Structural Dynamics of Community Gene Expression In a Freshwater Cyanobacterial Bloom Over a Day-Night Cycle

by

Jia Wang

B.S. Biomedical Engineering, Second Major in Economics
Johns Hopkins University (2009)

Submitted to the Department of Civil and Environmental Engineering in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE
in Civil and Environmental Engineering

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAR 0 8 2012
LIBRARIES

ARCHIVES

©2011 Massachusetts Institute of Technology. All rights reserved.

Signature of Author:_____
Jia Wang
Department of Civil and Environmental Engineering
August 5, 2011

Certified by:_____
Janelle R. Thompson
Assistant Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by:_____
Heidi M. Nepf
Chair of the Departmental Committee for Graduate Students

Structural dynamics of community gene expression in a freshwater Cyanobacterial bloom over a day-night cycle

by

Jia Wang

Submitted to the of Civil and Environmental Engineering
on August 5, 2011 in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Civil and Environmental Engineering

ABSTRACT

Studies of community gene expression, or metatranscriptomics, provide a powerful approach for quantifying changes in both the taxonomic composition (structure) and activity (function) of complex microbial systems in response to dynamic environmental conditions. We have used next-generation Illumina sequencing to characterize the metatranscriptome of a tropical eutrophic drinking water reservoir dominated by the toxigenic cyanobacterium *Microcystis aeruginosa* over a day-night cycle. Cyanobacterial blooms are a major problem in eutrophic lakes and reservoirs, negatively impacting the ecology of the water body through oxygen depletion upon bloom decay and in some cases through production of toxins. Waterborne Cyanobacterial toxins pose a public health risk through drinking and recreational exposure. The frequency of harmful Cyanobacterial blooms (CyanoHABs) is predicted to increase due to warming regional climates and increases in non-point source pollution due to urban expansion. CyanoHABs represent complex consortia of Cyanobacteria that live in association with diverse assemblages of heterotrophic and anoxygenic- photosynthetic bacteria, archaea, microbial Eukaryotes (algae, protozoa, and fungi) as well as viruses and zooplanktonic grazers.

Water sampling was carried out at six time points over a 24 hour period to capture variability associated with changes in the balance between phototrophic and heterotrophic activity. Total RNA was extracted and subjected to ribosomal depletion followed by cDNA synthesis, sequencing, and quality control, generating 493,468 to 678,064 95-101 bp reads per sample. Hierarchical clustering of transcription profiles supported sorting of samples into two clusters corresponding to "day" and "night" collection times. Annotation of reads through the MG-RAST pipeline (Metagenomics- Rapid Annotation using Subsystem Technology) revealed that the community taxonomic composition was relatively constant throughout the day-night cycle and was dominated by transcripts with highest identity to members of the phyla Cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes (in decreasing order) where *Microcystis* transcripts represented 15.3 to 25.6% of the total Bacterial transcriptomes ($E_{ave}=10^{-4.22}$). Community transcripts were enriched with genes from the Cyanobacterial photosynthetic KEGG pathway during the day ($p=0.004$). In contrast, Proteobacterial transcripts were enriched at night (20.4% of the total Bacterial transcriptome compared to 14.3% in the day, $p=0.039$). Metatranscriptomic quantification of microbial community gene expression in a Cyanobacterial bloom dominated by *M. aeruginosa* contributes to a fundamental understanding of nutrient and energy cycling over a day-night cycle. A better understanding of the structure, function, and interaction between members of the complex communities that support the proliferation of toxigenic Cyanobacteria will improve our ability to prevent and control CyanoHABs.

Thesis Supervisor: Janelle R. Thompson
Title: Assistant Professor of Civil and Environmental Engineering

Acknowledgements

This thesis would not have been possible without the guidance and help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, I would like to offer my sincerest gratitude to my advisor, Dr. Janelle Thompson, who has supported me throughout my thesis with her patience and knowledge whilst allowing me the room to work in my own way. I appreciate all her contributions of time, ideas, and funding to make my Master of Science experience productive and stimulating. The joy and enthusiasm she has for research was contagious and motivational for me.

I am also thankful for the assistance rendered by numerous colleagues from MIT and beyond. I would like to especially thank post-doctoral fellow, Dr. Samodha Fernando, for his help in RNA extraction and Illumina sequencing; Dr. Peter Shanahan for his assistance in making the field work in Singapore successful; my fellow graduate students, Kyle Peet and Adam Freedman, for their assistance in the population level study and sharing many happy moments working on projects together; Jean Pierre Nshimyimana for providing water samples in the preliminary study; Jessica Thompson for maintaining *Microcystis* pure culture; my lab mates Jia Yi Har, Timothy Helbig and post-doctoral fellow, Dr. Hector Hernandez, for their insightful discussions and making my lab experience enjoyable; Public Utilities Board (PUB) of Singapore for their generous support in both my undergraduate and graduate studies; the MIT-SMART Center for Environmental Sensing and Monitoring (CENSAM) for the support of this Master thesis work; my collaborators, Dr. Karina Yew-Hoong Gin and Shu Harn Te, as well as Dr. Pingping Gao, Dr. Caiping Feng and Ms. Yueat Tin Wong from PUB Water Hub for sharing their expertise in studying Cyanobacteria in Kranji Reservoir; Dr. Adrian Sharma and Dr. Elizabeth Ottesen for their stimulating discussions on metatranscriptomics analysis; members of the Polz, Alm, Chisholm and Delong lab for occasional use of their equipments; and everyone in Parsons Lab for providing a vibrant research environment.

Last but not least, I owe my deepest gratitude to my parents, Zhaokang Wang and Yuzhen Chen, my brother Xingsheng Wang and my grandparents for their love, encouragement and continual support in all my pursuits.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 What are Cyanobacteria and how can they lead to environmental problems?

Cyanobacteria (blue-green algae) are often a dominant phytoplankton group in eutrophic, shallow and warm freshwater bodies such as lakes and reservoirs. Cyanobacteria of the genera *Microcystis, Anabaena, Oscillatoria, Aphanizomenon*, and others can build up in large numbers to form loose, visible aggregates that may cover large areas. They can grow to form thick green scums that color the water, in the form of blooms (Fig. 1).

Cyanobacterial blooms are a long-standing significant problem in freshwater environments because of the many hazardous consequences they can lead to. For example, the production of potent neuro- and hepato-toxins by several bloom-forming species of Cyanobacteria such as *Microcystis aeruginosa* is a problem for humans and animals that drink untreated water and has been associated with respiratory problems as well as dermal irritation (Stewart I et al., 2009). Microcystin, a cyclic hepatotoxin, is one such cyanotoxin that is most commonly produced by Cyanobacteria (Chorus & Bartram, 1999). Microcystin is not only extremely toxic to fish, domestic animals and birds, but also poses as a threat to public health as it enters the food chain or becomes a direct risk when the bloom-forming freshwater body is used as a drinking water source (Fig. 2). Cyanobacterial blooms are a serious global problem and microcystins have been detected all over the world (Zurawell et al., 2005). Unfortunately, recreational exposure to Cyanobacteria has been associated with a variety of illnesses including gastroenteritis, skin rashes, allergic reactions, headaches, fever and over-exposure to Cyanobacterial toxins can be lethal. For this, the World Health Organization has acknowledged the concern and set a provisional guideline of 4 μg/L of microcystin for low-risk during recreational exposure (Hardy, 2008).

Even if a Cyanobacterial bloom is toxin-free, it can potentially be a serious environmental problem because it affects the oxygen holding capacity of the water body and leads to ecological imbalance. Surface scum-forming Cyanobacterial blooms interfere with re-aeration at the air-water interface and inhibit oxygenic photosynthesis at depth due to shading, thereby eliminating sources of oxygen to deeper waters. In addition, through bloom decay, Cyanobacterial blooms create excessive biological oxygen demand (BOD) and can lead to

hypoxic/anoxic conditions that have resulted in extensive fish kills in different parts of the world. Not only that, since Cyanobacteria play vital roles in aquatic food webs, their toxin production and accumulation also have allelopathic effects on phytoplankton, zooplankton and macrophytes, hence disrupting ecological balance. Moreover, dense scums affect lakes aesthetically and limit their usefulness as a drinking water source by causing other water-related problems, including foul odors and unpalatability.

Mechanisms to control Cyanobacterial blooms are of great interest from the perspective of water quality management. Multiple physical (e.g. hydrodynamics, temperature, light availability), chemical (nutrients), and biological (Cyanobacterial growth, grazing, competition) factors come into play that determine whether a bloom will occur (Fig. 3) and an understanding of these factors can be used to develop strategies to control Cyanobacterial cell density. For example, algicides such as copper sulfate are toxic to Cyanobacteria and have been used traditionally to control algal blooms in water supply storages and lakes. However, one caveat of this treatment is that when the algae cells are lysed by algicides, toxins released from disrupted cells will disperse throughout the water body and compromise the effectiveness of its removal by conventional filtration via flocculation. In a more severe case scenario, it has been reported that members of two local populations in Australia have suffered from an acute toxicity as a consequence of algicide treatment of water blooms (Falconer, 2001). Moreover, since copper is a broad-spectrum aquatic biocide, non-target species like zooplankton and fish will also be affected. This can have significant adverse environmental effects on the ecology of the water body as well as the sustainability of such water resource. Therefore, water management through other means is more desirable and can include nutrient source (nitrogen and phosphorus) reduction (Paerl et al., 2011a), improved water circulation (Mitrovic et al., 2011) which prevents persistent thermal stratification, and perhaps biological control by designing strategies that stimulate non-Cyanobacterial populations to inhibit growth and/or toxin production. For example, *Sphingomonas* is found to be capable of degrading microcystin toxins (Manage et al., 2010) and *Clostridium* might be involved in degradation of *Microcystis* scum (Xing et al., 2011).

Fig. 1: Dense scums covering the water surface can be observed in many Cyanobacterial bloom-forming lakes. < http://wallowavalleyonline.com/wvo/?p=1593>

Fig. 2: Cyanobacterial blooms can lead to extensive fish kills. <http://oh.water.usgs.gov/newsroom_2009.htm>

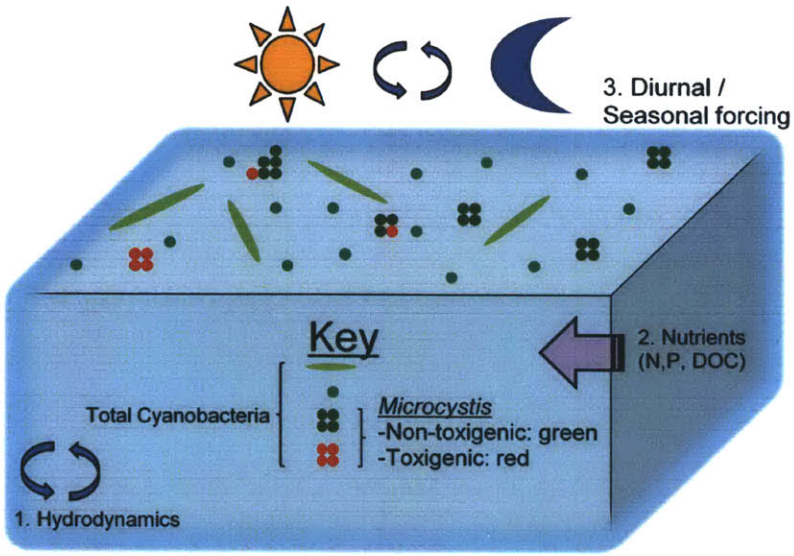


Fig. 3: Environmetnal forcings that may affect a Cyanobacterial bloom include water circulation and stratification (hydrodynamics), nutrient status and changes in environmental conditions such as irradiance, precipitation and temperature associated with diurnal (day/night) and seasonal changes (diurnal or seasonal forcing).

1.2 Structure and function of microcystin toxin

Cyanotoxins mediate part of the public-health risk associated with Cyanobacterial blooms. Microcystin toxins are produced by a diverse range of Cyanobacterial genera including *Microcystis, Anabaena, Planktothrix (Oscillatoria), Anabaenopsis, Nostoc, and Hapalosiphon* (Rinehart et al., 1994). They are secondary metabolites since they are not used for primary metabolism, growth, or cell division. The biological or ecological role of many toxins is unclear. Over 65 cyclic heptapeptides have been identified in the microcystin toxin family.

Structure of microcystin

Microcystins (MC) have a common structure of cyclo(Adda-D-Glu-Mdha-D-Ala-L-X-D-MeAsp-L-Z), where X and Z are variable L-amino acids, Adda is 3-amino-9-methoxy-2,6,8,-trimethyl-10-phenyl-4,6,0decadienoic acid, D-MeAsp is 3-methylaspartic acid, and Mdha is N-methyl-dehyroalanine (Fig. 4). Substitutions at these two positions give rise to at least 21 known primary microcystin analogues. For example, one of the most common and best studied microcystin is MC-LR where X and Z are leucine (L) and arginine (R) respectively. At pH 6 to 9, there is only limited passive diffusion of MC-LR directly from water into biota (De Maagd et al., 1999). Alterations to variable residues may or may not alter toxicity. For example, microcystins like MC-RR (arginine, arginine) is the least toxic while the most toxic and most studied microcystin variant is MC-LR.

Molecular synthesis of microcystin

Microcystin synthesis has a mixed polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) origin. Each cycle of polyketide or polypeptide chain elongation of a microcystin molecule is mediated by a module or clusters of enzymatic sites formed by large multifunctional protein complexes. The structure of the resulting microcystin can be determined by factors such as the order of modules or the number and type of catalytic domains within each module.

Neilan & Tillett (2000) identified microcystin synthetase, a 55kb gene cluster from *Microcystis aeruginosa PCC7806,* which is responsible for MC-LR synthesis. They showed that there are six large open reading frames (*mcyA, B, C, D, E and G*) of a mixed NRPS/PKS

nature and four more small open reading frames (*mcyF, H, I and J*) with putative precursor and microcystin tailoring function. McyH was later proposed as an ABC transporter gene (Pearson et al., 2004).

The Mcy gene cluster has a bidirectional operon structure (Fig. 5). The *mcyD-J* genes form the larger operon that encodes for PKS-NRPS modules catalyzing the formation of the pentaketide-derived β-amino acid moiety, Adda, and its linkage to D-glutamate. The smaller operon *mcyA-C* encodes for the NRPS modules for the extension of this dipeptidyl intermediate to a heptapeptide which subsequently is cyclized and becomes a mature microcystin (Neilan & Tillett, 2000).

Toxicological effects of microcystin

In vertebrates, the predominant toxic effects of microcystins are on the liver since microcystins are preferentially uptaken by hepatocytes. This toxicity is mediated through the active transport of microcystin into hepatocytes by the bile acid organic anion transport system. Some effects on the liver are induced redistribution of cytoskeletal components, separate of hepatocytes, breakdown of the sinusoidal endothelium and intraheptic heamorrhage. The most apparent effect in hepatocytes is morphological transformation of microtubules (Falconer & Yeung, 1992), a process that is initiated through inhibition of the protein phophatases PP1 and PP2A (MacKintosh et al., 1990). This disrupts the fine balance of phosphorylation/ dephosphorylation activities of kinase and phosphatase and hence intracellular signal transduction pathways. Microcystin also induces DNA oxidative damage (Zegura et al., 2003) and sublethal doses can result in hepatocellular apoptosis (Guzman and Solter, 2002).

Distribution of toxicity in a Cyanobacterial bloom

Toxin production and release by *Micocystis* cells is akin to a multi-level decision making process. At the genetic level, not all the *Microcystis* cells in a bloom event are potentially toxigenic, ie. some cells do not contain the toxin genes (*mcy* gene cluster) that confer the ability to produce toxins while other cells may have mutations in the operon (i.e. structural or regulatory) that prevent biosynthesis of the toxin. Thus, toxicity of many Cyanobacteria, including *Microcystis,* is cell-specific. Among the potentially toxigenic cells (i.e. containing a functional *mcy* gene cluster), expression of the *mcy* synthetase genes may be under some

regulatory controls. If the toxins are synthesized, whether they pose a contact-risk depends on whether they are released from the cell, for example, via active transport or cell lysis in the bloom or after skin contact or ingestion. Therefore, the toxicity of a Cyanobacterial bloom will be a function of the proportion of cells bearing toxin genes, the regulation of those toxin genes (i.e. whether the toxin is synthesized) and whether the toxins are released into the environment by active excretion or cell lysis.

Since toxin is cell-specific, the toxicity of a Cyanobacterial bloom depends on the proportion of toxigenic to nontoxic strains in Cyanobacteria communities. Thus, for a bloom event in a single water body, toxin production in the bloom may differ substantially depending on the distribution and expression of toxin genes. For example, 1000-fold differences in toxin concentrations have been observed in different strains of *Microcystis* (Zurawell et al., 2005).

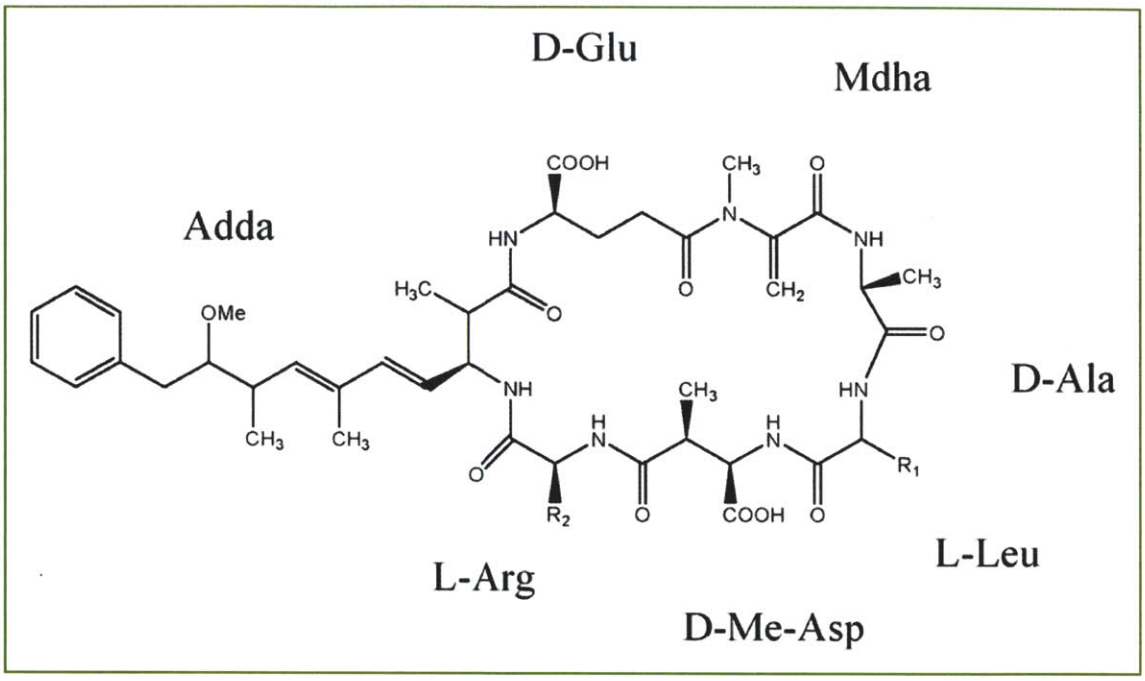


Fig. 4: Molecular structure of microcystin-LR with the seven amino acid modules.
< http://www.cyano-biotech.com/products/leadstructures/livertrans.html>

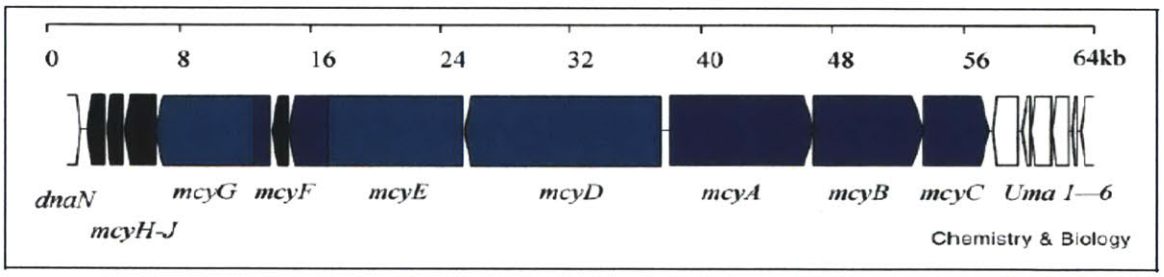


Fig. 5: Gene arrangement (*mcyA – mcyJ*) in microcystin synthetase, the mcy gene cluster. The relative size of open reading frames and direction of transcription are shown. (Neilan & Tillett, 2000)

1.3 How is microcystin production affected by environmental factors?

Blooms in the same water body can be toxic or non-toxic from one year to the next, depending on its strain composition. Different strains display varying toxicity levels under different laboratory conditions while the exact stimulus for toxin production is not well understood. Environmental parameters like light intensity, temperature, nutrient availability and trace metals have been simulated in lab studies to analyze their effects on microcystin production (Sivonen et al., 1999).

Light intensity

Studies on light intensity are highly variable in results, partly due to the different intensities and strains being experimented on. However, a low toxin production has been documented at low light intensities of 2-20 $\mu E/m^2/s$ while the highest level of toxin production occurs at 20 - 142 $\mu E/m^2/s$. (Kaebernick et al., 2000) showed that red light induces higher transcriptional levels of *mcyB* and *mcyD*. In addition, there seemed to be two light thresholds, one between dark and low light (16 $\mu E/m^2/s$), and one between medium (31 $\mu E/m^2/s$) and high light (68 $\mu E/m^2/s$), at which significant increases in transcription occurred. (1 $\mu E/m^2/s$ = 54 Lux if sunlight is the source).

Temperature

Different chemical forms of microcystin are found to be produced preferentially at different temperatures. For example, *Anabaena spp.* preferentially produces MC-LR at 25°C instead of MC-RR which is preferentially synthesized at higher temperatures (Hikmet et al., 2004). This can have implications for the toxicity of the bloom as MC-LR has been determined to be more toxic than MC-RR (Gupta et al., 2003). Ecological interactions can also be driven by temperature and affect microcystin concentrations by shifting the relative abundance of toxigenic and non-toxigenic populations. For example, water temperature was found to affect the relative abundance of two *Microcystis* species, *M. aeruginosa* and *M. wesenbergii* in both field (Furuike Pond in Japan) and laboratory settings (Imai et al., 2009). *M. aeruginosa* was more abundant at a relatively higher temperature (24.7 - 33.98 °C) while *M. wesenbergii* was more abundant during periods of lower temperature (19.6 - 28.68 °C) in the pond. Under a laboratory setting, the growth rate of *M. aeruginosa* was significantly higher at higher

temperatures (30 and 35 °C) while both species have similar growth rate at lower temperatures (20 and 25 °C).

Nutrient availability- nitrogen and phosphorus

Microcystin production from *Microcystis aeruginosa UTEX 2388* has been studied in a phosphorus-limited continuous culture (Oh et al., 2000). When phosphorus is limited, the specific growth rate (μ) of *M. aeruginosa* decreases from 0.8 to 0.1/day and the cellular phosphorus content decreases. As growth rate drops, the Nitrogen-to-Phosphorus atomic ratio of steady-state cells increases; carbon fixation rate at an ambient irradiance (160 microeinsteins $m^{-2}s^{-1}$) decreases; microcystin-producing rate also decreases linearly however microcystin content within cells increases per gram of dry weight. Therefore, the growth of *M. aeruginosa* was reduced under a phosphorus-limited condition due to a low carbon fixation rate, whereas the microcystin content was higher.

The degree of phosphorus limitation also appears to affect the type of microcystin produced (Oh et al., 2000). Under a severe phosphorus-limited condition (high N/P atomic ratio), the MC-LR content was lower than that of MC-RR, yet the ratio of MC-LR to MC-RR produced increased. Thus, the more toxic form, MC-LR, is preferentially produced under a phosphorus-limited condition.

Nitrogen affects the production of cyanotoxins differently in nitrogen-fixing and non-nitrogen-fixing Cyanobacteria. For instance, nitrogen fixers like *Anabaena* and *Nodularia* show highest levels of microcystin in a nitrogen-free medium while non-nitrogen fixers like *Oscillatoria* and *Microcystis* strains show highest levels of toxin at high levels of nitrogen (Rapala et al., 1997).

Heavy metals

Of the varitety of metals (iron, zinc, aluminium, cadmium, chromium, copper, manganese, nickel and tin) tested at non-lethal concentrations for their effects on the growth and microcystin production of *M. aeruginosa PCC7806*, significant effects only came from zinc and iron (Lukac & Aegerter, 1993). Zinc was found to be required for optimal growth and toxin production. When 0.01-0.25 μM zinc solution was added to trace-metal-depleted media,

growth rate increased to 1.5 times while microcystin production increased by 30%. However, cells were killed at very high concentration of zinc (10 μM).

Iron is essential for photosynthetic organisms like *Microcystis* because it is required for chlorophyll-a synthesis, respiration and photosynthesis. However, bioavailable ferrous iron is often limited in freshwater environments because it can be quickly oxidized to ferric iron which forms highly insoluble oxyhydroxides and becomes non-bioavailable. Iron depletion within cells can be detrimental because the modified phycobilisomes cannot make use of the extra light energy during the day, hence leading to oxidative stress in which harmful reactive oxygen species forms through the iron-mediated Fenton reaction (Michel & Pistorius, 2004; Latifi et al., 2008). Kosakowska et al. (2007) have shown that the threshold iron concentration required for maintaining chlorophyll-a synthesis and the growth rate of *M. aeruginosa* is 10μM.

In contrast to zinc, microcystin production and iron availability appeared to be negatively correlated with cells producing more microcystin toxin per dry weight under iron-limiting conditions. Low concentrations of iron (< 2.5 μM), which are implicated in slower cell growth, have led to higher microcystin yield: cells produced 20 to 40% more microcystin by dry weight (Lukac & Aegerter, 1993). (Lyck et al., 1996) also observed a negative relationship between microcystin and iron concentration: the ratio of toxin to chlorophyll or toxin to protein increased during iron starvation. Moreover, toxic *Microcystis aeruginosa CYA 228/1* strain maintained a greater tolerance to iron depletion or longer cell vitality than the non-toxic strain, *M. aeruginosa CYA 143* when iron availability falls low. (Utkilen & Gjolme, 1995) found that toxin-producing *M. aeruginosa* strains possess a more efficient iron uptake mechanism than nontoxic strains. He further hypothesized that microcystin production by a synthase is regulated by the amount of free Fe^{2+} present in the cell as microcystin might act as an intracellular chelator which keeps free cellular Fe^{2+} low. Thus, they suggested that microcystin may facilitate the uptake and storage of iron in toxic cells by keeping a low level of cellular Fe^{2+}.

Martin-Luna et al. (2006a) proposed that FurA, a ferric uptake regulator as well as a transcriptional regulator of genes involved in iron homeostasis and oxidative stress, might regulate the *mcy* gene cluster. The mechanism is such that when cellular Fe^{2+} iron is replete,

Fur proteins dimerize and bind to 'iron boxes' or conserved sequences in the promoter region to repress transcription of genes involved in iron uptake (Martin-Luna et al., 2006b). The group showed that Fur from *M. aeruginosa PCC7806* binds to in vitro promoter regions of several *mcy* genes which contain the putative 'iron box'. Thus, Fur acts as a sensor of iron availability and oxidative stress. This finding might help to explain the previous observation made by Utkilen & Gjolme (1995) that microcystin production increases under iron-deplete condition. This is because when the cellular iron concentration becomes low, transcriptional repression of *mcy* genes coming from FurA will be relieved, leading to higher microcystin synthesis.

Alexova et al., (2011) also found that microcystin production in *M. aeruginosa PCC7806* increases in an iron-dependent and FurA regulated manner. When the ability to produce microcystin is disrupted due to some mutations in the *mcy* gene cluster, processes such as photosynthesis, Fe^{2+} transport and transcription of the Fur family of transcriptional regulators are remodeled. Thus, he inferred that the presence of toxin might give an advantage to microcystin-producing Cyanobacteria as it protects the cell from reactive oxygen species-induced damage under iron-stressed condition.

1.4 Ecological interactions:

Microcystis live in a complex freshwater community where other autotrophs, heterotrophs, zooplankton, macrophytes and viruses (cyanophages) are present. Whether these ecological interactions influence the degree of toxin production in blooms is unclear. Some of the ecological interactions of *Microcystis* with other community members include competition for nutrients, light, and space; predation by zooplankton; and synergistic interactions like releasing fixed carbon (food) for heterotrophs or having the scums acting as habitats for heterotrophs.

Several studies touched on the allelopathic effects of *Microcystis* on growth inhibition of other algae. For example, it has been found that MC-LR inhibited the reproduction of the prasinophyte *Nephroselmis olivacea* (Chritoffersen, 1996). Engelke et al. (2003) found that *Microcystis aeruginosa* increased toxin production when a non-toxic culture of *Planktothrix agardhii*, or its spent medium, was added. Similarly, *Microcystis* sp. cells were observed to retard the growth and photosynthesis of the dinoflagellate *Peridinium gatunense* (Sukenik et al., 2002). Singh et al. (2001) also noted increased cell lysis of the Cyanobacteria *Nostoc muscorum* and *Anabaena* after exposure to MC-LR for six days. Kearns & Hunter (2001) reported that Cyanobacterial toxins could inhibit the motility of green algae.

One of the several ways that nature takes to keep the abundance of *Microcystis* in check is through predation by heterotrophs. Zooplankton and protozoan grazers can exert top-down control on *Microcystis* (Low et al., 2010). In some cases, higher abundance of larger zooplanktons preying on smaller zooplankton/protozoa have been correlated to higher densities of Cyanobacteria like *Microcystis* because they remove the smaller grazers that target *Micrcystis*. In other cases, larger zooplanktons feed directly on large aggregates of *Microcystis* cells in blooms. However, it has also been proposed that microcystin toxin may act as a grazing deterrant. Lampert (1981) studied the effects of *Microcystis aeruginosa* on survival, growth and food uptake of *Daphnia pulicaria*. He found that in pure suspensions of *Microcystis aeruginosa*, the daphnids did not survive more than 48 hours and growth was markedly reduced when only 50 µg carbon/l of *Microcystis* was added to the normal *Scenedesmus* food; and growth ceased at a concentration of 250 µg C/l.

Another control on the abundance of *Microcystis* in natural environments is lysis by viruses/ phages. Cyanobacteria viruses (cyanophages) have been observed to lyse *Microcystis* cells in the field (Fox et al., 1976) and in the lab (Tucker & Pollard, 2005; Yoshida et al., 2006). Yoshida et al. (2008) suggested that cyanophage dynamics may affect shifts in microcystin-producing and non-producing populations. Manage et al. (2001) also show that cyanophages are potentially an important factor of bloom dynamics as their study results show that *M. aeruginosa* are infected by cyanophage-like particles in a density-dependent manner: the densities of cyanophage-like particles became undetectable when the abundance of *M. aeruginosa* was low. Therefore, cyanophages are important agents in decomposing blooms as occasional peaks of density of cyanophage-like particles were detected in October, June, and August, when sharp declines in *M. aeruginosa* cell densities were also observed (Manage et al., 2001).

Microcystis are also found to be killed and degraded by other types of heterotrophs. Cytophaga-like bacteria are able to lyse Cyanobacteria by attachment and secretion of diffusible lytic substances (Rashidan et al., 2001) which include a variety of exoenzymes that are capable of hydrolyzing Cyanobacterial cell wall. Xing et al. (2011) found *Clostridium* (a Firmicute genus), an inhabitant of *Microcystis* bloom in Lake Taihu, China, is able to anaerobically degrade extracellular polymeric substances in *Microcystis* scum.

In addition, ecological interactions within the bloom may control the stability of microcystin toxin released into the water. *Sphingomonas* strains that co-occur with *Microcystis* in blooms have been found be capable of degrading microcystin toxins (Manage et al., 2010). A microcystin-degrading gene cluster, *mlr*, has been described in *Novosphingobium* (Jiang et al., 2011).

1.5 Study Site – Kranji Reservoir, Singapore

Kranji Reservoir, Singapore, the sampling place for this study, has been designated as a potential place for development of water sports and other recreational activity under the ABC (Active, Beautiful, Clean) Waters Program by the Singapore Public Utilities Board (PUB 2007). However, high abundances of Cyanobacteria dominated by *Microcystis* spp. have been observed in Kranji Reservoir and other connected Singaporean reservoirs and represent a management challenge (NTU, 2008). Initial work has been done to characterize the abundance and year-round occurrence of *Microcystis* in the reservoir system. Gin et al. (2005), measured bacterioplankton densities in five reservoirs in Singapore finding 1.14–7.41×10^6 Bacterial cells /ml. Phycoerythrin containing Cyanobacteria (e.g. *Microcystis*), were found to be dominant in four out of the five reservoirs at concentrations of 10^4 – 10^5 cells / ml. More recently QPCR was used to quantify *Microcystis aeruginosa* 16S rRNA gene biomarker in the Kranji Reservoir during a Cyanobacterial bloom (Te and Gin, 2011). The average concentration of *Microcystis* 16S rRNA genes was 4 × 10^6 16S rRNA gene copies/ml (equivalent to 2 x 10^6 cells/ml based on 2 rRNA operons per cell (Kaneko et al., 2007)) while approximately half of the *Microcystis* contained genes for toxin biosynthesis (Te and Gin, 2011). Moreover, local reservoirs in Singapore were found to be phosphorus limited and alkaline which favored the growth of phytoplankton and bacteria (Gin et al., 2005). Six species of *Microcystis* have been found in Singapore: *Microcystis aeruginosa, Microcystis flos-aquae, Microcystis robusta, Microcystis smithii, Microcystis wesenbergii* and *Microcystis zanardinii* (Pham et al., 2011). Low et al. (2010) found that zooplankton in tropical reservoirs in Singapore exert a top-down control on phytoplankton. They postulate that Cyanobacteria are able to proliferate because some genera like *Microcystis* are able to produce toxins which act as a potential defence mechanism against grazers like cladocerans and calanoids. Moreover, the sheer size of colony-forming *Microcystis* protects them from ingestion by zooplankton.

The Republic of Singapore is an island city-state located at 1° N (137 km north of the equator). Due to its tropical rainforest climate (high humidity and rainfall), it does not have four seasons and its temperature is relatively uniform (23 - 32 °C) all year round (Ministry of the Environment and Water Resources of Singapore, 2005). With a limited land area of 697 square kilometers and a large urban population size of 4.42 million, water scarcity has always

been a problem that the country has to face. The government body, Public Utilities Board (PUB), is responsible for water resource management in Singapore. They have developed a holistic approach to ensure a sustainable water supply for the country: for example, 50% of the total land area is transformed to water catchment areas; water source multiplication and diversification are made possible through the "4 National Taps Strategy" which consists of collecting rainwater from local catchments, importing water from Johor (Malaysia), reclaiming NEWater from wastewater and desalinating water from the sea.

Kranji Reservoir is one such local catchment. It has a size of approximately 647 hectares and is located within the Western Catchment of Singapore (1°25'N, 103°43'E, Fig. 6). In 1975, this reservoir was created by damming of an estuary which drained into the Johor Straits that separate Malaysia from Singapore. The reservoir catchment (6067 hectares) consists of four tributaries: Kangkar River, Tengah River, Pengsiang River and Pangsua River and has various land-uses, including forests, reserved areas, agriculture and residential areas (NTU 2008).

Kranji Reservoir has been selected to be developed as a spot for water-related recreational activities under the ABC (Active, Beautiful, Clean) Waters Program by PUB. This program aims to open up Kranji Reservoir as an aesthetically pleasing (the "Beautiful" element) and sufficiently high quality and safe (the "Clean" element) water gateway for recreational activities such as boating and fishing (the "Active" element). By doing so, PUB hopes to cultivate a sense of ownership and respect for water among Singaporeans. In preparation for this, PUB commissioned the MIT-SMART Center for Environmental Sensing and Monitoring (CENSAM) to carry out scientific studies on the evaluation of health risks related to recreational activities in Kranji Reservoir. This study, the metatranscriptomic quantification of microbial community gene expression in a Cyanobacterial bloom over a day-night cycle, will help PUB to strategize management direction and techniques in controlling and minimizing risks related to CyanoHABs.

Fig. 6: Map of Singapore. The location of Kranji Reservoir is indicated with A (Google map).

1.6 Project scope & Thesis focus:

Blooms of Cyanobacteria *(Microcystis)* constitute a threat to the safety and ecological quality of surface waters worldwide especially due to the production of hepatotoxin microcystin. The frequency of harmful Cyanobacterial blooms (CyanoHABs) is predicted to increase due to warming regional climates (Paerl et al., 2011b) and increases in non-point source pollution due to urban expansion (Novotny, 2011). CyanoHABs represent complex consortia of Cyanobacteria that live in association with diverse assemblages of heterotrophic and anoxygenic photosynthetic bacteria. A better understanding of the structure, function, and interaction between members of the complex microbial communities and environmental factors that support the proliferation of toxigenic Cyanobacteria will improve our ability to prevent and control CyanoHABs (e.g. ability to predict the occurrence of high toxicity blooms). This is a major concern in the Kranji Reservoir in Singapore as high abundances of Cyanobacteria dominated by *Microcystis* spp. have been observed (NTU, 2008). This also poses as a water management challenge as the reservoir is developed for future recreational usage.

The goal of this study is to describe the structural complexity within the toxigenic algal bloom of the Kranji Reservoir. While this thesis focuses on the structure (taxonomic composition) of the bloom, it paves the way for future analysis to explore how the structure and function of the bloom control the expression of toxin genes (i.e. functional dynamics). We are ultimately interested in whether interactions between *Microcystis*, toxin producing cells, and other members of the bloom community can be harnessed (in conjunction with ongoing programs directed at management of the bloom through nutrient source control and hydrodynamics) as a multi-pronged approach to control the persistent algal blooms in the Kranji Reservoir.

The exact mechanism controlling microcystin production, accumulation, release and the significance of microcystin to Cyanobacteria or its toxicity targets are still debated and remain obscure. Even though a few hypotheses have been suggested in terms of cellular (light regulation, iron transport, cell signaling) and ecological (anti-herbivory, allelopathy) functions, little is known about the natural function of microcystin, be it within populations of *Microcystis* or in a larger ecological context of a lake.

To address this complex problem, this study provides insights into Cyanobacterial dynamics in their natural habitats through an analysis of community gene expression patterns in a diel cycle. By elucidating the genetics basis of Cyanobacteria growth through a RNA-seq (sequencing of RNAs) approach, more effective water quality monitoring techniques such as novel ways to destroy toxins or prevent its production could be developed. Metatranscriptomic, or the study of community gene expression, provides a powerful approach for quantifying changes in both the taxonomic composition (structure) and activity (function) of complex microbial systems in response to dynamic environmental conditions.

Bacterioplankton sampling was carried out at six time points over a 24 hour period (day-night cycle) to capture variability associated with changes in the balance between phototrophic and heterotrophic activity. Total RNA was extracted and subjected to ribosomal depletion followed by cDNA synthesis and Next-generation Illumina sequencing. Short mRNA reads (~100bp each) that passed quality control measures were subjected to post-sequencing analysis. Read annotation was also carried out with the MG-RAST pipeline (Metagenomics-Rapid Annotation using Subsystem Technology) for additional analyses including organismal classification based on M5NR (M5 Non-redundant protein database). Community taxonomic composition based on transcripts at each sampling time point were scrutinized top down from domain, to phylum, class, order and sometimes all the way to the genus level in certain phyla. Results from this analysis provide an initial description of the active members of the community and at what time of the day they are active.

Functional classification was assessed by blasting (BLASTX) the reads against the NCBI database and imported into MEGAN (MEtaGenome Analyzer) for comparative analysis across six samples. Transcripts were assigned functional pathways based on their best protein hits from blastx to databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and SEED (Subsystems Technology). KEGG functional pathways or SEED subsystems that were enriched either during the day or night could be inferred from the abundance of reads that had hits to the genes falling within a particular pathway or subsystem. How and when pathways are utilized or regulated are important information for inferring community dynamics and interactions.

To analyze gene expression of *Microcystis aeruginosa*, transcripts that were mapped to this reference genome were normalized by gene length and total sample read count to give a representation of its abundance and hence expression level at each sampling time point. Hierarchical clustering and PCA (principal component analysis) were employed to separate samples into groups based on transcriptional profiles. Statistical tests including T-test and SAM (Significance Analysis for Microarrays) were applied to screen for genes with significant differential gene expression.

In conclusion, metatranscriptomic quantification of microbial community gene expression in a Cyanobacterial bloom dominated by *Microcystis aeruginosa* contributes to a fundamental understanding of nutrient and energy cycling over a day-night cycle. Ultimately, a better understanding of the structure, function and interactions within the complex communities that support toxigenic Cyanobacteria will enable better management and control of these CyanoHABs.

Chapter 2: Methods and Materials

2.1 Sample collection

The sampling field work was carried out at Kranji Reservoir in Singapore on the 14th and 15th January 2010. Six sampling time points were taken: 5pm, 8pm (14th Jan) 6am, 8am, 1pm and 3pm (15th Jan).

Plankton suspended in surface reservoir water (205 - 350ml) was filtered onto a 0.22μm Sterivex filter, and cells and nucleic acids were immediately preserved in 2.5ml of RNA*later* followed by storage at -80°C. The environmental parameters were measured using thermometers (for temperature) and light photometer (for light intensity).

The nutrient status of the reservoir was assessed by the Biological & Chemical Technology Division of a local company, Setsco Services Pte Ltd. The testing methods used for Dissolved Oxygen Carbon (DOC), Total Kjeldahl Nitrogen (TKN), Total Phosphorus (TP) are APHA: Pt 5310B, APHA: Pt 4500-N_{org} (D), APHA: Pt 4500-P (H) respectively. APHA is a Standard Method for the Determination of Water and Waste Water (APHA 21st Edition: 2005).

Microcystin toxin level in the surface samples (without lysing Cyanobacteria cells) was measured with QuantiTube™ Kit for Microcystin, Catalog number: ET 039. The calibration range for the kit is 0.4 - 2.5 ppb (parts per billion, ng/ml) and the Limit of Detection (LOD) is 0.18ppb. The kit works on the principle of a competitive Enzyme-Linked ImmunoSorbent Assay (ELISA), where microcystin toxin in the sample competes with horseradish peroxidase-labeled microcystin for a limited number of antibody binding sites on the inside surface of the test tubes. The outcome of competition is visualized with a color development. The darker is the color, the lower the concentration of microcystin in the sample. The optical density of the color was measured with a 450nm photometer.

After sample collection, Kranji Reservoir filter samples were returned to MIT. Extraction of RNAs, enrichment and amplification of prokaryotic mRNAs, cDNA synthesis and Illumina library preparation were performed by collaborator Dr. Samodha Fernando, a postdoctoral associate in the Thompson laboratory in Civil and Environmental Engineering at MIT

(sections 2.2 and 2.3). Sample collection and data analysis (sections 2.1, 2.4 to 2.7) were performed by the thesis author.

2.2 RNA extraction and depletion of ribosomal RNAs and Eukaryotic mRNAs

RNA extraction

Sterivex filter cartridges containing plankton samples were taken from the -80 °C freezer, thawed for 1 min and opened by removal of the cartridge tip. RNA*later* was removed by pushing through with a syringe. Then, the cartridge was wrapped with a sterile Whirl-Pak plastic bag and broken open by hammering its bottom portion close to its tip. A sterile forceps was used to pull out the inner core which had the filter attached. Without touching the inner core, a scalpel was used to cut the filter by making an incision along the top and bottom circumference, such that filter could be dislodged. The filter was placed in an eppendorf tube and kept on ice. To each tube, 1 ml TRIzol RNA purification reagent (Invitrogen) and sterile autoclaved zirconium beads were added followed by vortexing for 10 min at 4 °C. The greenish color on the filter due to Cyanobacteria disappeared during vortexing suggesting cells were lysed. After vortexing, tubes were incubated at room temperature for 5 min. The tubes were centrifuged to collect filter debris and supernatant was transferred to a new tube. 1 ml chloroform was added, vortexed thoroughly and incubated at room temperature for 10 min. Next, tubes were centrifuged at 9,000 x g for 30 min at 4 °C. The colorless upper aqueous phase (RNA) was transferred into a new eppendorf tube. 1/10 volume of 3 M Ammonium acetate, pH 5.2 was added to the tube of RNA, followed by adding 0.5 ml isopropanol to precipitate the total RNA. The tubes were capped and inverted a few times to mix. After incubating at -80 °C overnight, the tubes were centrifuged at 14,000 x g for 30 min at 4 °C. The supernatant was discarded and the pellets were washed with 500 µl 70 % ethanol, centrifuged at 14,000 x g for 15 min at 4 °C. The supernatant was discarded and the pellets were air-dried for about 5 min at room temperature. Extracted total RNAs (pellets) were resuspended in 50 µl DEPC treated RNase-free water and stored at -80 °C. The amount of extracted RNA is reflected in Appendix C.

Depletion of Eukaryotic mRNAs

To remove Eukaryotic mRNAs, PolyA subtraction was carried out using the Poly(A)Purist™ MAG Kit (Magnetic mRNA Purification Kit, Ambion Part No. AM1922). To the TRIzol

extracted RNA, 50 µl 2X Binding Solution was added and mixed thoroughly. The magnetic beads were prepared by placing tubes of 1 µl Oligo(dT) MagBeads suspension (1%) on the Magnetic Stand for 2 min. The buffer was aspirated off carefully and discarded. 30 µl Wash Solution 1 was added to the captured beads. Tubes were removed from the stand and beads were resuspended by inverting the tubes several times. The beads were then recaptured with the stand and the supernatant was discarded. This washing step was repeated once. Total RNA in 1X Binding solution was transferred to the washed Oligo(dT) MagBeads, sealed, inverted several times and incubated at 70 °C for 5 min which denatured secondary structure and maximized hybridization between the poly(A) sequences found on most mRNAs and the poly(T) sequences on the Oligo(dT) MagBeads. The mixture was then incubated at room temperature with gentle agitation for 30 min. Next, the tubes were placed back on the stand for 2 min to allow recapturing of beads. The supernatant, which had the Eukaryotic mRNA depleted, was transferred to a collection tube.

Depletion of Eukaryotic 18S, 28S rRNAs

rRNA frequently represents 53 % - over 90 % of the total RNA (Frias-Lopez et al., 2008, Stewart FJ et al., 2010). Thus, rRNA needs to be removed so that the majority of sequencing effort goes to mRNA. To remove Eukaryotic RNA (18S, 28S rRNAs and polyadenylated mRNAs), the MICROB*Enrich*™ Kit (Ambion Part No. AM1901) was used. 300 µl Binding Buffer was added to the supernatant from the Poly(A)Purist™ MAG Kit and vortexed gently to mix. 2 µl Capture Oligo Mix was added for every 5 µg of RNA and incubated at 70 °C for 10 min, then at 37 °C for 1 hr. Oligo MagBeads were prepared by withdrawing 25 µl Oligo MagBeads suspension for every 5 µg of input RNA to a 1.5 ml tube, capturing the beads with a magnetic stand, removing the supernatant, washing the beads with an equal volume of Nuclease-free Water and finally equilibrating the beads with an equal volume of Binding Buffer. Next, the RNA and Capture Oligo Mix were added to the prepared Oligo MagBeads and incubated at 37 °C for 15 min. The beads were captured with the magnetic stand and supernatant containing the enriched Bacterial RNA was recovered. Any remaining Bacterial RNA was recovered from the beads by washing them with 100 µl Wash Solution at 37 °C and recovering the wash.

Depletion of Bacterial 16S, 23S rRNAs

To remove Bacterial rRNAs (16S and 23S rRNAs), Bacterial mRNA enrichment was carried out using the MICROB*Express*™ Kit (Ambion Part No. AM1905). To the supernatant and wash from the MICROB*Enrich*™ Kit, 4 µl Capture Oligo Mix was added, vortexed and centrifuged briefly to get the mixture to the bottom of the tube. The mixture was incubated at 70 °C for 10 min, then 37 °C for 15 min to allow the hybridization between capture oligonucleotides and homologous regions of the 16S and 23S rRNAs. 50 µl Oligo MagBeads were prepared in the similar manner as the MICROB*Enrich*™ Kit and resuspended with an equal volume of Binding buffer at 37 °C. Next, the RNA and Capture Oligo Mix were added to the prepared Oligo MagBeads and incubated at 37 °C for 15 min. The beads were captured with the magnetic stand and supernatant containing the enriched mRNA was transferred to a collection tube. Any remaining mRNA was recovered from the beads by washing them with 100 µl Wash Solution at 37 °C and recovering the wash.

Depletion of tRNAs

To remove tRNAs and purify RNA from reactions above, MEGAclear™ Kit (Ambion Part No. AM1908) was used. 10 µl RNA was brought to 100 µl with Elution Solution and was mixed gently. 350 µl Binding Solution Concentrate was added and mixed gently by pipetting. 250 µl 100% ethanol was added and also mixed by pipetting. Next, the RNA mixture was pipetted onto a Filter Cartridge, centrifuged at 14,000 rpm for 1 min and the flow-through was discarded. The filter was washed twice with 500 µl Wash Solution. RNA was eluted by applying 50 µl Elution Solution to the center of the filter on a new tube which was capped, incubated at 70 °C for 10 min and then centrifuged at 12,000 x g for 1 min at room temperature.

RNA remaining from the four-step subtraction above was precipitated at -80 °C overnight with 0.1 volume 3 M NaAc, 0.02 volume glycogen and 3 volume ice cold 100% ethanol, with vortexing to mix. RNA was recovered by centrifugation at 13,000 rpm for 30 min. Supernatant was discarded and the pellet was twice washed with 750 µl 70% ethanol (centrifuge at 13,000 rpm for 5 min). Supernatant was discarded and RNA pellets were air-dried for 5 min, resuspended in 10 µl DEPC treated RNase-free water and stored at -80 °C.

Further depletion of Bacterial 16S, 23S rRNAs

To further remove any remaining 16S and 23S rRNA, the mRNA-ONLY™ Prokaryotic mRNA Isolation Kit (EPICENTRE Biotechnologies, Cat. Nos. MOP51010 and MOP51024) was used. The kit included a Terminator™ 5'→ 3' exonuclease that digests RNA having a 5' monophosphate. The reaction containing 2 µl mRNA-ONLY Prokaryotic 10X Reaction Buffer A, 0.5 µl RiboGuard RNase Inhibitor, 10 µl total RNA, 1 µl Terminator Exonuclease and 6.5 µl RNase-free water was incubated at 30 °C for 60 min and terminated by phenol extraction and ethanol precipitation. Specifically, RNase-free water was added to a total volume of 200 µl and then extracted once with buffer-saturated phenol. To the aqueous phase, 0.1 volume of 3 M sodium acetate and 2.5 volumes of ethanol were added and mixed thoroughly. The precipitation took overnight at -80 °C, then centrifuged at 14,000 rpm for 30 min at 4°C. The RNA pellet was washed with 70% ethanol and resuspended in 5 µl of RNase free water.

Amplifying subtracted RNAs

The subtracted RNA remaining from the above four kits were amplified using the MessageAmp™ II-Bacteria Kit (Ambion Part No. AM1790) which employed an in vitro transcription (IVT)-mediated linear amplification method. The 5 µl subtracted RNA was first incubated at 70 °C for 10 min, then placed on ice for 3 min. 5 µl Polyadenylation Master Mix (1.5 µl Nuclease-free water, 1.0 µl 10X Poly(A) Tailing Buffer, 1.0 µl RNase Inhibitor, 0.5 µl Poly(A) Tailing ATP and 1.0 µl Poly(A) Polymerase) was added and incubated at 37 °C for 15 min, then placed on ice. Next, 10 µl Reverse Transcription Master Mix (3 µl Nuclease-free water, 1 µl T7 Oligo(dT) VN, 1 µl 10X First Strand Buffer, 4 µl dNTP Mix, 1 µl ArrayScript) was added to synthesize first strand cDNA. Reaction was incubated at 42 °C for 2 hr. To synthesize second strand cDNA, 80 µl Second Strand Master Mix (63 µl Nuclease-free water, 10 µl 10X Second Strand Buffer, 4 µl dNTP Mix, 2 µl DNA Polymerase, 1 µl RNase H) was added, incubated at 16 °C for 2 hr, then placed on ice briefly. cDNA purification involved adding 250 µl cDNA Binding Buffer and passing mixture through a cDNA Filter Cartridge which was then washed with 500 µl Wash Buffer. Purified cDNA was eluted with 18 µl 55 °C Nuclease-free water. Antisense RNA (aRNA) was in vitro amplified by adding 24 µl IVT Master Mix (4 µl each of T7 ATP, T7 CTP, T7 GTP, T7 UTP, T7 10X Reaction Buffer and T7 Enzyme Mix), then incubating at 37 °C for 14 hr. 60 µl Nuclease-free water was added to

bring each sample to 100 µl. aRNA was purified by adding 350 µl aRNA Binding Buffer and mix, then adding 250 µl 100% ethanol and pipette thrice to mix. The mixture was passed through aRNA Filter Cartridge, washed with 650 µl Wash Buffer and eluted with 150 µl preheated Nuclease-free water. The amount of amplified RNA is reflected in Appendix C.

2.3 Illumina library preparation and sequencing

Libraries suitable for paired-end sequencing using the Illumina Genome Analyzer (Illumina, Inc.) were generated using a modified version of the standard Illumina GA protocol. cDNA fragments were blunt-ended, 'A'-tailed and ligated with "T" nucleotide overhang Illumina forked paired-end sequencing adapters (Illumina, Inc.) which contained custom barcodes for multiplex sequencing.

cDNA synthesis of amplified-subtracted RNAs

Amplified RNA was ethanol precipitated and resuspended in 9 µl DEPC-treated water. SuperScript® Double-Stranded cDNA Synthesis Kit (Invitrogen Catalog No. 11917-020) was used for cDNA synthesis. To 9 µl RNA, 2 µl Random Primer 6 (an oligo(dT) containing primer, NEB Catalog No. S1230S) was added and incubated at 70 °C for 10 min, then quick-chilled on ice. Next, 7 µl First Strand Master Mix (4 µl 5X First-Strand Reaction Buffer [250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$], 2 µl 0.1 M DTT, 1 µl 25 mM dNTP mix) was added, vortexed and centrifuged down briefly. The mixture was incubated at 25 °C for 3 min, then 42 °C for 2 min. 2 µl SuperScript® II RT (200 U/µl) was added, then incubated at 42 °C for 60 min, 70 °C for 15 min and held at 4 °C. Reaction was terminated by quick-chilling on ice. For the second strand synthesis, 130 µl master mix (92 µl DEPC water, 30 µl 5X Second-Strand Reaction Buffer [100 mM Tris-HCl (pH 6.9), 450 mM KCl, 23 mM MgCl2, 0.75 mM β-NAD+, 50 mM $(NH_4)_2SO_4$], 2 µl 25 mM dNTP mix, 1 µl E.coli DNA Ligase (10U/µl), 4 µl E.coli DNA Polymerase I (10U/µl, 1 µl E.coli RNase H) was added. The mixture was incubated at 16 °C for 2 hr. Then, 3 µl T4 DNA Polymerase (5U/µl) was added and incubated at 16 °C for 5 min. The mixture was quick-chilled on ice and 10 µl 0.5M EDTA was added. Chloroform precipitation was done next. 160 µl of phenol: chloroform:isoamyl alcohol (25:24:1) was added and vortexed thoroughly. After centrifugation at 14,000 x g for 5 min, about 140 µl upper aqueous layer was transferred to a new tube. 70 µl 7.5 M ammonium acetate and 500 µl 4 °C 100% ethanol were added and

vortexed thoroughly. The mixture was centrifuged at room temperature, 14,000 x g for 20 min and supernatant was discarded. The pellet was washed with 500 µl 4 °C 70% ethanol by centrifuging at 14,000 x g for 5 min and discarding the supernatant. cDNA pellet was air-dried for 10 min and re-suspended in 50 µl DEPC water.

Illumina library preparation

50 µl cDNA (~5 µg) was sheared using Adaptive Focused Acoustic technology on a Bioruptor®(Diagenode, Inc.) to generate fragments of 100-300bp in length. Sonication was carried out at Medium setting for 16 cycles of 30 sec burst and 30 sec pause. The bioruptor was filled with ice cold water and changed every 6 cycles in order to keep the environment cold.

End-repair was done with Quick Blunting™ Kit (NEB Catalog No. E1201L). To 50 µl sheared cDNA, 6.5 µl 10X Blunting Buffer (1X Blunting Buffer: 100 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 0.025% Triton X-100, 5 mM dithiothreitol, pH 7.5 at 25°C), 1 mM dNTP mix and 2.5 µl Blunting Enzyme mix (100 mM KCl, 10 mM Tris-HCl (pH 7.4), 0.1 mM EDTA, 1 mM dithiothreitol, 0.1% Triton X-100 and 50% Glycerol) were added. The mixture was incubated at room temperature for 30 min. The blunting reaction was purified using MinElute Reaction Cleanup Kit (Qiagen Catalog No. 28204) and eluted in 32 µl Elution Buffer.

An 'A'- base was added to the 32 µl blunt ends by adding 5 µl Klenow Buffer (NEBuffer 2 (10X)), 10 µl 1 mM dATP and 3 µl DNA Polymerase I, Large (Klenow) Fragment (NEB Catalog No. M0210S). The mixture was mixed well and incubated at 37 °C for 30 min. The mixture was then purified using MinElute Reaction Cleanup Kit (Qiagen Catalog No. 28204) and eluted in 10 µl Elution Buffer.

The adaptors were ordered as ss DNA and were ligated to the end-repaired "A"-tailed cDNAs. Six unique Adaptor mixes were prepared by mixing 50 µl Adaptor PEx.1 (100 µM), 50 µl Adaptor PEx.2 (100 µM), 20 µl Oligo Hybridization 10X Buffer [500 mM NaCl, 10 mM Tris-Cl pH 8.0, 1 mM EDTA pH 8.0] and 80 µl dH_2O. As each sample was uniquely barcoded with a distinct 6-nucleotide sequence, six pairs of adaptors were added to each reaction individually (Appendix D). The Adaptor mixes were incubated at 95 °C for 2 min

and then cooled to room temperature. To 10 µl of end-repaired "A"-tailed cDNAs, 15 µl 2X Quick Ligase Buffer [132 mM Tris-HCL, 20 mM MgCl2, 2 mM dithiothreitol, 2 mM ATP, 15% Polyethylene glycol (PEG 6000), pH 7.6 @25°C], 3 µl prepared Adaptor mix and 2 µl Quick Ligase (NEB Quick Ligase Kit, Catalog No. M2200L) were added and incubated at room temperature for 15 min. The mixture was then purified using QIAquick PCR Purification Kit (Qiagen Catalog No. 28106) and eluted in 30 µl Elution Buffer.

The ligated products were purified on a gel such that unligated adaptors and self-ligated adaptors were removed. A size-range of sequencing library appropriate for cluster generation was selected as well. 12 µl SYBR safe DNA stain was used instead of EtBr to visualize gel (120 ml, 1.7% agarose gel with distilled water and 1XTAE) under blue light. A gap of one empty lane was left between samples to prevent cross-contamination. 5 µl Orange loading dye (6X) was added and electrophoresis was run at 120V for 30 min. For each sample, a 2mm gel was cut at around 350 bp with a new UV-sterilized scalpel. Gel was transferred to a new tube, purified using NucleoSpin® Extract II Kit (MACHEREY-NAGEL Catalog No. 740609.50) and then eluted in 20 µl.

PCR was used to selectively enrich cDNA fragments that have adapter molecules on both ends as only these were able to attach to flow cells and generate clusters. The ligated products were PCR-amplified using 1 µl PE PCR1 and PE PCR2 primers (sequences in Appendix D). The PCR conditions were 95 °C for 3 min, 15 cycles of [95 °C for 30 sec, 65 °C for 30 sec and 72 °C for 30 sec], 72 °C for 5 min and held at 4 °C. The PCR amplified products were ran on a gel, gel purified using the same protocol as above and diluted to 2 ng/µl.

To ensure that multiplex sequencing was not biased towards any sample, a Real-time PCR in duplicate was carried out to find out the copy number or concentration of each sample which was essential for determining the amount of each sample required when samples were mixed in a multiplex sequencing. The primers (one pair per sample) targeted at the adaptor sequences which contained the unique barcodes for each sample. The reaction contained 7.5 µl 2X Master mix, 1 µl Primer mix (10 µM), 2 µl cDNA (2 ng/µl) and 4.5 µl dH_2O. The cycling conditions were 95 °C for 10 min followed by 40 cycles of [95 °C for 10 sec, 60 °C for 30 sec, 72 °C for 30 sec].

Illumina sequencing

Illumina sequencing was carried out on the six libraries using PE SEQ1 and PE SEQ2 primers (sequence in Appendix D), by personnel from the MIT BioMicro Center.

2.4 Quality control on Illumina reads

After sequencing, processing of this data needs to done to increase the quality of the sequence data. Post-processing of Illumina reads are carried out with a suite of in-house Perl scripts, with changes made to scripts written by Brian Knaus (Appendix A). The general processing was as follows: based on quality score information encoded in the original Illumina output "fastq" file (ie. located at the last line of a four-line coded read), poor quality Illumina reads were removed (Fig. 7). Any bases appearing after a "B" (inclusive) are truncated as they could not be identified accurately by the Illumina processing pipeline. Since the six samples were run in the same lane, they need to be demultiplexed or sorted based on their respective 6-base barcodes. Thus, only reads corresponding to a certain sample are pulled together into a single fasta file. Duplicates or identical reads are also removed as these are likely the result of technical error. Reads shorter than 20 bases are removed as well as they could be mapped to multiple regions of genome and hence are more difficult to interpret.

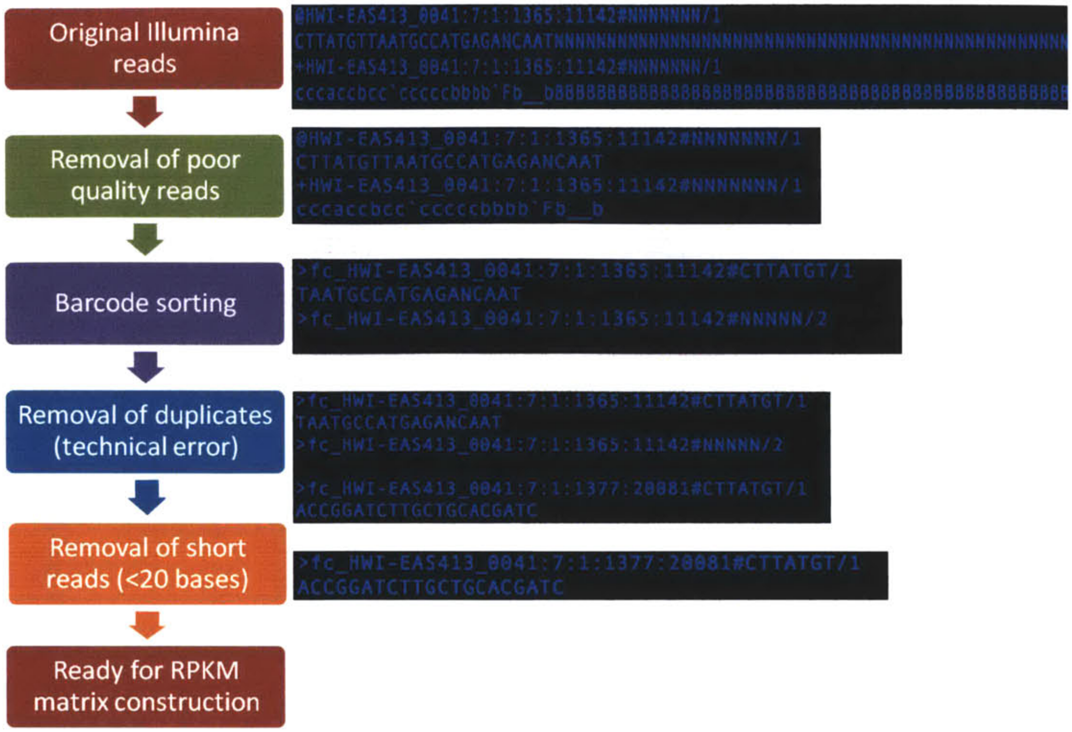


Fig. 7: Steps in processing original Illumina reads in preparation for analysis.

2.5 RPKM (Reads per kilobase of exon per million mapped reads) matrix construction

Following sequence processing, the commercial software, CLC Genomic Workbench is used to generate a matrix of transcript abundance in the samples so that the sequence information could be converted into expression data for use with MultiExperiment Viewer (MeV) for initial work (reported herein). The criteria in CLC for read assignment to the reference genome (*Microcystis aeruginosa* NIES-843) is a minimum alignment length of 90% of the bases in a read and a minimum 80% of identity in the aligned bases. Future work will be done at increasing stringencies. RPKM (Reads per kilobase of exon per million mapped reads) is the unit used for comparing transcript abundance. The formula described by Mortazavi et al. (2008) is as follows:

$$\text{RPKM} = \frac{\text{Total transcript reads for a gene}}{\text{Total mapped reads in a sample (millions)} \times \text{Gene length(KB)}}$$

RPKM is calculated by taking the total transcript reads for one gene and dividing that by both the total mapped reads of that sample (in millions) and by the exon length in kilobases. Performing this calculation for each transcript will provide expression values that are corrected for the varying numbers of reads in each data set (as they are all different and larger data sets will show more total counts) and corrected for gene lengths (as longer gene will results in more sequence counts of that gene). This was compared with the *Microcystis aeruginosa* reference genome so that only expression information from this organism or very closely related organisms was used (as the sampling was from an environment site which was not only this single organism). Hence, in this way, the transcript coverage for each gene and hence the gene expression level can be found.

2.6 *Microcystis* transcriptomic analysis

2.6.1 Grouping samples and genes by expression profiles

Hierarchical clustering was used to group both samples and genes while PCA was used to group samples only. All these three grouping methodologies were applied using MultiExperiment Viewer, MeV (Saeed et al., 2003, 2006). In MeV, the first step was to import expression data from RPKM matrix. The RPKM matrix was then log2 transformed to decrease the variance in data. Gaussian normalization across both samples and genes was done so that expression between samples and genes could be compared in the downstream analysis.

Hierarchical clustering was performed to cluster samples and genes based on similar expression patterns in an unsupervised way (i.e. without adding any human bias that would result from picking clusters manually based on what were thought to be similar). Pearson correlation and average linkage were applied as a distance measure to all the 6363 genes from the *Microcystis aeruginosa* reference genome.

PCA (Principal Component Analysis) was performed to capture the information embedded in thousands of dimensions from 6363 genes to only two dimensions for simpler visualization and interpretation of where the greatest variation came from. Samples were clustered based on medians and k-nearest neighbor (k=10).

2.6.2 Inferring biological roles -KEGG pathway enrichment of day-night samples (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined sets of genes (95 KEGG pathways for the 6363 genes in *Microcystis aeruginosa*) shows statistically significant, concordant differences between two biological states (e.g. phenotypes like day and night) (Subramanian et al., 2005; Mootha et al., 2003).

The background "Gene Set" input file is not readily available from the curated reference genome files on NCBI in the correct format required by GSEA. Therefore, it is manually

compiled by assigning all the 6363 genes to their respective KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. This results in a set of 95 KEGG pathways or "gene sets" and is used as the background reference set. All the analysis in GSEA is done with a minimum gene set size of 5 genes.

The phenotypic labeling method used in the GSEA test is categorical phenotype labeling to test for KEGG pathway enrichment of day-night differentially expressed genes. In this test, a text file which reflects the assignment of six samples to a phenotype of either day or night according to the grouping suggested by hierarchical clustering and PCA is supplied. The KEGG pathways that are significantly enriched ($p < 0.05$) with genes showing differential day-night expression are identified. The dataset tested with categorical phenotype labeling consists of the whole set of 6363 genes from the *Microcystis aeruginosa* reference genome.

This GSEA test uses a ranking metric of "signal to noise ratio" to measure a gene's correlation with a phenotype which results in a ranked list of all the tested genes. An enrichment score (ES) is calculated by walking down the ranked list of genes, increasing a running-sum statistic when a gene is present in the KEGG pathway and decreasing it when it is not. The magnitude of the increment depends on the correlation of the genes with the phenotype. A positive ES means that a KEGG pathway is enriched at the top of the ranked list and hence is a positively correlated pathway with the query gene. A negative ES means that a KEGG pathway is enriched at the bottom of the ranked list and is thus a negatively correlated pathway with the query gene.

2.6.3 Statistical tests for differential gene expression

Significantly differentially expressed genes are found via two statistical tests: T-test and SAM (Significance Analysis for Microarrays) A null hypothesis of no difference in gene expression under night and day conditions is tested with a two-class, unpaired T-test in MeV. Standard Bon-ferroni, adjusted Bon-ferroni and FDR tests are used to correct for multiple hypothesis testing. A two class unpaired SAM is done with 100 iteration and K-nearest neighbor of 10 using Euclidean distance measure. FDR of 0.05 has also been applied to SAM.

2.7 Whole community transcriptomic analysis

2.7.1 MG-RAST

Barcoded pair-end reads in all the six samples are concatenated into one file, compressed and submitted to the MG-RAST (Metagenomics- Rapid Annotation using Subsystem Technology) webserver (Meyer et al., 2008). MG-RAST carried out its own pipeline of quality control steps to remove bad quality reads and short reads via "sequence filtering". Identical reads are also removed via "dereplication". The parameters for dynamic trimming are set to default: i.e. quality threshold for low-quality bases =15, max number of low quality bases per read =5.

The taxonomic composition of the six samples is analyzed with the MG-RAST server using similarity to a large non-redundant protein database (M5NR, M5 Non-redundant protein database) and RDP (Ribosomal Database Project). Using M5NR, the affinities of reads for known metabolic function against both SEED subsystems and KEGG metabolic pathways are also tested. An E-value cutoff of < 1 has been used for initial taxonomic assignments and E-values from 1 to 1E-5 have been evaluated for trade-offs between the accuracy and coverage of gene annotation. Community domain distribution is found based on the taxonomic assignments of all the post-quality control reads from the all the annotation sources available in the MG-RAST server.

2.7.2 MEGAN

MEGAN (MEtaGenome ANalyzer) is a computer program that allows laptop analysis of large metagenomic data sets (Huson et al., 2007). In the preprocessing step, post-quality reads from in-house pipeline are blasted against the NCBI non-redundant database using BLASTX on the Darwin server. The parameters used in BLASTX are described in the Cluster_Blastx.csh file (Appendix B), as suggested by MEGAN manual http://ab.inf.uni-tuebingen.de/software/megan/how-to-use-blast/ and ftp://ftp.ncbi.nih.gov/blast/documents/blastall.htm. Specifically, the soft filtering strategy is used as it ensures that high scoring pairs (HSPs) containing low complexity regions will not break apart. The translation table used for the query sequence is code table 11 for Bacterial sequences that provides some alternative start codons. The word size is reduced to the minimum of 2 and the neighborhood word threshold score is reduced to 8 for BLASTX as suggested. The expectation value is set to 100, as suggested by the MEGAN manual that when comparing against large databases like NT or NR, such high amounts of expected random hits have to be accepted. Post-QC reads from each sample are split into 10 files for easy processing by BLASTX, and the results from 10 files are later concatenated into one file to be open and analyzed in MEGAN.

To compute and explore the taxonomical content of the data set, results from BLASTX are imported into MEGAN such that MEGAN employs the NCBI taxonomy to summarize and order the results. LCA (Lowest Common Ancestor) algorithm explicitly assigns every individual read, for which database hits are available, to some taxon in the NCBI taxonomy, regardless of the reads' suitability as a phylogenetic marker. Thus, MEGAN is primarily used to annotate reads either in terms of KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways or SEED (Subsystems Technology) hierarchy of functional classification.

MEGAN reports which KEGG pathways are present in a dataset as it captures all reads that are mapped to a given pathway of interest. Each read is mapped to a KEGG Orthology (KO) accession number, using best hit to reference sequences for which KO accession numbers are known. MEGAN then calculates the number of hits to each KEGG pathway.

Chapter 3: Results

3.1 Sampling site

Sampling was done in Kranji Reservoir, located in the northwest part of Singapore (1°25'N, 103°43'E, Fig. 6) (NTU, 2008). Six samples (krb1 – krb6) were collected at 5pm and 8pm of 14th January 2010 and 6am, 8am, 1pm and 3pm of 15th January 2010, respectively. The weather on these two days was tropical (humid and warm). A Cyanobacterial bloom was evident, as seen from a dense layer of scum on the water surface (Fig. 8).

Nutrient levels and dissolved microcystin concentrations were relatively constant over the experiment period (Fig. 9). The Total Kjeldahl Nitrogen (TKN) ranged from 1.01- 1.36 mg/L. The Total Phosphorus (TP) ranged from 0.086 - 0.1 mg/L. The Dissolved Organic Carbon (DOC) ranged from 5.48 – 7.74mg/L. Waterborne microcystin concentrations ranged from 0.3112 to 0.4888 ppb although cell-associated microcystin was not determined in this study and is likely to be higher. Light intensity varied from <0.01 lux at night to 21,670 lux during the day. The first water sample was collected at 5pm and was filtered within 1.5 hours of collection, which represented a "transition" between light and dark conditions. Subsequent water samples were filtered and immediately immersed in RNA*later* within 1 hour of water collection. Temperatures varied during the sampling period from 22.5 to 32.1°C with highest temperatures observed during the day (Fig. 9). Multivariate analysis of community gene expression profiles (described in section 3.4.1) revealed that samples from 5pm, 8pm and 6am shared the highest degree of similarity and are thus referred to as the "night" samples while samples from 8am, 1pm and 3pm shared highest similarity and are thus referred to as the "day" samples.

Fig. 8: A dense scum of Cyanobacteria was seen at the sampling site in Kranji Reservoir, Singapore, which was in the midst of a bloom event.

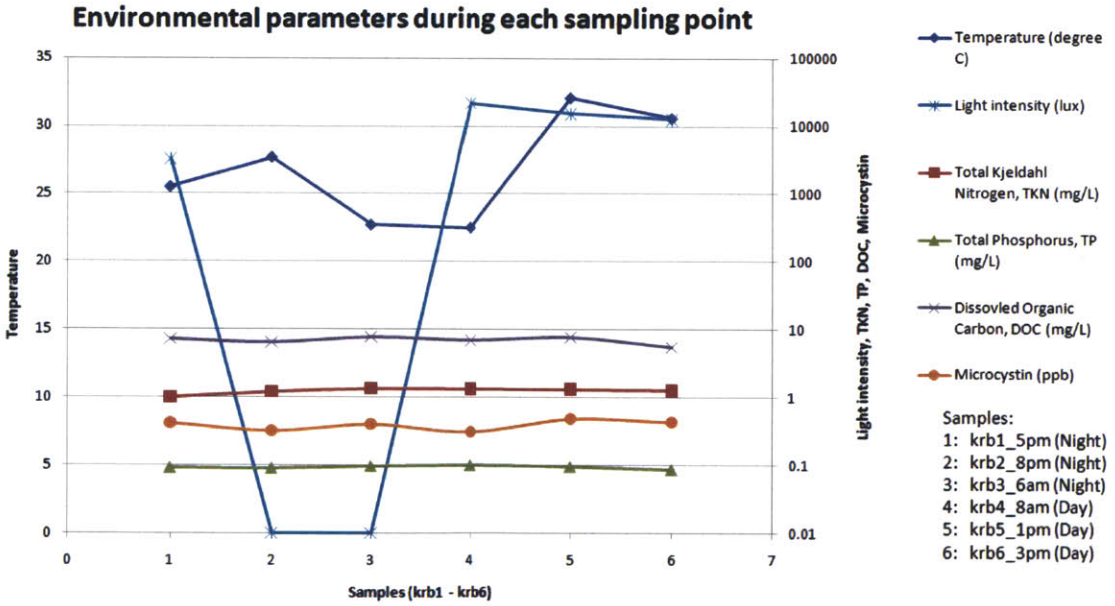


Fig. 9: Environmental parameters (Temperature, Total Kjehldahl Nitrogen, Total Phosphate, Dissolved Organic Carbon, Microcystin concentration and Light intensity) at the six sampling time points.

3.2 Sequencing outcome

Illumina sequencing of Kranji Reservoir bacterioplankton transcriptomes (samples krb1-krb6) yielded 8.24E5 to 1.31E6 reads per sample (Table 1). Poor quality reads were removed by demultiplexing, dereplication, removal of short reads and reads with low fastq quality scores. Quality control was performed using in-house Perl scripts (Appendix A) for all analysis based on MEGAN, while built-in pipelines for CLC Genomics workbench or the MG-RAST server were used for analyses based on these respective pipelines. Fig. 10 reflects that the number of reads that survive each quality-control (QC) step in read processing via our in-house scripts and a comparison of the three different quality control checks reveals similar performance of all methods across samples (Table 1).

Sequence statistics from MG-RAST (Table 2) show that the post-QC reads range from 91±19 to 94±16 bp in mean length for the six samples. Their mean GC content ranges from 47±9 to 51±11%. The distribution of Post-QC sequence lengths (bp) for the six samples are shown in Fig. 11. The peaks in the histograms (Fig. 11) reflect that the majority of the reads are in the 95-101 length range. In fact, 73.7% - 83.4% of the Post-QC reads in the six samples are between 95-101 bp long (Table 3).

In MG-Rast, reads are annotated by blasting them against a variety of public sequence databases (GenBank, IMG, KEGG, PATRIC, RefSeq, SEED, SwissProt, TrEMBL, eggnog, COG, KO, NOG, Subsystems, Greengenes, LSU, RDP and SSU). Fig. 12 shows the number of hits from each annotation source. RefSeq always give the largest number of annotation hits for all the six samples. In contrast, there are a low number of hits from ribosomal databases such as Greengenes, LSU, RDP and SSU (0.47 – 1.92% hits out of all the hits in all databases) indicating successful depletion of ribosomal sequences (16S rRNA, 18S rRNA, 23S rRNA, 28S rRNA etc.) that can represent 53% - over 90% of total sequences without depletion (Frias-Lopez et al., 2008, Stewart FJ et al., 2010).

Blasting the reads against all these databases has the advantage over blasting against a single database as this would mean that more reads could be annotated and hence a more extensive taxonomical and functional analysis of the community can be carried out. MG-RAST

combines the above databases into a single integrated database, named "M5NR", which will be the default annotation choice for all the MG-RAST analyses, unless otherwise stated.

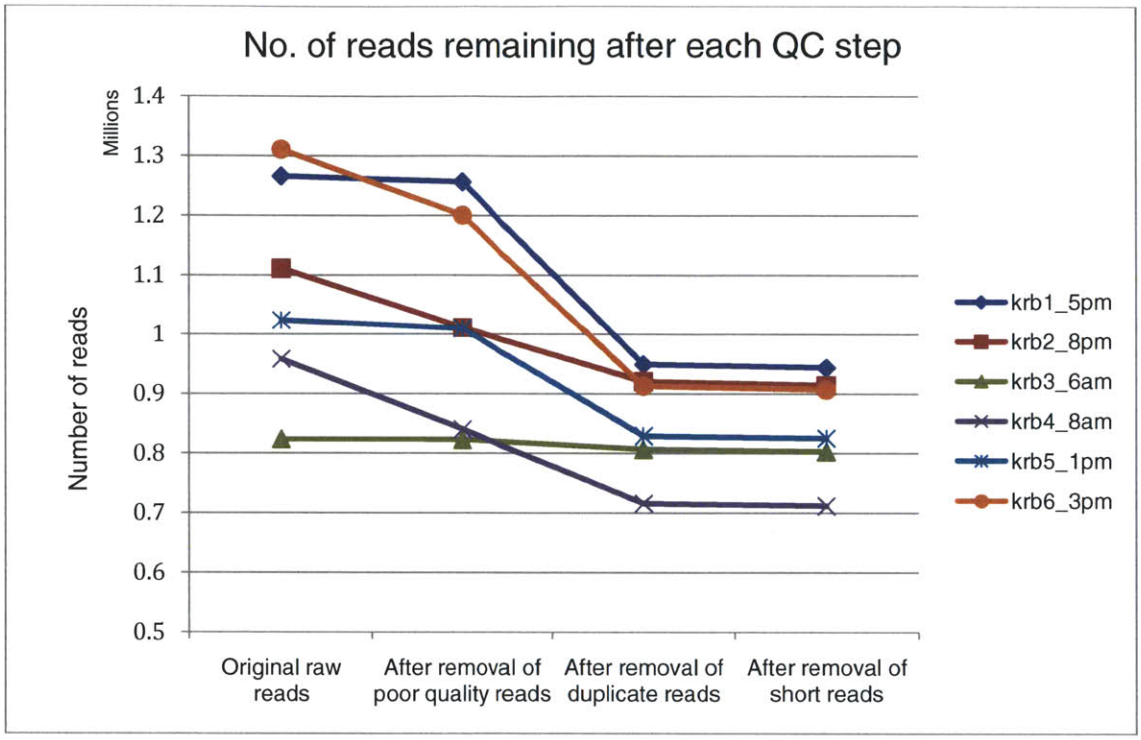


Fig. 10: Number of reads remaining after each quality control step in the in-house pipeline for quality control. Perl scripts used for each step are included in Appendix A.

Table 1: Pre- and Post-Quality Control reads produced by the different quality control (QC) pipelines employed in this study.

Sample	No. of reads (Pre-QC)	No. of reads (Post-QC)		
		In-house pipeline	MG-RAST built-in pipeline	CLC built-in pipeline
Krb1	1,265,955	944,045	894,579	977,552
Krb2	1,111,166	914,651	873,123	922,106
Krb3	824,021	803,400	785,797	743,396
Krb4	958,367	711,948	654,350	789,870
Krb5	1,023,722	825,565	746,516	947,426
Krb6	1,311,237	907,084	777,920	1,084,514

Table 2: Sequence statistics from MG-RAST.

Sample	Pre-QC		Post-QC	
	Mean sequence length (bp)	Mean GC content (%)	Mean sequence length (bp)	Mean GC content (%)
Krb1	100±0	50±11	91±19	49±12
Krb2	100±0	51±10	91±19	50±12
Krb3	100±0	48±10	94±15	47±10
Krb4	100±0	52±10	92±18	51±11
Krb5	100±0	48±8	94±16	47±9
Krb6	100±0	50±8	92±19	49±10

Table 3: Percentages of Post-QC reads that are 95-101bp in length (MG-RAST QC pipeline).

No. of reads	krb1	krb2	krb3	krb4	krb5	krb6
Total	894,579	873,123	785,797	654,350	746,516	777,920
95-100bp	365,327	352,228	347,366	264,724	328,397	321,669
100-101bp	312,737	291,356	303,105	228,744	293,836	277,823
% of 95-101bp reads	75.8	73.7	82.8	75.4	83.4	77.1

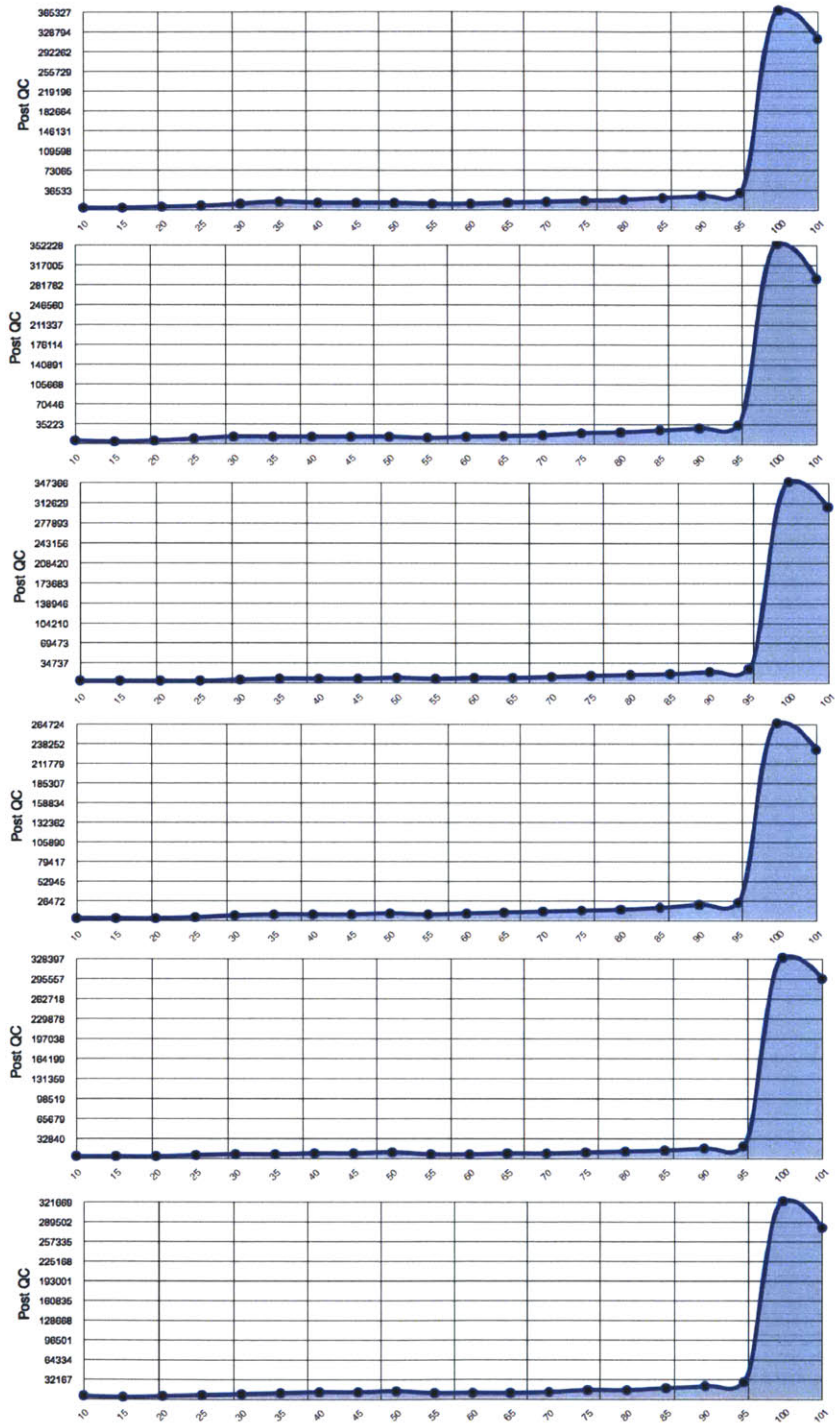


Fig. 11: Histograms showing the distribution of Post-QC sequence lengths (bp) for the six samples after processing in MG-RAST. Each position represents the number of sequences within a length range.

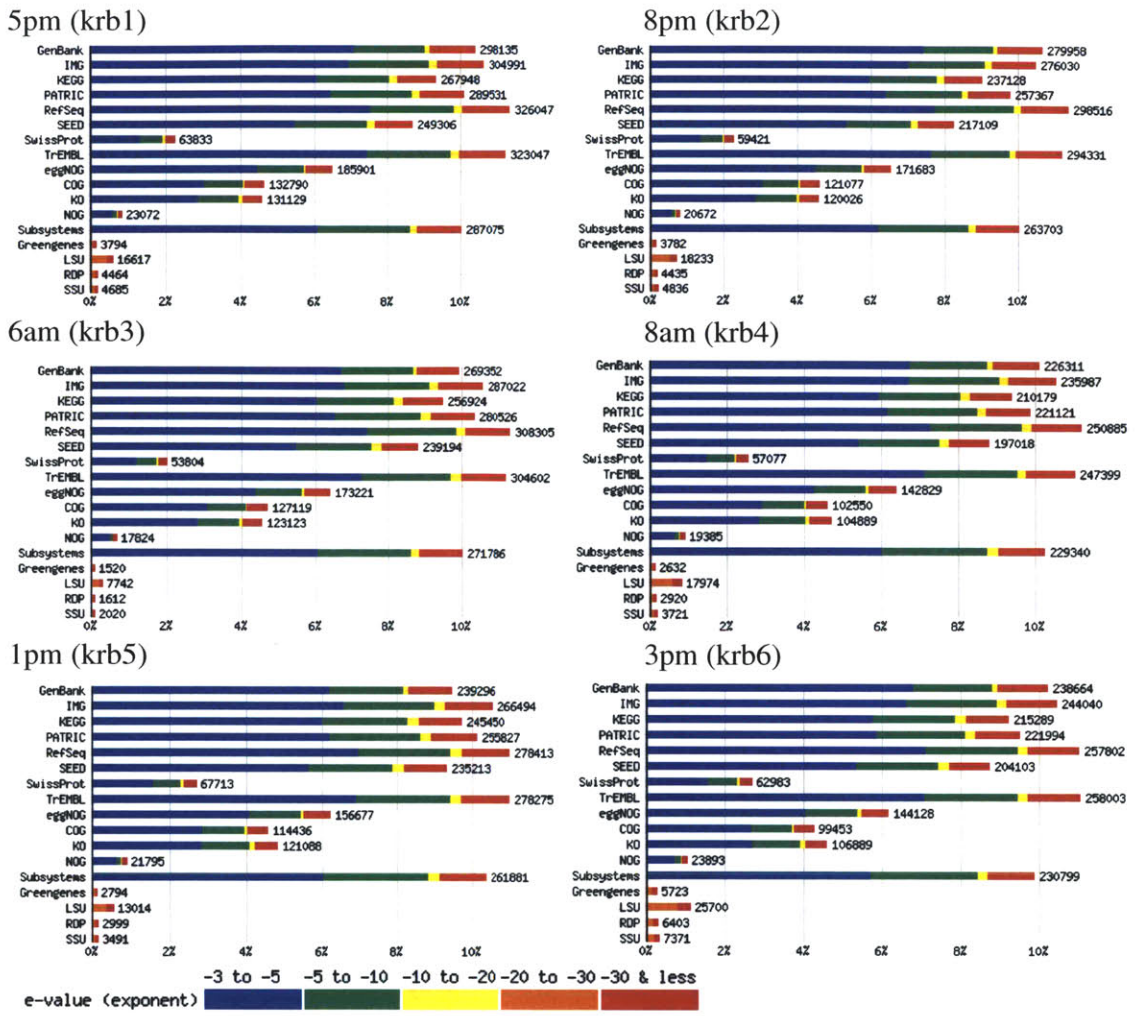


Fig. 12: Number of hits in a variety of annotation sources for the MG-RAST Post-QC reads in six samples. The hits are broken down by e-value ranges. Each bar segment represents a percentage of the total this across all sources, allowing for cross comparison of hits within different sources and e-value ranges.

3.3 Analysis of Kranji Reservoir plankton community structure

3.3.1 Community taxonomic composition by domain and phylum

The majority of transcripts from all Kranji Reservoir samples were identified as Bacterial by annotation with the MG-RAST server (86.55 - 90.36%). Eukaryotic transcripts represented 8.98 - 12.79%, followed by Archaeal transcripts 0.34 - 0.74% and viral transcripts 0.14 - 0.24% (Fig. 13). The community composition distribution by domain is relatively constant in the six water samples taken over the day/night cycle (Fig. 13).

Phylum-level assignment of transcripts based on classification by SEED subsystems and comparison to the M5NR protein database, both implemented via the MG-RAST server revealed that Bacteria, Eukaryote and Archaea domains included 28, 47-48 and 6 phyla respectively (Fig. 14). The relative abundance of transcripts from different phyla within each sample was estimated based on homology to sequences in all available annotation databases accessed through the MG-RAST pipeline and revealed a predominance of Cyanobacteria-like transcripts followed by Proteobacteria-, Actinobacteria-, Bacteroides-, and Firmicutes-like transcripts (Fig. 16).

Cyanobacteria are the dominant phylum based on the number of Illumina reads that have hits in the M5NR protein database (Fig. 15). At any point of time during the day, the number of transcripts of the Cyanobacterial origin (47.0 – 78.4 %) is the highest among all the phyla. The next phylum with the highest amount of transcript hits is Proteobacteria (10.7 – 23.2 %), followed by either Firmicutes (3.9 – 8.1 %) or Actinobacteria (2.7 – 8.8 %) and finally Bacteroidetes (1.6 – 6.3 %).

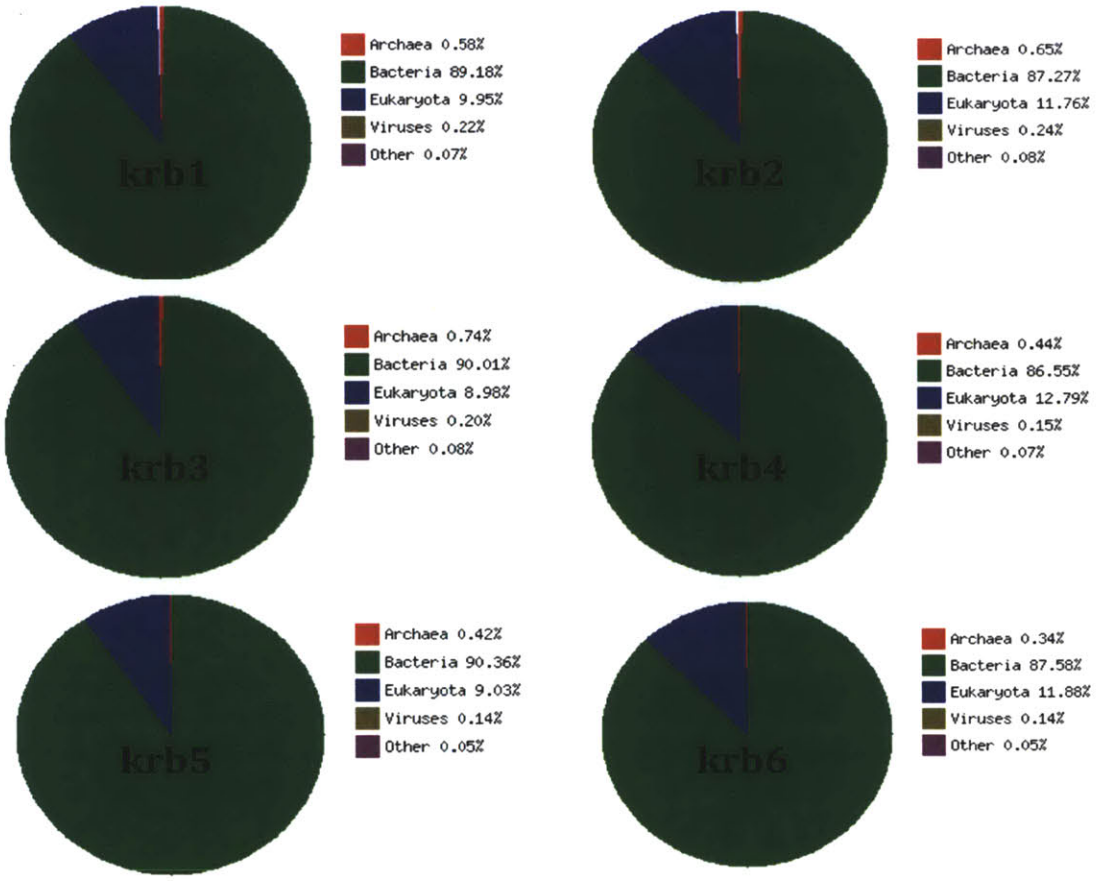


Fig. 13: Domain-level affiliation of transcripts based on MG-RAST assignment (E < 1). The pie charts are based on the combined taxonomic domain information of all the annotation source databases available in the MG-RAST server. The percentages are based only on assigned and classified reads. Transcriptomes Krb1-6 correspond to samples collected at 5pm, 8pm, 6am, 8am, 1pm and 3pm, respectively.

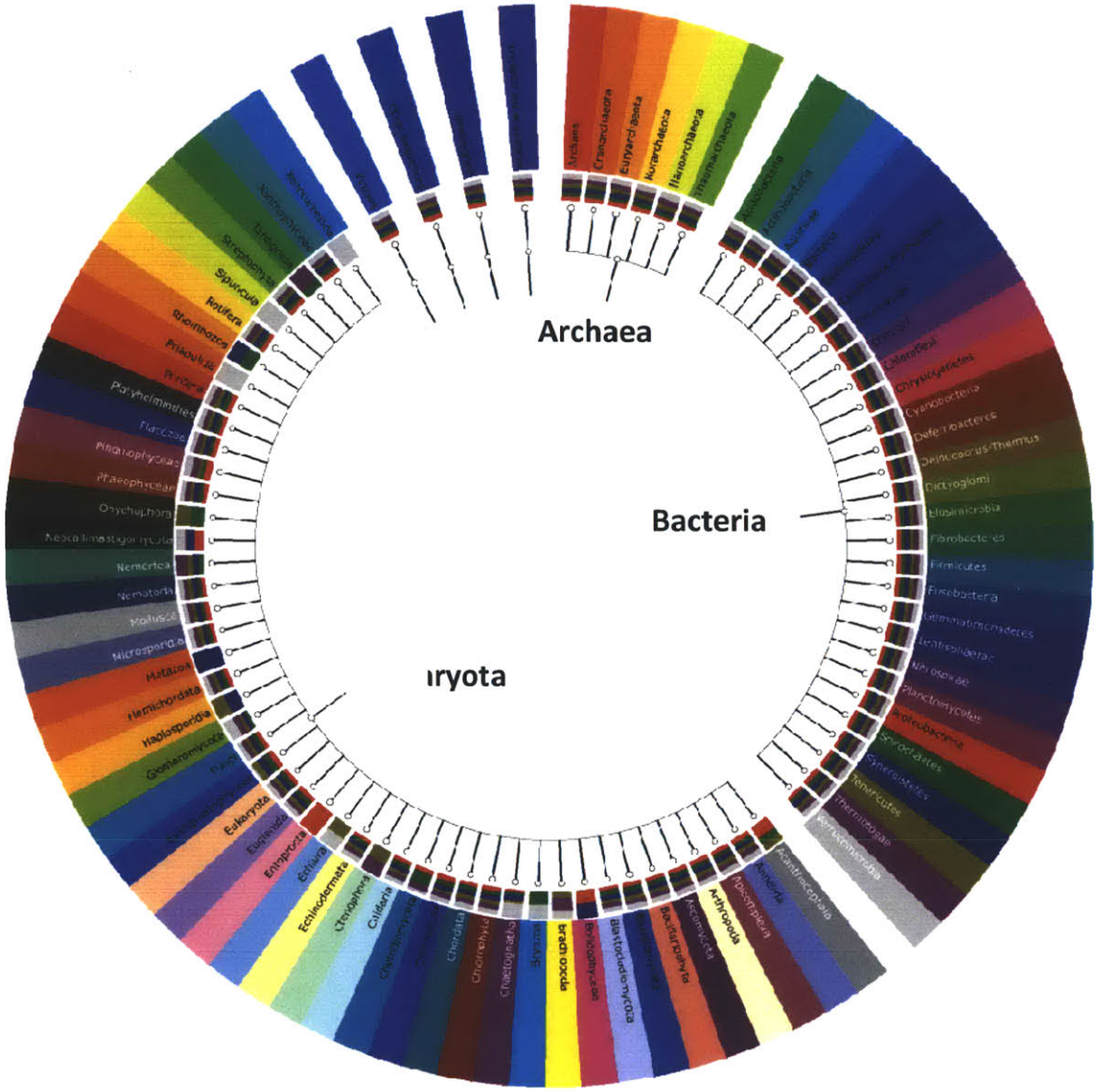

Archaea
Bacteria
ıryota
Arthropoda
Chordata
Echinodermata
Mollusca
Nematoda
Platyhelminthes
Annelida
Cnidaria
Porifera
Ascomycota
Basidiomycota
Euglenida
Ciliophora
Rotifera
Bacteroidetes
Proteobacteria
Cyanobacteria
Firmicutes
Actinobacteria
Chlorophyta
Rhodophyta
Phaeophyceae
Bryozoa
Brachiopoda
Hemichordata
Tardigrada
Nematomorpha
Nemertea
Chaetognatha
Echiura
Sipuncula
Entoprocta
Kinorhyncha
Loricifera
Priapulida
Gastrotricha
Cycliophora
Xenacoelomorpha
Haplosporidia
Glomeromycota
Microsporidia
Metazoa
Nematoda
Euglenozoa
Cercozoa
Amoebozoa
Fungi
Haptophyceae
Cryptophyceae
Dinophyceae
Chlorarachniophyceae
Eustigmatophyceae
Xanthophyceae
Bacillariophyta
Oomycetes
Thaumarchaeota
Euryarchaeota
Crenarchaeota
Korarchaeota
Nanoarchaeota
Thermoplasmata
Methanomicrobia
Halobacteria
Thermococci
Methanobacteria
Methanococci
Archaeoglobi
Thermoprotei
Aquificae
Thermotogae
Deinococcus-Thermus
Chloroflexi
Dictyoglomi
Elusimicrobia
Fibrobacteres
Fusobacteria
Gemmatimonadetes
Lentisphaerae
Nitrospinae
Planctomycetes
Spirochaetes
Synergistetes
Tenericutes
Verrucomicrobia
Chlamydiae
Cyanobacteria
Deferribacteres
Spirochaetes
Acidobacteria

Fig. 14: MG-RAST organism classification by SEED, at the phylum level. Different colors in the outer wheel correspond to different phyla, while the inner ring of stacked blocks shows the abundance of a particular phylum in the six samples (grey: krb1 (5pm); purple: krb2 (8pm); olive yellow: krb3 (6am); blue: krb4 (8am); green: krb5 (1pm); red: krb6 (3pm)).

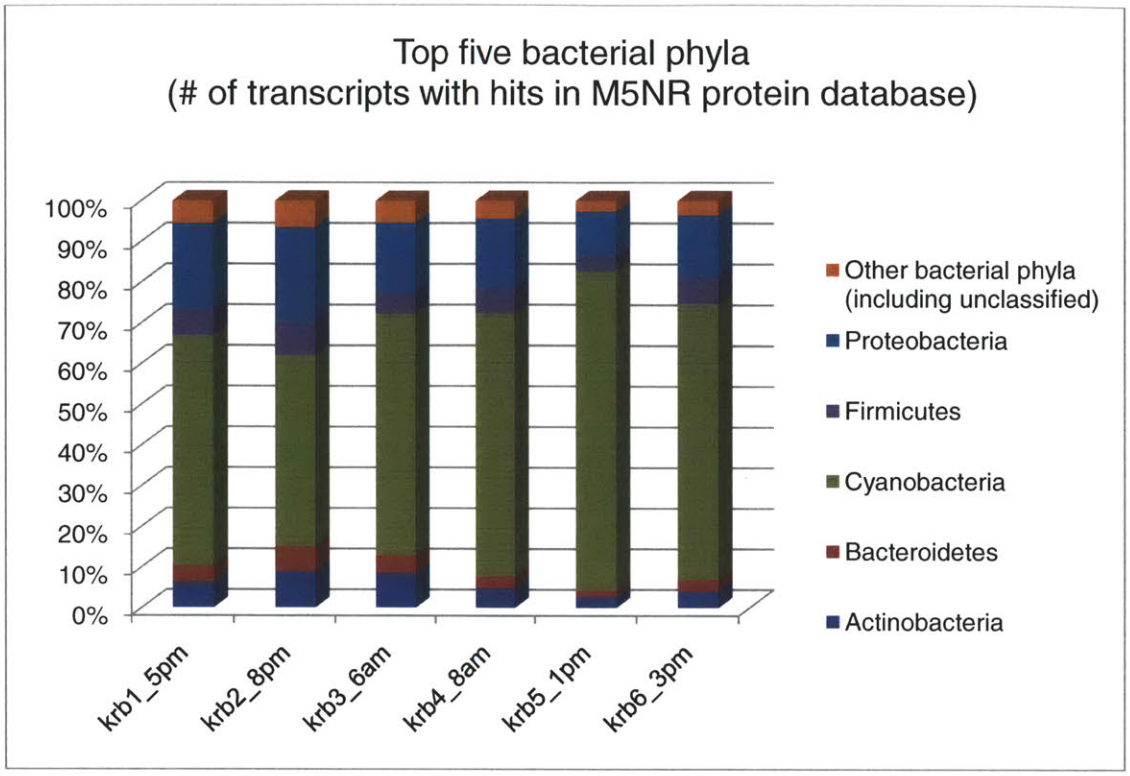


Fig. 15: Top five Bacterial phyla in each sample by the abundance of transcripts with hits in the M5NR protein database (E-value < 1).

Table 4: Community taxonomic composition similarity measure by Pearson correlation, based on protein hits in the M5NR integrated protein database for all bacteria at the phylum level.

	krb1	krb2	krb3	krb4	krb5	krb6
krb1	1.000	0.991	0.996	0.994	0.972	0.989
krb2		1.000	0.980	0.972	0.932	0.961
krb3			1.000	0.997	0.984	0.994
krb4				1.000	0.991	0.999
krb5					1.000	0.995
krb6						1.000

3.3.1.1 Bacterial/ Archaeal phyla of Kranji Reservoir

Annotation of transcripts in MEGAN (against the NCBI Blastx protein database) and MG-RAST (against the M5NR protein database, e-value < 1) revealed that the top five phyla in all samples were Bacterial, as expected due to depletion of Eukaryotic ribosomal and mRNA during preparation of Illumina libraries. They are Cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes in a decreasing order of abundance (Fig. 16). A predominance of Cyanobacterial transcripts is consistent with the observation that the sampling site was suffering from a dense Cyanobacterial bloom when the samples were collected.

The number of MG-RAST annotated transcripts for Cyanobacteria ranges from 256,103 to 609,539 in the six samples representing 47.0 to 78.4 % of the total Bacterial transcriptomes (E_{ave} ranges from 1E-3.26 to 1E-2.77, Fig. 17). Proteobacterial transcripts are second in terms of abundance, ranging from 10.7 to 23.2 % of the total Bacterial transcriptomes (E_{ave} ranges from 1E-0.78 to 1E-0.62, Fig. 17). The abundance for Firmicutes 3.9 to 8.1 % (E_{ave} ranges from 1E0.29 to 1E0.56, Fig. 17) and Actinobacteria 2.7 to 8.8 % (E_{ave} ranges from 1E-0.66 to 1E0.52, Fig. 17) are similar in order representing the third or fourth most abundant phyla depending on the sample. The abundance for Bacteroidetes transcripts ranges from 1.6 to 6.3 % (E_{ave} ranges from 1E-1.18 to 1E-0.61, Fig. 17) which place it the fifth most abundant phyla in the freshwater community transcripts. Transcripts from other phyla range from 36 to 7,453 reads per sample, representing 0.03 to 1.04 % of the total community gene expression with E_{ave} ranges from 1E-2.02 to 1E1.89. In total, they constitute < 4.7 % of Bacterial assignments. Transcripts associated with the dominant Cyanobacterial phyla are annotated with the highest confidence (i.e. lowest average E values) of all phyla.

Only five Archaeal phyla are represented in transcript annotations and correspond to a minor portion of Kranji Reservoir plankton transcripts. The most abundant Archaeal phylum is Euryarchaeota (75.0 to 77.2 %; E_{ave} ranges from 1E1.07 to 1E1.2, Fig. 18) followed by Crenarchaeota (17.0 to 20.9 %; E_{ave} ranges from 1E1.74 to 1E2.13, Fig. 18). The rest of the phyla constitutes < 4.5 % of Archaeal assignments.

Relative abundance found by blasting reads against the SEED subsystem database and the M5NR database (cutoff E-value < 1) shows that the proportions of transcripts from the different Bacterial and Archaeal phyla remains relatively constant over the course of the day-night cycle (Fig. 14, 17, 18). This shows that the prokaryotic community is relatively stable over the course of the day since each phylum is equally represented by the six samples taken over a day/night cycle.

The Bacterial community in Kranji Reservoir is highly skewed towards Cyanobacteria as indicated by the slopes of the six rank abundance curves (Fig. 16). A steep gradient indicates low evenness as the high ranking phyla have much higher abundances than the low ranking phyla. A shallow gradient indicates high evenness as the abundances of different phyla are similar. Although Fig. 16 is in logarithmic scale, it is still easy to discern that each sample has a highly uneven Bacterial community, skewed by the highly abundant Cyanobacteria.

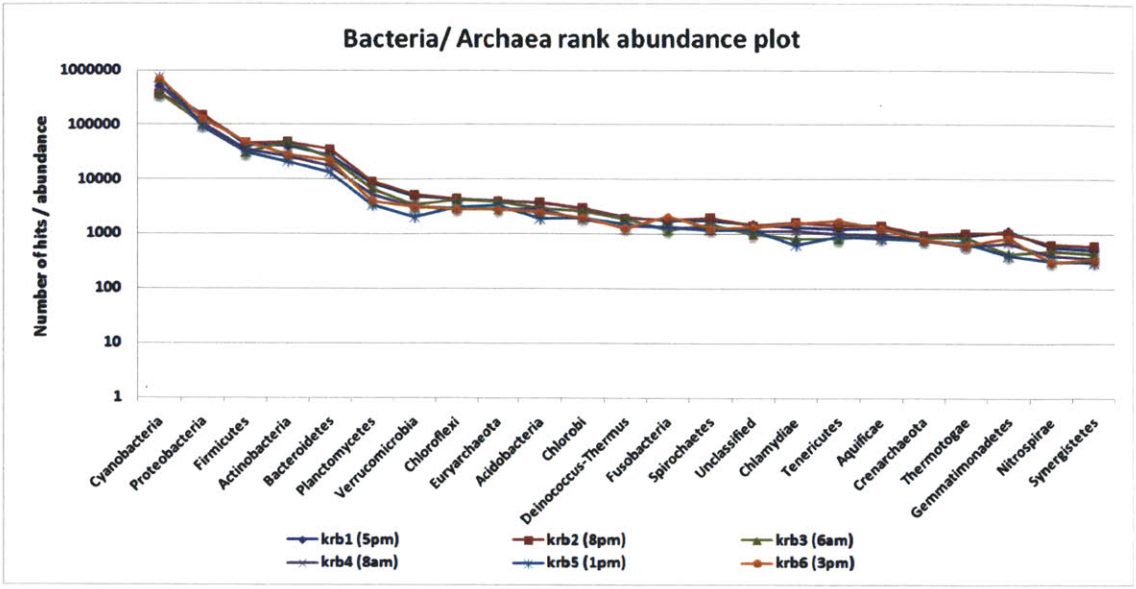


Fig. 16: Bacteria/ Archaea rank abundance plot of the six samples in Kranji Reservoir transcriptomes annotated with total database in MG-RAST. The vertical axis represents the number of hits in all the available databases on a logarithmic scale.

Acidobacteria | (-0.29|80.69|19.67) (-0.23|79.6|19.75) (-0.27|80.17|19.82) (-0.05|80.23|19.27) (-0.39|80.36|20.08) (-0.2|78.95|20.01)
Actinobacteria | (0.19|81.55|18.95) (-0.44|82.07|19.87) (-0.66|81.4|20.38) (0.3|81.97|18.64) (0.52|82.35|18.21) (0.32|82.23|18.59)
Aquificae | (0.13|79.14|19.54) (-0.1|81.7|19.04) (0.45|79.35|18.79) (0.36|78.03|19.6) (0.31|82.11|18.43) (0.64|80.31|18.39)
Bacteroidetes | (-0.67|80.59|20.39) (-1.18|80.55|21.33) (-0.87|80.1|20.87) (-0.73|81.35|20.27) (-0.61|81.36|20.14) (-0.82|81.28|20.57)
Candidatus Poribacteria | (-0.11|79.76|19.05) (1.68|90.8|14.62) (-0.48|79.09|19.77) (-0.23|81.91|19.61) (0|83.84|18.12) (-1.14|82.14|19.95)
Chlamydiae | (0.31|82.81|18.42) (-0.2|82.46|19.36) (-0.42|78.38|21.03) (-0.13|82.17|19.3) (-0.28|85.54|18.61) (0.64|82.17|17.59)
Chlorobi | (-0.7|79.35|20.79) (-1|79.59|21.25) (-0.75|80.1|20.79) (-0.51|79.88|20.28) (-0.73|79.83|20.73) (-0.78|80.91|20.46)
Chloroflexi | (-0.06|79.07|19.96) (-0.31|80.46|20.33) (-0.42|79.49|20.58) (-0.35|80.27|20.09) (-0.28|80.71|20.02) (-0.19|80.3|20.04)
Chrysiogenetes | (0.77|82.51|17.92) (0.86|85.06|17.12) (0.39|84.14|18.23) (0.01|87.12|17.96) (1.89|91.69|14.55) (1.68|90.54|15.09)
Cyanobacteria | (-3.26|88.24|23.19) (-2.77|86.43|22.74) (-3.01|86.38|23.21) (-3.09|87.02|22.98) (-3.14|87.11|23.18) (-3.24|88.38|23.17)
Deferribacteres | (-0.27|80.29|19.52) (-0.57|81.43|20.02) (0.22|80.48|18.83) (0.4|79.89|18.71) (0.19|77.88|19.86) (0.37|81.23|18.21)
Deinococcus-Thermus | (0.49|80.74|18.5) (-0.18|78.75|20.27) (0.22|78.57|19.38) (0.33|79.12|19.45) (0.56|81.62|18.09) (0.5|82.13|18.19)
Dictyoglomi | (-0.32|79.23|20.46) (-0.33|79.08|20.21) (0.02|79.3|19.52) (0.01|80.45|19.27) (-0.5|78.6|20.76) (-0.12|76.64|20.63)
Elusimicrobia | (0.09|82.62|18.52) (0|82.34|18.63) (0.83|80.26|18.14) (-0.18|85.4|18.24) (-0.12|84.97|18.25) (0.15|85.34|17.85)
Fibrobacteres | (0.58|83.2|17.88) (0.53|80.36|18.64) (0.43|82.75|17.88) (-0.17|80.72|20.24) (0.83|83.89|17.33) (1.65|81.74|16.4)
Firmicutes | (0.48|81.46|18.41) (0.3|81.86|18.64) (0.47|81.51|18.46) (0.46|81.61|18.38) (0.56|82.81|17.91) (0.29|82.72|18.48)
Fusobacteria | (0.89|81.91|17.52) (0.74|81.08|17.99) (0.92|82.1|17.49) (1.01|82.25|17.21) (0.67|81.96|18.11) (0.75|81.26|18.12)
Gemmatimonadetes | (-1.67|78.67|22.52) (-0.69|79.73|20.48) (0.63|81.67|17.88) (-1.23|82.14|20.91) (-0.07|83.89|18.67) (0.23|86.71|17.63)
Lentisphaerae | (0.55|80.41|18.4) (0.14|80.15|19.22) (0.32|80.73|18.71) (0.24|79.41|19.13) (0.85|81.9|17.53) (0.53|82.92|17.58)
Nitrospirae | (0.31|83.3|18.05) (0.14|81.97|18.49) (0.52|83.08|18.08) (0.55|81.82|17.99) (0.93|82.96|17.29) (0.65|83.72|17.44)
Planctomycetes | (-0.6|79.83|20.12) (-1.13|78.6|21.72) (-1.24|80.21|21.38) (-2.02|80.02|22.67) (-0.26|81.75|19.31) (-0.88|79.58|20.99)
Proteobacteria | (-0.78|83|20.16) (-0.62|82.53|19.99) (-0.74|82.62|20.15) (-0.73|82.53|20.12) (-0.71|83.61|19.88) (-0.66|83.64|19.76)
Spirochaetes | (0.18|81.89|18.87) (-0.18|83.17|18.84) (0.63|81.11|18.41) (0.29|83.72|18) (0.62|82.42|18.2) (0.86|81.2|17.73)
Synergistetes | (0.91|81.5|17.55) (0.78|80.49|18.13) (0.58|80.29|18.56) (0.78|81.11|17.96) (1.27|82.51|16.82) (0.43|80.29|18.95)
Tenericutes | (1.2|82.64|16.96) (0.85|81.04|17.8) (1.17|81.96|17.06) (1.28|83.63|16.47) (1.19|81.18|17.43) (1.03|82.62|17.26)
Thermotogae | (0.53|81.39|18.29) (0.3|81.58|18.36) (0.22|80.96|19.01) (0.27|82.63|18.07) (0.27|83.62|18.05) (0.17|81.96|18.6)
Verrucomicrobia | (-0.84|79.88|20.76) (-0.49|79.92|20.21) (-0.51|80.16|20.25) (-0.29|81.1|19.48) (-0.14|82.12|19.01) (-0.52|80.11|20.42)
unclassified (derived from Bacteria) | (-0.7|80.28|20.7) (-0.42|80.01|20.01) (0.02|80.9|19.26) (-0.61|78.93|20.81) (-0.4|82.85|19.41) (-0.32|81.78|19.34)

Phylum (ave. log E value | % identity | protein alignment length) for 5pm, 8pm, 6am, 8am 1pm and 3pm (krb1-6)

Fig. 17: Distribution of Bacterial phyla in Kranji Reservoir transcriptomes. Values represent percent of total Bacterial transcripts that match the indicated phylum in the M5NR database (e-value <1). krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

Crenarchaeota | (1.78|83.34|15.67) (1.74|82.77|16.13) (1.81|81.73|16.25) (2.11|81.01|16.02) (2.13|84.29|15.02) (2.05|82.86|15.64)
Euryarchaeota | (1.09|81.4|17.34) (1.08|81.21|17.29) (1.11|82.12|17.25) (1.2|81.17|17.28) (1.14|81.64|17.15) (1.07|82.52|17.16)
Korarchaeota | (2.17|86.22|14.54) (2.02|83.68|15.57) (0.92|80.02|18.59) (1.73|82.73|16.03) (2.21|91.24|13.47) (2.61|86.93|14.03)
Nanoarchaeota | (2.66|88.42|12.98) (2.22|80.23|16.52) (1.91|84.47|14.4) (-2.75|80.38|25.65) (1.3|75.94|17.99) (3.58|93.59|11.25)
Thaumarchaeota | (1.94|85.12|15.23) (1.31|80.7|17.23) (2.11|81.6|15.21) (1.59|82.16|16.35) (1.24|83.8|16.89) (1.95|81.04|15.8)
unclassified (derived from Archaea) | (1.25|81.06|16.65) (-1|80.66|20.72) (0.59|78.58|19.33) (-0.02|81.23|19.08) (-0.08|77.91|20.3) (0.96|80.79|17.5)

Phylum (ave. log E value | % identity | protein alignment length) for 5pm, 8pm, 6am, 8am 1pm and 3pm (krb1-6)

Fig. 18: Distribution of Archaeal phyla in Kranji Reservoir transcriptomes. Values represent percent of total Archaeal transcripts that match the indicated phylum in the M5NR database (e-value <1).

3.3.1.2 Eukaryota phyla of Kranji Reservoir

Eukaryotic phyla were represented in a minority of transcripts from the Kranji Reservoir samples taken in this study (8.98 - 12.79 %), as measured by the rank abundance hits in the total database in MG-RAST (Fig. 19). This can be explained in part by the strategy used to prepare mRNAs from the reservoir which included a poly-T hybridization step to remove polyA-tailed Eukaryotic mRNAs (Section 2.2). However the residual distribution of Eukaryotic mRNA in the samples provides insight into the taxa that reside in the bloom community of the Kranji Reservoir. The average E-value (E_{ave}) of transcripts assigned to Eukaryotic phyla was 1E-11 to 1E4 based on annotation to the M5NR protein database. Higher E-values (e.g. E_{ave} >1) indicate highly tentative assignments, possibly reflecting a relatively limited database of Eukaryotic genome sequences for comparison. The most abundant Eukaryotic phylum was Streptophyta (9,413 - 22,135 reads per sample or 18.1 - 28.9 % of the Eukaryotic transcriptomes with an E_{ave} ranging from 1E-1.68 to 1E-0.38, Fig. 20). Streptophyta includes several orders of green algae and land plants, consistent with previous work showing green algae are dominant phytoplankton in Kranji Reservoir (Gin et al., 2005). Transcripts from the Streptophyta represent ~27-fold fewer transcripts than the dominant Bacterial phylum, Cyanobacteria. Other Eukaryotic phyla represented in Kranji Reservoir Eukaryotic transcriptomes included Unclassified phyla (11.9 -16.2 %), Chordata (animal vertebrate and some closely related invertebrates, with an average of 11,793 reads, (E_{ave} ranging from 1E1.51 to 1E2.08, Fig. 20)); Chlorophyta (another division of green algae or aquatic photosynthetic Eukaryotes, with an average of 2,840 reads (E_{ave} ranging from 1E-2.46 to 1E-1.7, Fig. 20)); Arthropoda (invertebrates with an exoskleton, segmented body and jointed appendages, with an average of 5,855 reads (E_{ave} ranging from 1E0.55 to 1E1.72, Fig. 20)); Ascomycota (sac fungi, with an average of 7,539 reads (E_{ave} ranging from 1E1.53 to 1E1.77, Fig. 20)); Bacillariophyta (diatoms, with an average of 1,480 reads (E_{ave} ranging from 1E-5.85 to 1E-3.97, Fig. 20)); Apicomplexa (protists or parasites of animals, with an average of 3,246 reads (E_{ave} ranging from 1E0.32 to 1E1.16, Fig. 20)); Nematoda (roundworms, with an average of 2,141 reads (E_{ave} ranging from 1E0.67 to 1E1.64, Fig. 20)); Rotifera (wheel microscopic animals common in freshwater environment, with an average of 11 reads (E_{ave} ranging from 1E-0.71 to 1E2.17, Fig. 20)); Basidiomycota (bracket fungi and mushrooms, with an average of 1,617 reads (E_{ave} ranging from 1E1.28 to 1E1.94, Fig. 20)); Cnidaria (sessil Anthozoa like sea anemones, corals, jellyfish and freshwater Hydrozoa or

hydra-like animals, with an average of 1,540 reads (E_{ave} ranging from 1E-0.84 to 1E1.45, Fig. 20)); Mollusca (Gastropods like snails and slugs, with an average of 75 reads (E_{ave} ranging from 1E-0.74 to 1E1.48, Fig. 20)); Phaeophyceae (brown algae, with an average of 619 reads (E_{ave} ranging from 1E-4.33 to 1E-3.22, Fig. 20)); and lastly Platyhelminthes (flatworms, with an average of 482 reads (E_{ave} ranging from 1E1.06 to 1E2.1, Fig. 20)). The rest of the phyla constitute < 13.6 % of Eukaryotic assignments.

While the Bacterial/Archeal phyla are present in similar proportion across samples, the Eukaryotic phyla are more patchy in occurrence, possibly reflecting their ecology, and/or uncertainty regarding the annotation of low-abundance transcripts (Fig. 14). For example, transcripts sharing the highest identity with Xenoturbellida, Sipuncula, Priapulida, Metazoa and Entoprocta phyla are only found in one sample each. Similarly, transcripts sharing the highest identity with Tardigrada, Rhombozoa, Onychophora, Haplosporidia, Echiura, Ctenophora, Bryozoa and Bolidophyceae phyla are present in two samples each. In contrast, transcripts from the phyla Streptophyta, Chlorophyta and Bacillariophyta are present in all samples, which also reflect the Eukaryotic phyla annotated with a relatively higher confidence than the annotations associated with the other Eukaryotic phyla (E_{ave}< 1E-0.38). Eukaryota likely play an important role in an ecological context of the Kranji Reservoir plankton community for example via predation (grazing) of bacterioplankton or by serving as primary producers (e.g. green algae).

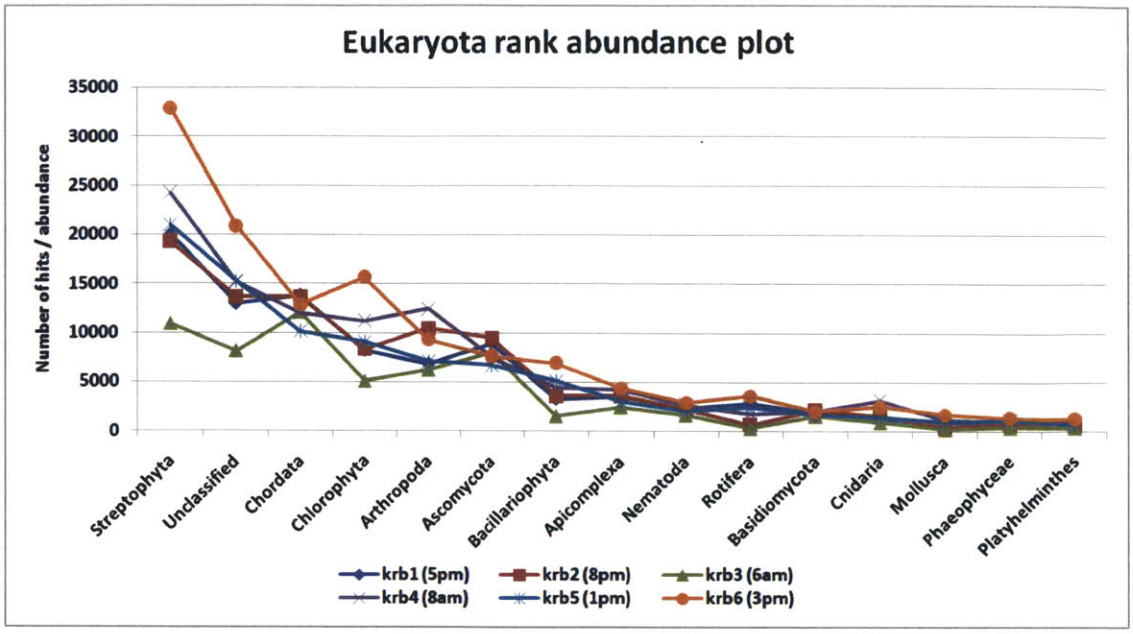


Fig. 19: Eukaryota rank abundance plot in Kranji Reservoir transcriptomes annotated in MG-RAST (E value <1) in the six samples. The vertical axis represents the number of hits in all the available databases on a logarithmic scale.

Phylum (ave. log E value | % identity | protein alignment length) for 5pm, 8pm, 6am, 8am 1pm and 3pm (krb1-6)

Fig. 20: Distribution of Eukaryotic phyla in Kranji Reservoir transcriptomes. Values represent percent of total Eukaryotic transcripts that match the indicated phylum in the M5NR database (e-value <1). krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.3 Diversity within the top five phyla

To resolve Bacterial diversity at a finer scale than Phylum, taxa composition within each of the phyla was explored using the M5NR protein database annotation available in the MG-RAST server. The total number of transcripts assigned to be of bacteria origin range from 544,759 to 777,489 reads for the six samples.

The top five phyla (Cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes) have been broken further down to the Order, Family and Genus levels in terms of taxonomic assignment. The percent of total Bacterial transcripts that match the indicated taxa are calculated for each of the finer scale taxon within each of the six samples. Spectra of percentage values that directly correspond to transcript abundance levels are reflected in the form of heatmaps (Fig. 21- Fig. 26) for each of the five phyla in the subsection that follow below.

3.3.1.3.1 Cyanobacteria

Within the Cyanobacteria phylum, there is a wide range of genera found in the community samples from Kranji Reservoir (Fig. 21). Out of the total Bacterial community reads, the most dominant genus is *Microcystis* representing 15.3 - 25.6 % of all transcripts (E_{ave} = 1E-4.22). 10 additional genera of Cyanobacteria, each representing >1 % of the total transcripts across the six samples, contribute higher to the total community gene expression than most other Bacterial genera, families, or orders (Fig. 23-26). The major Cyanobacterial genera in the Kranji transcripts, in order of abundance are *Microcystis* (20.41 ± 3.52 %), *Cyanthece* (10.58 ± 1.92 %), *Synechococcus* (6.21 ± 1.42 %), *Anabaena* (5.57 ± 1.38 %), *Synechocystis* (4.80± 1.18 %), *Nostoc* (4.09 ± 0.82 %), *Trichodesmium* (2.32 ± 0.41 %), *Crocosphaera* (1.29 ± 0.22 %), *Arthrospira* (1.26 ± 0.14 %), *Thernosynechococcus* (1.21 ± 0.28 %), and *Microcoleus* (1.04 ± 0.16 %). The distribution of Cyanobacterial genera is relatively stable throughout the day (i.e. across the six samples taken at six time points of a day). Comparing across the whole spectrum of genera within each sample in Fig. 21, the dominant genera like *Microcystis*, *Cyanothece* and *Synechococcus* remain as the dominant ones within each sample. Of microcystin-producing Cyanobacteria, *Microcystis*, *Planktothrix*, *Anabaena*, and *Nostoc* are the most common. This distribution of Cyanobacterial genera agrees well with analysis of residual ribosomal RNAs in the transcriptomes (Table 6) that could be annotated with higher confidence ($E < 10^{-20}$).

Comparison on the abundance of transcripts from each genus across the six day/night samples reveals preliminary evidence of enrichment of Cyanobacterial transcripts during the day (Fig. 21). For instance, *Cyanothece* transcripts appear to be enriched during the day relative to at night as there is an average increase of 32.7 % in the proportion of transcripts recovered from the day than at night. This difference is significant (p = 0.0323). Similarly, average increases of day-time transcript abundance relative to night-time abundance for *Synechococcus*, *Synechocystis*, *Anabaena* and *Nostoc* (48.2 %, 54.8 %, 49.4 % and 37.1 % respectively) are all statistically significant (p = 0.011, 0.0043, 0.0388 and 0.0382 respectively). In *Microcystis*, there is also an average increase of 12.8 % in the number of transcripts recovered from the day than at night, but it is not statistically significant (p = 0.3886). However, a general trend of enriched transcript abundance for photosynthetic

Cyanobacteria during the day is consistent with higher metabolic rates while sunlight is available in the day.

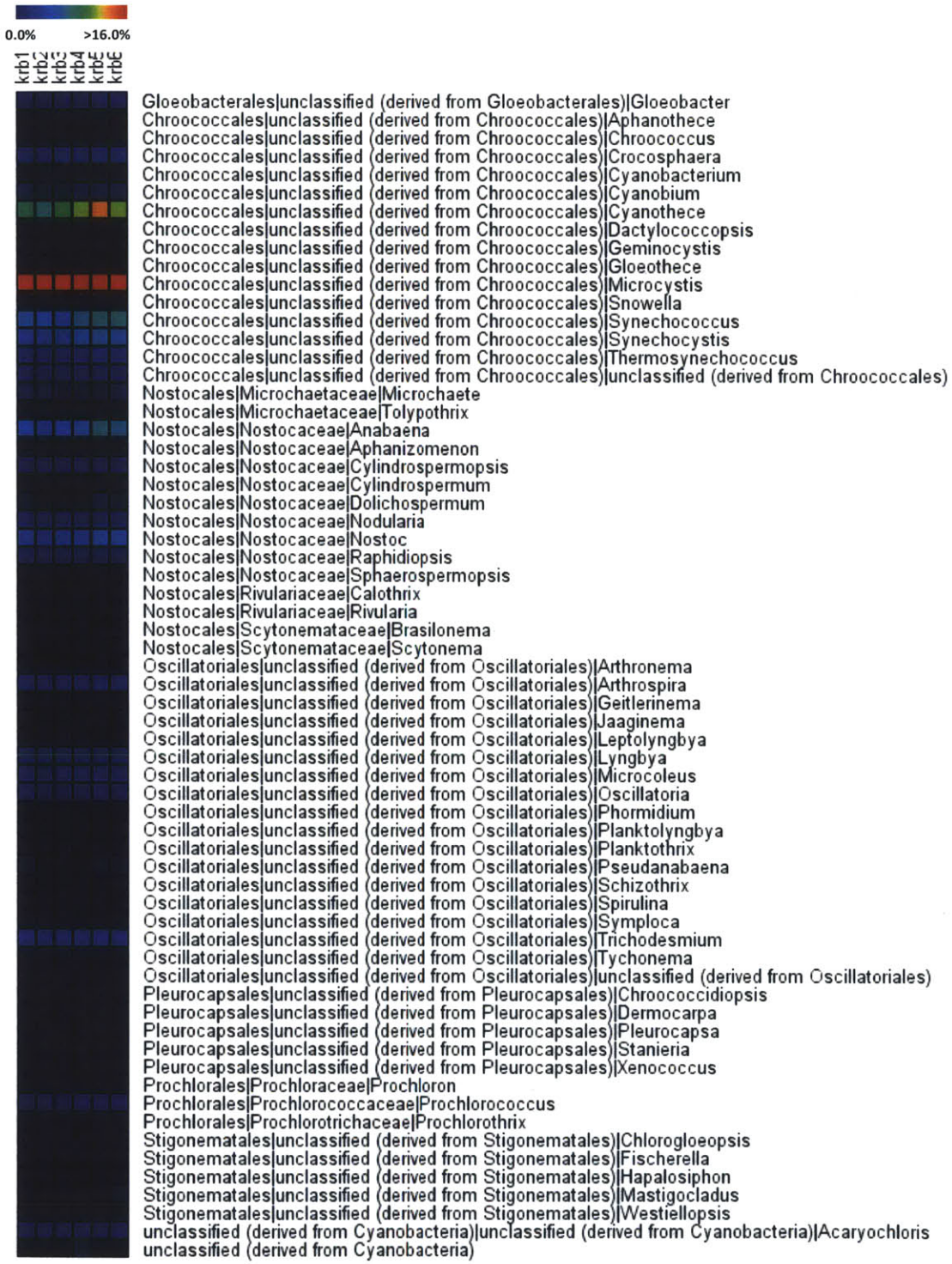


Fig. 21: Distribution of Cyanobacteria taxa (Order | Family | Genus) in Kranji Reservoir samples at E-value <1. Values represent percent of total Bacterial transcripts that match the indicated taxa in the M5NR protein database. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.3.2 Proteobacteria

Transcripts from the Proteobacateria phylum, were the second most abundant in Kranji Reservoir (10.7 – 23.2 % of total Bacterial transcriptomes), as annotated by MG-RAST (E-value < 1). The dominant classes of Proteobacteria observed in transcript annotations were Alpha-proteobacteria, Beta-proteobacteria and Gamma-proteobacteria (Fig. 22). Therefore, in the rest of the discussion on Proteobacteria, the focus will be placed on taxa within these three classes.

The most abundant order in Proteobacteria is Burkholderiales (a Beta-proteobacteria) (3.08 ± 0.90%; E_{ave} ranges from 1E-2.04 to 1E-1.43, Fig. 23). The other prominent orders in decreasing order of abundance are Rhizobiales (an Alpha-proteobacteria) (2.08 ± 0.60%; E_{ave} ranges from 1E-0.99 to 1E-0.41), Rhodobacaterales (an Alpha-proteobacteria) (1.33 ± 0.41%; E_{ave} ranges from 1E-0.03 to 1E0.39), Enterobacteriales (a Gamma-proteobacteria) (1.29 ± 0.28%; E_{ave} ranges from 1E-1.87 to 1E-0.81), Pseudomonadales (a Gamma-proteobacteria) (0.88 ± 0.21%; E_{ave} ranges from 1E-0.53 to 1E0.03) and Myxococcales (a Delta-proteobacteria) (0.66 ± 0.25%; E_{ave} ranges from 1E-0.08 to 1E0.2). The richness and distribution of different orders of Proteobacteria is similar in all the six samples as the dominant orders stay dominant throughout the day/night cycle (Fig. 23).

Gene expression in some Proteobacterial classes appear to be enriched at night relative to during the day – this is in contrast to apparent enrichment of Cyanobacterial gene expression during the day and likely reflects an enrichment of heterotrophic processes in the bloom at night. For example, in Rhizobiales, there is an average decrease of 34.6 % in the proportion of transcripts recovered from the day than at night and this is statistically significant ($p = 0.0403$). By the same token, the average decrease for Burkholderiales, Rhodobacterales, Myxococcales, Pseudomonadales, Enterobacteriales are 30.7, 33.8, 41.9, 32.0 and 19.8 % respectively. However, only the decrease for Pseudomonadales is statistically significant ($p = 0.0266$).

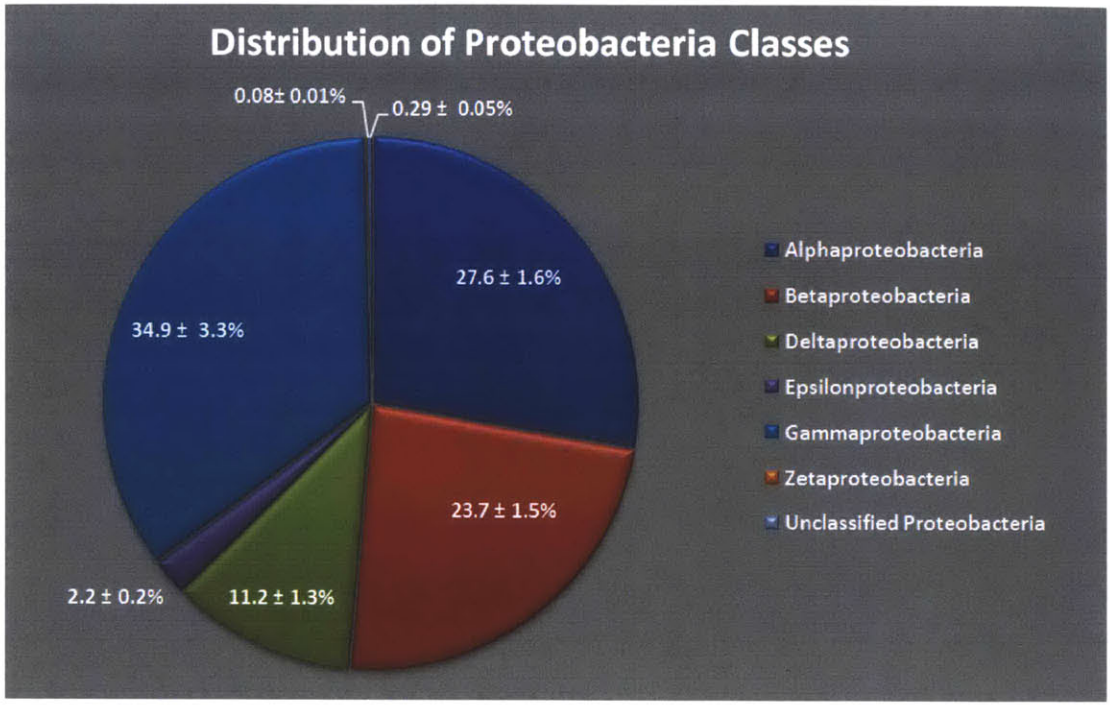


Fig. 22: Distribution of Proteobacteria Classes in Kranji Reservoir transcriptomes (Krb1-6).

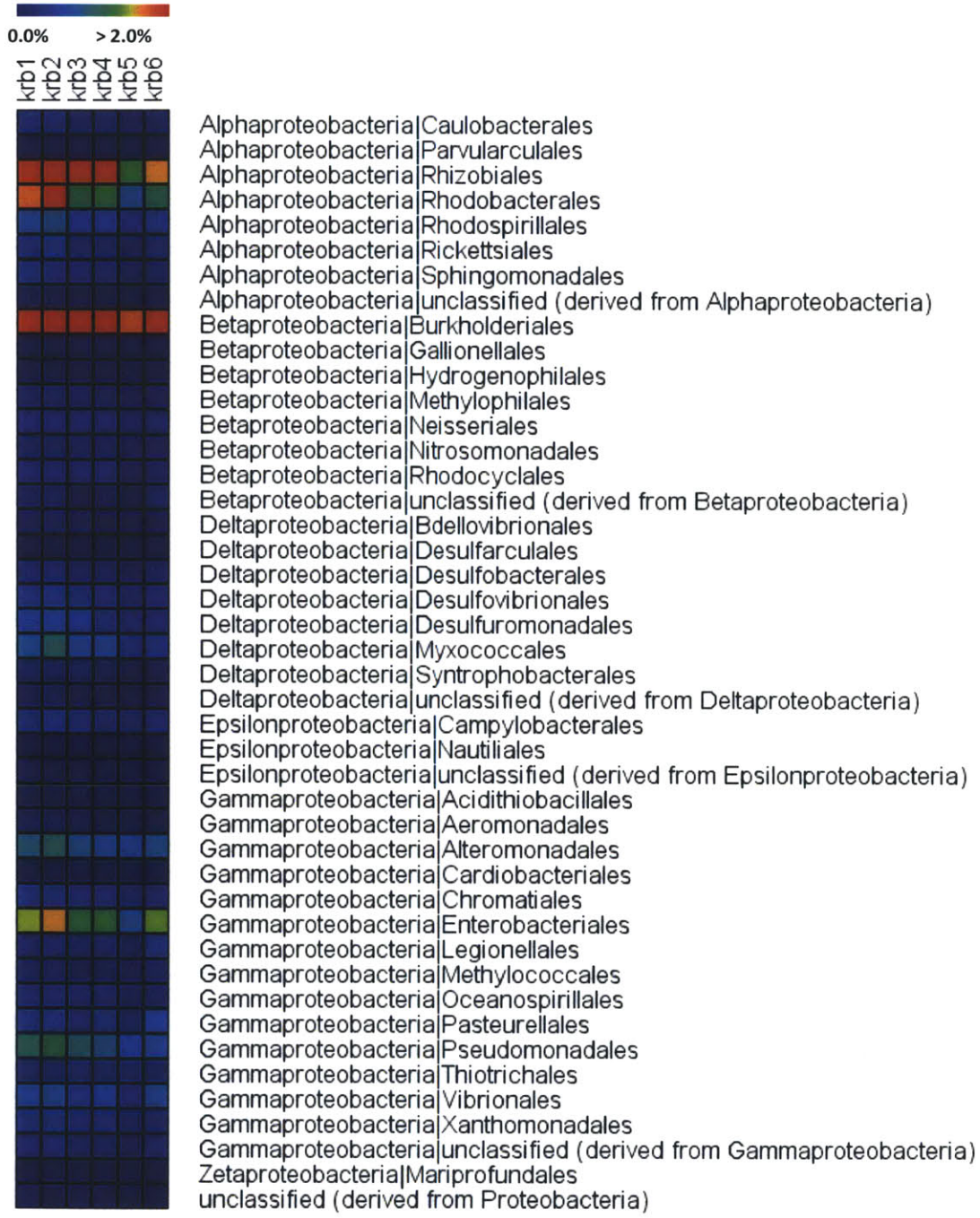


Fig. 23: Distribution of Proteobacteria taxa (Class | Order) in Kranji Reservoir samples at E-value <1. Values represent percent of total Bacterial transcripts that match the indicated taxa in the M5NR protein database. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.3.3 Firmicutes

The most abundant order of Firmucutes in Kranji Reservoir transcriptomes is Clostridiales (2.80 ± 0.73 % of the total transcripts; E_{ave} ranges from 1E0.52 to 1E0.28, Fig. 24). The other prominent orders in decreasing order of abundance are Bacillales (1.68 ± 0.39 %; E_{ave} ranges from 1E-0.25 to 1E0.28), Lactobacillales (0.91 ± 0.21 %; E_{ave} ranges from 1E0.35 to 1E0.85), Thermoanaerobacterales (0.34 ± 0.07 %; E_{ave} ranges from 1E-0.37 to 1E0.19) and Selenomonadales (0.24 ± 0.06 %; E_{ave} ranges from 1E0.47 to 1E0.99). Firmicutes play an important role in the ecology of Kranji Reservoir. For example, studies have shown that certain species of the genus *Clostridium* (within the Clostridiales family) might be degraders of *Microcystis* bloom scum (Xing et al., 2011). Comparing across the whole spectrum of orders within each sample in Fig. 24, dominant taxa in one sample stay dominant in the other samples and there is no consistent trend of enrichment of transcripts in either "day" or "night" samples, in contrast to apparent enrichment of Cyanobacterial transcripts during the day and Proteobacterial transcripts at night.

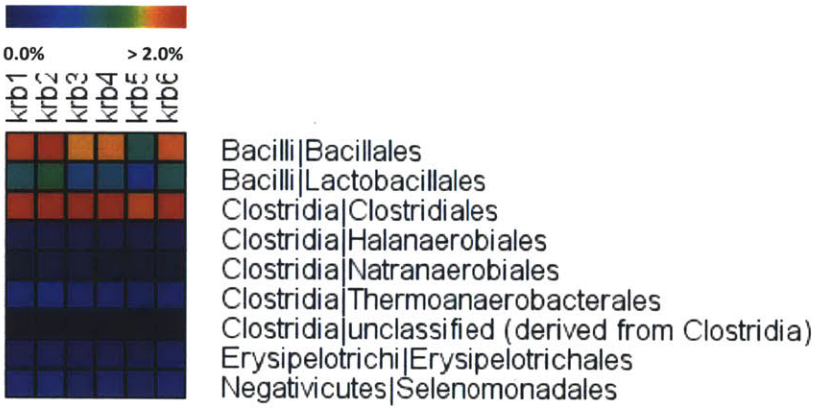


Fig. 24: Distribution of Firmicute taxa (Class | Order) in Kranji Reservoir samples at E-value <1. Values represent percent of total Bacterial transcripts that match the indicated taxa in the M5NR protein database. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.3.4 Actinobacteria

The most abundant family in the Phylum Actinobacteria is the Streptomycetaceae (1.23 ± 0.48 % of the total transcripts; E_{ave} ranges from 1E-1.75 to 1E1.32, Fig. 25). Its genus *Streptomyces* produce over two-thirds of the clinically useful antibiotics of natural origin (Berdy 1980). The other prominent families in decreasing order of abundance are Mycobacteriaceae (0.53 ± 0.24 %; E_{ave} ranges from 1E-1.62 to 1E-0.15), Kineosporiaceae (0.44 ± 0.16 %; E_{ave} ranges from 1E-0.76 to 1E1.04), Corynebacteriaceae (0.38 ± 0.12 %; E_{ave} ranges from 1E-0.45 to 1E1.52), Frankiaceae (0.31 ± 0.15 %; E_{ave} ranges from 1E-0.58 to 1E0.59) and Micrococcaceae (0.25 ± 0.13 %; E_{ave} ranges from 1E-0.2 to 1E1.05).

Comparing across the whole spectrum of orders within each sample in Fig. 25, dominant taxa in one sample may not stay dominant in the other samples. For example, Mycobacteriaceae are the second most abundant family in 5pm, 8pm samples, but they become the 4th in the 3pm sample.

Comparing each order across their abundance levels in the six samples in Fig. 25, just like Cyanobacteria and Proteobacteria, the day-night separation in terms of transcript abundance is also evident in Actinobacteria. The average decrease in the number of transcripts recovered from the day than at night for Frankiaceae, Mycobacteriaceae, Nocardiaceae, Nocardioidaceae, Nocardiopsaceae, Pseudonocardiaceae and Streptomycetaceae (57.1 %, 54.8 %, 59.4 %, 66.7 %, 68.3 %, 59.7 % and 47.9 % respectively) are all statistically significant ($p = 0.0119, 0.0164, 0.0084, 0.0361, 0.0470, 0.0208$ and 0.0211 respectively).

Fig. 25: Distribution of Actinobacteria taxa (Order | Family) in Kranji Reservoir samples at E-value <1. Values represent percent of total Bacterial transcripts that match the indicated taxa in the M5NR protein database. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.3.5 Bacteroidetes

The most abundant family represented in transcripts from the Bacteroidetes phylum is Flavobacteriaceae (1.08 ± 0.51 %; E_{ave} ranges from 1E-2.08 to 1E-1.05) (Fig. 26). The other prominent families in decreasing order of abundance are Bacteroidaceae (0.58 ± 0.21 %; E_{ave} ranges from 1E-1.06 to 1E-0.31), Cytophagaceae (0.56 ± 0.28 %; E_{ave} ranges from 1E-2.6 to 1E-0.67), Porphyromonadaceae (0.33 ± 0.11 %; E_{ave} ranges from 1E-1.4 to 1E0), Prevotellaceae (0.26 ± 0.11 %; E_{ave} ranges from 1E-0.51 to 1E0.03) and Sphingobacteriaceae (0.26 ± 0.14 %; E_{ave} ranges from 1E-2.51 to 1E1).

Comparing each family across their transcript abundance levels in the six samples in Fig. 26, just like Cyanobacteria, Proteobacteria and Actinobacteria, the day-night separation in terms of transcript abundance is also evident in Bacteroidetes. The average decrease in the number of transcripts recovered from the day than at night for Bacteroidaceae, Cytophagaceae, Flavobacteriaceae and Sphingobacteriaceae are (42.2 %, 58.8 %, 52.5 % and 62.1 % respectively). Only the decrease for Cytophagaceae and Sphingobacteriaceae are statistically significant ($p = 0.0227$ and 0.0295 respectively).

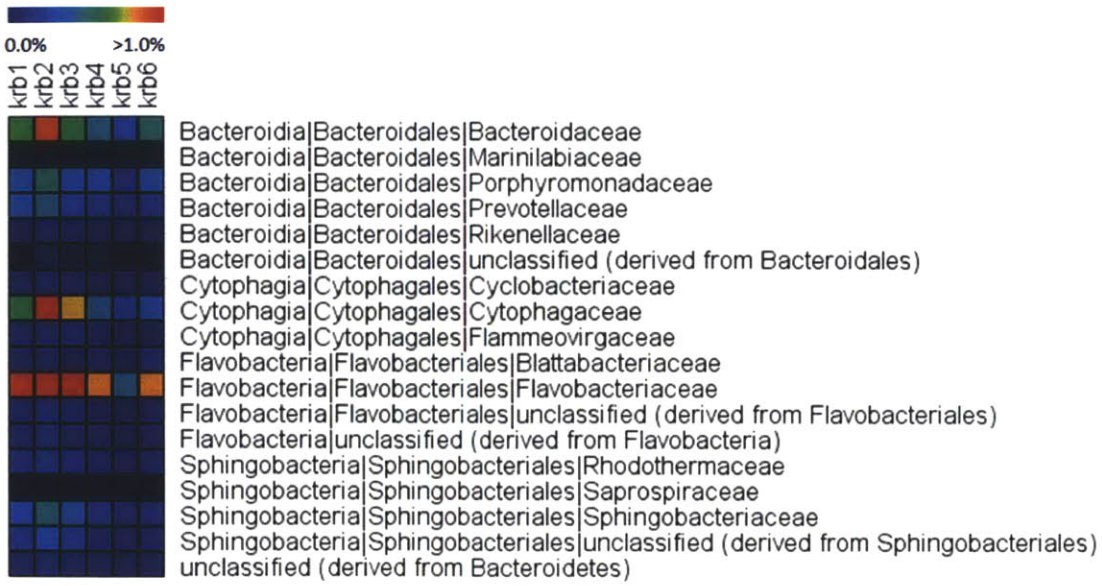


Fig. 26: Distribution of Bacteroidetes taxa (Class | Order | Family) in Kranji Reservoir samples at E-value <1. Values represent percent of total bacteria transcripts that match the indicated taxa in the M5NR protein database. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

3.3.1.4 Diversity within *Microcystis*

At an E-value < 1, the most abundant species in *Microcystis* genus is *Microcystis aeruginosa* (19.93 %), although there may be other *Microcystis* species based on transcripts (each has < 0.22 % abundance).

Comparing each genus across their abundance levels in the six samples in Fig. 27, the day-night separation in terms of transcript abundance does not seem to follow strictly to the 24-hour clock. The average transcript abundance of *Microcystis aeruginosa* are 18.4 %, 14.9 %, 22.8 %, 19.4 %, 24.8 % and 19.4 % for the six samples krb1_5pm, krb2_8pm, krb3_6am, krb4_8am, krb5_1pm, krb6_3pm respectively. The difference in transcript abundance in the day-night category is not statistically significant ($p = 0.43$). There is an increase of 53.0 % in abundance from krb2_8pm to krb3_6am sample. At both of these times, the surrounding light index all has gone down to 0 lux, or complete darkness. However, *Microcystis aeruginosa* seem to show peak abundance in the morning (krb3_6am, 22.8 %) and at noon (krb5_1pm, 24.8 %). Thereafter, its abundance decreases sharply from 24.8 % to 19.4 % and lastly to the day-minimum at around night time (krb2_8pm:14.9%).

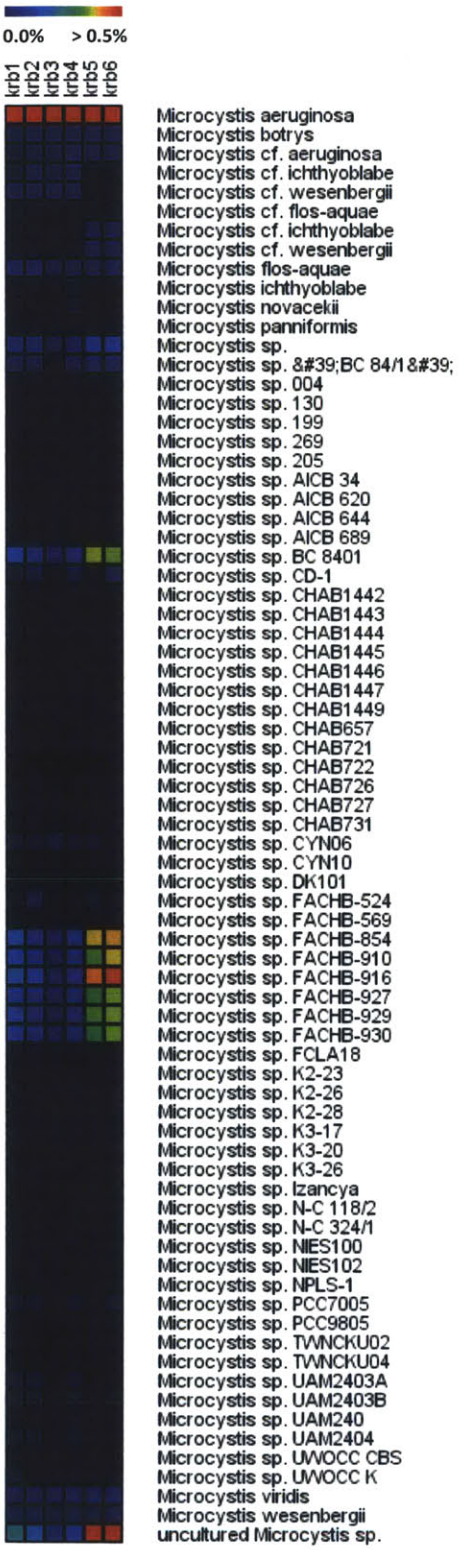


Fig. 27: Distribution of *Microcystis* at the Species level at E-value <1. Values represent the percentage of sequences with a hit in the M5NR protein database of a particular species. krb1 – krb6: 5pm, 8pm, 6am, 8am, 1pm and 3pm.

Table 5: Observed richness of total community (including Bacteria, Eukaryote, Archaea) at genera level.

	5pm (krb1)	8pm (krb2)	6am (krb3)	8am (krb4)	1pm (krb5)	3pm (krb6)
RDP genera $E < 10^{-20}$	455	482	292	371	357	565
M5NR genera E < 1	2,361	2,380	2,229	2,147	2,154	2,207
M5NR genera E < 0.001	654	676	593	632	684	714

Table 6: Major genera in expressed ribosomal sequences from Kranji Reservoir. Matches to the Ribosomal database (RDP) had an average E-value < 1E-20; an average % identity >99%; and an average alignment length >50nt.

5pm (Krb1)			8pm (Krb2)			6am (krb3)			8am (Krb4)			1pm (Krb5)			3pm (Krb6)		
phylum	genus	No.	phylum	genus	No.	phylum	genus	No.	phylum	genus	No.	phylum	genus	No.	phylum	genus	No.
Cyanobacteria	Microcystis	18337	Cyanob	Microcystis	12197	Cyanob	Microcystis	5926	Cyanob	Microcystis	16423	Cyanob	Microcystis	21302	Cyanob	Microcystis	44663
Cyanobacteria	Synechococcus	856	Cyanob	Synechococcus	752	Cyanob	Synechococcus	153	Cyanob	Synechococcus	334	Cyanob	Lyngbya	584	Cyanob	Synechococcus	1663
Proteobacteria	Acetobacter	375	Proteot	Acetobacter	375	Cyanob	Lyngbya	113	Cyanob	Lyngbya	257	Cyanob	Synechococc	352	Cyanob	Lyngbya	1104
Cyanobacteria	Lyngbya	373	Cyanob	Planktothrix	296	Cyanob	Anabaenopsis	95	Cyanob	Anabaenopsis	142	Cyanob	Anabaenopsi	256	Proteot	Vibrio	678
Cyanobacteria	Planktothrix	354	Cyanob	Lyngbya	279	Cyanob	Stanieria	50	Proteot	Acetobacter	142	Cyanob	Dolichosperm	187	Proteot	Acetobacter	634
Cyanobacteria	Dolichospermum	226	Cyanob	Dolichospermum	239	unassig	unassigned	48	Cyanob	Crocosphaera	118	Proteot	Acetobacter	185	Cyanob	Anabaenopsis	567
Cyanobacteria	Anabaenopsis	217	unassig	unassigned	142	Cyanob	Dolichospermum	43	Cyanob	Dolichospermum	115	Cyanob	Spirulina	102	Cyanob	Planktothrix	481
Cyanobacteria	Synechocystis	139	Cyanob	Cyanothece	133	Proteot	Acetobacter	38	Cyanob	Planktothrix	97	Cyanob	Synechocysti	93	Cyanob	Dolichospermum	407
Cyanobacteria	Cyanothece	136	Cyanob	Leptolyngbya	118	Cyanob	Phormidium	36	Cyanob	Stanieria	97	Cyanob	Planktothrix	89	Cyanob	Phormidium	358
Proteobacteria	Polynucleobacter	133	Cyanob	Phormidium	114	Cyanob	unclassified (deriv	34	unassig	unassigned	83	unassig	unassigned	69	unassig	unassigned	293
unassigned	unassigned	127	Cyanob	Anabaenopsis	107	Cyanob	Leptolyngbya	33	Proteot	Vibrio	66	Cyanob	Oscillatoria	68	Cyanob	Synechocystis	231
Cyanobacteria	Phormidium	114	Cyanob	Stanieria	106	Cyanob	Oscillatoria	30	Cyanob	Spirulina	54	Cyanob	Crocosphaera	66	Cyanob	Prochlorococcus	229
Cyanobacteria	Spirulina	111	Cyanob	Spirulina	102	Cyanob	Planktothrix	28	Cyanob	Cyanothece	53	Cyanob	Anabaena	56	Cyanob	Spirulina	220
Cyanobacteria	Stanieria	110	Cyanob	Synechocystis	96	Cyanob	Synechocystis	25	Cyanob	Synechocystis	53	Cyanob	Phormidium	52	Cyanob	Pseudanabaena	209
Cyanobacteria	Anabaena	98	Cyanob	Anabaena	90	Cyanob	Anabaena	25	Cyanob	Pseudanabaena	53	Cyanob	Nostoc	47	Cyanob	Cyanothece	191
Cyanobacteria	Leptolyngbya	93	Cyanob	Nostoc	75	Cyanob	Pseudanabaena	21	Cyanob	Phormidium	47	Cyanob	Pseudanabae	46	Cyanob	Stanieria	186
Cyanobacteria	Prochlorococcus	78	Proteot	Polynucleobacter	74	Cyanob	Gloeothece	20	Cyanob	Leptolyngbya	45	Plancto	Pirellula	38	Cyanob	Leptolyngbya	165
Cyanobacteria	Oscillatoria	77	Cyanob	Oscillatoria	73	Cyanob	Nostoc	19	Cyanob	Prochlorococcus	41	Cyanob	Calothrix	35	Cyanob	Nostoc	161
Cyanobacteria	Crocosphaera	70	Cyanob	Prochlorococcus	64	Cyanob	Prochlorococcus	19	Cyanob	Anabaena	35	Cyanob	Cylindrosperr	34	Cyanob	Anabaena	155
Cyanobacteria	Pseudanabaena	53	Cyanob	Aphanizomenon	59	Bactero	Cytophaga	16	Cyanob	unclassified (deriv	32	Firmicu	Lactobacillus	34	Cyanob	Oscillatoria	147
Cyanobacteria	Cylindrospermopsis	44	Proteot	Burkholderia	59	Cyanob	Spirulina	16	Cyanob	Nostoc	30	Cyanob	Prochlorococ	32	Cyanob	Prochlorothrix	145
Cyanobacteria	Arthrospira	44	Proteot	Vibrio	57	Proteot	Polynucleobacter	16	Cyanob	Aphanizomenon	28	unclass	Paulinella	29	unclass	Paulinella	120
Proteobacteria	Burkholderia	44	Cyanob	Sphaerospermopsis	54	Proteot	Vibrio	15	Proteot	Sorangium	27	Cyanob	Sphaerosperr	28	Cyanob	Cylindrospermop	106
Cyanobacteria	Cyanobacterium	41	Plancto	Planctomyces	52	Plancto	Planctomyces	13	Firmicu	Clostridium	26	Firmicu	Clostridium	28	Cyanob	unclassified (deriv	100
Cyanobacteria	Nostoc	41	Plancto	Pirellula	48	Proteot	Burkholderia	12	unclass	Paulinella	24	Cyanob	Cyanothece	27	Cyanob	Sphaerospermop	93
Firmicutes	Clostridium	39	unclass	Paulinella	46	Cyanob	unclassified (deriv	10	Cyanob	Tolypothrix	22	Cyanob	Aphanizomer	26	Proteot	Polynucleobacter	91
unclassified (Eukaryota)	Paulinella	39	Cyanob	Cylindrospermopsis	44	Proteot	Herbaspirillum	10	Cyanob	Cylindrospermops	20	Cyanob	unclassified (	24	Cyanob	Aphanizomenon	80
Cyanobacteria	Aphanizomenon	34	Firmicu	Clostridium	42	Proteot	Sorangium	10	Cyanob	Arthrospira	20	Cyanob	Leptolyngbya	22	Proteot	Burkholderia	72
Bacteroidetes	Cytophaga	31	Proteot	Sorangium	42	Cyanob	Cyanobacterium	9	Proteot	Burkholderia	19	Bactero	Cytophaga	17	Proteot	Pseudomonas	65
Planctomycetes	Planctomyces	31	Bactero	Cytophaga	38	Cyanob	Crocosphaera	8	Strepto	Festuca	18	Cyanob	Stanieria	17	Cyanob	Nodularia	63
Cyanobacteria	Prochlorothrix	30	Cyanob	unclassified (derived fr	35	Cyanob	Cyanothece	8	Cyanob	Sphaerospermops	17	Cyanob	Cyanobacteri	16	Bactero	Chitinophaga	59
Bacteroidetes	Chitinophaga	27	Verruco	Opitutus	35	Cyanob	Arthrospira	8	Plancto	Planctomyces	17	Proteot	Burkholderia	16	unclass	unclassified (deriv	58
Cyanobacteria	unclassified (derived	26	Bactero	Chitinophaga	32	Cyanob	unclassified (deriv	8	Cyanob	Oscillatoria	16	unclass	unclassified (	16	Bactero	Terrimonas	54
Proteobacteria	Rubrivivax	26	Cyanob	Pseudanabaena	32	Plancto	Pirellula	8	Proteot	Pseudomonas	16	Strepto	Angiopteris	16	Cyanob	Crocosphaera	54
Cyanobacteria	Sphaerospermopsis	25	unclass	unclassified (derived fr	32	Proteot	Acidovorax	7	Proteot	Listonella	16	Plancto	Planctomyces	15	Cyanob	Arthrospira	52
Cyanobacteria	Chroococcidiopsis	24	Proteot	Methylomonas	30	Proteot	Oxalicibacterium	7	Plancto	Pirellula	15	Cyanob	Prochlorothri	14	Cyanob	unclassified (deriv	46
Proteobacteria	Sorangium	24	Cyanob	Prochlorothrix	27	Verruco	unclassified (deriv	7	Cyanob	Nodularia	14	Cyanob	Fischerella	14	Cyanob	Cyanobacterium	45
unclassified (Bacteria)	unclassified (derived	24	Cyanob	Cyanobacterium	26	Strepto	Angiopteris	7	Cyanob	unclassified (deriv	14	Strepto	Staurastrum	14	Cyanob	Calothrix	44
Proteobacteria	Pseudomonas	23	Cyanob	Arthrospira	25	Bactero	Flavobacterium	6	Plancto	Blastopirellula	14	Proteot	Sorangium	13	Proteot	Herbaspirillum	44
Verrucomicrobia	Opitutus	23	Cyanob	unclassified (derived fr	25	Cyanob	Gloeocapsa	6	Verruco	Opitutus	14	Strepto	Festuca	13	Cyanob	Trichodesmium	43
Various other genera		1609	Various other genera		1811	Various other genera		486	Various other genera		1017	Various other genera		956	Various other genera		2998
Total ribosomal hits		24326	Total ribosomal hits		18083	Total ribosomal hits		7449	Total ribosomal hits		19691	Total ribosomal hits		25048	Total ribosomal hits		57174

Table 7: Taxonomic affiliation of transcripts. Top 80 Genera with E-value < 0.001 and minimum protein alignment of 17 amino acids (or 51 nt) against M5NR protein database.

5pm (Krb1)			average	8pm (Krb2)			average	6am (Krb3)			average
phylum	genus	No.	E value |%| len.	phylum	genus	No.	E |%| len	phylum	genus	No.	E |%| len
Cyanobacteria	Microcystis	116189	-6.05|95.46|26.45	Cyanob	Microcystis	80823	-5.42|93.58|25.48	Cyanob	Microcystis	128917	-6.63|95.98|27.12
Cyanobacteria	Synechocystis	24381	-4.51|85.48|26.11	Cyanob	Synechocystis	18069	-4.91|89.24|25.5	Cyanob	Synechocystis	20693	-3.96|84.97|24.79
Cyanobacteria	Nostoc	13584	-4.99|83.53|27.13	unassig	unassigned	11936	-4.47|84.76|25.79	unassig	unassigned	17945	-4.42|85.46|25.48
Cyanobacteria	Synechococcus	8728	-5.09|88.83|25.91	Cyanob	Nostoc	10590	-5.01|84.07|27.69	Cyanob	Anabaena	14957	-5.19|83|27.3
unassignec	unassignec	7454	-4.82|83|27.11	Cyanob	Cyanothece	6989	-3.11|83.48|24.7	Cyanob	Cyanothece	12313	-3.03|83.28|23.87
Cyanobacteria	Anabaena	3896	-5.25|87.3 26.22	Cyanob	Synechococcus	6623	-5.44|88.35|26.52	Cyanob	Nostoc	11388	-4.78|82.96|26.77
Cyanobacteria	Trichodesmium	2288	-4.35|82.98|27.61	Cyanob	Thermosynechoc	4845	-3.72|84.89|24.2	Cyanob	Thermosynechococ	5070	-3.59|84.21|24.2
Proteobacteria	Escherichia	1288	-3.88|92.81|23.09	Cyanob	Anabaena	2789	-5.23|84.91|27.13	Cyanob	Synechococcus	4402	-5.19|85.51|26.93
Proteobacteria	Burkholderia	1215	-3.47|85.14|24.41	Cyanob	Trichodesmium	1688	-4.15|80.69|28.06	Cyanob	Trichodesmium	1684	-6.2|78.34 30.84
Cyanobacteria	Prochlorococcus	856	-4.78|85.28|26.59	Actinob	Streptomyces	799	-4.96|78.46|28.31	Proteob	Escherichia	1261	-4.37|93.15|23.75
Cyanobacteria	Arthrospira	836	-5.41|81.77|29.14	Proteob	Burkholderia	652	-3.79|86.1 24.28	Actinob	Streptomyces	1078	-5.09|79.35|28.87
Cyanobacteria	Dolichospermum	534	-6.71|92.77|28.61	Cyanob	Prochlorococcus	504	-4.63|86.72|25.93	Proteob	Burkholderia	604	-3.72|85.32|24.79
Proteobacteria	Shigella	359	-3.75|92.84|22.7	Bactero	Bacteroides	472	-4.8|79.45 28.52	Proteob	Shigella	528	-4.03|93.41|23.05
Cyanobacteria	Mastigocladus	290	-5.61|89.85|25.87	Cyanob	Acaryochloris	464	-3.84|72.86|27.29	Bactero	Flavobacterium	257	-3.96|85.31|23.97
Bacteroidetes	Bacteroides	278	-3.11|79.87|24.81	Bactero	Flavobacterium	406	-4.17|83.67|26.13	Cyanob	Arthrospira	250	-5.53|81.29|28.96
Bacillariophyta	Phaeodactylum	244	-4.31|91.71|24.02	Proteob	Escherichia	362	-3.51|89.63|23.13	Bactero	Cytophaga	213	-6.01|88|28.6
Bacillariophyta	Odontella	236	-5.32|91.76|25.18	Bactero	Cytophaga	326	-3.54|82.58|24.93	Actinob	Acidothermus	211	-4.11|83.74|25.61
unclassified (de	Galdieria	214	-7.62|90.54|28.86	Cyanob	Mastigocladus	291	-3.98|79.02|25.51	Actinob	Propionibacterium	211	-3.51|85.89|24.03
Bacillariophyta	Thalassiosira	210	-4.63|87.91|25.62	Cyanob	Arthrospira	288	-5.61|81.8 28.53	Cyanob	Prochlorococcus	179	-4.82|87.16|26.23
unclassified (de	Cyanidium	210	-6.26|87.9 26.68	Cyanob	Dolichospermum	282	-5.74|86.93|28.96	Cyanob	Planktothrix	170	-5.51|81.6 28.76
unclassified (de	Guillardia	210	-3.46|88.47|22.94	Actinob	Mycobacterium	248	-5.64|85.58|28.2	Actinob	Kineococcus	159	-3.68|82.4 25.8
Proteobacteria	Ralstonia	207	-7.12|86.99|29.54	Bacillari	Thalassiosira	216	-6.67|89.68|27.75	Actinob	Nocardioides	158	-3.19|82.01|25.19
Cyanobacteria	unclassified (deriv	206	-3.56|85.94|24.41	Proteob	Ralstonia	204	-4.15|82.4 26.92	unclass	Cyanidium	133	-6.15|93.26|25.97
Bacteroidetes	Cytophaga	196	-3.35|82.08|24.71	Bacillari	Odontella	203	-3.58|91.11|22.16	unclass	unclassified (derive	120	-3.95|89.92|23.63
Proteobacteria	Neisseria	195	-3.42|85.28|23.93	unclass	Guillardia	203	-3.39|87.64|23.08	Bactero	Gramella	116	-3.26|78.73|25.56
unclassified (de	unclassified (deriv	195	-4.98|78.08|28.82	Bactero	Gramella	202	-3.84|81.87|25.86	Proteob	Ralstonia	111	-3.35|80.8 24.96
unclassified (de	Porphyra	192	-6.57|92.02|27.16	Plancto	Rhodopirellula	194	-4.53|82.22|27.25	Cyanob	Geminocystis	110	-4.02|89.65|23.2
Proteobacteria	Acidovorax	190	-5.03|86.26|26.07	unclass	unclassified (der	181	-4.84|78.58|27.99	Cyanob	Dolichospermum	109	-5.77|83.65|28.78
Cyanobacteria	Calothrix	184	-5.17|91.66|23.83	Actinob	Acidothermus	169	-3.31|81.4 24.3	Bactero	Parabacteroides	105	-3.13|81.57|24.54
Actinobacteria	Streptomyces	160	-5.11|77.02|29.57	Proteob	Acidovorax	168	-3.73|86.44|25.25	Proteob	Geobacter	104	-3.35|77.73|26.43
Cyanobacteria	Prochlorothrix	145	-3.22|91.59|21.13	Proteob	Acinetobacter	150	-7.08|88.65|29.1	Proteob	Methylobacillus	99	-5.69|79.07|29.03
unclassified (de	unclassified (deriv	145	-4.83|88.35|25	unclass	Paulinella	143	-3.33|82.84|24.5	Proteob	Nitrosomonas	98	-3.29|80.1 25.12
Proteobacteria	Brucella	140	-3.15|81.64|24.06	Bacillari	Phaeodactylum	142	-4.7|91.36 24.58	Cyanob	unclassified (derive	97	-4.91|86.06|25.96
Proteobacteria	Salmonella	138	-3.83|84.96|24.9	Actinob	Frankia	141	-3.18|83.19|24.46	unclass	Paulinella	91	-3.04|83.93|23.84
unclassified (de	Microvirus	127	-7.84|98.53|28.21	Cyanob	Fischerella	140	-4.11|83.35|25.34	unclass	Guillardia	90	-3.02|89.73|21.89
Proteobacteria	Methylobacillus	119	-3.46|84.21|24.18	Bactero	Parabacteroides	139	-3.47|81.66|25.16	Cyanob	Mastigocladus	89	-4.73|85.36|25.7
unclassified (de	Pulvinaster	119	-6.38|92.43|28.12	unclass	Galdieria	136	-3.66|88.46|22.55	Proteob	Salmonella	88	-4.95|89.7 25.57
unclassified (de	Durinskia	116	-4.86|90.59|24.66	Actinob	Arthrobacter	118	-3.86|78.69|27.31	Proteob	Bradyrhizobium	85	-5.16|80.81|28.19
Proteobacteria	Cupriavidus	111	-3.33|83.07|24.58	unclass	unclassified (der	115	-4.77|85.33|26.4	Actinob	Saccharopolyspora	83	-3.28|82.67|24.72
Cyanobacteria	Pseudanabaena	107	-4.17|85.74|26.19	Proteob	Methylococcus	114	-3.37|85.73|23.43	Bacillari	Odontella	77	-4.43|88.87|24.19
Bacteroidetes	Flavobacterium	106	-3.97|80.75|26.35	Proteob	Methylobacillus	112	-3.55|83.23|24.97	Proteob	Chromobacterium	73	-5.18|80.28|28.24
Proteobacteria	Dechloromonas	106	-3.19|84.86|23.86	Proteob	Nitrosospira	105	-3.21|88.32|22.56	Actinob	Frankia	72	-3.39|83.53|24.4
Phaeophyceae	Dictyota	100	-6.7|90.03 27.57	Cyanob	Calothrix	103	-8.38|93.27|30.12	Proteob	Brucella	71	-3.38|79.68|25.42
Bacteroidetes	Gramella	98	-3.4|80.63 25.31	Proteob	Neisseria	103	-3.45|81.5 25.55	Actinob	Beutenbergia	69	-3.47|78.69|26.28
unclassified (de	Kryptoperidinium	98	-3.52|88.53|22.93	Cyanob	Microchaete	101	-3.75|83.24|24.05	Bacillari	Thalassiosira	69	-5.14|89.84|25.71
Cyanobacteria	Microchaete	97	-3.11|83.5 24.11	Proteob	Methylibium	98	-4.44|85.37|25.91	Cyanob	Leptolyngbya	68	-6.93|82.11|30.73
Cyanobacteria	Planktothrix	97	-5.93|88.66|28.17	Proteob	Chromobacteriu	98	-3.27|83.61|24.62	unclass	Pulvinaster	66	-10 100|33
unclassified (de	unclassified (deriv	93	-5.72|91.89|26.18	Proteob	Xanthomonas	98	-4.66|84.35|26.64	Proteob	Cupriavidus	64	-3.88|80.03|26.83
Cyanobacteria	Fischerella	92	-7.36|85.96|30.5	Strepto	Pinus	94	-7.6|97.23 27.2	Strepto	Pinus	64	-6.04|94.84|25.26
Proteobacteria	Pseudomonas	88	-4.53|77.63|28.78	unclass	Cyanidioschyzon	91	-3.29|86.52|23.02	unclass	unclassified (derive	62	-7.26|98.18|28.93
unclassified (de	unclassified (deriv	87	-4.31|78.42|26.77	Actinob	Salinispora	86	-3.29|84.18|24.14	Spiroch	Leptospira	58	-4.01|68.74|30.23
Cyanobacteria	Leptolyngbya	85	-5.1|80.79 27.91	Cyanob	Planktothrix	86	-6.52|86.38|28.79	unclass	unclassified (derive	56	-5.27|78.42|29.2
Cyanobacteria	Thermosynechoc	82	-3.85|87.26|23.59	unclass	Gracilaria	82	-6.22|83.88|29.55	unclass	unclassified (derive	53	-7.4|94.37 27.69
Proteobacteria	Methylibium	82	-3.76|83.75|25.29	Cyanob	Prochlorothrix	77	-4.58|88.03|24.04	Bacillari	Phaeodactylum	52	-4.37|87.74|25.07
Bacteroidetes	Parabacteroides	81	-3.38|80.11|25.54	Cyanob	unclassified (der	69	-3.3|84.72 23.78	Chlorop	Chlamydomonas	52	-5.01|83.32|28.22
Actinobacteria	Nocardioides	79	-3.38|83.1 24.88	Phaeop	Dictyota	69	-6.85|88.8 27.78	Actinob	Kocuria	47	-3.1|82.69 24.29
Proteobacteria	Rhodospirillum	79	-3.03|84.36|23.56	Proteob	Rhizobium	65	-4.07|77.84|27.51	Strepto	Oryza	43	-5.91|91.31|26.31
Bacteroidetes	Porphyromonas	73	-3.1|80.45 24.62	Proteob	Haemophilus	64	-3.42|87.35|23.71	unclass	Galdieria	43	-4.57|91.31|23.57
Streptophyta	Flaveria	71	-4.5|80.31 27.26	Proteob	Herminiimonas	63	-6.76|91.32|28.72	Cyanob	Prochlorothrix	39	-4.62|88.07|25.01
unclassified (de	Heterosigma	71	-4.61|90.57|24.53	Proteob	Leptothrix	63	-3.6|84.54 25.05	Bactero	Porphyromonas	37	-3.81|81.25|26.27
Chromerida	unclassified (deriv	69	-4.57|84.6 24.39	Firmicu	Bacillus	62	-4.15|76.18|28.22	Proteob	Neisseria	37	-3.66|73.39|28.01
Actinobacteria	Kineococcus	65	-3.4|83.46 24.87	unclass	Rhodomonas	62	-7.05|91.5 28.3	Strepto	Megaleranthis	37	-4.1|98.33 21.5
Proteobacteria	Azoarcus	64	-5.73|86.38|27.19	Strepto	Flaveria	60	-4.13|86.57|25.43	Actinob	Micrococcus	36	-3.39|78.8 25.91
Proteobacteria	Rickettsia	63	-4.05|87.13|24.75	unclass	Durinskia	60	-3.71|92.8 22.23	Strepto	Mesostigma	36	-4.58|84.04|26.51
Streptophyta	Triticum	57	-3.38|90.7 21.82	Strepto	Spinacia	59	-3|82.4 23.52	Proteob	Rickettsia	35	-3.73|92.16|22.83
Cyanobacteria	Acaryochloris	56	-4.02|88.42|22.5	Chrome	unclassified (der	56	-5.58|92.9 25.15	Proteob	Rhodobacter	33	-3.58|78.02|26.2
Streptophyta	Mesostigma	56	-4.19|84.57|25.39	Chlorop	Leptosira	55	-3.73|94.34|22.09	Proteob	Acinetobacter	33	-5.33|81.97|27.5
Actinobacteria	Mycobacterium	54	-6.05|83.53|30.34	unclass	Emiliania	54	-4.57|87.01|25.09	Proteob	Vibrio	33	-4|86.55|24.14
Firmicutes	Bacillus	54	-4.79|77.17|28.26	Proteob	Bordetella	53	-4.53|83.84|27.01	unclass	unclassified (derive	32	-5.09|83.5 27.83
Proteobacteria	Leptothrix	54	-3.66|82.64|25.04	Proteob	Pseudomonas	52	-4.96|78.22|28.51	Strepto	Zea	31	-4.57|80.14|27.62
Proteobacteria	Candidatus Pelagi	53	-3.35|83.82|24.2	Spiroch	Leptospira	52	-4.09|77.82|26.82	Cyanob	Acaryochloris	28	-3.92|85.06|23.68
Streptophyta	Pisum	52	-3.1|87.09 22.74	Strepto	Mesostigma	52	-3.97|85.86|24.76	Cyanob	Pseudanabaena	27	-6.07|84.94|31.67
unclassified (de	unclassified (deriv	52	-4.48|78.89|27.32	Proteob	Shigella	50	-3.24|88.93|22.77	Proteob	Leptothrix	27	-3.12|82.47|24.61
unclassified (de	Emiliania	51	-3.95|88.4 23.9	unclass	Microvirus	49	-5.37|97.28|24.32	unclass	Kryptoperidinium	27	-4.28|85.81|24.61
unclassified (de	T4-like viruses	50	-3.8|81.19 26.21	Firmicu	Clostridium	48	-3.98|76.64|28.12	Actinob	Mycobacterium	26	-5.99|83.68|29.36
Chlorophyta	Scenedesmus	47	-5.91|82.7 28.72	Proteob	Bradyrhizobium	48	-6.55|81.74|30.15	Proteob	Haemophilus	26	-4.35|77.98|28.63
Streptophyta	Platanus	46	-3.35|93.8 22.59	Proteob	Maricaulis	48	-3.2|83 24.83	Actinob	Clavibacter	25	-4.43|81.18|26.87
Streptophyta	Marchantia	44	-5.69|91.95|24.98	Strepto	Physcomitrella	47	-4.19|88.85|24.17	Cyanob	Fischerella	25	-7.97|91.36|28.39
Streptophyta	Chaetosphaeridiu	43	-6.27|90.56|26.15	Actinob	Beutenbergia	44	-3.28|83.37|24.93	Proteob	Shewanella	25	-3.65|83.19|24.89
unclassified (de	Paulinella	42	-3.3|85.64 24.37	Strepto	Solanum	44	-3.49|93.72|22.25	Actinob	Corynebacterium	24	-4.69|73.26|29.56
Chlorophyta	Chlorella	41	-4.89|89.21|25.11	Strepto	Psilotum	43	-3|87.64|21.92	Cyanob	Lyngbya	24	-5.08|80.3 28.82

8am (Krb4)			average	1pm (Krb5)			average	3pm (Krb6)			average
phylum	genus	No.	E |%| len	phylum	genus	No.	E |%| len	phylum	genus	No.	E |%| len
Cyanobacteria	Microcystis	107154	-6.75|95.3 27.16	Cyanobacteria	Microcystis	162555	-6.55|94.8 27.07	Cyanobacteria	Microcystis	130095	-6.77|94.97|27.26
Cyanobacteria	Synechocystis	29279	-3.93|86.04|24.72	Cyanobacteria	Synechocystis	47520	-4.29|87.86|25.1	Cyanobacteria	Synechocystis	36449	-4.78|86.29|25.13
unassignec	unassignec	22762	-4.22|83.51|25.88	unassignec	unassignec	13899	-4.8|85.72 26.41	Cyanobacteria	Nostoc	24433	-4.88|81.76|27.79
Cyanobacteria	Nostoc	14857	-4.51|81.49|27.04	Cyanobacteria	Synechococcu	12504	-5.17|88.77|26.12	unassignec	unassignec	22591	-4.43|84.62|26
Cyanobacteria	Cyanothece	14072	-3.05|84.59|23.52	Cyanobacteria	Anabaena	8235	-5.24|86.73|26.44	Cyanobacteria	Synechococcus	13939	-5.44|90.76|25.78
Cyanobacteria	Trichodesmium	13106	-3.86|82.25|25.54	Cyanobacteria	Thermosynech	5352	-3.3|86.15 23.07	Cyanobacteria	Anabaena	7032	-6.17|87.63|27.84
Cyanobacteria	Synechococcus	9617	-6.14|89.3 27.5	Cyanobacteria	Trichodesmiun	4562	-5.45|77.13|29.07	Cyanobacteria	Thermosynechoco	4247	-3.96|86.58|24.17
Cyanobacteria	Anabaena	5226	-5.35|85.47|26.97	Cyanobacteria	Nostoc	3318	-5.21|81.59|28.17	Cyanobacteria	Trichodesmium	3685	-3.17|80.89|25.5
Cyanobacteria	Thermosynechoc	2982	-3.12|87.21|22.84	Cyanobacteria	Acaryochloris	1251	-3.28|83.66|24.04	Proteobacteria	Burkholderia	1117	-3.57|85.71|24.4
Cyanobacteria	Prochlorococcus	1552	-4.19|88.54|24.29	Cyanobacteria	Cyanobium	1110	-3.09|85.06|23.63	Bacillariophyta	Thalassiosira	949	-5.56|92.35|25.06
Proteobacteria	Burkholderia	1545	-3.63|86.62|24.21	Proteobacteria	Burkholderia	725	-3.94|87.26|24.8	Cyanobacteria	Acaryochloris	886	-3.02|78.76|23.72
Cyanobacteria	Arthrospira	591	-4.94|82.69|27.02	Cyanobacteria	Arthrospira	648	-4.54|80.09|27.34	Cyanobacteria	Arthrospira	616	-5.23|82.74|28.63
Cyanobacteria	Microchaete	496	-3.22|86.95|23.56	Bacillariophyta	Thalassiosira	610	-7.09|93.22|27.94	Bacillariophyta	Odontella	591	-4.24|92.81|22.47
Proteobacteria	Cupriavidus	438	-3.66|81.07|25.87	unclassified (de	Guillardia	580	-3.98|92.3 22.53	unclassified (de	unclassified (deriv	542	-7.71|92.41|29.04
unclassified (de	Cyanidium	325	-8.15|94.41|28.21	Cyanobacteria	Prochlorococc	461	-4.4|87.41 25.51	unclassified (de	Cyanidium	537	-5.5|93.26 24.24
Cyanobacteria	unclassified (deri	315	-6.1|87.88 26.94	unclassified (de	unclassified (d	456	-5.2|95.49 24.19	unclassified (de	Pulvinaster	536	-5.05|95.8 24.65
Cyanobacteria	Dolichospermum	301	-7.05|94.25|29.25	Proteobacteria	Escherichia	435	-3.86|92.43|23.24	Cyanobacteria	Dolichospermum	473	-6.89|90.24|29.31
Proteobacteria	Bordetella	296	-4.11|83.31|25.96	unclassified (de	Pulvinaster	384	-4.09|93.22|23.91	Bacillariophyta	Phaeodactylum	472	-4.33|91.43|23.29
Bacillariophyta	Thalassiosira	289	-6.92|91.9 27.54	Cyanobacteria	unclassified (d	375	-4.38|84.84|27.37	unclassified (de	Guillardia	432	-4.49|89.8 23.96
Streptophyta	Pinus	253	-5.12|96.68|22.8	Streptophyta	Pinus	370	-4.23|92.37|22.43	Cyanobacteria	Mastigocladus	429	-4.01|79.85|25.55
unclassified (de	Guillardia	242	-3.4|87.69 22.98	Proteobacteria	Cupriavidus	366	-3.35|80.47|25.66	Phaeophyceae	Dictyota	404	-5.38|92.37|24.16
Proteobacteria	Ralstonia	237	-3.23|82.99|24.34	unclassified (de	Cyanidium	357	-5.56|91.17|25.07	unclassified (de	unclassified (deriv	394	-5.82|82.36|28.99
unclassified (de	Pulvinaster	231	-5.83|95.75|25.67	Bacillariophyta	Phaeodactylun	348	-4.22|93.34|23.01	Streptophyta	Pinus	375	-6.58|93.92|25.57
unclassified (de	unclassified (deri	204	-7.46|94.74|28.07	Phaeophyceae	Dictyota	330	-4.91|87.75|26.65	Cyanobacteria	unclassified (deriv	369	-4.38|87.98|25.47
Cyanobacteria	Prochlorothrix	198	-3.9|88.46 23.09	Bacillariophyta	Odontella	319	-5.08|92.75|23.81	Proteobacteria	Escherichia	331	-3.47|92.96|22.41
Bacillariophyta	Odontella	181	-4.29|90.78|23.34	unclassified (de	unclassified (d	315	-5.17|84.2 26.65	Cyanobacteria	Prochlorothrix	313	-3.62|89.44|22.21
unclassified (de	Galdieria	173	-6.51|92.14|26.15	Cyanobacteria	Mastigocladus	311	-3.12|71.92|26.74	Proteobacteria	Polynucleobacter	262	-3.28|86.65|23.36
Proteobacteria	Neisseria	163	-3.37|83.77|24.1	Cyanobacteria	Prochlorothrix	291	-4.31|89.71|23.33	unclassified (de	Durinskia	244	-4.7|92.84 23.83
Chlorophyta	Leptosira	150	-3.27|89.82|21.78	unclassified (de	Galdieria	262	-6.01|89|26.38	unclassified (de	Galdieria	237	-4.86|91.16|24.27
unclassified (de	unclassified (deri	148	-4.37|88.03|25.1	Chlorophyta	Oedogonium	226	-5.64|93.96|25.05	Chromerida	unclassified (deriv	216	-4.81|92.02|23.04
Proteobacteria	Acidovorax	144	-3.48|86.03|24.43	Proteobacteria	Acidovorax	211	-3.23|85.58|23.89	Cyanobacteria	Fischerella	213	-4.79|84.9 26.69
Bacteroidetes	Cytophaga	134	-3.52|77.99|27.8	Chlorophyta	Leptosira	210	-4.29|90.9 23.71	unclassified (de	T4-like viruses	213	-4.77|87.24|24.61
Bacillariophyta	Phaeodactylum	131	-4.67|91.95|24.06	Streptophyta	Xylolejeunea	209	-3.26|96.05|21.47	unclassified (de	Porphyra	195	-7.49|92.25|28.15
Cyanobacteria	Geminocystis	123	-3.69|85.37|24.1	unclassified (de	Rhodomonas	208	-4.82|89.34|25.06	unclassified (de	Kryptoperidinium	194	-3.74|89.07|22.8
unclassified (de	unclassified (deri	121	-4.85|76.79|29.09	unclassified (de	Gracilaria	204	-4.13|88.28|24.07	Cyanobacteria	Leptolyngbya	185	-4.99|88.47|26.66
unclassified (de	T4-like viruses	119	-3.35|87.79|22.68	Chromerida	unclassified (d	201	-5.39|92.61|24.34	unclassified (de	unclassified (deriv	180	-5.61|88.53|25.84
unclassified (de	unclassified (deri	115	-4.97|83.37|27.55	unclassified (de	unclassified (d	199	-4.44|92.79|24.45	Cyanobacteria	Geminocystis	173	-3.84|88.61|23.88
Streptophyta	Marchantia	113	-4.71|86.96|25.79	Streptophyta	Marchantia	188	-3.49|89.53|21.75	Cyanobacteria	Planktothrix	173	-6.14|87.6 29.09
Proteobacteria	Escherichia	110	-3.94|89.33|23.51	Chlorophyta	Stigeoclonium	174	-3|90.64|22.5	Proteobacteria	Ralstonia	172	-4.37|87.01|24.95
unclassified (de	Rhodomonas	108	-5.1|87.48 25.26	unclassified (de	T4-like viruses	164	-4.35|89.28|23.44	Bacteroidetes	Flavobacterium	170	-5.41|89.7 26.07
Planctomycetes	Rhodopirellula	104	-5.3|76.21 29.64	Streptophyta	Solanum	160	-4.82|86.88|26.35	Firmicutes	Bacillus	169	-3.72|85.12|24.86
unclassified (de	Porphyra	98	-4.39|86.67|25.13	unclassified (de	Durinskia	160	-5.18|94.18|24.01	Chlorophyta	Nephroselmis	153	-3.6|89.69 22.1
Proteobacteria	Methylobacillus	96	-3.55|80.58|26.09	Cyanobacteria	Calothrix	155	-7.25|92.01|28.38	Cyanobacteria	Phormidium	152	-7.78|87.42|31.16
Proteobacteria	Chromobacterium	96	-3.34|84.96|23.76	Proteobacteria	Pseudomonas	153	-5.68|82.68|29.43	unclassified (de	Heterosigma	142	-5.93|88.21|28.3
Proteobacteria	Bradyrhizobium	92	-3.18|79.88|25.08	Cyanobacteria	Planktothrix	146	-7.15|88.44|30.18	Bacteroidetes	Bacteroides	139	-4.26|83.11|26.22
unclassified (de	Microvirus	90	-4.61|98.38|22.58	Proteobacteria	Neisseria	134	-3.74|83.8 25.12	Proteobacteria	Thiomonas	135	-4.91|86.21|26.34
Phaeophyceae	Dictyota	87	-6.07|84.89|29.73	Cyanobacteria	Microchaete	125	-3.7|82.66 25.2	Proteobacteria	Pseudomonas	134	-4.31|81.16|26.55
Proteobacteria	Thiobacillus	86	-3.11|83.86|24.18	unclassified (de	Kryptoperidini	124	-3.71|87.29|23.02	Cyanobacteria	Prochlorococcus	127	-6.91|90.57|28.14
Cyanobacteria	Planktothrix	83	-5.18|84.13|27.78	Streptophyta	Mesostigma	120	-3.63|87.63|22.89	unclassified (de	Gracilaria	127	-4.99|80.18|27.78
Chlorophyta	Oedogonium	83	-6|92.66|26.06	Cyanobacteria	Dolichospermu	114	-6.94|90.45|29.28	Chlorophyta	Chlamydomonas	126	-5.81|81.68|29.44
Proteobacteria	Comamonas	82	-4.31|85.73|25.26	Bacteroidetes	Cytophaga	111	-3.58|82.77|24.99	Chlorophyta	Scenedesmus	125	-4.97|85.68|27.57
Proteobacteria	Methylibium	77	-4.04|84.75|25.68	Cyanobacteria	Fischerella	106	-5.68|89.43|27.5	Phaeophyceae	Rugulopteryx	125	-5.36|97.69|23.57
Cyanobacteria	Leptolyngbya	74	-6.17|87.66|28.22	Bacteroidetes	Bacteroides	104	-3.77|80.99|25.86	Bacteroidetes	Cytophaga	122	-3.21|81.97|24.61
Proteobacteria	Nitrosospira	74	-4.66|85.43|28.23	Proteobacteria	Ralstonia	100	-3.47|84.39|24.42	Streptophyta	Mesostigma	121	-3.92|85.94|24.18
unclassified (de	Durinskia	74	-6.64|91.96|26.94	Proteobacteria	Shigella	97	-3.26|90.66|22.8	Streptophyta	Populus	104	-3.38|79.05|26.62
unclassified (de	Gracilaria	74	-4.6|85.43 24.71	unclassified (de	Dinophysis	95	-6.44|96.44|25.86	Proteobacteria	Neisseria	99	-3.51|83.27|24.26
Cyanobacteria	Calothrix	73	-7.1|95 28.2	Bacteroidetes	Porphyromona	93	-3.45|81.97|24.72	unclassified (de	Cyanidioschyzon	95	-6.35|88.59|28.27
Proteobacteria	Azoarcus	71	-3.89|80.9 26.5	Proteobacteria	Xanthomonas	93	-5.61|85.94|27.78	Proteobacteria	Shigella	90	-3.7|90.24 23.56
Streptophyta	Staurastrum	68	-4.7|89.05 24.29	Chlorophyta	Chlamydomon	93	-5.23|84.08|27.19	unclassifiec {de	unclassifiec {deriv	86	-5.61|83.62|28.11
Proteobacteria	Verminephrobac	67	-3.2|78.87 25.94	unclassified (de	unclassified (d	93	-5.43|93.48|24.68	unclassifiec {de	unclassifiec {deriv	86	-5.82|82.19|29.39
Bacteroidetes	Gramella	66	-3.11|80.49|24.65	unclassified (de	Emiliania	89	-3.51|88.36|22.87	Streptophyta	Cryptomeria	82	-3.83|90.98|21.67
unclassified (de	Cyanidioschyzon	66	-5.59|91.66|24.91	Proteobacteria	Bordetella	88	-5.44|87.64|26.99	Streptophyta	Oryza	81	-3.34|95.27|21.49
Chlorophyta	Chlorella	65	-6.67|84.39|25.42	Firmicutes	Bacillus	86	-3.51|81.57|25.53	Proteobacteria	Rhodospirillum	80	-3.86|80.82|26.39
Streptophyta	Glycine	65	-4.02|88.31|22.86	Cyanobacteria	Pseudanabaen	85	-6.69|90.38|28.25	unclassified (de	Aureococcus	79	-3.82|93.81|22.23
Bacteroidetes	Flavobacterium	64	-3.66|87.99|24.16	Streptophyta	Calycanthus	85	-6.38|95.24|25.99	Proteobacteria	Methylobacillus	76	-3.04|86.42|23.21
Chromerida	unclassified (deri	64	-4.34|91.7 21.38	Proteobacteria	Nitrosomonas	81	-3.26|84.2 24.24	Chlorophyta	Chlorella	73	-5.2|88.58 25.21
Streptophyta	Oryza	64	-3.25|94.37|21.78	Cyanobacteria	Prochloron	79	-4.75|82.44|27.11	Bacteroidetes	Gramella	71	-3.48|83.66|24.55
Proteobacteria	Rhodopseudomo	63	-3.64|83.67|24.87	Streptophyta	Glycine	79	-3|84.47|22.18	unclassified (de	unclassified (deriv	71	-4.86|95.84|23.44
Proteobacteria	Acidiphilium	63	-3.09|77.96|25.19	unclassified (de	Teleaulax	79	-6.61|99.81|24.64	Bacteroidetes	Porphyromonas	66	-3.44|82.25|25.38
unclassified (de	Emiliania	62	-5.22|86.2 26.6	Streptophyta	Oryza	77	-5.56|93.58|25.36	Proteobacteria	Acidovorax	65	-5.06|76.85|29.25
Actinobacteria	Mycobacterium	60	-4.7|80.72 28.68	Proteobacteria	Campylobacte	76	-3.93|85.66|24.81	Proteobacteria	Methylibium	65	-4.17|83.25|26.33
Proteobacteria	Xylella	59	-3.16|82.7 22.59	Chlorophyta	Nephroselmis	76	-3.38|89.22|22.54	Streptophyta	Staurastrum	64	-6.75|92.7 26.78
Cyanobacteria	Phormidium	58	-4.99|84.1 27.25	Streptophyta	Trifolium	75	-3.47|96.61|19.67	unclassified (de	Emiliania	64	-3.22|84.91|23.22
Bacillariophyta	Skeletonema	57	-8.92|86.65|32.11	Chlorophyta	Scenedesmus	74	-3.74|79.09|25.98	Proteobacteria	Nitrosospira	63	-3.38|84.98|24.21
unclassified (de	Paulinella	57	-3.05|86.26|23.49	Streptophyta	Triticum	74	-3.25|90.08|22.83	Bacillariophyta	Skeletonema	63	-4.38|92.55|24.25
Chlorophyta	Scenedesmus	55	-4.16|80.85|26.12	Streptophyta	Staurastrum	71	-4.43|91.07|23.56	Proteobacteria	Rickettsia	62	-4.07|87.32|24.54
Proteobacteria	Brucella	54	-3.37|82.2 25.33	Chlorophyta	Chlorella	70	-5.83|91.42|26.24	Bacteroidetes	Parabacteroides	61	-3.5|81.49 25.7
Streptophyta	Mesostigma	54	-3.24|82.89|24.41	Proteobacteria	Brucella	69	-4.47|90.72|25.19	Streptophyta	Platanus	61	-3.58|95.13|21.37
Streptophyta	Cupressus	52	-4.93|88.25|25.66	Phaeophyceae	Rugulopteryx	68	-4.73|97.39|22.8	Streptophyta	Hernandia	58	-10 92.84|32.33
Cyanobacteria	Mastigocladus	51	-6.31|88.12|27.83	Bacteroidetes	Flavobacterium	65	-5.02|82|27.38	Proteobacteria	Rhodopseudomon	57	-3.3|83.02 24.22
Cyanobacteria	Prochloron	50	-4.47|83.53|26.91	Proteobacteria	Methylibium	65	-3.87|87.27|24.52	Proteobacteria	Azoarcus	57	-5.61|89.54|26.68

3.3.1.5 Selection of appropriate E-value cutoff for taxonomic and functional annotation in MG-RAST

E-value is a metric that describes the likelihood of recovering a match of one sequence to another by chance alone. The smaller the E-value (or the more negative the log of the E-value) is for a given annotation, the more confidence there is in saying that the match is due to a similar evolutionary origin of the two sequences rather than due to chance. Therefore, when annotating environmental sequences against a particular database, choosing an appropriate E-value cutoff is important as it determines the confidence level in assigning taxonomic affiliations or functional roles to the sequences. Hence, if one sequence has previously been confidently annotated in a database, then the same annotation can be applied to the other matching "unknown" sequences from an environmental sample with a level of confidence constrained by the magnitude of the E-value.

Making high-confidence annotation for short-read sequences such as those used in this study is a challenge because E-value can be strongly influenced by the sequence length: shorter sequences have a higher probability of false-positive matches. In addition, the confidence of annotation is also affected by the availability of reference genomes or sequences in a database. For this reason previous metagenomic/metatranscriptomic studies have used E-value cutoff thresholds of <1E-3 (metatranscriptomic study with average length of 272 bp; Gilbert et al., 2010) and <1E-5 (Poroyko et al., 2010).

Since our sequences are short (90-100 bp) and the environment has not been characterized genomically before (i.e. few relevant reference genomes), we wanted to explore the range of reasonable confidence estimates that would return annotation of the bacterioplankton community that agreed well with the well-established observation that the Cyanobacterium *Microcystis aeruginosa* makes up the majority of Kranji Reservoir plankton. For example, Te & Gin (2011) have shown that within Cyanobacteria, *Microcystis* can be 100 fold higher in abundance than the Cyanobacterium *Anabaena.*

RDP (Ribosomal Database Project) is a well-curated large database of ribosomal reads and the hits in this database have very high quality annotation (E-value < 1E-20, Fig. 12). Even though the samples have undergone rRNA subtraction, the depletion protocol is not 100%

efficient and assuming that the remaining rRNA is reflective of the community richness (although distribution might be skewed due to biases in depletion), RDP can be used as a gauge of the percentage of *Microcystis* to be expected from MG-RAST (M5NR) taxonomic assignment. Table 8 shows that the majority (75.4%) of rRNA reads are from *Microcystis*, thus it is reasonable to extrapolate that the majority of total RNA are also likely to be from *Microcystis*.

Annotation outcomes from CLC Genomics Workbench, MG-RAST (RDP) based on annotation of ribosomal sequences, and MG-RAST (M5NR) based on annotation of total transcripts were compared at different confidence thresholds to evaluate the percentage of reads being annotated as *Microcystis* in the Kranji Reservoir metatranscriptome sample krb1 (5pm) (Table 8). MG-RAST (RDP) analysis reveals that 75.4 % of the residual ribosomal sequences are from *Microcystis* (E < 1E-20) while 38.9 % of total transcripts were aligned to the reference genome of *Microcystis aeruginosa* NIES-843 in CLC Genomics Workbench based on an alignment criteria of > 80 % nucleotide identity and > 90 % alignment length (Table 8). It would be justifiable to pick an E-value which will result in a similar percentage of reads being assigned as *Microcystis*.

An E-value cutoff of < 0.001, and a minimum of 17 amino acids alignment by MG-RAST (M5NR) resulted in 60.7% of total annotated transcripts being assigned as *Microcystis* reads. In contrast, increasing the E-value cutoff to < 1 resulted in 15.8% of total annotated transcripts being assigned as *Microcystis* reads (Table 8).

Thus, under the guidelines of 75.4% and 38.9% of *Microcystis* reads in RDP and CLC respectively, E < 0.001 is likely to be a good parameter for taxonomic or functional analysis focusing on *Microcystis*. However, such threshold (E < 0.001) only results in 20.6% of total post-QC reads to be assigned to domains (Table 9). This means that around 80% of the reads from the 5pm sample would not be used for inferring community taxonomy. On the other hand, a threshold of E-value of < 1 with a minimum of 17 amino acids alignment by MG-RAST (M5NR) resulted in 72.1% of total post-QC reads to be assigned to domains.

Therefore, there needs to be a good balance between annotation confidence (E-value choice) and annotation efficiency (total number of annotated reads). Most of the reads lost in a more

stringent E-value of < 0.001 than E-value of < 1 are from non-*Microcystis* reads as the number of hits assigned to domains dropped from 708,114 to 183,846 while the number of hits assigned as *Microcystis* only dropped slightly from 119,341 to 116,189 (Table 9).

To obtain a more accurate estimate of the expected percentage of *Microcystis* reads in Kranji Reservoir metatranscriptomes, a possible future approach will be using BLAST to determine the relative abundance of all *Microcystis* reads for housekeeping genes that 1) can be confidently assigned and 2) have not been subjected to subtraction protocols like rRNA (e.g. rpoB, rpoA, hspbO or recA). Appropriate E-values that yield similar percentage as this more accurate percentage can therefore be determined and used for total metatranscriptome taxonomic and functional annotation. Moreover, as part of the future study, additional reference sequences would be needed for high-confidence annotation of the remaining Kranji transcriptomes. To get preliminary insights we used both E-values cutoffs of < 1 and < 0.001.

Table 8: Percent of reads annotated as *Microcystis* for sample krb1 (5pm) using different annotation databases and thresholds.

	MG-RAST (RDP) *	CLC Genomics Workbench *	MG-RAST (M5NR) e-value < 1*	MG-RAST (M5NR) e-value < 0.001*
No. of reads assigned as *Microcystis*	18,337	380,151	119,459	116,189
Total no. of reads annotated	24,313	977,552	757,110	191,377
% reads assigned as *Microcystis*	75.4	38.9	15.8	60.7

* The thresholds for MG-RAST (RDP) is e-value < 1E-20, minimum of 50nt aligned; for CLC Genomics Workbench is 90% alignment length, 80% identity; for MG-RAST (M5NR) is e-value < 1 and e-value < 0.001 with minimum of 17 amino acids aligned respectively.

Table 9: Taxonomic read assignment for sample krb1 (5pm) at different e-values, different minimum alignment lengths and no threshold for minimum identity.

E-value	Min. amino acid alignment length	No. of *Microcystis* hits	No. of hits assigned to domains	Total no. of hits at this threshold	Total Post-QC reads without threshold	% of hits assigned to *Microcystis*	% of hits assigned to domains	% total Post-QC reads assigned to domains
1	17	119,341	708,114	757,110	894,579	15.8	85.2	72.1
0.1	17	119,310	462,083	556,437	894,579	21.4	83.0	51.7
0.01	17	118,893	386,232	468,788	894,579	25.4	82.4	43.2
0.001	17	116,189	183,846	191,377	894,579	60.7	96.1	20.6
0.0001	17	1,447	6,878	7,443	894,579	19.4	92.4	0.8
0.00001	17	512	3,148	3,305	894,579	15.5	95.2	0.4
1	0	119,459	708,114	823,568	894,579	14.5	86.0	79.2

3.3.2 Preliminary analysis of community-level functional capacity

Community taxonomic analysis from the previous section attempts to answer the question of "what types of microorganisms are present in the Kranji Reservoir?" However, to address questions like 1) What are the activities of these organisms? And 2) How do microbial activities change during a day-night cycle? The function of the expressed genes can be studied. Sequence analysis of expressed mRNA thus provides both taxonomic information of the microbial community (section 3.3.1) and functional information (preliminary analysis in this section). Changes in expressed gene profiles between samples may be attributed to changes in the structure (taxonomic composition) of the community or to changes in the activities of a resident set of taxa.

3.3.2.1 KEGG pathways

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an integrated bioinformatic database resource that is used as a reference knowledge base for biologic interpretation of large-scale data sets generated by sequencing and other high-throughput experimental technologies (Kanehisa et al., 2004). The KEGG pathway database records molecular interaction and reaction networks for the six Level 1 pathways: Metabolism (66,794 hits), Genetic information processing (15,506 hits), Environmental information processing (7,625 hits), Cellular processes (2,884 hits), Organismal systems (1,842 hits) and Human diseases (2,790 hits) (Fig. 28). Each of these Level 1 pathways braches into multiple Level 2 pathways (Fig. 29) while the Level 2 pathways are further branched to include 159 Level 3 pathways (Fig. 30). Hence, a hierarchical tree of pathways is formed. Every Level 3 pathway contains enzymes and proteins that are involved in a certain cellular process. Level 3 pathways are progressively grouped together if the processes that they represent come under one network which is responsible for a particular cellular function. Thus, the higher up the pathway is in the hierarchy, the more general the pathway represents in terms of its function.

A very high proportion of reads (72 – 82% of the total) in our study are either not assigned or with no hits. "No hits" reads are reads that are not found in the KEGG annotated database. "Not assigned" means that even though the reads can be found in the database, they do not belong to any of the 159 KEGG pathways.

Number of hits in the three-level hierarchy of KEGG pathways:

Among the KEGG Level 1 pathways, Metabolism has the largest number of hits and contains the highest number of Level 3 pathways (64% of all Level 3 pathways). Within Metabolism, the Level 2 pathways with the highest number of hits are Carbohydrate metabolism (13,175), Energy metabolism (31,061), Nucleotide metabolism (7,701), Amino acid metabolism (12,189) and Cofactor and vitamins metabolism (9,115) (Fig. 29).

The second highest Level 1 pathway in terms of number of hits is Genetic information processing which includes Level 2 pathways like Transcription (1,275), Translation (6,760), Folding, sorting and degradation (4,669) and Replication and repair (2,802). The third highest Level 1 pathway is Environmental information processing which includes Level 2 pathways like Membrane transport (5,207), Signal transduction (2,582) and ECM-receptor interaction (46) (Fig. 29).

Among the 159 KEGG level 3 pathways, the top ten pathways with the largest number of total hits in six samples in decreasing order are Photosynthesis (15,411), Purine metabolism (5,423), Ribosome (5,088), Pyrimidine metabolism (4,823), Oxidative phosphorylation (4,589), ABC transporters (4,047), Photosynthesis-antenna proteins (3,987), Glycolysis/ Gluconeogenesis (3,966), Carbon fixation in photosynthetic organisms (3,858) and Porphyrin and chlorophyll metabolism (3,101).

Fig. 28: MEGAN KEGG Level 1 pathways. Each node represents a KEGG term and contains a pie which shows the comparative distribution of the six samples. The size of the pie is proportional to the number of hits at that node.

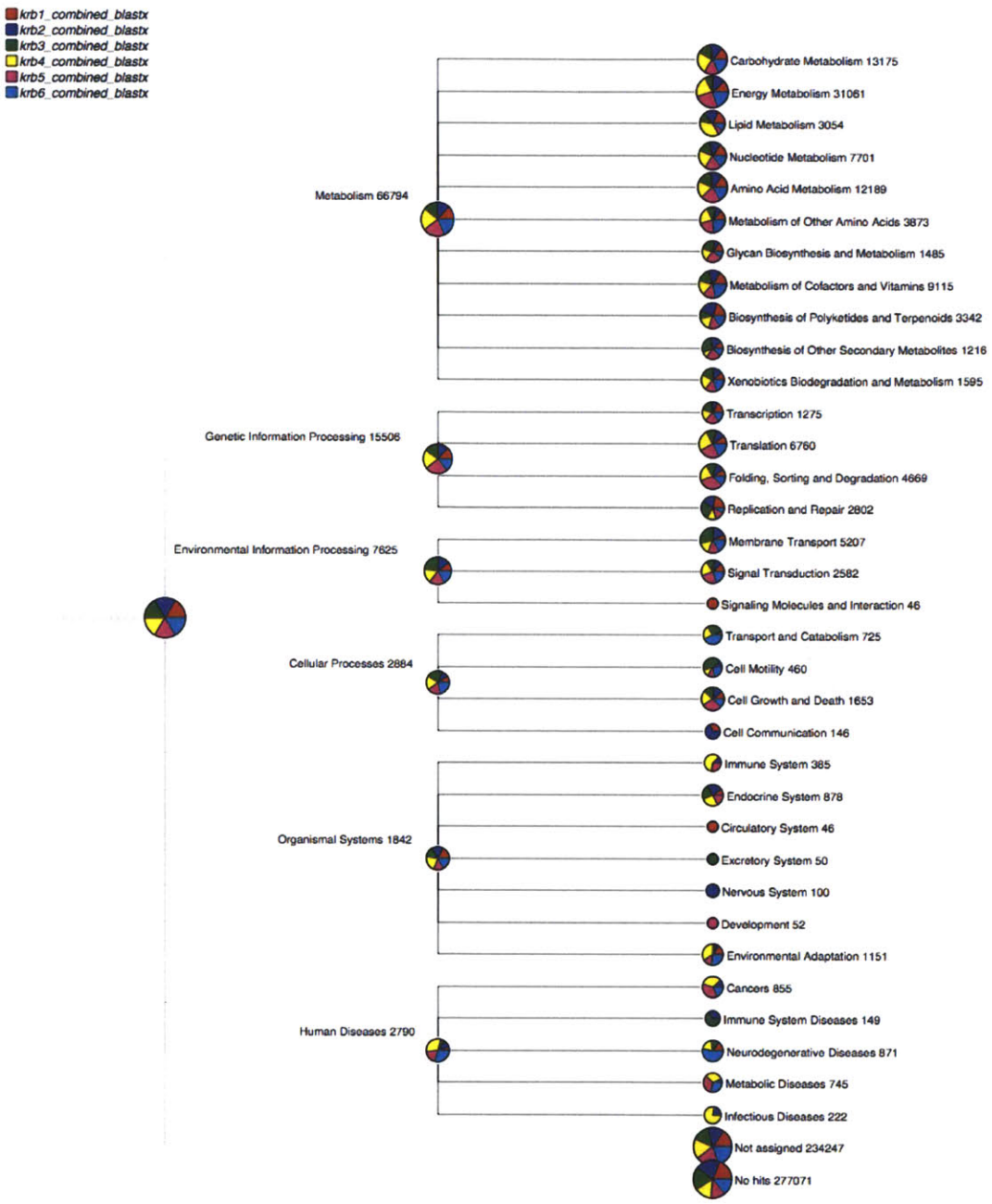


Fig. 29: MEGAN KEGG Level 1-2 pathways. Each node represents a KEGG term at a particular hierarchical level. The higher the level, the more general the term represents. Each node contains a pie which shows the comparative distribution of the six samples at a particular pathway and KEGG Level. The pies at the terminal leaves represent the KEGG level 2 pathways. The size of the pie is proportional to the number of hits at that node.

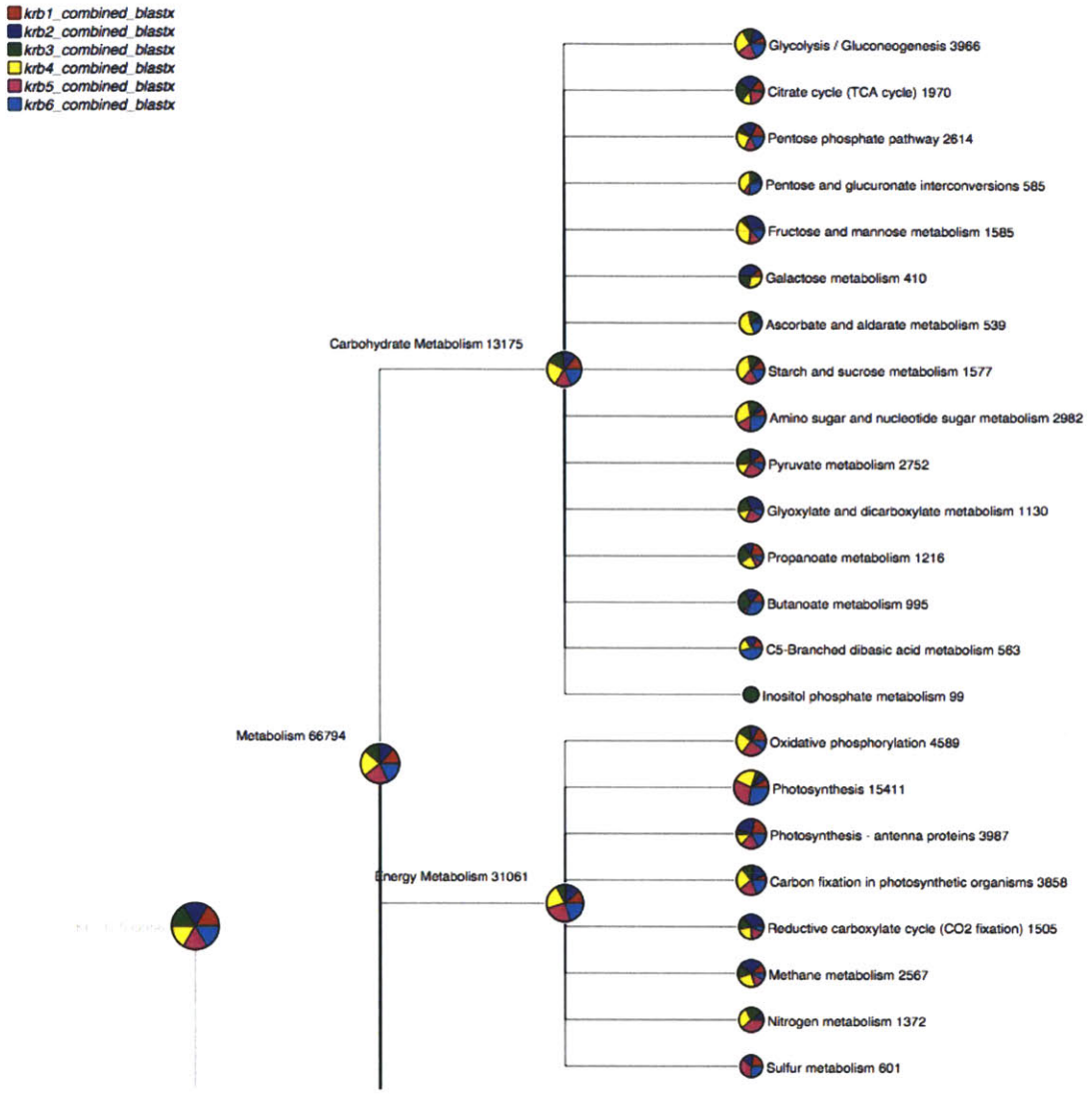


Fig. 30: MEGAN KEGG Level 1-3 pathways (partial list). Each node represents a KEGG term at a particular hierarchical level. The higher the level, the more general the term represents. Each node contains a pie which shows the comparative distribution of the six samples at a particular pathway and KEGG Level. The pies at the terminal leaves represent the KEGG level 3 pathways. The size of the pie is proportional to the number of hits at that node.

Community day-night enrichment of KEGG Level 3 pathways:

Energy metabolism pathways:

Since KEGG Level 1 pathway Metabolism contain the highest abundance of hits, singling out its Level 3 pathways enables their variation across six time points of a day to be put into perspective (Fig. 31).

During the day, the Photosynthesis pathway is up-regulated as many photosynthetic organisms including Cyanobacteria in Kranji Reservoir are actively photosynthesizing. Among all the reads that have hits in the KEGG database, 3.15 %, 4.17 % and 3.82 % of them are mapped to the Photosynthesis pathway in the three day samples (8am, 1pm, 3pm) respectively. The percent of reads mapped drops progressively for the three night samples as it gets deeper into the night (1.23 %, 0.97 % and 0.68 % for 5pm, 8pm, 6am respectively). Hence, the average percent of reads in the Photosynthesis pathway is significantly higher in the day ($3.72 \pm 0.52\%$) than at night ($0.96 \pm 0.27\%$) ($p = 0.004$) (Fig. 31).

Carbon fixation (Calvin cycle) is another pathway that is very active in the day as it is the stage in Photosynthesis that utilizes the energy and reducing power (ATP and NADPH) generated from light harvesting to produce the terminal product, G3P (Glyceraldehyde 3-phosphate). These so called "dark reactions" of photosynthesis can occur in the absence of light. A higher percent of reads are assigned to this pathway in the day samples ($0.75 \pm 0.10\%$) than the night samples ($0.42 \pm 0.20\%$), although this is not statistically significant ($p = 0.08$) (Fig. 31).

The overall metabolic rate is higher in the day as indicated by a greater number of transcripts being assigned to Glycolysis/gluconeogenesis and Oxidative phosphorylation pathways in the day samples. The average percent of reads in Glycolysis/gluconeogenesis pathway is significantly higher in the day (0.80 ± 0.16 %) than at night (0.40 ± 0.09 %) ($p = 0.03$). The average percent of reads in Oxidative phosphorylation pathway decreases from $0.83 \pm 0.35\%$ in the day to $0.56 \pm 0.15\%$ at night, although this is not statistically significant ($p = 0.32$). However, the average percent of reads in Citrate (TCA) cycle is higher at night (0.39 ± 0.12 %) than in the day (0.21 ± 0.17 %), although this is also not statistically significant ($p = 0.21$). The metabolic rate is expected to increase during the day for autotrophs like

Cyanobacteria that gain energy from sunlight, but not necessarily for heterotrophs that gain energy from metabolism of organic compounds. Further work is needed to examine changes separately in these two groups in order to tease apart the functional response of different populations of autotrophs and heterotrophs.

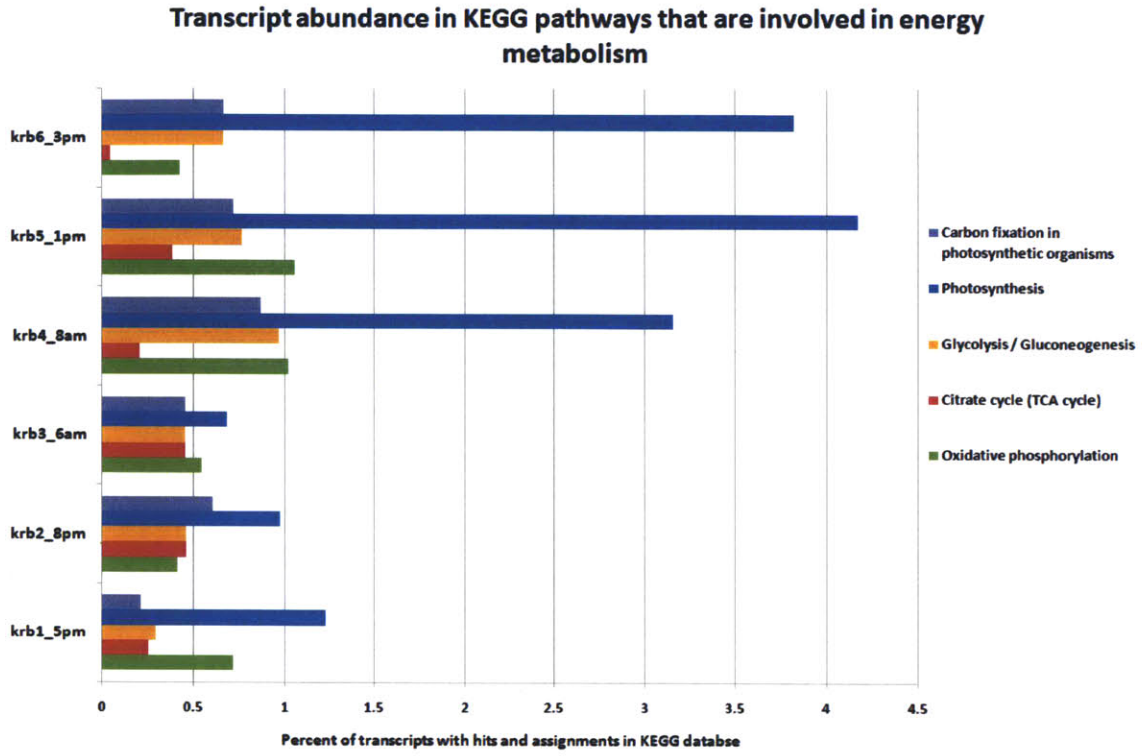


Fig. 31: Six-sample comparison of community level transcript abundance in selected KEGG energy metabolic pathways.

KEGG pathways enriched in the day:

KEGG Level 3 pathways for Ribosome and Sucrose and starch metabolism are up-regulated in the day (Fig. 32a). A significantly higher percent of reads are assigned to the Ribosome pathway in the day samples (1.12 ± 0.28 %) than the night samples (0.42 ± 0.15 %) ($p = 0.02$). For the Starch and sucrose metabolism pathway, there are (0.34 ± 0.14 %) transcripts in the day samples than the night samples (0.13 ± 0.07 %), although this is not statistically significant ($p = 0.09$). The Ribosome pathway is indicative of basic cellular function which is expected to be more active when metabolism rate increases during the day. Moreover, since photosynthesis can only occur when sunlight is available, we would also expect more starch and sucrose being made in the day, hence resulting in higher level of transcripts found in the Sucrose and starch metabolism pathway.

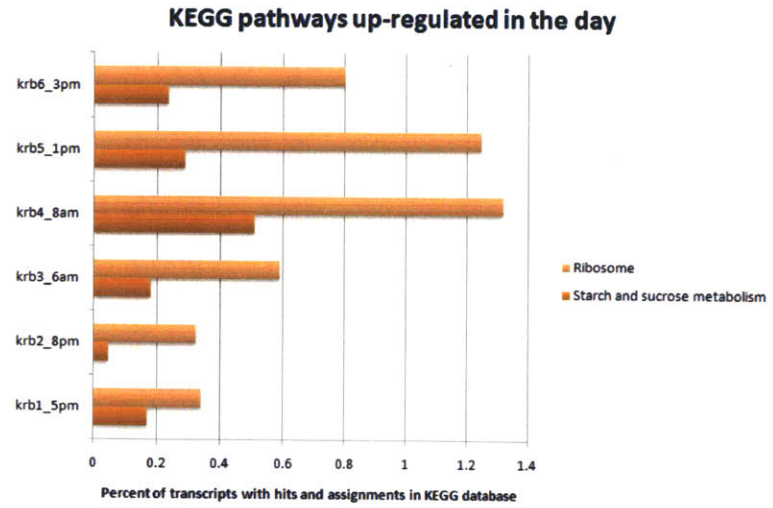
KEGG pathways enriched at night- DNA repair:

KEGG pathways for Nucleotide excision repair, Base excision repair and Mismatch repair are up-regulated at night (Fig. 32b). Comparison between the day and night samples are 0.27 ± 0.04 % versus 0.03 ± 0.06 % ($p = 0.005$); 0.16 ± 0.02 % versus 0.00 ± 0.00% ($p = 0.0003$); 0.26 ± 0.12 % versus 0.13 ± 0.07 % ($p = 0.18$) respectively for these three pathways. In particular, no reads in any of the three day samples are assigned for the Base excision repair pathway. These results support the fact that dark reactions (excision repair) are particularly high at night.

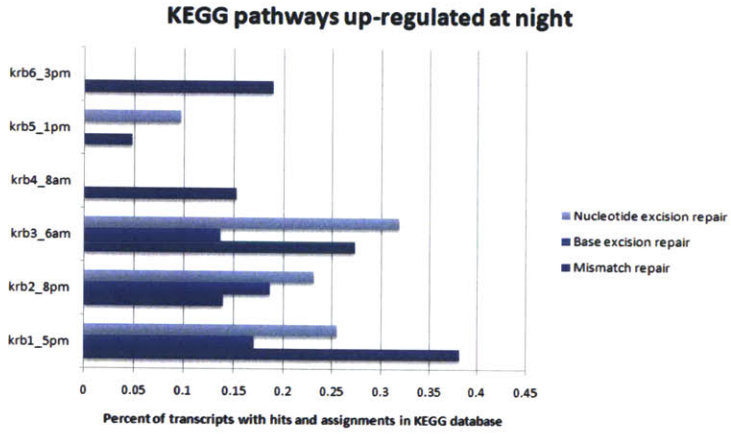
KEGG pathways peaked in the 6am sample:

KEGG pathways for Bacterial secretion system, ABC transporters and Protein export peaked in the 6am sample (Fig. 32c). The peaked and average number of transcripts for each pathway are 0.32 % versus 0.18 ± 0.09 %; 1.09 % versus 0.62 ± 0.28 %; 0.32 % versus 0.19 ± 0.11 % respectively. There is no significant variation of these mechanisms in the day and night samples, indicating that export of substances such as minerals, organic ions, oligosaccharides, peptides and amino acids are occurring around the clock. However, simultaneously peaking at 6am might suggest that great amount of substance export are actually done just before sunrise, perhaps in preparation for the vast array of activities in the day.

a)


KEGG pathways up-regulated in the day
krb6_3pm
krb5_1pm
krb4_8am
krb3_6am
krb2_8pm
krb1_5pm
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Ribosome
Starch and sucrose metabolism
Percent of transcripts with hits and assignments in KEGG database

b)


KEGG pathways up-regulated at night
krb6_3pm
krb5_1pm
krb4_8am
krb3_6am
krb2_8pm
krb1_5pm
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Nucleotide excision repair
Base excision repair
Mismatch repair
Percent of transcripts with hits and assignments in KEGG database

c)

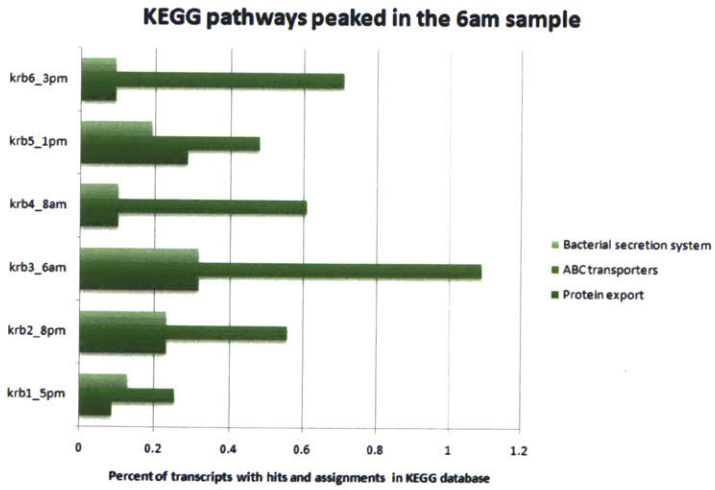

KEGG pathways peaked in the 6am sample
krb6_3pm
krb5_1pm
krb4_8am
krb3_6am
krb2_8pm
krb1_5pm
0
0.2
0.4
0.6
0.8
1
1.2
Bacterial secretion system
ABC transporters
Protein export
Percent of transcripts with hits and assignments in KEGG database

Fig. 32: Six-sample comparison of community level transcript abundance in selected KEGG pathways. a) KEGG pathways for Ribosome and Sucrose and starch metabolism are up-regulated in the day. b) KEGG pathways for Nucleotide excision repair, Base excision repair and Mismatch repair are up-regulated at night. c) KEGG pathways for Bacterial secretion system, ABC transporters and Protein export peaked in the 6am sample.

3.3.2.2 SEED functional classification

Besides KEGG (where transcripts are structured into pathways by their presence in certain pathways), MEGAN also uses the SEED classification (where transcripts are classified into one or more subsystems based on its functional roles) for functional analysis. Transcripts are assigned to functional roles based on the highest scoring gene in a BLAST comparison against a protein database. Then, the different functional roles are grouped into subsystems.

Transcripts in our study are grouped into 569 SEED functional roles. A very high proportion of transcripts (79 – 87% of the total) are either not assigned or with no hits. The top five SEED functional roles with the highest number of transcripts assigned are Photosystem II protein D1 PsbA (0.66 ± 0.67 %), Central carbohydrate metabolism (0.59 ± 0.16 %), Chaperone protein DnaK (0.20 ± 0.25 %), Allophycocyanin beta chain (0.18 ± 0.14 %) and Phycocyanin beta chain (0.15 ± 0.07 %).

In agreement with analysis based on KEGG, the Photosynthesis subsystem contains the highest number of transcripts (13,138) assigned. Transcriptional level of 23 photosynthesis genes in this subsystem is shown in Fig. 33. Variation along the x-axis reflects the set of photosynthesis genes that are particularly active in a certain sample or during a certain time of the day. For example, among the set of 11 genes (*PsaB, PsaF, PsbA, PsbD, PsbC, PsbB, PsbU, PsbH, PetG, Cyt b6, PetD*) that have transcripts present in the krb5 (1pm) sample, 4 genes (*PsbA, PsbD, PsbC, PsbB*) are particularly transcriptionally active or has the highest number of transcripts.

Variation along the y-axis reflects the transcriptional level of a certain gene across the six time points of the day. For example, looking at the six transcriptional levels for a photosynthetic electron transport chain gene, *petD*, no transcripts are found in the 6am sample when it is completely dark. Its transcriptional level starts to pick up in the 8am sample and increases steadily in the 1pm sample, until the 3pm sample, then subsequently in the 5pm sample, it drops again until no further transcripts are found in 8pm. In this way, a 24-hour variation in transcriptional level of photosynthesis genes in the community can be traced around the clock.

There are three photosynthesis genes that have a significantly higher average transcriptional level in the day samples (8am, 1pm, 3pm) than the night samples (5pm, 8pm, 6am). They are *psaF* (encoding for photosystem I subunit III) (0.09 ± 0.03 % versus 0.00 ± 0.00 %, p = 0.005); *psbA* (encoding for photosystem I P700 chlorophyll a apoprotein A1) (1.22 ± 0.41 % versus 0.09 ± 0.09 %, p = 0.009) and *petD* (encoding for cytochrome b6-f complex subunit 4) (0.16 ± 0.05 % versus 0.03 ± 0.05 %, p = 0.034).

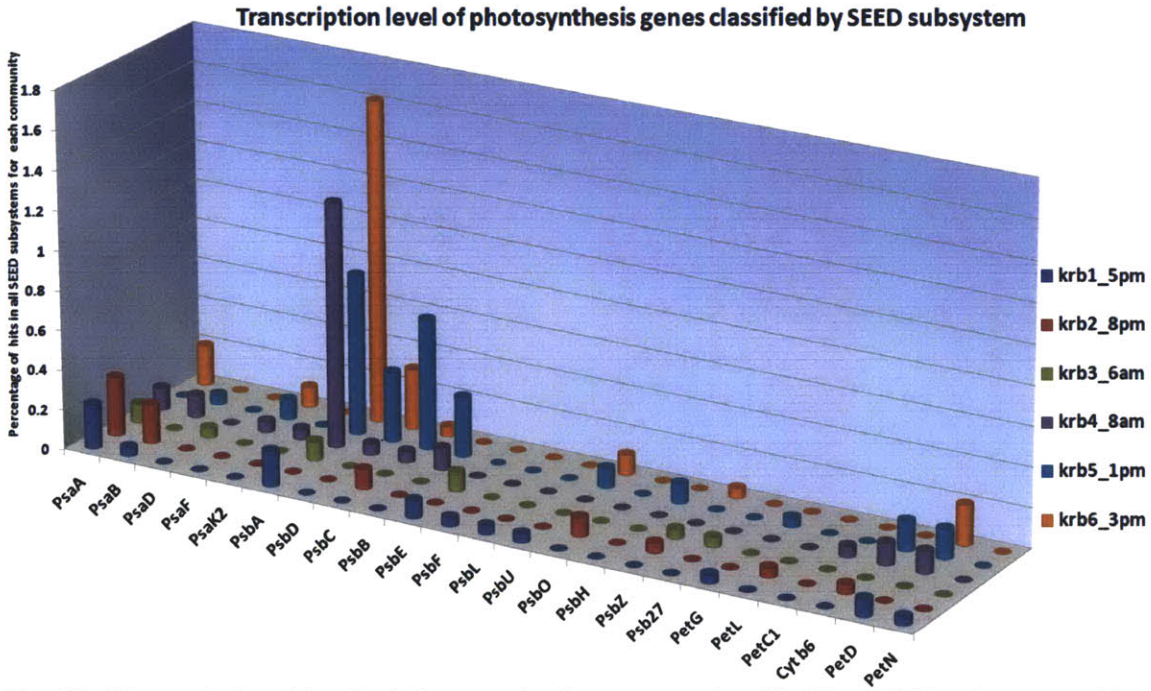


Fig. 33: Transcriptional level of photosynthesis genes as classified by SEED subsystems. Bar length represents the percentage of hits in all SEED subsystems for each community.

3.4 Analysis of gene expression in *Microcystis*-like populations

3.4.1 Sample and gene classification by *Microcystis* gene expression

All reads from each of the six transcriptomes that met the criteria of > 90% alignment of bases and > 80% identity in the aligned bases when mapped to the reference genome (*Microcystis aeruginosa* NIES-843) are assigned as *Microcystis aeruginosa*-like reads. There is a possibility that other closely-related Cyanobacteria might be aligned to the *Microcystis aeruginosa* genome at these criteria, so these analyses may represent the dynamics of *Microcystis aeruginosa*-like populations in the reservoir.

Based on the similarity in gene expression profiles of 6363 genes in the reference genome, hierarchical clustering separated the six transcriptomes into two groups: 5pm, 8pm and 6am samples into one group (Night group) while the 8am, 1pm and 3pm samples into the other (Day group) (Fig. 34). The 1pm and 3pm samples are most similar in overall gene expression and the 5pm and 8pm samples are the next closest pair in terms of similarity and are grouped together, followed by the 6am sample. Such unsupervised classification conforms to the grouping suggested by the timing when samples are collected. This result suggests that the overall transcriptomic profile of *Microcystis aeruginosa* and closely related organisms that match the reference genome follows a diel pattern. Exactly which genes may account for this day-night separation is explored further using statistical tests.

PCA (Principal Component Analysis) of Fig. 35 appears to provide further support for the day-night grouping in hierarchical clustering as the first principal component axis, which explains the highest variance (26%), again separates the six samples into two distinct groups by timing (red circled). The 5pm, 8pm and 6am samples are lying further to the left of the PCA plot and hence constituting the "night" samples. The 8am, 1pm and 3pm samples are separated far away from the night samples as they are lying more towards the right of the PCA plot, hence they constitute the "day" samples. The second principal component axis explains 20% of the variation which further separates the three samples within each day-night grouping. Results from PCA thus show that most of the separation shown in hierarchical clustering comes from the time of the day (i.e. by PCA component 1 axis).

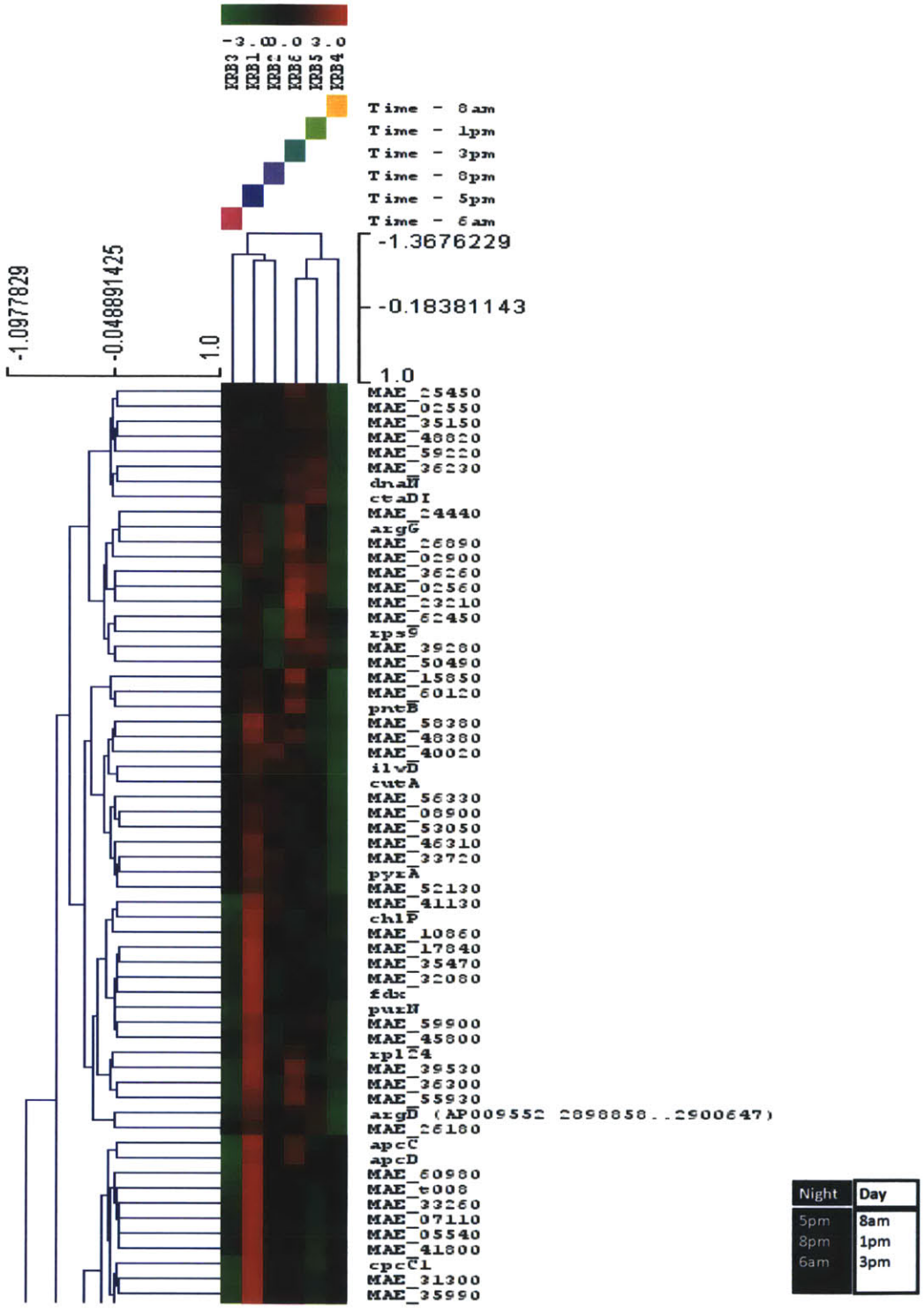


Fig. 34: Hierarchical clustering of six samples into day and night categories (partial list of genes).

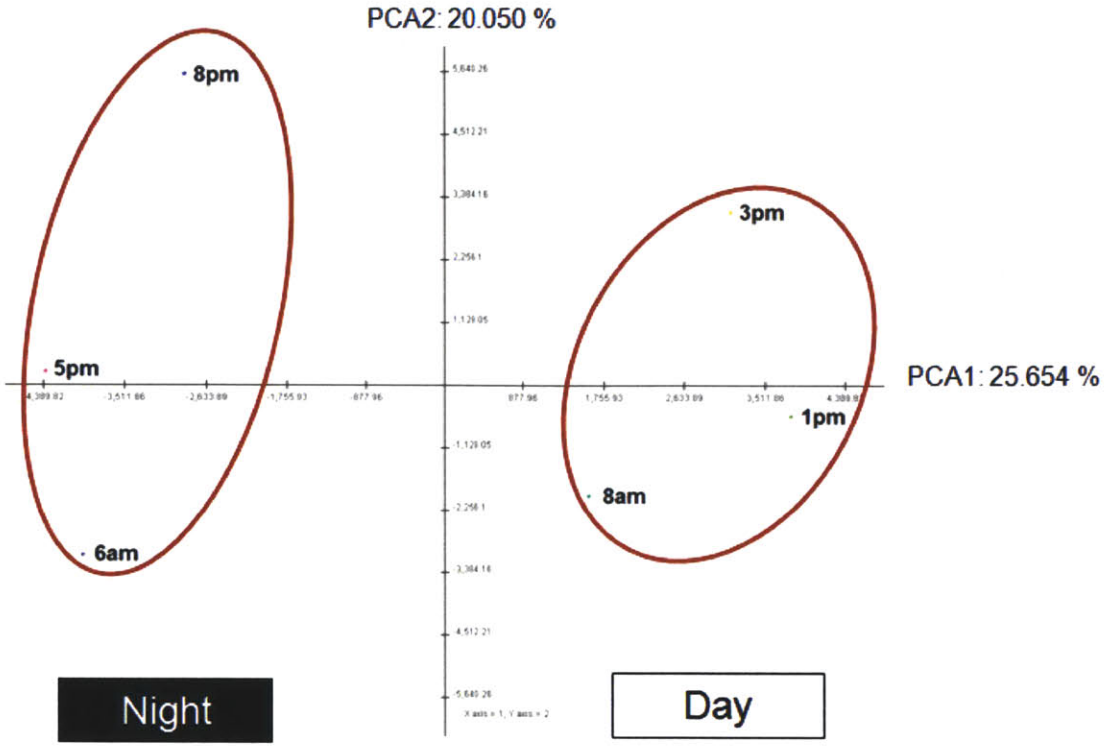


Fig. 35: Principal Component Analysis of the six samples, as classified based on transcription level of all the 6363 genes in *Microcystis aeruginosa*.

3.4.2 KEGG pathway enrichment by genes with day-night differential expression

To gain a better understanding on the functions of the 6363 *Microcystis* genes in terms of the KEGG pathways that they represent, GSEA (Gene Set Enrichment Analysis) is tested on the gene expression profiles for their enrichment of any of the 95 KEGG pathways in day or night conditions.

At a nominal $p < 0.05$, two KEGG pathways are shown to be significant in the day (Table 10). They are "Amino sugar and nucleotide sugar metabolism" (nominal $p = 0.00$) and "Carbon fixation in photosynthetic organism" (nominal $p = 0.017$). However, when the statistical test is corrected for the probability of making one or more false discoveries when performing multiple hypothesis testing, both pathways are shown to be insignificant (q-value = 0.062 and 0.414 respectively) at a false discovery rate (FDR) of 0.05 and having a Familywise error rates (FWER) of 0.047 and 0.474 respectively. The "Amino sugar and nucleotide sugar metabolism" KEGG pathway is still significant with a low FWER, but it is rejected by a high FDR.

At a nominal $p < 0.05$, two KEGG pathways are shown to be significant at night. They are "Mismatch repair" (nominal $p = 0.017$) and "Terpenoid backbone biosynthesis" (nominal $p = 0.049$). However, when a FDR of 0.05 is imposed, both pathways are shown to be insignificant (q-value = 0.624 and 0.995 respectively). The Familywise error rates (0.510 and 0.889 respectively) also reveal a high likelihood of false discovery.

Transcript abundance variation by KEGG functional categories such as photosynthesis pathway genes, TCA cycle genes, RuBisCO genes, ATP synthase genes are shown in Fig. 36-39). The gene representation levels correspond to the values in RPKM matrix. Fold change is calculated by dividing the individual gene representation level at each time point by the average level in the six samples if the individual time point level is higher; if otherwise, division is reversed.

Table 10: Preliminary analysis of functional enrichment of differential gene expression over a day-night cycle based on categorical labels in GSEA. Tested reads match > 90% alignment and > 80% identity with the reference genome (*Microcystis aeruginosa* NIES-843)

KEGG pathway enriched (Day)	NOM p-val	FDR q-val	FWER p-val
Amino sugar and nucleotide sugar metabolism	0.000	0.062	0.047
Carbon fixation in photosynthetic organisms	0.017	0.414	0.474

KEGG pathway enriched (Night)	NOM p-val	FDR q-val	FWER p-val
Mismatch repair	0.017	0.624	0.510
Terpenoid backbone biosynthesis	0.049	0.995	0.889

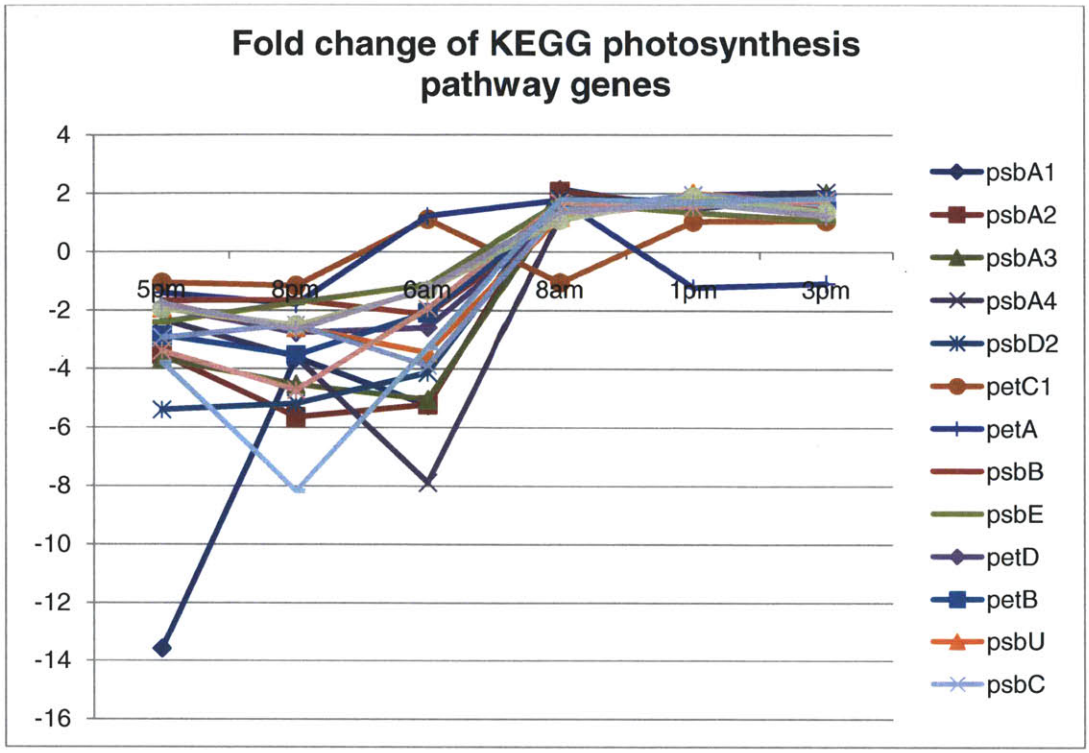


Fig. 36: Variations in the transcript abundance of genes involved in the KEGG photosynthesis pathway.

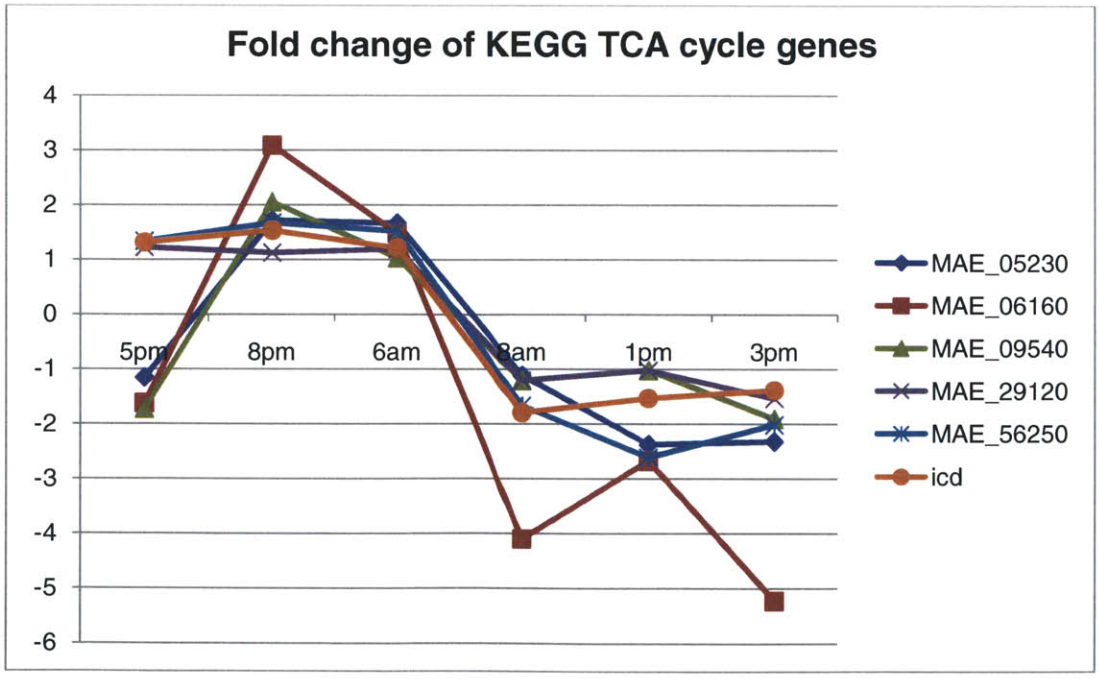


Fig. 37: Variations in the transcript abundance of genes involved in the KEGG TCA cycle pathway.

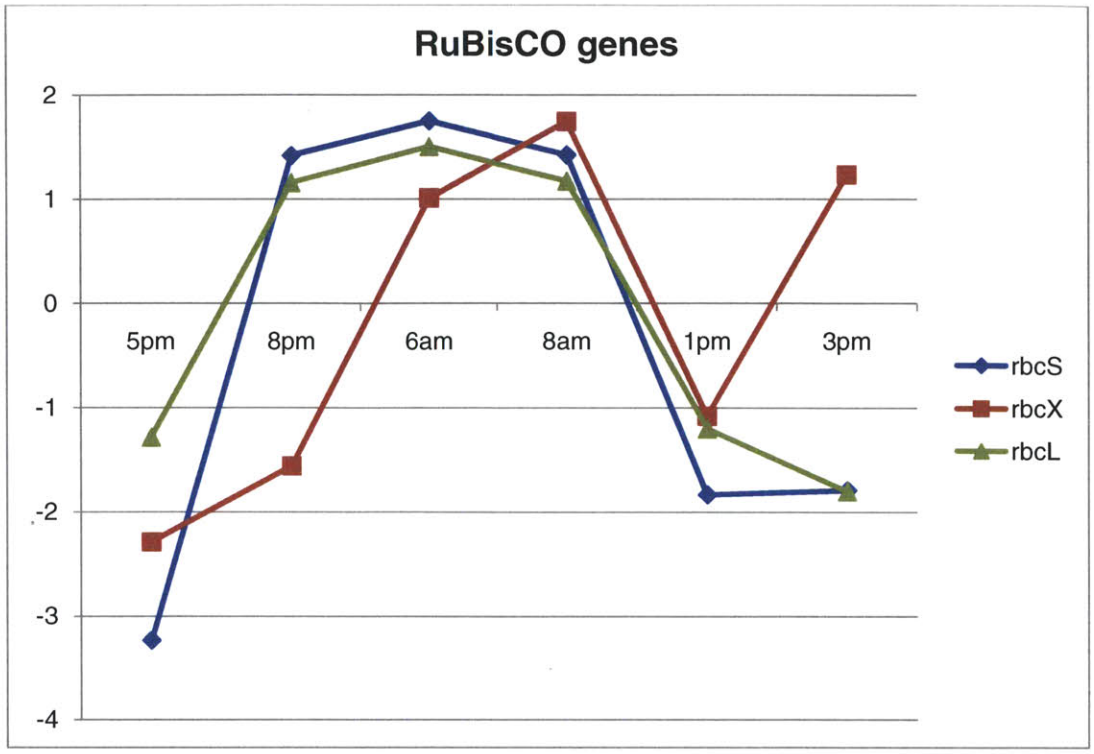


Fig. 38: Variations in the transcript abundance of RuBisCO (Ribulose bisphosphate carboxylase) genes.

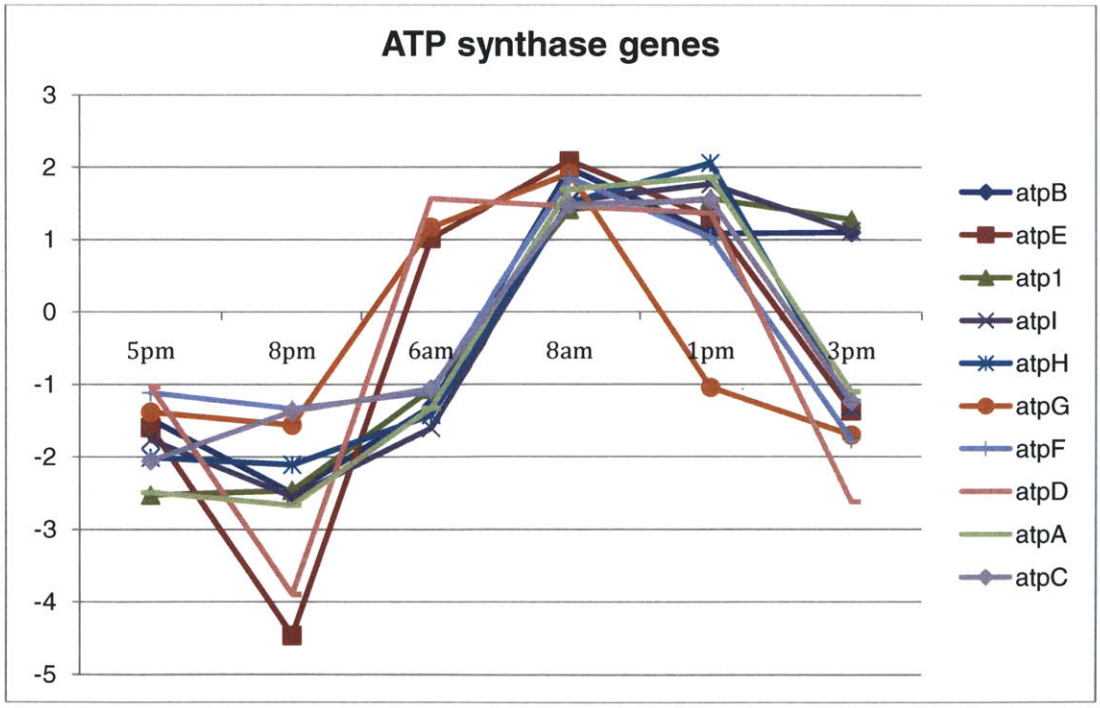


Fig. 39: Variations in the transcript abundance of ATP synthase genes.

3.4.3 Identification of *Microcystis* genes with differential representation in day-night transcriptomes

While the cluster and gene set enrichment based approach gave a number of different pathways that are important for understanding changes that *Microcystis aeruginosa* undergoes during a diel cycle, it is important to know what are the most differentially represented genes during night and day, and if the genes responsible for microcystin toxin production are statistically significantly differentially represented during day and night.

Two statistical tests: a T-test and Significance Analysis for Microarrays (SAM) are conducted to determine the most significantly differentially expressed genes during day and night. The T-test resulted in 86 differentially represented genes ($p < 0.05$) (Fig. 40). However, when using either a standard or adjusted Bon-ferroni correction, or FDR (False discovery rate of 0.05) to correct for multiple hypothesis testing, none of the genes are able to pass the more stringent threshold level of significance. One possible reason is that this test is limited by the number of samples.

The SAM analysis with a FDR of 0.05 resulted in a set of 20 differentially represented genes in the transcriptomes (Fig. 41). Four of which (Photosystem II reaction center protein H, cytochrome b6, photosystem II reaction center D2 protein and ATP synthase CF0 A chain Atpl) are involved in photosynthesis, indicating a shift in photosynthetic gene expression as light availability changes. This result lends some credibility to these genes being true positives for differential expression in day-night condition. However, out of the 20 genes, 4 genes encode for either hypothetical or unknown protein. More research needs to be done on these genes in order to have a better understanding on the differential gene expression in *Microcystis aeruginosa*.

In addition, in both statistical tests, none of the 10 genes in the mcy synthetase gene cluster is found to be differentially represented in transcriptomes during day and night conditions.

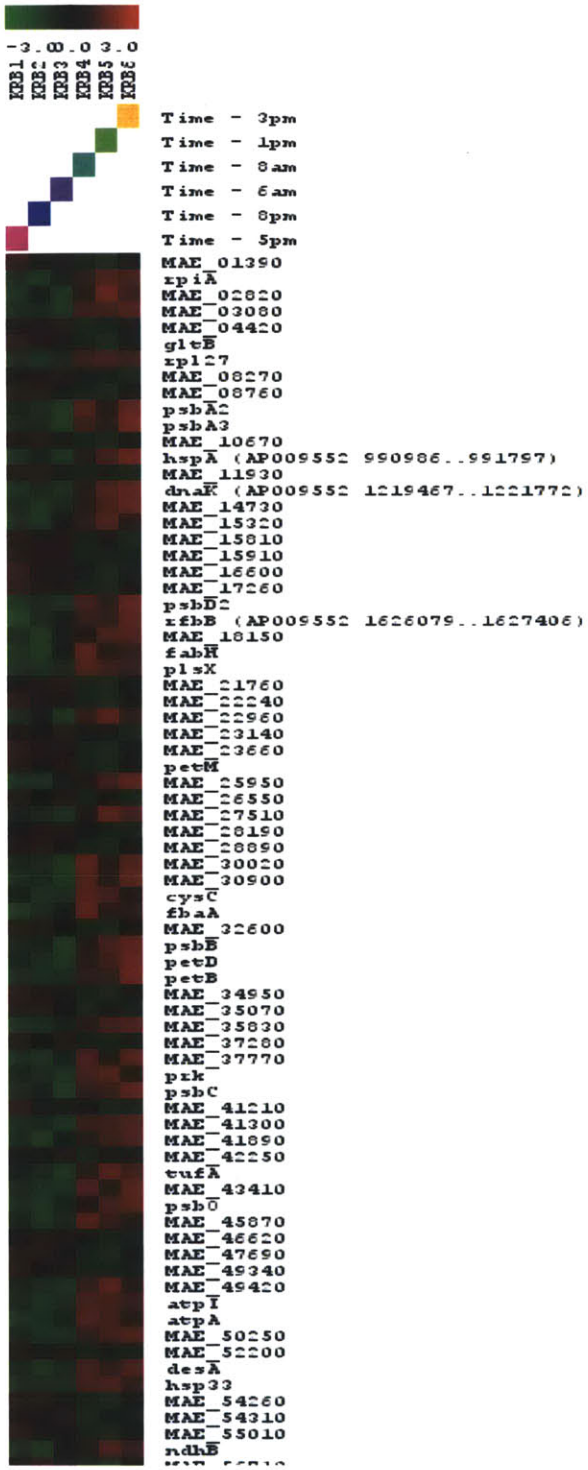


Fig. 40: 86 genes have significant differential gene expression during day and night by T-test ($p < 0.05$) before Bon-ferroni correction (partial heatmap shown). Each row shows the level of representation of a particular gene in the *Microcystis aeruginosa* reference genome.

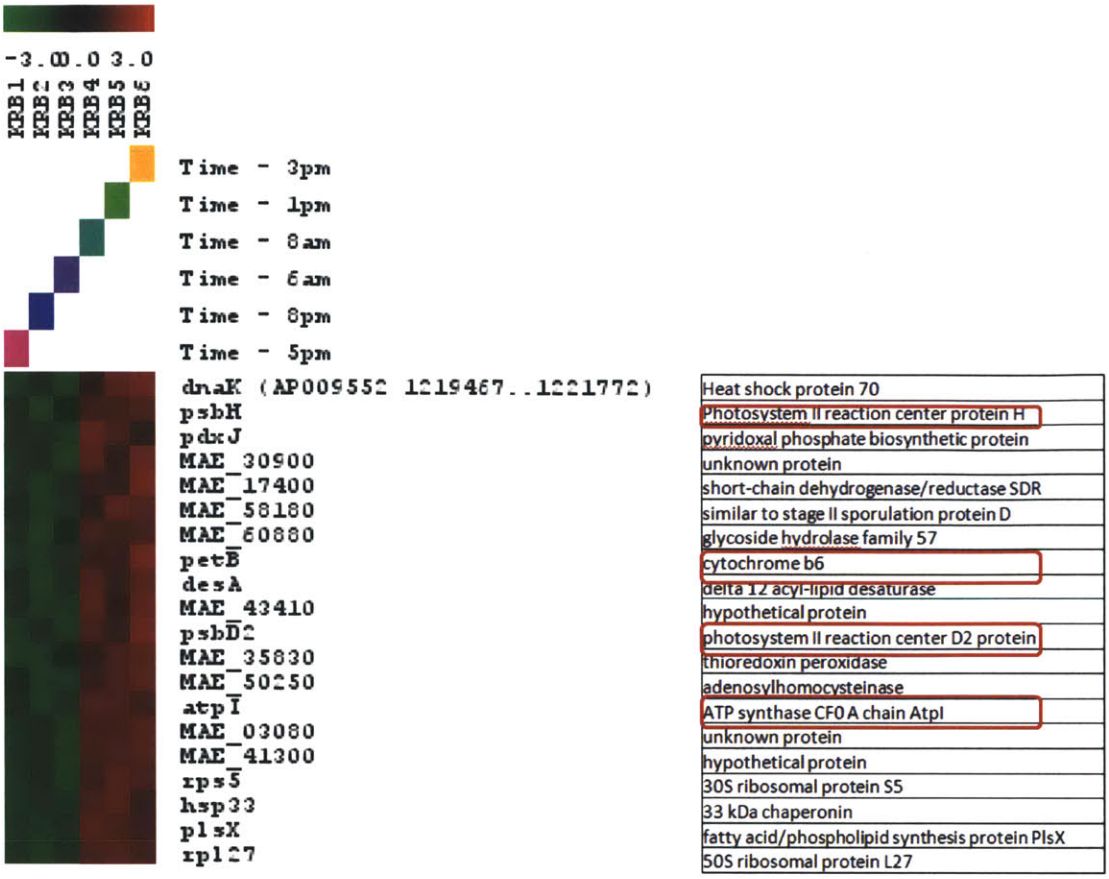


Fig. 41: 20 genes have significant differential gene representation during day and night by SAM (Significance Analysis for Microarrays). Circled proteins in red are involved in photosynthesis.

Chapter 4: Discussion

4.1 The role of bacteria in the freshwater ecology of Kranji Reservoir

Community taxonomic composition (Section 3.3.1) points us to the importance of the Bacterial community in structuring the overall dynamics of Kranji Reservoir. Similarity by community composition across the six different timepoints as measured by Pearson correlation in terms of the abundance of all Bacterial phyla shows that Bacterial community has a stable taxonomic distribution throughout the course of a day (Fig. 14, Table 4). The correlation index for all pairs of samples is above 0.9, indicating high positive correlation among all pairs of samples based on Bacterial community composition at the Phylum-level. Specifically, the maximum correlation (0.999) is found between samples krb4 (8am) and krb6 (3pm), while the minimum correlation (0.932) is found between samples krb5 (1pm) and krb2 (8pm). This finding implies that even though samples are taken in either complete light (day, 1pm) or darkness (night, 8pm), they are very similar in Bacterial community composition.

High similarity in Bacterial composition among six samples is also evident by the phyla rank abundance distribution in Fig. 16. Such consensus points to the fact that the community composition in terms of Bacterial phylum abundance is very stable throughout the day. The dominant players always dominate, even though there are minor fluctuations in terms of abundance and rank swapping among the less dominant phyla. Thus, the richness of each sample or community is similar, as the number of phyla present in each community is consistent. At the genus level, the observed richness among the five samples is also relatively constant. There are 593 – 714 annotated genera by the M5NR database ($E < 0.001$) (Table 5).

The bloom community is comprised of autotrophs and heterotrophs, with the autotrophs being mainly Cyanobacteria and Eukaryotic algae. The ecological role of autotrophs is to fix CO_2 and act as the primary producers. The main Bacterial primary producer in Kranji Reservoir appears to be the Cyanobacterium *Microcystis* by both RDP and M5NR genera annotation (Table 6- 7).

4.2 Structure and function of autotrophic assemblage of the Kranji Reservoir plankton

Since Cyanobacteria are autotrophs, they are expected to be very active during the day. Different genera of Cyanobacteria (like *Microcystis, Synechococcus, Lyngbya*) always rank at the top of the list of all community genera by abundance of transcripts annotated by both RDP and M5NR (Table 6-7). However, Cyanobacteria are expected to also express many genes at night. For instance, Cyanobacteria continue to carry out many metabolic processes at night including DNA repair and carbon fixation. Since Cyanobacteria is the most abundant Bacterial phylum in Kranji Reservoir, it is no surprise that Cyanobacterial transcripts are also highly abundant at night even though some of their photosynthetic mechanisms are shut down at night.

In general, the read abundance for the Cyanobacterium *Microcystis*, the transcript abundance is lower for the samples in the "night" category compared to those in the "day" category (Fig. 21). However, the abundance level for the 6am sample (krb3) is actually unusually high (22.9 %). Recent studies have shown that some bacteria e.g. *Procholorococcus* ramp up transcription level before sunrise, in preparation for a higher metabolic rate in the day (Zinser et al., 2009). Similar mechanism might have been operating in *Microcystis*.

Another possible explanation is that during the planktonic phase, *Microcystis* cells are organized into colonies that carry out daily vertical migrations in the water column. In the early morning, colonies accumulate to very high abundance in the top few centimeter of the water column where samples were taken (i.e. krb3 (6am), 22.9 %). During the rest of the morning, they progressively disappear from the surface. Indeed, the transcript abundance just two hours after (ie. krb4 (8am)), dropped by 13.5 % (from 22.9 % to 19.8 %). However, the abundance for the rest of the sunlit morning stays high, likely due to the increased cellular metabolism such as photosynthesis which made more growth-promoting nutrients available. After that, transcript abundance decreases in the evening (18.8 %, krb1 (5pm)) until they reach a low abundance of 15.3 % near the end of the day (krb2 (8pm)).

PCA of expressed *Microcystis*-like genes may imply the existence of a Bacterial circadian rhythm in *Microcystis aeruginosa* in a complex plankton community (Fig. 35). A previous study (Straub et al., 2011) on a cultured single strain of *Microcystis*, undergoing a simulated

24-hour light-dark cycle, showed that cultured samples traced a circular path just like a 24-hour clock when plotted with a PCA.

Even though the diel transcriptomics study by Straub et al. provides a background for comparison with our study, it is important to take note of the differences in the approach taken. For example, they use a microarray designed for 96% of the total predicted protein-coding genes in the reference genome, instead of using community RNA-seq or metatranscriptomics while mapping back to the reference genome in our case. Moreover, a different *Microcystis aeruginosa* strain is used as the reference genome: strain PCC 7806 in their case as opposed to strain NIES-843 in our study. Thus, some genes present in one strain might not be present in the other or their gene name IDs may not be the same. Thus, care should be taken when comparing expression profiles directly.

In addition, a pure culture in a simulated 24 hours of alternating dark-light condition is used in their study instead of an environmental sampling in a natural environment in our study. Although a simulated dark-light condition does not provide a progressive increase or decrease in light intensity during transitions which might affect expression of some genes, their approach of using strictly controlled lab conditions might help to magnify the effects due to light which might be masked by noise in a complex natural environment. However, our study using environmental samples provides insight on how *Microcystis* behaves in a natural environment including potential interactions with other species in the complex bloom community.

4.3 Cyanotoxin production

Cyanobacterial toxins are a health risk in eutrophic freshwaters. Based on transcript abundance, the dominant microcystin-producing genus in the reservoir is *Microcystis* (20.41 ± 3.52 % of total transcripts). Other cyanotoxin producers like *Anabaena*, *Nostoc* and *Planktothrix* have 5.57 ± 1.38 %, 4.09 ± 0.82 % and 0.04 ± 0.01 % of total transcripts respectively (Fig. 21). Simply based on this read abundance, it might be suggested that there is a high probability that at any time, toxins found in Kranji Reservoir are most likely produced by *Microcystis*, much less likely by *Anabaena* and *Nostoc*, and the least likely by *Planktothrix*. However, in a previous study carried out in Lake Erie, where *Microcystis* has been thought to be the primary toxin producer, observations from 2003 to 2004 suggested that the microcystin-producing community may not be composed entirely of *Microcystis*, as only 0.03 to 4.7 % of these cells contained the toxin biosynthetic machinery (Rinta-Kanto et al., 2005). In another study done by the same group of scientists on the same lake, it was found through a phylogenetic analysis on the *mcyA* gene that 100 % of the sequences from one sampling site clustered with toxigenic Cyanobacterium *Planktothrix aghardii* (Rinta-Kanto et al., 2006). Therefore, even though *Microcystis* population might be most abundant, they might not be the major toxin producers when other lower-abundant toxigenic strains are present. Further study on the sources of toxin needs to be done before any conclusion on the distribution of toxin producers in Kranji Reservoir can be drawn.

A recent study by (Te & Gin, 2011) on Kranji Reservoir showed that about half of the *Microcystis* found contain toxin genes by QPCR. Such high percentage compared to 0.03 - 4.7 % Lake Erie might suggest that tropical climate provides a conducive environment for toxigenic strains of *Microcystis*. Preliminary data from this study shows that toxin gene transcripts (*mcyA – mcyJ*) are not differentially represented over a day-night cycle although further analysis and additional studies are required to confirm this result.

4.4 Structure and function of heterotrophic assemblage of the Kranji Reservoir plankton

Heterotrophs are consumers of autothonous and allocthonous organic matter. Heterotrophs include many species of Proteobacteria, Firmicutes, Bacteroides, Actinobacteria – although in some cases heterotrophs may have light-harvesting pigments enabling mixotrophic growth. Gene expression profiles reveal signatures of interactions between the autotrophs and the heterotrophs.

The most abundant Alphaproteobacterial class in Kranji Reservoir is Rhizobiales (2.08 ± 0.60 %; E_{ave} ranges from 1E-0.99 to 1E-0.41). Alphaproteobacteria play an important role in the ecology of freshwater as many of them are nitrogen-fixers (e.g. members of Rhizobiales) (MacLean et al., 2007) which provides a selective advantage in ecosystems where nitrogen is limiting. They also contribute to the remineralization of organic matter. Quantitative assays based on FISH (Fluorescence In Situ Hybridization) have consistently shown that Alphaproteobacteria are resistant to predation and some types are able to form filaments and aggregates when they are exposed to increased microeukaryote grazing as an escape from predation.

Alphaproteobacteria genera, *Sphingomonas* (Sphingomonadales order, Sphingomonadaceae family; average read abundance of 0.066 %) and *Brevundimonas* (Caulobacterales order, Caulobacteraceae family; average read abundance of 0.031 %) are the most common genera isolated from purification plants. It has been reported that different strains of these two genera can either promote or inhibit the growth of a coexisting Cyanobacterial bloom through various types of interactions like nutrient cycling and production of growth-inhibiting and cell-lysing compounds (Berg et al., 2009). For example, *Sphingomonas* is found to be capable of degrading microcystin toxins (Manage et al., 2010). Another genus in the Sphingomonadaceae family, *Novosphingobium* (average abundance of 0.041 %), is found to contain *mlr*, a microcystin-degrading gene cluster (Jiang et al., 2011).

It has been reported that Gammaproteobacteria are not particularly abundant in freshwater lakes. However, this class of Proteobacteria emerged as the most abundant (Fig. 22) among all classes in Kranji Reservoir and the most abundant Gammaproteobacteria is

Enterobacteriales (1.29 ± 0.28%). One hypothesis that could explain this phenomenon is suggested by (Zavarzin et al., 1991) which is that Gammaproteobacteria are copiotrophs which flourish in environments that are rich in nutrients. Since Kranji Reservoir is an eutrophic water source, Gammaproteobacteria would be able to grow significantly faster than the average lake bacterioplankton and hence show up in high abundance. In addition, vast amount of literature have used members of the Enterobacteriales in the source tracking of pollutants in surface waters (Stoeckel et al., 2007). Their presence of high abundance might reflect Kranji Reservoir's role as a receiving body for urban non-point source pollution.

Betaproteobacteria are also often numerically abundant in freshwater lakes as they are opportunists who can grow quickly when exposed to nutrient pulses in mesocosms. The most abundant Betaproteobacteria in Kranji Reservoir is Burkholderiales (3.08 ± 0.90%). Species *Polynucleobacter necessarius* (average read abundance of 0.081%) belong to the PnecC tribe which is part of the four tribes (PnecA- PnecD) in the Pnec clade. This clade is the best-studied freshwater lake clade within the Betaproteobacteria (Newton et al., 2010). However, in all the six samples taken from Kranji Reservoir, PnecC is the only taxon with hits. It is known that PnecC prefers more acidic habitats and they are particularly competitive in environments rich in humic acids. Also, PnecC is readily enriched upon exposure to elevated levels of allochthonous dissolved organic carbon which is highly possible in Kranji Reservoir as it receives water from various neighboring watersheds.

Members of Betaproteobacteria are often cocultured with algae and are associated with Cyanobacteria. Thus, they are very competitive in photosynthetically active planktonic systems, for instance, in a Cyanobacterial bloom. Their competitiveness in freshwaters could be attributed to their ability to balance vulnerability to grazing with fast growth rates. Therefore, their abundance is positively associated with watercolor metrics that are proxies for low-molecular-weight alga-derived substrates (Simek et al., 2010).

Clostridiales is the most abundant order of Firmucutes (2.80 ± 0.73 % of total transcripts; E_{ave} = 1E0.52 - 1E0.28, Fig. 24). Studies have shown that certain species of the Firmicute genus *Clostridium* (within the Clostridiales family) dominate the Bacterial communities during anaerobic degradation of *Microcystis* scum, suggesting theses microbes' involvement in the degradation process (Xing et al., 2011). *Clostridia* are metabolically versatile organisms.

Some species are able to utilize various substrates, ranging from high molecular weight and complex structures to low molecular weight and simple structures. During the anaerobic incubation of the *Microcystis* scum, *Clostridium* species appear to have the collaborative ability to hydrolyze substrates and produce low molecular weight intermediates like acetate and ethanol, and even directly generate biohydrogen gas (Chang et al., 2006).

The relatively high abundance of Actinobacteria-like transcripts (2.7 - 8.8 % of total Bacterial transcripts) in the Kranji Reservoir could be explained by a few factors. One factor is that Actinobacteria are small in size which allows them to escape from size-selective grazers. Their cell walls contain a S layer which make them less "edible" by predators. Another factor is that Actinobacteria have high GC content and are pigmented such that they are better protected from UV damage especially in the epilimnion. There is also evidence that actinorhodopsin genes are found in them which might suggest a potential mechanism for a supplemental mode of energy generation during daylight. These factors came together in making Actinobacteria one of the most persistent lineages in freshwater lakes. In our study, it has been found that Actinobacteria genera such as *Streptomyces* are particularly enriched at night: up to 1078 transcripts in the "night" samples as compared to less than 65 transcripts in the "day" samples (M5NR annotation at $E < 0.001$, Table 7). There are also more varieties of Actinobacteria genera emerged as one of the top genera in the "night" samples (4 - 13 genera) as opposed to "day" samples (0 - 1 genera) (Table 7).

It has been reported that increased nutrient concentrations generally select against Actinobacteria. Even though they have no problem in taking up nutrient substrates, their growth rates are often below average compared to bacteria of other phyla (Simek et al., 2006). The ratio of allochthonous to autochthonous carbon can affect the distribution of members in the Actinobacteria phyla as different classes of Actinobacateria have different preferences for substrate sources. Kranji is an eutrophic reservoir and the large amount of nutrient inflows from watersheds might result in lower abundance of Actinobacteria compared to other lakes.

Within the Bacteroidetes, there are three distinct classes: Bacteroidia, Cytophagia, Flavobacteria, and Sphingobacteria. In the epilimnion layer of lakes, Bacteroidetes may consist of a large proportion of particle-associated bacteria which participate in the degradation of complex biopolymers. Thus, Bacteroidetes tend to occur during periods or at

sites characterized by high external dissolved organic carbon (DOC) loading or alga-derived DOC inputs (Zeder et al., 2009).

Freshwater lake Bacteroidetes are often found in high abundance during periods following Cyanobacterial blooms. In a study by Eiler et al., the majority of the Bacterial community measured was made up of the Bacteroidetes *Flavobacterium*-like lineages following the senescence and decline of an intense Cyanobacterial bloom (Eiler & Bertilsson, 2007). This study hence showed that many *Flavobacterium*-like populations are favored during periods of high heterotrophic activity and enhanced growth. Zeder et al. found that certain groups of the Flavobacteria indeed have high growth rates during and after a spring phytoplankton bloom, indicating that these bacteria are adapted to a copiotroph life-style or high nutrient conditions.

Cytophaga-like bacteria are able to lyse Cyanobacteria by attachment and secretion of diffusible lytic substances (Rashidan et al., 2001) which include a variety of exoenzymes that is capable of hydrolyzing Cyanobacterial cell wall. *Cytophagaceae* strains have been isolated from the polysaccharide capsule of *Microcystis* cells and even from water around *Microcystis* blooms. The lysis of Cyanobacteria may be an important contributor for the sudden disappearance of Cyanobacterial blooms as they control the Cyanobacteria population dynamics.

4.5 Day-night difference in community functional capacity

Photosynthesis and metabolism

Out of the top ten KEGG pathways with the largest number of total hits in the six transcriptomes, there are four pathways that deal with photosynthesis and its related mechanisms: antenna protein involvement, chlorophyll metabolism as well as carbon fixation. The photosynthesis KEGG pathway is statistically significantly up-regulated in the day.

An additional 4 of the top 10 KEGG pathways that deal with energy production and utilization are Glycolysis/ Gluconeogenesis, Purine metabolism, Oxidative phosphorylation and Carbon fixation. The Glycolysis/ Gluconeogenesis pathway is significantly up-regulated in the day samples. Glycolysis/ Gluconeogenesis involve the break-down and generation of glucose which will be involved in processes like oxidative phosphorylation for energy generation.

Pathways that are represented evenly in both the "day" and "night" samples include Purine metabolism, Oxidative phosphorylation and Carbon fixation. Purine metabolism regulates the conversion among ATP (Adenosine Triphosphate), ADP (Adenosine Diphosphate) and AMP (Adenosine Monophosphate), as well as among GTP (Guanosine Triphosphate), GDP (Guanosine Biphosphate) and GMP (Guanosine Monophosphate). Oxidative phosphorylation uses energy released by the oxidation of nutrients (where electrons are transferred from electron donors to acceptors like oxygen) to produce ATP. Carbon fixation in the Calvin cycle is also an energy-expensive process as each product of G3P (Glyceraldehyde-3-Phosphate) requires energy input from 9 ATP and reducing power of 6 NADPH (Nicotinamide Adenine Dinucleotide Phosphate).

High rate of DNA repair mechanisms at night

During the day, UV radiation can cause intra-strand linkage of adjacent pyrimidines, usually thymines (thymine dimers), which prevents DNA replication from proceeding. In bacteria, there are four main DNA repair mechanisms: photoreactivation (light repair), excision repair (dark repair), postreplicative (recombinational) translesion bypass repair and SOS response. The three KEGG pathways in Fig. 32b all fall under the excision repair method which typically occurs in the dark (Brock et al., Biology of Microorganisms, 10th edition; Clancy

2008). Among the three pathways, the Mismatch repair pathway (which normally accounts for 99% of all repairs) is detected with reads in all six samples while the other two pathways (nucleotide and base excision repair) almost exclusively have reads in the night samples (except for krb5 (1pm)).

High rate of Glycolysis/gluconeogenesis in the day

Out of the three KEGG pathways involved in cellular respiration or ATP/energy production, both Glycolysis/gluconeogenesis and Oxidative phosphorylation pathways are up-regulated in the day with the former being statistically significant. The Citrate (TCA) cycle pathway has more reads at night although not statistically significant. These three pathways are thought to be the core processes in cellular respiration. Glycolysis first initiates cellular respiration by converting glucose to pyruvate which is fed into the Citrate (TCA) cycle for the generation of NADPH which carries the reducing power to the Electron transport chain (ETC) in Oxidative phosphorylation. The proton gradient built up in ETC then allows for ATP synthesis.

4.6 Day-night difference in *Microcystis aeruginosa* functional capacity

Day-night difference in *Microcystis aeruginosa* functional capacity is analyzed with GSEA which looks into one whole dataset of 6363 genes to find KEGG pathways that are enriched with genes that show day-night differential expression. In Table 10, even though none of the pathways that passed the nominal $p < 0.05$ also passed the FDR test, it is interesting to note that the "Mismatch repair" pathway is enriched at night under nominal $p < 0.05$. This result coincides with results from MEGAN that transcripts encoding for Mismatch repair are particularly abundant at night.

Photosynthesis gene transcripts show a two-fold enrichment in the "day" samples and approximately a four-fold reduction in the "night" samples (Fig. 36). TCA cycle gene transcripts show a general trend of enrichment at night and reduction in the day which corresponds well to the fact that cellular respiration makes up for a higher proportion of overall energy generation at night (Fig. 37). RuBisCO gene trascripts also show a general trend of higher representation at night than during the day which indicates that carbon fixation occurs at a higher level at night (Fig. 38). One explanation of this might be that the RuBisCO enzyme is oxygen-sensitive, so carbon fixation can occur more efficiently at night when oxygen production by photosyntheis is absent (Brock et al., Biology of Microorganisms, 10th edition; Clancy 2008). ATP synthase gene transcripts, however, are generally enriched in the day (Fig. 39) which is reasonable since photosynthesis, another mode for ATP generation besides cellular respiration, occurs during the day.

Although the oberved trends in transcript abundance over a day-night cycle agree well with patterns expected due to changes in microbial activity, additional work to constrain biases associated with preparing transcriptomes (e.g. cDNA amplification, RNA depletion and sequencing) is currently in progress to link differential enrichment of transcripts (e.g. Fig. 36-39) to changes in gene expression in the environment.

Chapter 5: Conclusions and future directions

This metatranscriptomics study is set out to understand the varying community taxonomic composition (structure) and ultimately functional capacity as different microbes take on active roles at different time point of a day in a bloom event. This thesis has attempted to address the questions "who is there?" and "when are they active?" which provide insights into the structural dynamics of the plankton community in the Kranji Reservoir. This work on the structural dynamics, paves the way for future analysis to address the question "what are they doing?" Such information may be useful for water quality management to effectively control Cyanobacterial blooms and hence minimize public risk associated with cyanotoxins. The bloom in the Kranji Reservoir is a complex environment with approximately 593 to 714 genera co-occurring in the surface waters (based on M5NR annotation with $E < 0.001$). Thus, in order to understand the dynamics or biology of toxin producers, it is important to consider the whole community. Results from this study show that community taxonomic composition stays relatively constant throughout a day in terms of the abundance of the most active members in the community. The five most abundant Bacterial phyla were: Cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes in a decreasing order. The most dominant genus is *Microcystis*, representing 15.3 to 25.6 % of all transcripts (E_{ave} = 1E-4.22).

We have demonstrated the usefulness of metatranscriptomics as a method to track changes in the environment since bacteria can be considered as sensors which respond rapidly to environmental cues. In this study, we have carried out a preliminary analysis of photosynthesis and metabolism gene expression changes because they give clear signal and obvious adjustment to environmental changes. In the future, we will be looking at other biological functional correlations with the environmental changes.

Preliminary analysis shows that the overall transcriptional profiles of *Microcystis aeruginosa*, and closely related Cyanobacteria, differ and follow a seemingly Bacterial circadian rhythm. This difference enables the grouping of six samples into "day" and "night" categories. Further analysis such as K-means clustering and Gene Set Enrichment Analysis may help to single out groups of genes that show similar gene expression pattern and thus are potentially co-regulated (e.g. the toxin genes). Understanding the pathways that affect microcystin toxin

production is important for managing bloom events and minimizing public risks associated with this toxin.

Further analysis on functional capacity via MG-RAST and MEGAN may increase our understanding on the ecological interactions between *Microcystis* and other players in the reservoir, thus shedding light on mechanisms that triggers toxin release and production. Moreover, reference genomes of the other major set of taxa in Kranji Reservoir can be sequenced and *de novo* assembled in order to allow for better annotation of the transcripts and hence determination of their functional significance since a large proportion of reads cannot be annotated with confidence at a more stringent E-value. Increasing sampling time points (such as before, during and after the bloom in a seasonal setting or for every hour in a diel setting), spatial distribution (such as multiple sites in the reservoir or along a reservoir depth profile) and analysis of increased sample replicates at each time point in future experiments would allow for much greater statistical power tests and would thus better at distinguishing true gene expression signals from sample noise.

Other future work includes an analysis on the functional dynamics within individual taxa like *Microcystis* which are the main microcystin producers; *Clostridium* which may be degraders of *Microcystis* bloom scum and *Sphingomonas* which may be capable of degrading microcystin toxin through the activity of *mlr* genes. Furthermore, reads that are mapped to other reference genomes (besides *Microcystis aeruginosa*) in Kranji samples can be singled out such that their RPKM or expression values can be analyzed statistically for any difference in day-night expression or correlation to gene expression in other populations (e.g. *Microcystis*). If the reference genome for some major taxa in Kranji Reservoir does not exist in the current database, genome sequencing and *de novo* assembly should be carried out in the future. This will help in getting more reads in the community annotated such that it will facilitate our study on the dynamics of genes and subsystems involved in nutrient cycling and toxin biosynthesis.

We anticipate that such study will ultimately help us to efficiently manage the water quality and minimize loss and damage in a Cyanobacterial bloom.

Reference

Alexova R, Fujii M, Birch D, Cheng J, Waite TD, Ferrari BC, Neilan BA (2011) Iron uptake and toxin synthesis in the bloom-forming *Microcystis aeruginosa* under iron limitation. Environ Microbiol. 13(4):1064-77.

Berdy, J (1980) Recent advances in and prospects of antibiotic research. Process Biochemistry. Vol. 15, pp. 28+.

Berg KA, Lyra C, Sivonen K, Paulin L, Suomalainen S, Tuomi P, Rapala J. (2009) High diversity of cultivable heterotrophic bacteria in association with Cyanobacterial water blooms. ISME J. 1:532-544.

Brian Knaus Short read tool box. http://brianknaus.com/software/srtoolbox/shortread.html

Brock, Madigan, Martino and Parker, Biology of Microorganisms, 10th edition. Prentice Hall

Chang JJ, Chen WE, Shih SY, Yu SJ, Lay JJ, Wen FS, Huang CC. (2006). Molecular detection of the clostridia in an anaerobic biohydrogen fermentation system by hydrogenase mRNA-targeted reverse transcription-PCR. Appl Microbiol Biotechnol 70: 598-604.

Chorus I, Bartram J. (1999) Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management. E and FN Spon, An imprint of Routledge, London, 416 pp.

Chritoffersen K (1996) Effect of microcystin on growth of single species and on mixed natural populations of heterotrophic nanoflagellates. Natural Toxins 4:215-220.

Clancy, S. (2008) DNA damage & repair: mechanisms for maintaining DNA integrity. Nature Education 1(1)

De Maagd PGJ, Hendriks AJ, Seinen W and Sijim, DTHM (1999) pH-dependent hydrophobicity of the Cyanobacteria toxin microcystin-LR. Water Res. 33: 677-680.

Eiler A, Bertilsson S. (2007) Flavobacteria blooms in four eutrophic lakes: linking population dynamics of freshwater bacterioplankton to resource availability. Appl. Environ. Microbiol. 73: 3511-3518.

Engelke CJ, Lawton LA, Jaspars M (2003). Elevated microcystin and nodularin levels in Cyanobacteria growing in spent medium of *Planktothrix agardhii*. Arch. Hydrobiol. 158, 541-550.

Falconer IR and Yeung DSK (1992) Cytoskeletal changes in hepatocytes induced by *Microcystis* toxins and their relation to hyperphosphorylation of cell proteins. Chem. Biol. Interact. 81, 181–196.

Falconer IR. (2001) Toxic Cyanobacterial bloom problems in Australia waters: risks and impacts on human health. Phycologia 40: 228233

Fox JA, Booth SJ, Martin EL. (1976) Cyanophage SM-2: a new blue-green algal virus. Virology. 73(2):557-60.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW et al. (2008). Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci USA 105: 3805–3810.

Gilbert JA, Meyer F, Schriml L, Joint IR, Mühling M, Field D. (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. Stand Genomic Sci. 3(2):183-93.

Gin KYH, Neo SY (2005) Microbial populations in tropical reservoirs using flow cytometry. Journal of Environmental Engineering, Vol.131, No. 8 1187–1193

Gupta N, Pant SC, Vijayaraghavan R, Rao PV. (2003) Comparative toxicity evaluation of Cyanobacterial cyclic peptide toxin microcystin variants (LR, RR, YR) in mice. Toxicology.188(2-3):285-96.

Guzman RE, Solter PF. (2002) Characterization of sublethal microcystin-LR exposure in mice. Vet Pathol.39(1):17-26.

Hardy J. (2008) Washington State Recreational Guidance for Microcystins (Provisional) and Anatoxin-a (Interim/Provisional). Washington State Department of Health <http://www.doh.wa.gov/ehp/oehas/pubs/334177recguide.pdf>

Hikmet Katırcıoğlu1, Beril S. Akın and Tahir Atıcı (2004) Review: Microalgal toxin(s): characteristics and importance. African Journal of Biotechnology Vol. 3 (12), 667-674

Huson DH, Auch AF, Qi J, Schuster SC. (2007) MEGAN analysis of metagenomic data. Genome Research 17(3):377-86.

Imai H, Chang KH, Kusaba M and Nakano SI (2009) Temperature-dependent dominance of *Microcystis (Cyanophyceae)* species: *M. aeruginosa and M. wesenbergii.* Journal of Plankton Research Vol. 31(2), 171-178.

Jiang Y, Shao J, Wu X, Xu Y, Li R. (2011) Active and silent members in the mlr gene cluster of a microcystin-degrading bacterium isolated from Lake Taihu, China. FEMS Microbiology Letters. 1-7.

Kaebernick M, Neilan BA, Borner T and Dittmann E. (2000) Light and the transcriptional response of the microcystin biosynthesis gene cluster. Appl. Environ. Microbiol. 66: 3387-3392.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res. 32(Database issue):D277-80.

Kaneko T, Nakajima N, Okamoto S, Suzuki I, Tanabe Y, Tamaoki M, Nakamura Y, Kasai F, Watanabe A, Kawashima K, Kishida Y, Ono A, Shimizu Y, Takahashi C, Minami C, Fujishiro T, Kohara M, Katoh M, Nakazaki N, Nakayama S, Yamada M, Tabata S, Watanabe MM. (2007) Complete genomic structure of the bloom-forming toxic Cyanobacterium Microcystis aeruginosa NIES-843. DNA Res.14(6):247-56. Epub 2008 Jan 11.

Kearns KD, Hunter MD (2001) Toxin-producing *Anabaena flos-aquae* settling of *Chlomydomonas reinhardtii*, a competing motile alga. Microb. Ecol. 42: 80-86.

Kosakowska A, Nedzi M, Pempkowiak J (2007) Responses to the toxic Cyanobacterium *Microcystis aeruginosa* to iron and humic substances. Plant Physiol Biochem 45: 365-370

Lampert W (1981) Inhibitory and toxic effects of blue-green algae on Daphnia. Internationale Revue der gesamten Hydrobiologieund Hydrographie, 66(3): 285-298.

Latifi A, Ruiz M, Zhang CC (2008) Oxidative stress in Cyanobacteria. FEMS Microbiol Rev 33: 258-278.

Low EW, Clews E, Todd PA, Tai YC, Ng PKL (2010) Top-down control of phytoplankton by zooplankton in tropical reservoirs in Singapore 58(2): 311-322

Lukac M, Aegerter R (1993) Influence of trace metals on growth and toxin production of *Microcystis aeruginosa*. Toxicon 31(3): 293-305.

Lyck S, Gjolme N, and Utkilen H. (1996) Iron starvation increases toxicity of M. aeruginosa CYA 228/1 (Chroococcales, Cyanophyceae). Phycologia 35: 120-124.

MacKintosh C, Beattie KA, Klumpp S, Cohen P, Codd GA. (1990) Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. FEBS Lett. 264(2):187-92.

MacLean AM, Finan TM, Sadowsky MJ (2007) Genomes of the symbiotic nitrogen-fixing bacteria of legumes. Plant Physiol. 144(2):615-22.

Manage PM, Kawabata Z, Nakano S. (2001) Dynamics of cyanophage-like particles and algicidal bacteria causing Microcystis aeruginosa mortality. Limnology 2(2): 73-78

Manage PM, Edwards C, Lawton LA (2010) Bacterial Degradation of Microcystin. Interdisciplinary Studies on Environmental Chemistry — Biological Responses to Contaminants, 97–104.

Martin-Luna B, Sevilla E, Hernandez JA, Bes MT, Fillat MF, Peleato ML. (2006a) Fur from *Microcystis aeruginosa* binds in vitro promoter regions of the microcystin biosynthesis gene cluster. Phytochemistry 67: 876-881.

Martin-Luna B, Hernandez JA, Bes MT, Fillat MF, Peleato ML. (2006b) Identification of a ferric uptake regulator from *Microcystis aeruginosa* PCC7806. FEMS Microbiol Lett 254: 63-70.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386

Michel KP, Pistorius EK (2004) Adaptation of the photosynthetic electron transport chain in Cyanobacteria to iron deficiency: the function of IdiA and IsiA. Physiol Plant 120: 36-50.

Ministry of the Environment and Water Resources of Singapore. (2005) Towards Environmental Sustainability, State of the Environment 2005 Report.

Mitrovic SM, Hardwick L, Dorani F (2011) Use of flow management to mitigate Cyanobacterial blooms in the Lower Darling River, Australia. J. Plankton Res. 33(2): 229-241

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. (2003) Nat Genet 34(3), 267-273.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 621-628

Neilan BA, Tillett D. (2000) Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide-polyketide synthase system. Chemistry & Biology .Vol 7 No 10, 753-764.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2010) A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev. 75(1):14-49.

Novotny V. (2011) The Danger of Hypertrophic Status of Water Supply Impoundments Resulting from Excessive Nutrient Loads from Agricultural and Other Sources. Journal of Water Sustainability. Vol. 1, Issue 1, 1–22

NTU (2008) Water quality monitoring, modeling and management for Kranji Catchment/Reservoir system – Phases 1 and 2 ,May 2004 to December 2007. Division of Environmental and Water Resources Engineering, School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

Oh HM, Lee SJ, Jang MH and Yoon BD (2000) Microcystin production by *Microcystis aeruginosa* in a phosphorus-limited chemostat. Appl. Environ. Microbiol. 66, 176-179

Paerl HW, Xu H, McCarthy MJ, Zhu G, Qin B, Li Y, Gardner WS. (2011a) Controlling harmful Cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): the need for a dual nutrient (N & P) management strategy. Water Res. 45(5):1973-83

Paerl HW, Hall NS, Calandrino ES (2011b) Controlling harmful Cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change. Sci Total Environ. 409(10):1739-45.

Pearson LA, Hisbergues M, Börner T, Dittmann E, Neilan BA. (2004) Inactivation of an ABC transporter gene, mcyH, results in loss of microcystin production in the Cyanobacterium *Microcystis aeruginosa* PCC 7806. Appl Environ Microbiol. 70(11):6370-8.

Pham MN, Tan HTW, Mitrovic S, Yeo HHT (2011) A checklist of the algae of Singapore. Raffles Museum of Biodiversity Research. National University of Singapore ISBN 978-981-08-9284-5 (online)

Poroyko V, White JR, Wang M, Donovan S, Alverdy J, Liu DC, Morowitz MJ (2010) Gut microbial gene expression in mother-fed and formula-fed piglets. PLoS One. 5(8):e12459.

PUB (2007) "ABC Water Masterplan Western Catchment". ABC Waters Programme. Public Utilities Board. <http://www.pub.gov.sg/abcwaters/ABCWaterMasterPlan/Documents/FMP4fastviewing_Sept2007.pdf >

Rapala J, Sivonen K, Lyra C, Niemelä SI (1997) Variation of microcystins, Cyanobacterial hepatotoxins, in Anabaena spp. as a function of growth stimuli. Appl. Environ. Microbiol. 63(6): 2206-2212

Rashidan KK, Bird DF. (2001) Role of predatory bacteria in the termination of a Cyanobacterial bloom. Microb Ecol 41: 91-105.

Rinehart KL, Namikoshi M and Choi BW (1994) Structure and biosynthesis of toxins from blue-green algae (Cyanobacteria). J Appl Phycol, 6, 159-176.

Rinta-Kanto JM, Ouellette AJ, Boyer GL, Twiss MR, Bridgeman TB, Wilhelm SW. (2005) Quantification of toxic *Microcystis* spp. during the 2003 and 2004 blooms in western Lake Erie using quantitative real-time PCR. Environ. Sci. Technol. 39:4198–4205.

Rinta-Kanto JM, Wilhelm SW. (2006) Diversity of Microcystin-Producing Cyanobacteria in Spatially Isolated Regions of Lake Erie. Appl Environ Microbiol 72(7):5083-5.

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. (2006) TM4 microarray software suite. Methods in Enzymology. 411:134-93.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. (2003) TM4: a free, open-source system for microarray data management and analysis. 34(2):374-8.

Simek K, Hornák K, Jezbera J, Nedoma J, Vrba J, Straskrábová V, Macek M, Dolan JR, Hahn MW. (2006) Maximum growth rates and possible life strategies of different bacterioplankton groups in relation to phosphorus availability in a freshwater reservoir. Environ Microbiol. 8(9):1613-24.

Simek K, Kasalicky V, Jezbera J, Jezberová J, Hejzlar J, Hahn MW. (2010) Broad habitat range of the phylogenetically narrow R-BT065 cluster, representing a core group of the Betaproteobacterial genus Limnohabitans. Appl Environ Microbiol. 76(3):631-9.

Singh DP, Tyagi MB, Kumar A, Thakur JK, Kumar A (2001) Antialgal activity of a hepatotoxin-producing Cyanobacterium *Microcystis aeruginosa*. World J. Microbiol. Biotechnol. 17, 15-22.

Sivonen K, Jones G. (1999) Cyanobacterial toxins. Toxic Cyanobacteria in water. A guide to their public health consequences, monitoring and management (Chorus, I. and Bartram, J., Eds.), 41-111.

Stewart I, Carmichael WW, Sadler R, McGregor GB, Reardon K, Eaglesham GK, Wickramasinghe WA, Seawright AA, Shaw GR. (2009) Occupational and environmental hazard assessments for the isolation, purification and toxicity testing of Cyanobacterial toxins. Environ Health. 8:52.

Stewart FJ, Ottesen EA, DeLong EF. (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J. 4(7):896-907.

Stoeckel DM, Harwood VJ (2007) Performance, design, and analysis in microbial source tracking studies. Appl Environ Microbiol. 73(8):2405-15.

Straub C, Quillardet P, Vergalli J, de Marsac NT, Humbert JF. (2011) A day in the life of *Microcystis aeruginosa* strain PCC 7806 as revealed by a transcriptomic analysis. PLoS One. 6(1):e16208.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102, 15545-15550.

Sukenik A, Eshkol R, Livne A, Hadas O, Rom M, Tchernov D, Vardi A, Kaplan A (2002) Inhibition of growth and photosynthesis of the dinoflagellate *Peridinium gatunense* by *Microcystis* sp. (Cyanobacteria): a novel allelopathic mechanism. Limnol. Oceanogr. 47, 1656-1663.

Te SH, Gin KYH (2011) The dynamics of Cyanobacteria and microcystin production in a tropical reservoir of Singapore. Harmful Algae 10(3), 319-329.

Tucker S, Pollard P. (2005) Identification of cyanophage Ma-LBP and infection of the Cyanobacterium *Microcystis aeruginosa* from an Australian subtropical lake by the virus. Appl Environ Microbiol. 71(2):629-35.

Utkilen H, Gjolme N (1995) Iron-stimulated toxin production in *M. aeruginosa*. Appl. Environ. Microbiol. 61: 797-800.

Xing P, Guo L, Tian W, Wu QL.(2011) Novel *Clostridium* populations involved in the anaerobic degradation of *Microcystis* bloom. ISME 5, 792-800.

Yoshida T, Takashima Y, Tomaru Y, Shirai Y, Takao Y, Hiroishi S, Nagasaki K. (2006) Isolation and characterization of a cyanophage infecting the toxic Cyanobacterium *Microcystis aeruginosa*. Appl Environ Microbiol. 72(2):1239-47.

Yoshida M, Yoshida T, Kashima A, Takashima Y, Hosoda N, Nagasaki K, Hiroishi S. (2008) Ecological dynamics of the toxic bloom-forming Cyanobacterium *Microcystis aeruginosa* and its cyanophages in freshwater. Appl Environ Microbiol. 74(10):3269-73.

Zavarzin GA, Stackebrandt E, Murray RG. (1991) A correlation of phylogenetic diversity in the Proteobacteria with the influences of ecological forces. Can. J. Microbiology. 37: 1-6.

Zeder, M., S. Peter, T. Shabarova, and J. Pernthaler. (2009) A small population of planktonic Flavobacteria with disproportionally high growth during the spring phytoplankton bloom in a prealpine lake. Environ. Microbiol. 11: 2676-2686.

Zegura B, Sedmak, B, and Filipic M (2003) Microcystin-LR induces oxidative DNA damage in human hepatoma cell line HepG2. Toxicon 41, 41-8.

Zinser ER, Lindell D, Johnson ZI, Futschik ME, Steglich C, Coleman ML, Wright MA, Rector T, Steen R, McNulty N, Thompson LR, Chisholm SW. (2009) Choreography of the trancriptome, photophysiology, and cell cycle of a minimal photoautotroph , Prochlorococcus. PLos One Vol 4, Issue 4, e5135.

Zurawell RW, Prepas EE. (2005) Hepatotoxic Cyanobacteria: A review of the biological importance of microcystins in freshwater environments. J. Toxicology and Environmental Health, Part B, 8:1-37.

Appendix

Appendix A: Perl Code for processing of raw illumina reads

```
perl fastq_btrim.pl -a 101222_Solexa-JT
Lib2_samodha_Thompson_L7_1_sequence.txt -t T
perl fastq_btrim.pl -a 101222_Solexa-JT-
Lib2_samodha_Thompson_L7_2_sequence.txt -t T


perl bcsort_fastq_pe_v3.pl -a 101222_Solexa-JT-
Lib2_samodha_Thompson_L7_1_sequence_Btrim.fq -b 101222_Solexa-JT-
Lib2_samodha_Thompson_L7_2_sequence_Btrim.fq -c barcodes.txt


perl -e '$count=0; $len=0; while(<>) {s/\r?\n//; s/\t/ /g; if
(s/^>//) { if ($. != 1) {print "\n"} s/ |$/\t/; $count++; $_ .=
"\t";} else {s/ //g; $len += length($_)} print $_;} print "\n"; warn
"\nConverted $count FASTA records in $. lines to tabular
format\nTotal sequence length: $len\n\n";' krb1.fa > krb1.tab
perl -e '$column = 2; $unique=0; while(<>) {s/\r?\n//; @F=split /\t/,
$_; if (! ($save{$F[$column]}++)) {print "$_\n"; $unique++}} warn
"\nChose $unique unique lines out of $. total lines.\nRemoved
duplicates in column $column.\n\n"' krb1.tab > krb1_unique.tab
perl -e '$len=0; while(<>) {s/\r?\n//; @F=split /\t/, $_; print
">$F[0]"; if (length($F[1])) {print " $F[1]"} print "\n"; $s=$F[2];
$len+= length($s); $s=~s/.{60}(?=.)/$&\n/g; print "$s\n";} warn
"\nConverted $. tab-delimited lines to FASTA format\nTotal sequence
length: $len\n\n";' krb1_unique.tab > krb1_unique.fa


perl -e '$count=0; $len=0; while(<>) {s/\r?\n//; s/\t/ /g; if
(s/^>//) { if ($. != 1) {print "\n"} s/ |$/\t/; $count++; $_ .=
"\t";} else {s/ //g; $len += length($_)} print $_;} print "\n"; warn
"\nConverted $count FASTA records in $. lines to tabular
format\nTotal sequence length: $len\n\n";' krb6_unique.fa >
krb6_ls.tab
perl -e '$col = 2;' -e 'while (<>) { s/\r?\n//; @F = split /\t/, $_;
$len = length($F[$col]); print "$_\t$len\n" } warn "\nAdded column
with length of column $col for $. lines.\n\n";' krb6_ls.tab >
krb6_ls_length.tab
perl -e '$col=3; $limit=20; while(<>) {BEGIN {$count=0} s/\r?\n//;
@F=split /\t/, $_; if ($F[$col] > $limit) {$count++; print "$_\n"}}
warn "\nChose $count lines out of $..\n\n"' krb6_ls_length.tab >
krb6_only_long.tab
perl -e '@cols=(0, 1, 2); while(<>) {s/\r?\n//; @F=split /\t/, $_;
print join("\t", @F[@cols]), "\n"} warn "\nJoined columns ", join(",
", @cols), " for $. lines\n\n"' krb6_only_long.tab >
krb6_only_long_three_cols.tab
perl -e '$len=0; while(<>) {s/\r?\n//; @F=split /\t/, $_; print
">$F[0]"; if (length($F[1])) {print " $F[1]"} print "\n"; $s=$F[2];
$len+= length($s); $s=~s/.{60}(?=.)/$&\n/g; print "$s\n";} warn
"\nConverted $. tab-delimited lines to FASTA format\nTotal sequence
```

```
length: $len\n\n";' krb6_only_long_three_cols.tab >
krb6_only_long.fa
perl -e ' $split_seqs=100000; $out_template="krb6_NUMBER.fasta";
$count=0; $filenum=0; $len=0; while (<>) { s/\r?\n//; if (/^>/) { if
($count % $split_seqs == 0) { $filenum++; $filename = $out_template;
$filename =~ s/NUMBER/$filenum/g; if ($filenum > 1) { close SHORT }
open (SHORT, ">$filename") or die $!; } $count++; } else { $len +=
length($_) } print SHORT "$_\n"; } close(SHORT); warn "\nSplit
$count FASTA records in $. lines, with total sequence length
$len\nCreated $filenum files like $filename\n\n"; '
krb6_only_long.fa
```

Appendix B: "Cluster_Blastx_all.csh" file containing BLASTX parameters

```
#! /bin/sh
#$ -S /bin/bash
#$ -t 1-140
#$ -cwd
export CLASSPATH=/opt/Bio/ncbi/bin
OUTBASE=/home/jiawang/krb/
SEEDFILEF=${OUTBASE}/fasta_list_all.txt
SEEDF=$(sed -n -e "$SGE_TASK_ID p" $SEEDFILEF)
/opt/Bio/ncbi/bin/blastall -p blastx \
    -i ${OUTBASE}/${SEEDF} \
    -d /data/genomics/ncbi/blast/Jan11/nr \
    -b 10 -v 10 \
    -W 2 \
     -e 100 \
    -F "m S" \
     -Q 11 \
     -f 8 \
  -o ${OUTBASE}/${SEEDF}.blastx
```

Appendix C: Amount of RNA extracted and amplified after multiple subtractions.

Samples	Volume of reservoir water filtered (ml)	Total RNA extracted (ng)	Total RNA/water filtered (ng/ml)	Amplified RNA (ng)
5pm (krb1)	275	435	1.58	3,150
8pm (krb2)	215	240	1.12	3,450
6am (krb3)	215	265	1.23	2,700
8am (krb4)	350	465	1.33	6,300
1pm (krb5)	350	660	1.89	16,800
3pm (krb6)	350	1,305	3.73	25,650

Appendix D: Barcodes, Adaptors, PCR primers and Sequencing primers used in Illumina library preparation and sequencing

Barcodes:

krb1: AAGTCA
krb2: AGAGGT
krb3: CATAAG
krb4: CCTGAT
krb5: CTAGTA
krb6: GTCATC

Adaptors:

krb1:
PE02.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACTTT
PE02.2: /5Phos/AAGTCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

krb2:
PE08.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCTCTT
PE08.2: /5Phos/AGAGGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

krb3:
PE16.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTATGT
PE16.2: /5Phos/CATAAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

krb4:
PE19.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCAGGT
PE19.2: /5Phos/CCTGATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

krb5:
PE24.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACTAGT
PE24.2: /5Phos/CTAGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

krb6:
PE37.1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGACT
PE37.2: /5Phos/GTCATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

PCR primers for amplification of ligated products:

PE PCR1:
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE PCR2:
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Illumina sequencing primers:

PE SEQ1: ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE SEQ2: CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT