MU1625	syn_MU1644	85.921	89.802	67	99.527	7
syn_MU1625	syn_MU1648	96.8407	99.848	1		
syn_MU1625	syn_MU1651	86.0704	90.411	63	99.325	10
syn_MU1625	syn_MU1652	96.8453	99.848	1		
syn_MU1625	syn_MU1655	77.1473	85.246	27	99.055	14
syn_MU1625	Syn_PH4-1509	95.0898	91.324	57		
syn_MU1625	Syn_PROS-U-1	89.421	91.629	55	99.865	2
syn_MU1642	syn_MU1642	100	NA	NA		
syn_MU1642	syn_MU1643	94.4427	92.694	48		
syn_MU1642	Syn_BL107_IVa	78.0012	83.714	107		
syn_MU1642	Syn_CC9311_Ia	77.0493	84.627	101		
syn_MU1642	Syn_HB1133	85.6349	89.498	69		
syn_MU1642	Syn_HBA1120	85.6394	89.498	69		
syn_MU1642	syn_MU1611	85.6453	88.28	77		
syn_MU1642	syn_MU1625	85.6595	88.128	78		
syn_MU1642	syn_MU1644	92.956	92.694	48		
syn_MU1642	syn_MU1648	85.6321	88.128	78		
syn_MU1642	syn_MU1651	90.4526	90.563	62		
syn_MU1642	syn_MU1652	85.7117	88.128	78		
syn_MU1642	syn_MU1655	77.0758	85.792	26		
syn_MU1642	Syn_PH4-1509	85.668	89.498	69		
syn_MU1642	Syn_PROS-U-1	85.5758	90.259	64		

syn_MU1642	Syn_WH8102_IIIa	80.1719	86.758	87		
syn_MU1642	Syn_WH8109_IIa	90.2454	91.02	59		
syn_MU1643	syn_MU1643	100	NA	NA	NA	
syn_MU1643	Syn_BL107_IVa	78.0347	85.388	96	99.46	8
syn_MU1643	Syn_CC9311_Ia	77.4123	85.54	95	98.515	22
syn_MU1643	Syn_HB1133	85.804	91.172	58		
syn_MU1643	Syn_HBA1120	85.784	91.172	58		
syn_MU1643	syn_MU1611	85.7804	90.715	61	99.527	7
syn_MU1643	syn_MU1625	85.8267	90.868	60	99.527	7
syn_MU1643	syn_MU1642	94.5366	92.694	48		
syn_MU1643	syn_MU1644	93.2054	95.434	30	100	0
syn_MU1643	syn_MU1648	85.7802	90.868	60		
syn_MU1643	syn_MU1651	90.5087	93.455	43	99.662	5
syn_MU1643	syn_MU1652	85.8514	90.868	60		
syn_MU1643	syn_MU1655	77.2006	86.339	25	98.582	21
syn_MU1643	Syn_PH4-1509	85.6747	91.172	58		
syn_MU1643	Syn_PROS-U-1	85.7936	91.629	55		
syn_MU1643	Syn_WH8102_IIIa	80.3246	88.28	77	99.055	14
syn_MU1643	Syn_WH8109_IIa	90.38	94.216	38	99.73	4
syn_MU1644	syn_MU1644	100	NA	NA		
syn_MU1644	Syn_BL107_IVa	77.996	84.475	102	99.46	8
syn_MU1644	Syn_CC9311_Ia	77.358	84.17	104	98.515	22
syn_MU1644	Syn_HB1133	85.9518	90.411	63		
syn_MU1644	Syn_HBA1120	85.9782	90.411	63		
syn_MU1644	syn_MU1611	85.9902	89.65	68	99.527	7
syn_MU1644	syn_MU1625	86.1148	89.802	67	99.527	7
syn_MU1644	syn_MU1642	93.0266	92.694	48		
syn_MU1644	syn_MU1643	93.1624	95.434	30	100	0
syn_MU1644	syn_MU1648	85.9893	89.802	67		
syn_MU1644	syn_MU1651	90.754	93.151	45	99.662	5
syn_MU1644	syn_MU1652	86.0166	89.802	67		
syn_MU1644	syn_MU1655	77.0887	85.246	27	98.582	21
syn_MU1644	Syn_PH4-1509	85.7809	90.411	63		
syn_MU1644	Syn_PROS-U-1	85.8709	91.324	57	99.392	9
syn_MU1644	Syn_WH8102_IIIa	80.2995	87.367	83	99.055	14
syn_MU1644	Syn_WH8109_IIa	90.6954	93.76	41	99.73	4
syn_MU1648	syn_MU1648	100	NA	NA		
syn_MU1648	Syn_BL107_IVa	77.8135	86.758	87		
syn_MU1648	Syn_CC9311_Ia	76.9867	85.236	97		
syn_MU1648	Syn_HB1133	95.4654	91.324	57		
syn_MU1648	Syn_HBA1120	95.5134	91.324	57		
syn_MU1648	syn_MU1611	99.0797	99.848	1		
syn_MU1648	syn_MU1625	96.8828	99.848	1		
syn_MU1648	syn_MU1642	85.6511	88.128	78		

syn_MU1648	syn_MU1643	85.7612	90.868	60		
syn_MU1648	syn_MU1644	85.8498	89.802	67		
syn_MU1648	syn_MU1651	85.9888	90.411	63		
syn_MU1648	syn_MU1652	99.5721	100	0		
syn_MU1648	syn_MU1655	77.2553	85.246	27		
syn_MU1648	Syn_PH4-1509	95.447	91.324	57		
syn_MU1648	Syn_PROS-U-1	89.4116	91.629	55		
syn_MU1648	Syn_WH8102_IIIa	79.8883	86.301	90		
syn_MU1648	Syn_WH8109_IIa	85.4434	89.954	66		
syn_MU1651	syn_MU1651	100	NA	NA		
syn_MU1651	Syn_BL107_IVa	78.3841	84.779	100	99.527	7
syn_MU1651	Syn_CC9311_Ia	76.952	84.627	101	98.582	21
syn_MU1651	Syn_HB1133	86.0772	91.02	59	98.65	20
syn_MU1651	Syn_HBA1120	86.0456	91.02	59	99.392	9
syn_MU1651	syn_MU1611	86.0133	90.259	64	99.325	10
syn_MU1651	syn_MU1625	86.0202	90.411	63	99.325	10
syn_MU1651	syn_MU1642	90.3487	90.563	62		
syn_MU1651	syn_MU1643	90.543	93.455	43	99.662	5
syn_MU1651	syn_MU1644	90.6893	93.151	45	99.662	5
syn_MU1651	syn_MU1648	86.0043	90.411	63		
syn_MU1651	syn_MU1652	86.1455	90.411	63		
syn_MU1651	syn_MU1655	77.0635	85.792	26	98.65	20
syn_MU1651	Syn_PH4-1509	85.8874	91.02	59		
syn_MU1651	Syn_PROS-U-1	86.0709	91.629	55	99.325	10
syn_MU1651	Syn_WH8102_IIIa	80.6232	87.823	80	98.852	17
syn_MU1651	Syn_WH8109_IIa	93.7059	95.738	28	99.932	1
syn_MU1652	syn_MU1652	100	NA	NA		
syn_MU1652	Syn_BL107_IVa	78.0174	86.758	87		
syn_MU1652	Syn_CC9311_Ia	77.1227	85.236	97		
syn_MU1652	Syn_HB1133	95.4581	91.324	57		
syn_MU1652	Syn_HBA1120	95.4774	91.324	57		
syn_MU1652	syn_MU1611	99.0213	99.848	1		
syn_MU1652	syn_MU1625	96.8918	99.848	1		
syn_MU1652	syn_MU1642	85.5433	88.128	78		
syn_MU1652	syn_MU1643	85.753	90.868	60		
syn_MU1652	syn_MU1644	85.9523	89.802	67		
syn_MU1652	syn_MU1648	99.5821	100	0		
syn_MU1652	syn_MU1651	86.0902	90.411	63		
syn_MU1652	syn_MU1655	77.0534	85.246	27		
syn_MU1652	Syn_PH4-1509	95.3347	91.324	57		
syn_MU1652	Syn_PROS-U-1	89.5027	91.629	55		
syn_MU1652	Syn_WH8102_IIIa	79.9553	86.301	90		
syn_MU1652	Syn_WH8109_IIa	85.4855	89.954	66		
syn_MU1655	syn_MU1655	100	NA	NA		

syn_MU1655	Syn_BL107_IVa	77.7485	84.153	29	98.717	19
syn_MU1655	Syn_CC9311_Ia	90.8567	95.082	9	99.932	1
syn_MU1655	Syn_HB1133	77.4681	86.885	24	99.122	13
syn_MU1655	Syn_HBA1120	77.5784	86.885	24	99.122	13
syn_MU1655	syn_MU1611	77.2757	84.699	28	99.055	14
syn_MU1655	syn_MU1625	77.4229	85.246	27	99.055	14
syn_MU1655	syn_MU1642	77.0982	85.792	26		
syn_MU1655	syn_MU1643	77.3447	86.339	25	98.582	21
syn_MU1655	syn_MU1644	77.2157	85.246	27	98.582	21
syn_MU1655	syn_MU1648	77.061	85.246	27		
syn_MU1655	syn_MU1651	77.2648	85.792	26	98.65	20
syn_MU1655	syn_MU1652	77.1354	85.246	27		
syn_MU1655	Syn_PH4-1509	77.2075	86.885	24		
syn_MU1655	Syn_PROS-U-1	77.5274	87.432	23		
syn_MU1655	Syn_WH8102_IIIa	77.4456	87.432	23	98.649	20
syn_MU1655	Syn_WH8109_IIa	77.5756	84.699	28	98.582	21
syn_HB1133	Syn_HB1133	100	NA	NA		
syn_HB1133	Syn_BL107_IVa	78.2173	84.323	103	99.595	
syn_HB1133	Syn_CC9311_Ia	77.3756	84.17	104	99.055	
syn_HB1133	Syn_HBA1120	99.9856	100	0		
syn_HB1133	syn_MU1611	95.5276	91.172	58	99.932	1
syn_HB1133	syn_MU1625	95.1065	91.324	57	99.932	1
syn_HB1133	syn_MU1642	85.4414	89.498	69	99.46	
syn_HB1133	syn_MU1643	85.8819	91.172	58	99.46	8
syn_HB1133	syn_MU1644	85.8593	90.411	63	99.392	8
syn_HB1133	syn_MU1648	95.4369	91.324	57		
syn_HB1133	syn_MU1651	86.0332	91.02	59	99.122	9
syn_HB1133	syn_MU1652	95.4926	91.324	57		
syn_HB1133	syn_MU1655	77.3262	86.885	24	99.932	13
syn_HB1133	Syn_PH4-1509	99.8066	100	0		
syn_HB1133	Syn_PROS-U-1	89.3469	92.085	52	99.932	1
syn_HB1133	Syn_WH8102_IIIa	80.0726	86.758	87	99.19	12
syn_HB1133	Syn_WH8109_IIa	85.6247	90.411	63	99.325	10
syn_HBA1120	Syn_HBA1120	100	NA	NA		
syn_HBA1120	Syn_BL107_IVa	78.3232	84.323	103	99.595	6
syn_HBA1120	Syn_CC9311_Ia	77.3673	84.17	104	99.055	14
syn_HBA1120	Syn_HB1133	99.9888	100	0		
syn_HBA1120	syn_MU1611	95.5506	91.172	58	99.932	1
syn_HBA1120	syn_MU1625	95.1061	91.324	57	99.932	1
syn_HBA1120	syn_MU1642	85.5613	89.498	69		
syn_HBA1120	syn_MU1643	85.932	91.172	58	99.46	8
syn_HBA1120	syn_MU1644	85.8524	90.411	63	99.46	8
syn_HBA1120	syn_MU1648	95.4761	91.324	57		
syn_HBA1120	syn_MU1651	85.9513	91.02	59	99.392	9

syn_HBA1120	syn_MU1652	95.5439	91.324	57		
syn_HBA1120	syn_MU1655	77.2318	86.885	24	99.122	13
syn_HBA1120	Syn_PH4-1509	99.8406	100	0		
syn_HBA1120	Syn_PROS-U-1	89.3392	92.085	52	99.932	1
syn_HBA1120	Syn_WH8102_IIIa	79.9218	86.758	87	99.19	12
syn_HBA1120	Syn_WH8109_IIa	85.6284	90.411	63	99.325	10
syn_PH4-1509	Syn_PH4-1509	100	NA	NA		
syn_PH4-1509	Syn_HB1133	99.8507	100	0		
syn_PH4-1509	Syn_HBA1120	99.8508	100	0		
syn_PH4-1509	syn_MU1611	95.5589	91.172	58		
syn_PH4-1509	syn_MU1625	95.0838	91.324	57		
syn_PH4-1509	syn_MU1642	85.506	89.498	69		
syn_PH4-1509	syn_MU1643	85.728	91.172	58		
syn_PH4-1509	syn_MU1644	85.693	90.411	63		
syn_PH4-1509	syn_MU1648	95.435	91.324	57		
syn_PH4-1509	syn_MU1651	85.7113	91.02	59		
syn_PH4-1509	syn_MU1652	95.4873	91.324	57		
syn_PH4-1509	syn_MU1655	76.9626	86.885	24		
syn_PH4-1509	Syn_PROS-U-1	89.2099	92.085	52		
syn_PROS-U-1	Syn_PROS-U-1	100	NA	NA		
syn_PROS-U-1	syn_MU1652	89.5303	91.629	55		
syn_PROS-U-1	Syn_BL107_IVa	78.1215	85.084	98		
syn_PROS-U-1	Syn_CC9311_Ia	77.8989	84.018	105		
syn_PROS-U-1	Syn_HB1133	89.1857	92.085	52		
syn_PROS-U-1	Syn_HBA1120	89.2494	92.085	52		
syn_PROS-U-1	syn_MU1611	89.3032	91.476	56	99.865	2
syn_PROS-U-1	syn_MU1625	89.3566	91.629	55	99.865	2
syn_PROS-U-1	syn_MU1642	85.619	90.259	64		
syn_PROS-U-1	syn_MU1643	85.6066	91.629	55	99.392	9
syn_PROS-U-1	syn_MU1644	85.8791	91.324	57	99.392	9
syn_PROS-U-1	syn_MU1648	89.3696	91.629	55		
syn_PROS-U-1	syn_MU1651	86.1625	91.629	55	99.325	10
syn_PROS-U-1	syn_MU1655	77.3094	87.432	23	99.055	14
syn_PROS-U-1	Syn_PH4-1509	89.1169	92.085	52		
syn_PROS-U-1	Syn_WH8102_IIIa	80.2544	87.519	82	99.122	13
syn_PROS-U-1	Syn_WH8109_IIa	85.6552	91.324	57	99.257	11
syn_MU1611	Syn_CC9311_Ia	77.4858	85.084	98	98.987	15
syn_MU1611	Syn_BL107_IVa	78.4571	86.606	88	99.527	7
syn_MU1611	Syn_WH8102_IIIa	80.0446	86.454	89	99.122	13
syn_MU1611	Syn_WH8109_IIa	85.5768	89.802	67	99.392	9
syn_MU1625	Syn_BL107_IVa	78.318	86.758	87	99.527	7
syn_MU1625	Syn_CC9311_Ia	77.2419	77.2419	96	99.055	14
syn_MU1625	Syn_WH8102_IIIa	80.0754	86.301	89	99.122	13
syn_MU1625	Syn_WH8109_IIa	85.5469	89.954	67	99.392	9

Appendix VII: Genomic islands in *Synechococcus* HB1133 predicted using Alien_Hunter-1.7

Island location	Threshold	Score
1..10000	19.302	24.075
240000..245000	19.302	21.799
265000..282500	19.302	45.744
407500..412500	19.302	21.391
642500..647500	19.302	20.335
752500..770000	19.302	68.093
872500..925000	19.302	66.004
1105000..1117500	19.302	48.798
1327500..1335000	19.302	39.568
1420000..1432500	19.302	26.695
1560000..1567500	19.302	21.471
1787500..1795000	19.302	23.498
1812500..1817500	19.302	30.300
1907500..1912500	19.302	24.159
1975000..1992500	19.302	35.956
2122500..2135000	19.302	44.684

Appendix VIII: List of genes unique to clade IIh found in at least three isolates *Synechococcus* spp. HB1133, HBA1120, PH41509 and Pros-U-1. The genes located in genomic islands are highlighted in yellow.

Product	Locus tag, gene identifier
putative membrane protein	HB1133_00005
putative membrane protein	HB1133_01765
GAF domain protein	HB1133_00262
Capsular polysaccharide biosynthesis protein	HB1133_01021
possible phosphatase	HB1133_01205
cyclic nucleotide-binding domain protein	HB1133_01219
possible EnvZ-OmpR osmosensory histidine kinase	HB1133_01697
4'-phosphopantetheinyl transferase	HB1133_01751
possible rsbU Serine Phosphatase	HB1133_02391
putative multidrug efflux ABC	HB1133_02487
response regulator receiver (CheY-like)	HB1133_02505

Hypothetical proteins	HB1133_00002, HB1133_00371, HB1133_00405, HB1133_00539, HB1133_00716, HB1133_00824, HB1133_01117, HB1133_01545, HB1133_01546, HB1133_01594, HB1133_01596, HB1133_01603, HB1133_01606, HB1133_01611, HB1133_01641, HB1133_01756, HB1133_02566, HB1133_00323, HB1133_00527, HB1133_00530, HB1133_00544, HB1133_01460, HB1133_01461, HB1133_01483, HB1133_01505, HB1133_01606, HB1133_01747, HB1133_01922, HB1133_02420, HB1133_02426, HB1133_02486