EXPLORING MICROBIAL COMMUNITY DYNAMICS

EXPLORING MICROBIAL COMMUNITY DYNAMICS: POSITIVE SELECTION FOR GAIN OF RPOS FUNCTION IN *ESCHERICHIA COLI* & MICROBIAL PROFILING OF THE NIAGARA REGION

By STEVEN R. BOTTS, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University

MASTER OF SCIENCE (2016) McMaster University

(Biology) Hamilton, Ontario

TITLE: Exploring microbial community dynamics: Positive selection for gain of RpoS function in *Escherichia coli* & microbial profiling of the Niagara Region

AUTHOR: Steven R. Botts, B.Sc. (McMaster University, Hamilton, ON)

SUPERVISOR: Dr. Herb E. Schellhorn

NUMBER OF PAGES: xii,128

LAY ABSTRACT

The effect of changing environmental conditions on microbial population structure can be observed at both the species and community level. Within the *Escherichia coli* species, we investigated reversion of loss of function mutations in the RpoS protein regulator in high salt conditions and identified RpoS restoration under selective pressure. At the community level, we examined the microbial DNA of water samples from the Niagara Region under select environmental conditions and assessed the viability of next-generation sequencing in augmenting traditional water quality monitoring methods. Both within-species and community-wide analyses offer insight into how microbial populations respond and adapt to environmental fluctuations.

ABSTRACT

The effect of changing environmental conditions on microbial population structure can be observed at both the species and community level. Within the *Escherichia coli* species, null mutations in the RpoS stationary phase regulator are commonly selected by growth on poor carbon sources. In contrast, mutations which restore RpoS function may provide a selective advantage for cells exposed to environmental stress. The loss and subsequent restoration of RpoS form a population-level switch for adaptation within poor carbon and high stress environments. To investigate selection for RpoS reversion, we exposed *rpoS*-deficient *E. coli* to high salt concentrations and assessed the phenotype of presumptive mutants. 3-9% of salt-resistant mutants contained reversion mutations within *rpoS*, while in 91-97% the loss of RpoS function was maintained and mutations at alternative gene loci were identified. These results show that RpoS function can be restored in deficient *E. coli* under selective pressure. At the community level, the application of next-generation sequencing (NGS) technology to characterize environmental microbial diversity can potentially augment traditional water quality monitoring methods. To investigate the use of NGS in identifying microbial taxa within the Niagara Region, we collected water samples from Lake Erie, Lake Ontario, and nearby areas and examined the metagenome of microbial communities. A QIIME (Quantitative Insights Into Microbial Ecology) analysis of sequence data identified significant differences in relative microbial abundance with respect to sample metadata (e.g. location and subtype), significant correlations between relative abundance and quantitative parameters (e.g. *Escherichia coli* counts and fecal DNA markers), and detected pathogen-containing taxa at a relative

abundance of 0.1-1.5%. These results show that sequence-based analyses can be used in conjunction with traditional identification methods to profile the metagenomic community of environmental samples and predict water quality. Both within-species and community-wide analyses thus offer insight into how microbial populations respond and adapt to environmental fluctuations.

ACKNOWLEDGEMENTS

I would first like to thank my supervisor and mentor Dr. Herbert Schellhorn, who helped foster my passion for scientific research. As a second year undergraduate student, he introduced me to the world of academia, refined my technical skills, and allowed me the freedom to pursue my own scientific interests. His guidance and encouragement in life and research have been invaluable during my undergraduate and graduate studies. I will always look fondly upon our discussions of history, literature, and science, those of which have shaped myself as a competent researcher and individual.

In addition, I would like to thank my co-supervisor Dr. Brian Golding, who taught me the foundations of bioinformatics and provided a unique perspective on computational techniques within the field.

Finally, I would like to thank Sarah Chiang for supporting my transition into the Schellhorn Lab and Athanasios Paschos for providing guidance during my final graduate years, as well as past and present lab members Mohammad Mohiuddin, Mohammad Howard-Azzeh, and Shirley Wong for contributing to a productive and enjoyable research environment.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1:

Positive selection for gain of RpoS function in *Escherichia coli*

1.1 ABSTRACT

RpoS, while an important stationary phase regulator in the enteric bacterium *Escherichia coli*, is attenuated by loss of function mutation in a considerable fraction of natural populations. Null mutations within *rpoS* are commonly selected by growth on poor carbon sources or during extended incubation (likely due to increases in nutrient transport and/or oxidative metabolism- both of which are traits of *rpoS* mutants). In contrast, mutations which restore RpoS function may provide a selective advantage for cells exposed to environmental stress. The loss and subsequent restoration of RpoS form a population-level switch for adaptation within poor carbon and high stress environments. To investigate selection for RpoS reversion, we exposed *rpoS*-deficient *E. coli* to high salt concentrations and assessed the phenotype of presumptive mutants. 3-9% of salt-resistant mutants contained reversion mutations within *rpoS*, while in 91-97% the loss of RpoS function was maintained and mutations at alternative gene loci were identified. These results show that RpoS function can be restored in deficient *E. coli* under selective pressure. As osmotic stress is often encountered in the environment, this selection may serve as an adaptive mechanism for competing cells.

1.2 INTRODUCTION

This literature review will focus on the growth and adaptation of *Escherichia coli* within the natural environment. It will investigate the role of the RpoS transcriptional regulator in responding to environmental stress and introduce the concept of an RpoS mutational switch- that which facilitates selective loss and gain of RpoS function at the population level. Finally, it will discuss recent advances in DNA sequencing technology which allow for comprehensive mutational analyses.

1.2.1 *Escherichia coli*

1.2.1.1 A model organism

The gamma-proteobacterium *Escherichia coli*, a facultative anaerobe and well-studied human commensal, colonizes the gastrointestinal tract during infancy and consequently thrives in a probiotic, mutualistic relationship with its host (Altenhoefer et al., 2004). Isolated in 1922 and easily amenable to genetic manipulation (e.g. conjugation (Lederberg, 1946) and generalized transduction (Lennox, 1955)), the bacterium serves as a prokaryotic model organism in numerous microbial studies. The *E. coli* pangenome is estimated to contain more than 13,000 genes and include a core set of approximately 2,200 (Rasko et al., 2008). Through analogous proteins and regulatory pathways, the wealth of knowledge acquired from experiments involving *E. coli* has helped elucidate gene function in a substantial number of prokaryotic species, as well as within many eukaryotes.

1.2.1.2 *E. coli* in the environment

As an enteric bacterium, *E. coli* is frequently introduced in the environment from animal reservoirs. Concentrations of harmful and/or commensal *E. coli s*trains reach up to 10^9 cells per gram of sewage from humans and animals (Edberg et al., 2000). The primary reservoir of pathogenic *E. coli*, cattle, can harbour concentrations of enterohaemorrhagic *E. coli* O157:H7 as high as 10^7 cells per gram, presenting an important concern for public health (Chase-Topping et al., 2007). *E. coli* grows optimally in the warm and nutrient-rich environment of the gastrointestinal tract and minimally outside of a host, but persistence of cells can be high. For example, *E. coli* cells in marine sediment remain culturable over 68 days without an significant difference in cell count (Davies et al., 1995), which is similarly seen in water (Fish et al., 1995). *E. coli* may therefore persist in diverse environmental, and potentially stressful, conditions (e.g. osmotic stress encountered during the desiccation of urine (Putnam, 1971)).

1.2.2 RpoS and the global stress response

1.2.2.1 The RpoS sigma factor

The *rpoS* gene and RpoS protein denote a prokaryotic transcriptional regulator synthesized in *Escherichia coli* and many proteobacteria (Chiang et al., 2010). As a sigma factor, RpoS complexes with the RNA polymerase (RNAP) core enzyme, facilitates binding of the resultant holoenzyme to downstream promoters, and initiates the transcription of corresponding genes (Ishihama, 2000). After transcription initiation, sigma factors generally dissociate from the holoenzyme (Ishihama, 2000) or shift into a weakly

bound conformation (Kapanidis et al., 2005). The structural change produced in the holoenzyme complex allows for RNAP-mediated transcription elongation and generation of an RNA product (Ishihama, 2000).

The RpoS sigma factor is associated with gene regulation upon entry into the bacterial stationary phase following exponential growth (Dong, 2008) and exposure to diverse environmental stresses- including near-UV radiation (Sammartano et al., 1986), acid exposure (Small et al., 1994), heat shock (Hengge-Aronis et al., 1991), oxidative stress (Sammartano et al., 1986), and nutrient deprivation (Lange et al., 1991). Downstream targets of RpoS thus have an essential role in shaping bacterial adaptation and survival within the environment.

1.2.2.2 Regulation of *rpoS*

Both the *rpoS* gene and RpoS protein are regulated at several levels of gene and protein organization. Prior to transcription, *rpoS* is directly regulated through binding sites flanking the major promoter (e.g. as observed in *arcA*-induced repression (Mika, 2005)). Following transcription, regulatory elements facilitate translation by mRNA stabilization (e.g. the *cspC/cspE* complex (Phadtare et al., 2001)), alteration of mRNA secondary structure (e.g. *hfq*; *rprA* (Brown et al., 1997; Majdalani et al., 2002)), or anti-sense mechanism (e.g. d*srA* (Majdalani et al., 1998)). At the post-translational level, the RpoS protein may be stabilized (e.g. *dnaK*; *iraP* (Rockabrand et al., 1998; Bougdour et al., 2006)) or targeted for protease-mediated degradation (e.g. *ClpXP* and *rssB* (Zhou et al., 2001)).

Together, these studies outline an intricate regulatory network for controlling the expression and function of the RpoS protein.

1.2.2.3 The RpoS regulon

The genes and/or operons under regulatory control of RpoS are collectively referred to as the RpoS regulon and account for nearly 10% of the *E. coli* K-12 genome in both stationary phase and under environmental stress (Patten et al., 2004). Consequently, *rpoS* expression manifests within a multitude of cellular pathways and functions, including stress resistance (e.g. DNA damage repair by *xthA*-encoded exonuclease (Demple, 1983)), morphology (e.g. the *osmB*-encoded outer membrane lipoprotein involved in cell aggregation (Jung et al., 1990)), metabolism (e.g. the *gadX*-encoded AraC transcriptional activator of the *E. coli* glutamic acid decarboxylase pathway (Tramonti et al., 2002)), and molecular transport (e.g. the *potF*-encoded ABC superfamily putrescine transporter (Kabir et al., 2004)).

RpoS regulation of *E. coli* virulence genes is also of particular concern for microbial pathogenesis and public health. Previous microarray analyses have elucidated RpoS-regulation of O-island virulence factors within the enterohaemorrhagic *E. coli* O157:H7 strain EDL933 (e.g. the LEE island elements *ler*, *cesF*, and *Z5139*) (Dong et al., 2009).

Considerable diversity in RpoS regulon expression has been observed among *E. coli* isolates from natural environments (Chiang et al., 2011). Assays of these naturally occurring RpoS mutants (0.3% of 2,040 environmental isolates) have revealed differential

expression of regulon genes *KatE* and *AppA*, in addition to reduced catalase activity and growth on alternative carbon sources (i.e. succinate).

1.2.2.4 The RpoS population-level switch

RpoS is frequently selected for loss in *E. coli* by growth on poor carbon (Chen et al., 2004; Dong et al., 2009), limited carbon and nitrogen sources (Notley-McRobb et al., 2002), and during extended incubation (King et al., 2004), likely because this loss allows for increased expression of tricarboxylic cycle enzymes (Patten et al., 2004). In contrast, mutations which restore RpoS function may provide a selective advantage for cells exposed to environmental stress. The conflicting pleiotropy in RpoS expression was first discussed in a 2003 review by Ferenci and later proposed as an *rpoS* population-level switch using a laboratory model (Chen et al., 2004). The previous model plated RpoS-dependent *lacZ* fusion strains on lactose so that exclusively RpoS revertants would grow. It is not clear, however, if natural conditions will result in similar selection for the *rpoS* wild type. The potential for reversion in natural environments would allow for mutations to serve as an adaptive mechanism in nature.

1.2.3 DNA sequencing

1.2.3.1 Next-generation techniques

Within the past several decades, an evolving need for the efficient, low-cost sequencing of large templates has driven the development of high-thoughput, next-generation sequencing technologies. These methods reduce the cost of traditional

technologies by fragmenting samples and conducting massively parallel sequencing-by-synthesis reactions, generating as many as 6 billion paired-end reads per run (Peng et al., 2013).

Next-generation sequencing techniques utilize parallelization to reduce the time associated with processing large metagenomic or whole-genome samples. For example, 454-sequencing, relies on the parallel amplification of bound adapter-ligated DNA fragments within emulsion micro-chambers and detection of pyrophosphate release upon nucleotide incorporation (Rothberg et al., 2008). The 454 GS FLX+^{TM} system generates an output of up to 1,000,000 reads and total throughput of up to 700 Mb (Schellhorn et al., 1998). Alternatively, Illumina HiSeq^{TM} technology relies on the parallel "bridge amplification" of adapter-ligated fragments bound to a solid surface and successive nucleotide washes of fluorescent reversible terminators. The Hi-Seq^{TM} 2000 system outputs up to 3 billion single reads or 6 billion paired-end reads, resulting in a total throughput of up to 600 Gb at a read length of 2x100 bp (Peng et al., 2013).

1.2.3.2 Genome assembly

Due to limitations in the read length of modern sequencing technologies, larger DNA samples must first be fragmented into libraries of overlapping fragments, or contigs, prior to sequencing events. This bottom-up approach to generating sequence data requires the assembly of contigs by comprehensive alignment of overlapping regions- an analysis typically conducted by computer programs or sequence assemblers. The computational and temporal requirements of this analysis are further reduced through the use of an existing

reference sequences for direct alignment of contigs- a process referred to as mapping assembly. Sequence assembly by mapping contrasts more laborious *de-novo* methods, which require more in-depth contig analysis in the absence of reference DNA (Pevzner, 2000).

Several open-source and commercial sequence assemblers are currently available for data analysis. Many of these programs have been specifically designed integrate data generated by multiple sequencing technologies. For instance, the hybrid version of the MIRA assembler was published by Chevreux and colleagues as the first open-source assembler capable of assembling combinations of 454 and Sanger sequencing reads (Chevreux et al., 2004). Other assemblers have been engineered to meet the continual demand for assembly of greater read volumes. The SHARCGS assembler published by Dohm and colleagues, for example, was the first assembler used to compile Illumina sequence data contained 100 million reads of 300-400 base-pairs (Dohm et al., 2007).

More recently, the Bowtie 2 aligner was designed to map high-throughput sequence data of about 50 up to 100s or 1,000s of characters per read (Langmead et al., 2012). Its alignment algorithm operates in the following four stages- first, seed substrings are extracted from reads and corresponding reverse compliments. Each substring is then aligned to the provided reference without gap consideration by a full-text minute index. The alignments are subsequently prioritized and assessed for position along the reference. Finally, seeds are extended into full-length alignments by SIMD-accelerated programming (Langmead et al., 2012).

1.2.3.3 Galaxy

The Galaxy project was initiated in 2005 to develop an open, web-based platform for genomic research (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010). Galaxy serves as a bioinformatics workflow management system, allowing for the execution of a series of data manipulation tasks by graphical user interface. The platform is available as a public web server or open-source software for local or cloud operation. The standard Galaxy interface is depicted in Figure 1.

a)

b) c)

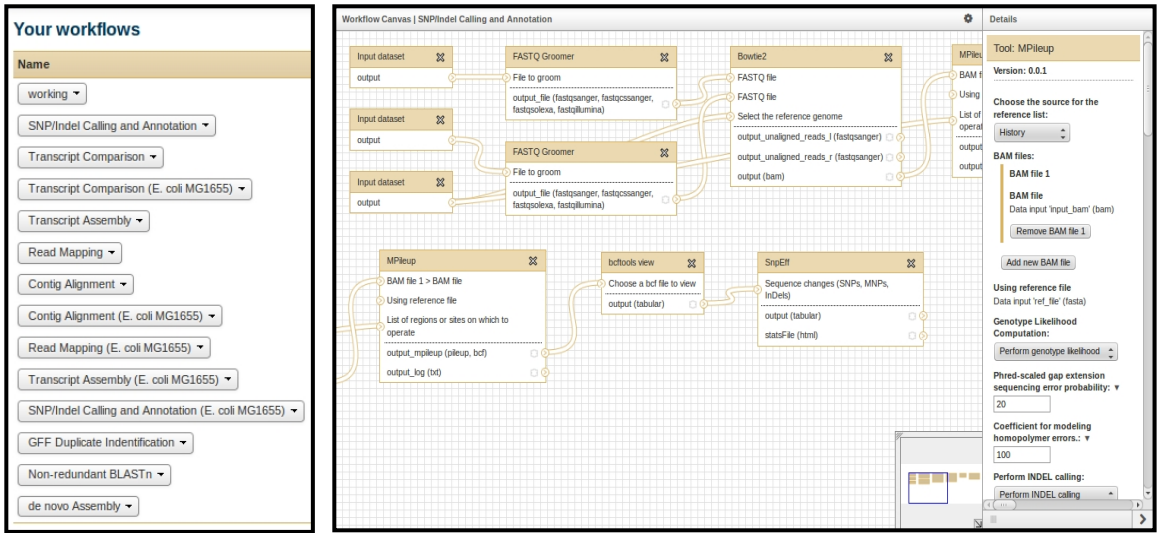


Figure 1. Galaxy implementation for data integration and analysis.

a) Galaxy functions as a bioinformatics workflow management system for the manipulation and analysis of biomedical data. Workflows for desired bioinformatic tasks (e.g. read mapping and SNP calling) are b) selected from a prebuilt database or c) designed using a custom editor.

1.3 PROJECT OUTLINE

1.3.1 Objectives

1. Selection for the gain of RpoS function in *E. coli* under natural conditions (i.e. high salt). *rpoS*$^{+}$ and *rpoS*$^{-}$ *E.coli* will be first assayed for growth in 6% NaCl to confirm the role of RpoS in adaptation under osmotic stress. *rpoS*$^{-}$ *E. coli* will then be plated on 6% NaCl in the presence of X-gal to select for salt-resistent mutants. Colonies with restored RpoS function will manifest a blue phenotype due to an *osmY-lacZ* fusion. This methodology is furthered described in Section 1.4.6.
2. Analysis of second-site mutations at alternative gene loci which may confer salt resistance. White colony isolates from salt viability experiments will be sequenced and aligned with the reference genomes of the corresponding *rpoS*$^{-}$ strains to identify mutations.
3. Development of a local Galaxy server for lab distribution and execution of relevant bioinformatics tasks (e.g. *de novo* assembly and contig alignment, BLAST search, read mapping, and transcript assembly). An annotated *E. coli* pangenome will be developed for use as a reference during read or contig alignment.

1.3.2 Significance

This project studies RpoS-dependent growth and RpoS reversion of *E. coli* in a natural condition (i.e. high salt). Previous studies have demonstrated selection for the loss of RpoS function in poor carbon environments (Chen et al., 2004; Dong et al., 2009). The potential for RpoS reversion would therefore support the hypothesis of an RpoS switch for natural

E. coli populations. Mutations resulting in this reversion may then serve as an adaptive mechanism for *E. coli* within high stress environments.

1.4. METHODS

1.4.1 Bacterial strains

Strains used in this study are listed in Table 1. The *Escherichia coli* K-12 strain MG1655 and previously constructed *rpoS*$^-$ derivative HS2210 (Patten et al., 2004) were assayed for growth under osmotic stress. *rpoS*$^-$ derivatives of *E. coli* GC4468 (Carlioz et al., 1986) containing an *osmY-LacZ* fusion (Schellhorn et al., 1998) were assayed for gain of function mutation in high salt.

Table 1. Strains used in this study.

Strain	Genotype	Source/reference
MG1655	F- λ- *ilvG- rfb-50 rph-1*	CGSC, Yale University
HS2210	as MG1655 but *ΔrpoS, precise deletion*	(Patten et al., 2004)
GC4468	F- Δ(*lac-argF*)U169 *rpsL179*	(Carlioz et al., 1986)
HS1091	as GC4468 but *rsd91-lacZ+*	(Schellhorn et al., 1998)
HS1091p	as HS1091 but *rpoS::Tn10*	(Schellhorn et al., 1998)
SS13	as HS1091 but *rpoS* (L234*)	Schellhorn Lab- unpublished
SS35	as HS1091 but *rpoS* (95_96insTTAGTAGA)	Schellhorn Lab- unpublished
SS53	as HS1091 but *rpoS* (E265*)	Schellhorn Lab- unpublished

1.4.2 Media and growth conditions

For salt growth assays, *E. coli* strains were grown at 37°C with shaking in 0.5% glucose M9 minimal media (Miller, 1992) to an optical density at 600 nm (OD_{600}) of ~1. Cultures were used to inoculate 0.5% glucose M9 minimal media supplemented with 0-8% NaCl at an OD_{600} of ~0.05 in a 96-well microplate. Microplates were then incubated at 37°C with shaking. For salt viability assays, *E. coli* strains were grown at 37°C with shaking in Luria-Bertani (LB) media (Miller, 1992) to an OD_{600} of ~1. Cultures were resuspended

in 0.5% glucose M9 minimal media, plated on LB supplemented with 6% NaCl and 25 μg/mL X-gal, and incubated overnight at 37°C.

1.4.3 Phenotype microarray

E. coli MG1655 (*rpoS*$^+$) and HS2211 (*rpoS*$^-$) were previously submitted for analysis by Biolog Phenotype MicroArray^{TM}. Strains were subjected to nearly 2,000 metabolic and chemical sensitivity tests and assayed for differential growth.

1.4.4 Salt growth assays

E. coli MG1655 (*rpoS*$^+$) and HS2210 (*rpoS*$^-$) were incubated in-microplate at 37°C with shaking and assayed for OD_{600} every 30 min from 0 to 4.5 h using a Thermo Scientific Multiskan® Spectrum. The spectrophotometer was calibrated with 0.5% glucose M9 minimal media. *E. coli* strains were grown and assayed in triplicate.

1.4.5 Calculation of generation time

Generation times for salt growth assays were determined by taking the average slope of the linear range of $\log_2$-transformed OD_{600} values with time among replicates. Statistical analyses (i.e. unpaired samples T-tests) were conducted using Microsoft Excel® Data Analysis software.

1.4.6 Salt viability assays

E. coli GC4468 derivatives were centrifuged at 4,000 x *g* for 15 min and resuspended in 0.5% glucose M9 minimal media in two successive washes. Strains were diluted to ~10^7 cells/ml with M9 minimal media, plated on LB X-gal with 6% NaCl, and incubated overnight at 37°C to select for gain of function mutations. Strains were also diluted and plated on LB X-gal with 0% NaCl to determine total cell density and mutation frequency. RpoS reversion was predicted among blue colonies by increased expression of an RpoS-regulated *osmY-lacZ* fusion. Second-site mutations at alternative gene loci were predicted among white colonies. A proportion of *rpoS*$^-$ cells which did not undergo RpoS reversion or second-site mutation was identified by a smaller white colony phenotype. These smaller colonies were observed with an order of magnitude larger frequency and were not considered salt-resistant mutants in subsequent analyses. The viability assay and selection method are illustrated in Figure 2. Differential colony phenotypes are depicted in Figure 3.

The optimal dilution for salt viability assays was determined in previous experiments in which salt-resistant mutants were not observed at lower concentrations and a dense background of smaller white colonies prevented isolation of mutants at higher concentrations.

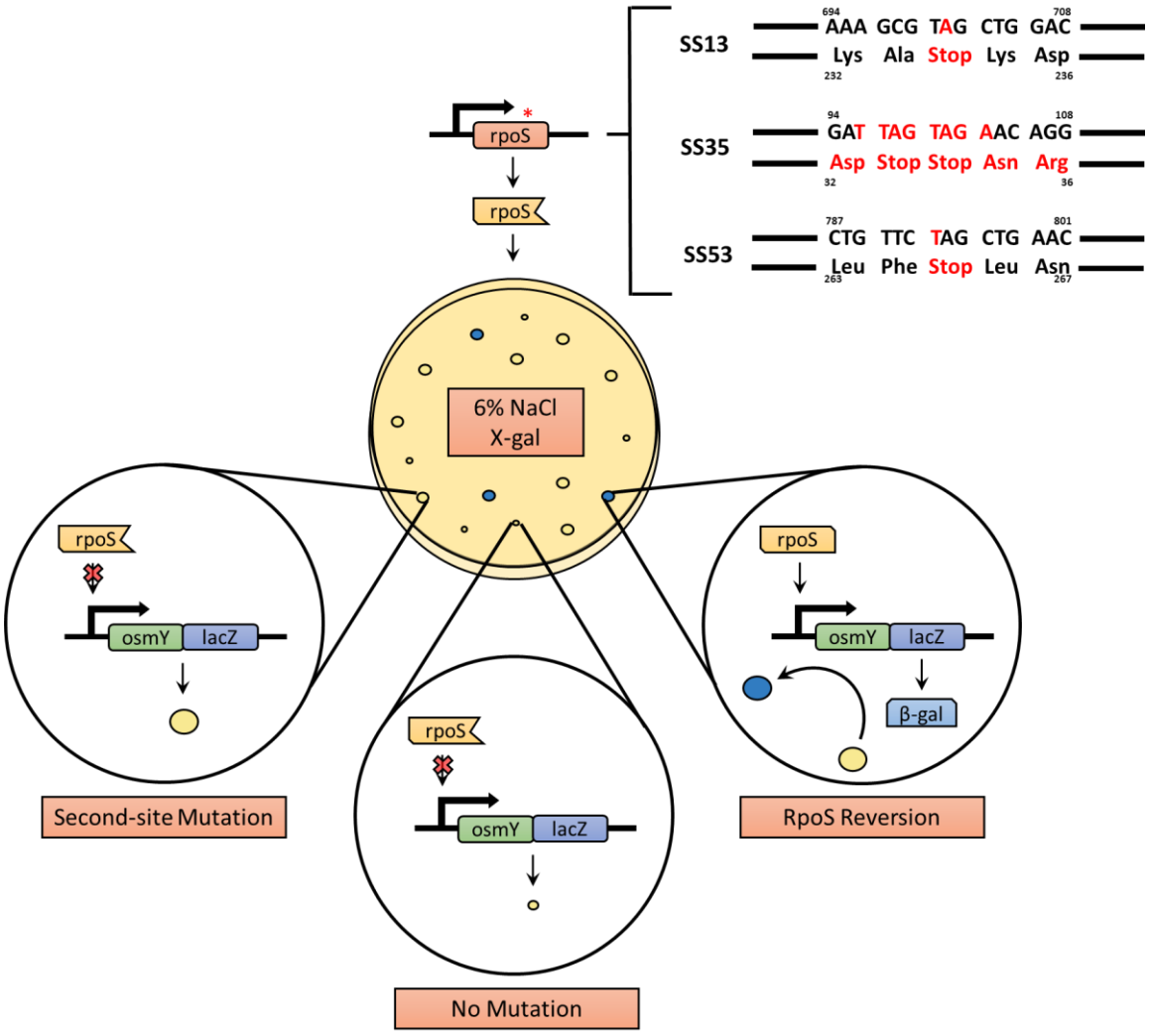


Figure 2. Isolation of RpoS revertants in high salt.

rpoS^{-} *E. coli* strains were plated on LB X-gal with 6% NaCl. Colonies display a blue or white phenotype due to an RpoS-dependent *osmY-lacZ* fusion. Those displaying a blue phenotype indicate RpoS reversion while those displaying a white phenotype indicate second-site mutations at alternative gene loci. Cells which did not undergo RpoS reversion or second-site mutation manifest with a smaller white colony phenotype and were not considered salt-resistant mutants in subsequent analyses.

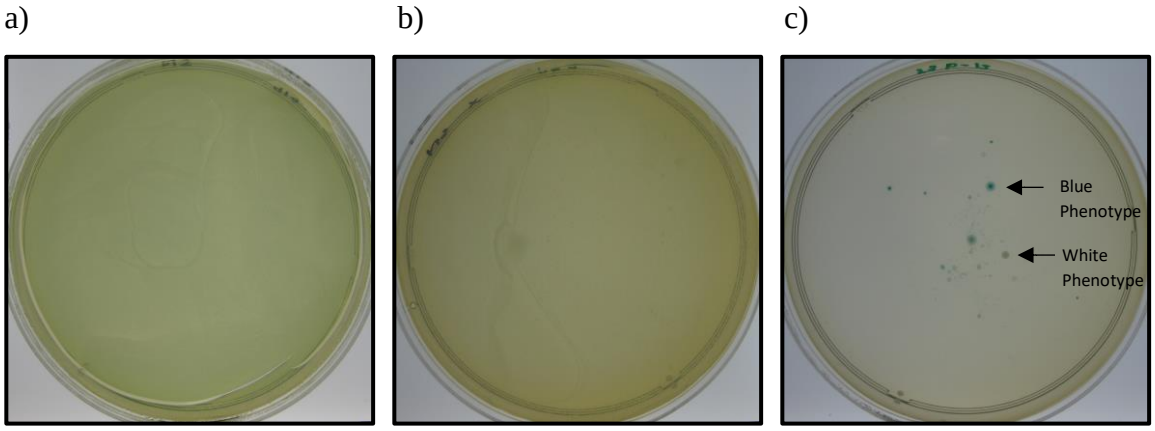


Figure 3. RpoS-dependant colony phenotypes on LB X-gal.

E. coli HS1091 (*rpoS*$^{+}$) and HS1091p (*rpoS*$^{-}$) were plated on LB X-gal with 0% NaCl. HS1091 and HS1091p respectively display a) blue and b) white phenotypes due to an RpoS-dependent *osmY-lacZ* fusion. c) *E. coli* SS53 (*rpoS*-) was plated on LB X-gal with 6% NaCl. Blue and white colony mutants respectively indicate RpoS reversion and second-site mutations at alternative loci.

1.4.7 Sequencing of salt-resistant mutants

One blue colony and 30 white colony mutants (20 from SS53, 5 from SS13, and 5 from SS35) were prepared with a Nextera XT Library preparation kit and sequenced with Illumina© HiSeq paired-end technology (2 x 100 bp). Genomes were sequenced with 100-fold converage to distinguish bona fide second-site mutations from sequencing errors.

1.4.8 RpoS reversion analysis

One blue colony and one white colony sequence file were aligned to *E. coli* str. K-12 substr. MG1655 (accession # NC_000913) within Galaxy (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010) using Bowtie 2 (Langmead et al., 2012) and

visualized with Broad Institute© Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

1.4.9 Galaxy

1.4.9.1 Implementation

A local installation of the Galaxy platform was created for distribution and data analysis within Linux operating systems. The directory structure for the server is depicted in Figure 4. Database, index, tool repository, and visualization folders allow for future expansion of function within the Galaxy distributable. The shell script fetch_taxonomy.sh imports taxonomic data for sequence annotation while setup.sh and start.sh respectively install and initialize the Galaxy server. Instructions for server configuration and data manipulation are included within a readme text file. The readme and shell scripts have been appended (File S1-S4).

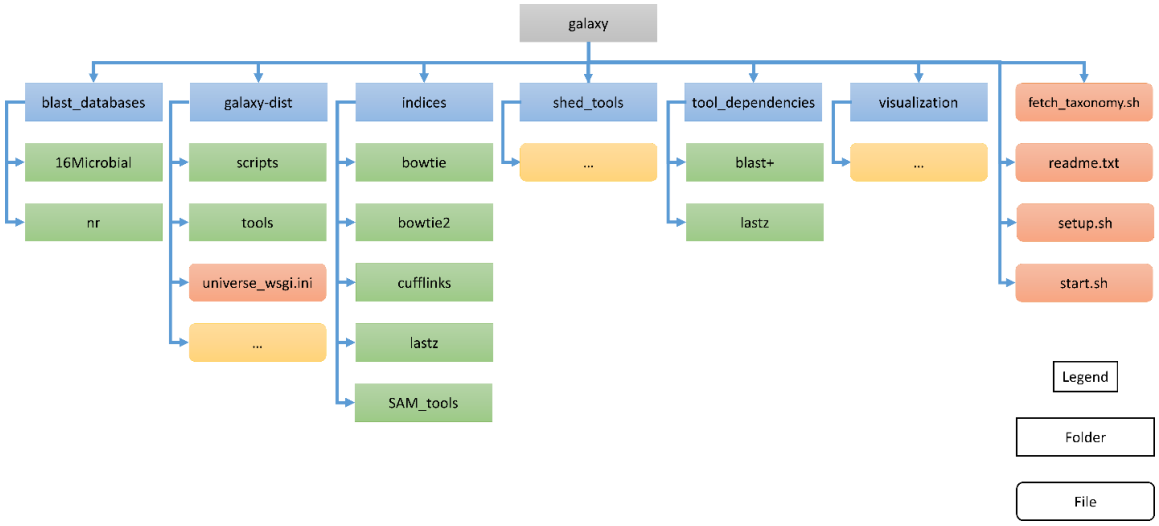


Figure 4. Directory structure of Galaxy server.

The local directory is comprised of database, index, and tool repositories, the galaxy distributable, as well as readme and shell files for installing and running the Galaxy server.

1.4.9.2 Workflow creation

A series of workflows were configured within the Galaxy server to allow efficient execution of common bioinformatics tasks (e.g. *de novo* assembly and contig alignment, BLAST search, read mapping, and transcript assembly). The complete list of workflows is detailed in Table 2.

Table 2. Workflows available within Galaxy server.

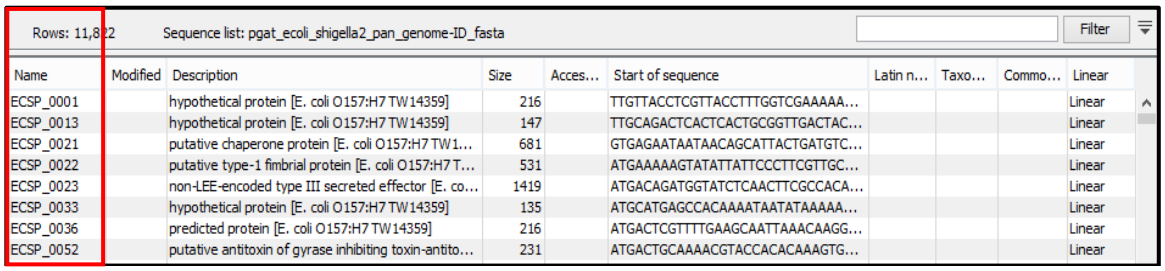
Workflow	Description	Programs
Contig Alignment (General & *E. coli* MG1655)	Alignment of assembled contigs to reference sequence	Lastz; SAM-to-BAM
De novo Assembly	Contig assembly from Illumina sequence reads	FastQC; FASTQ Groomer; FASTQ Interlacer; velveth; velvetg
GFF Duplicate Identification	Identification of duplicate entries within GFF files	Select; Group; Filter
Non-redundant BLASTn	Nucleotide BLAST against local database and taxonomic assignment	NCBI BLAST+ blastn; Trim; Fetch taxonomic representation
Read Mapping (General & *E. coli* MG1655)	Read mapping for Illumina sequence reads	FASTQ Groomer; Bowtie2
SNP/Indel Calling and Annotation (General & *E. coli* MG1655)	SNP/Indel identification and annotation from Illumina read alignment	FASTQ Groomer; Bowtie2; Mpileup; bcftools view; SnpEff
Transcript Assembly (General & *E. coli* MG1655)	Transcript assembly from Illumina sequence reads	FASTA Width; FastQC; FASTQ Groomer; Tophat for Illumina; Cufflinks
Transcript Comparison (General & *E. coli* MG1655)	Comparison between Illumina transcript assemblies	FASTA Width; Cuffcompare; Cuffdiff

1.4.10 Pangenome annotation

A pangenome file was generated from 54 *E. coli* and 9 *Shigella* genomes using PGAT, a prokaryotic genome analysis tool (Brittnacher et al., 2011). The pangenome contains 11,822 ORF entries and serves as a reference for read mapping and alignment. Entries were annotated with gene names using the Perl program pangenome_annotation.pl, which assigned gene names to respective sequence identifiers by annotation table. Input

and output files are depicted in Figure 5. The Perl program written for pangenome annotation has been appended (File S5).

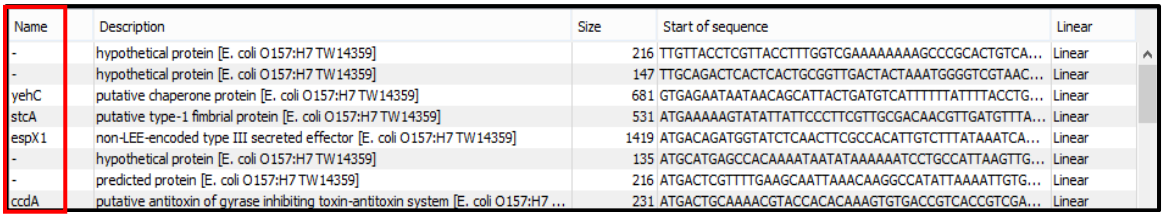
a)



Rows: 11,822 Sequence list: pgat_ecoli_shigella2_pan_genome-ID_fasta Filter

Name	Modified	Description	Size	Acces...	Start of sequence	Latin n...	Taxo...	Commo...	Linear
ECSP_0001		hypothetical protein [E. coli O157:H7 TW14359]	216		TTGTTACCTCGTTACCTTTGGTCGAAAAA...				Linear
ECSP_0013		hypothetical protein [E. coli O157:H7 TW14359]	147		TTGCAGACTCACTCACTGCGGTTGACTAC...				Linear
ECSP_0021		putative chaperone protein [E. coli O157:H7 TW1...	681		GTGAGAATAATAACAGCATTACTGATGTC...				Linear
ECSP_0022		putative type-1 fimbrial protein [E. coli O157:H7 T...	531		ATGAAAAAGTATATTATTCCCTTCGTTGC...				Linear
ECSP_0023		non-LEE-encoded type III secreted effector [E. co...	1419		ATGACAGATGGTATCTCAACTTCGCCACA...				Linear
ECSP_0033		hypothetical protein [E. coli O157:H7 TW14359]	135		ATGCATGAGCCACAAAATAATATAAAAA...				Linear
ECSP_0036		predicted protein [E. coli O157:H7 TW14359]	216		ATGACTCGTTTTGAAGCAATTAAACAAGG...				Linear
ECSP_0052		putative antitoxin of gyrase inhibiting toxin-antito...	231		ATGACTGCAAAACGTACCACACAAAGTG...				Linear

b)



Name	Description	Size	Start of sequence	Linear
-	hypothetical protein [E. coli O157:H7 TW14359]	216	TTGTTACCTCGTTACCTTTGGTCGAAAAAAAAGCCCGCACTGTCA...	Linear
-	hypothetical protein [E. coli O157:H7 TW14359]	147	TTGCAGACTCACTCACTGCGGTTGACTACTAAATGGGGTCGTAAC...	Linear
yehC	putative chaperone protein [E. coli O157:H7 TW14359]	681	GTGAGAATAATAACAGCATTACTGATGTCATTTTTTATTTTACCTG...	Linear
stcA	putative type-1 fimbrial protein [E. coli O157:H7 TW14359]	531	ATGAAAAAGTATATTATTCCCTTCGTTGCGACAACGTTGATGTTTA...	Linear
espX1	non-LEE-encoded type III secreted effector [E. coli O157:H7 TW14359]	1419	ATGACAGATGGTATCTCAACTTCGCCACATTGTCTTTATAAATCA...	Linear
-	hypothetical protein [E. coli O157:H7 TW14359]	135	ATGCATGAGCCACAAAATAATATAAAAAATCCTGCCATTAAGTTG...	Linear
-	predicted protein [E. coli O157:H7 TW14359]	216	ATGACTCGTTTTGAAGCAATTAAACAAGGCCATATTAAAATTGTG...	Linear
ccdA	putative antitoxin of gyrase inhibiting toxin-antitoxin system [E. coli O157:H7 ...	231	ATGACTGCAAAACGTACCACACAAAGTGTGACCGTCACCGTCGA...	Linear

Figure 5. Gene annotation of *E. coli* pangenome.

a) Sequence identifiers in pangenome FASTA file were replaced with b) gene names by Perl program and associated annotation table.

1.4.11 Second-site mutation analysis

SS53, SS13, and SS35 *rpoS* null mutants were aligned with an annotated reference genome (*E. coli* str. K-12 substr. MG1655; accession # NC_000913) using CLC® Genomics Workbench. Annotated consensus sequences were extracted from these alignments for use as reference genomes during the second-site mutation analysis.

White colony isolates with sufficient mapping coverage were aligned to their repective parent genomes using CLC® Genomics Workbench to identify mutations at

alternative gene loci which confer salt resistance. These mutations were filtered according to the following criteria:

1. A frequency of greater than 75% among sequence reads.
2. Minimal nucelotide (nt) conflict at the corresponding site in the reference genome (i.e. less than 1 nt conflict in 100). In many cases, a high degree of sequence variation within each reference would manifest as false-positive mutations in isolates (e.g. if a given site consists of 40% G and 60% T in the reference sequence, a consensus base of T is selected and a T>G false-positive mutation is identified in downstream analyses).

The Perl program written to annotate mutations with nucleotide conflicts at the corresponding position in the reference genome (see criterion 2 above) has been appended (File S6).

1.5. RESULTS

1.5.1 Analysis of RpoS-dependent growth phenotypes

rpoS^{+} and *rpoS*$^{-}$ *E. coli* strains were previously tested for growth in nearly 2,000 nutrient conditions by Biolog© Phenotype MicroArray. Increased growth of *rpoS*$^{+}$ *E. coli* relative to *rpoS*$^{-}$ was observed under osmotic stress in media supplemented with 5-6.5% NaCl (Figure 6). Based on this comparison, the effect of RpoS on adaptation to high salt was further examined.

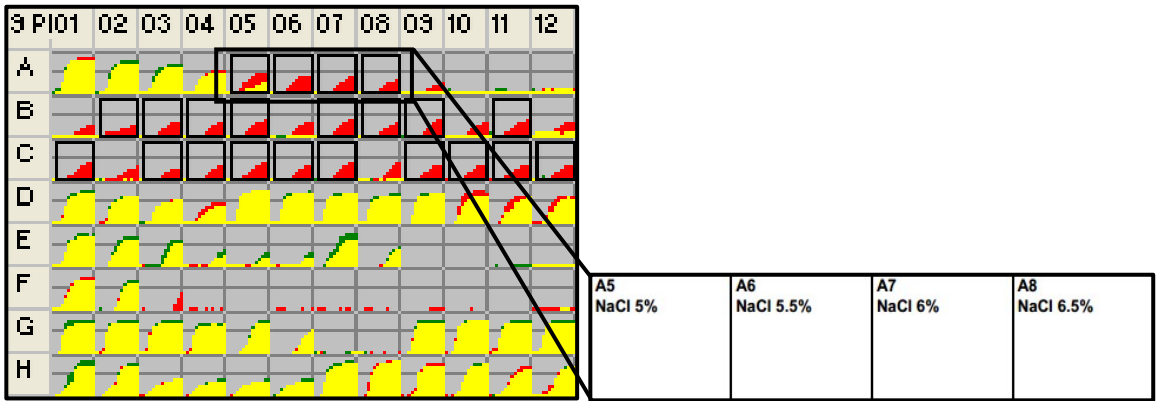


Figure 6. Growth of *rpoS*$^{+}$ and *rpoS*$^{-}$ *E. coli* in an array of nutrient conditions.

E. coli MG1655 (*rpoS*$^{+}$) and HW2211 (*rpoS*$^{-}$) were analyzed by Biolog Phenotype MicroArray^{TM}. Panels show overlaid growth curves of MG1655 (red) and HS2211 (green) in 96 environmental conditions. Areas of overlap are coloured yellow. Significant differential growth of *E. coli* strains in 5-6.5% NaCl is indicated. Images were obtained from microarray report.

1.5.2 Determination of plate reader linearity.

The linearity of the Thermo Scientific Multiskan® Spectrum was assayed to determine an effective range for monitoring growth of *E. coli* in liquid culture. *E. coli*

MG1655 was diluted from an initial OD_{600} of 3.69 and assayed for absorbance at 600 nm. Linearity between absorbance and concentration was observed from an OD_{600} of 0.074 to 2.21 (Figure 7).

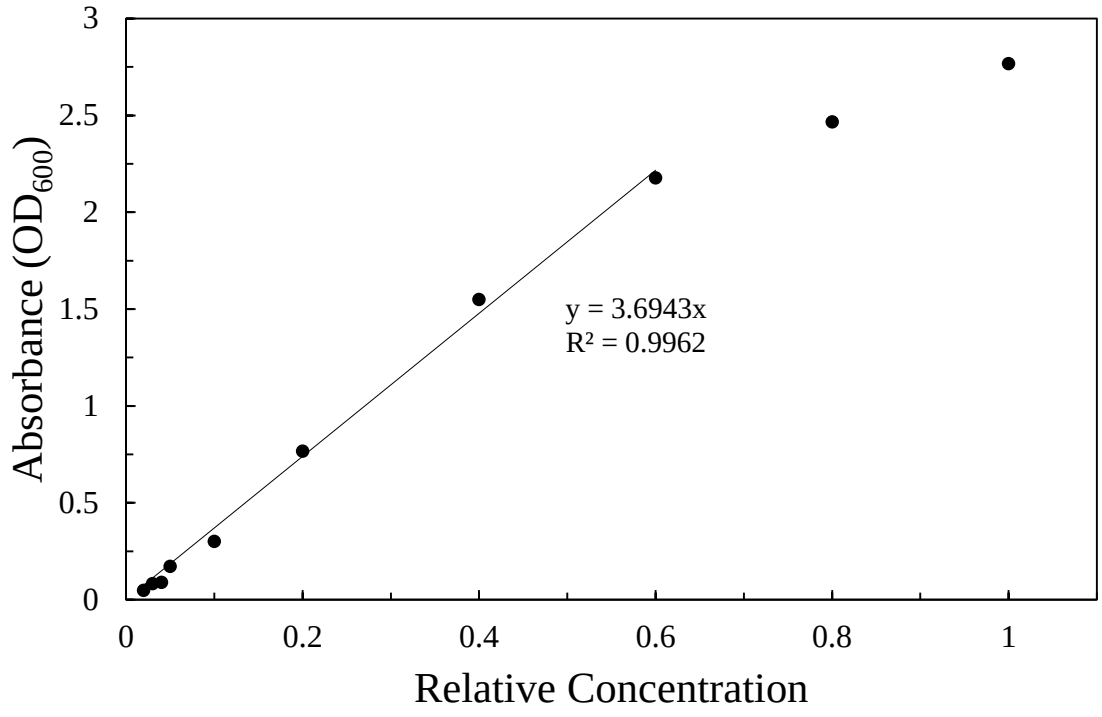


Figure 7. Assay of plate reader linearity.

E. coli MG1655 was diluted from an initial OD_{600} of 3.69 and assayed for absorbance at 600 nm using a Thermo Scientific Multiskan® Spectrum.

1.5.3 Effect of RpoS on growth in high salt

rpoS$^+$ and *rpoS*$^-$ *E. coli* strains were assayed for growth in 0-8% NaCl (Figure 8). Generation times for *rpoS* genotypes were similar in 0% NaCl (68 minutes for *rpoS*$^+$ and 72 minutes for *rpoS*$^-$ strains). A significant difference in generation time was observed

between genotypes in 2, 4, and 6% NaCl ($p < 0.05$, unpaired samples t-test). The greatest difference occurred in 6% NaCl, where the generation time of *rpoS*$^-$ *E. coli* (290 minutes) was 37% greater than *rpoS*$^+$ (212 minutes). Growth of both genotypes was comparable in 8% NaCl (315 minutes for *rpoS*$^+$ and 337 minutes for *rpoS*$^-$ strains) (Figure 9).

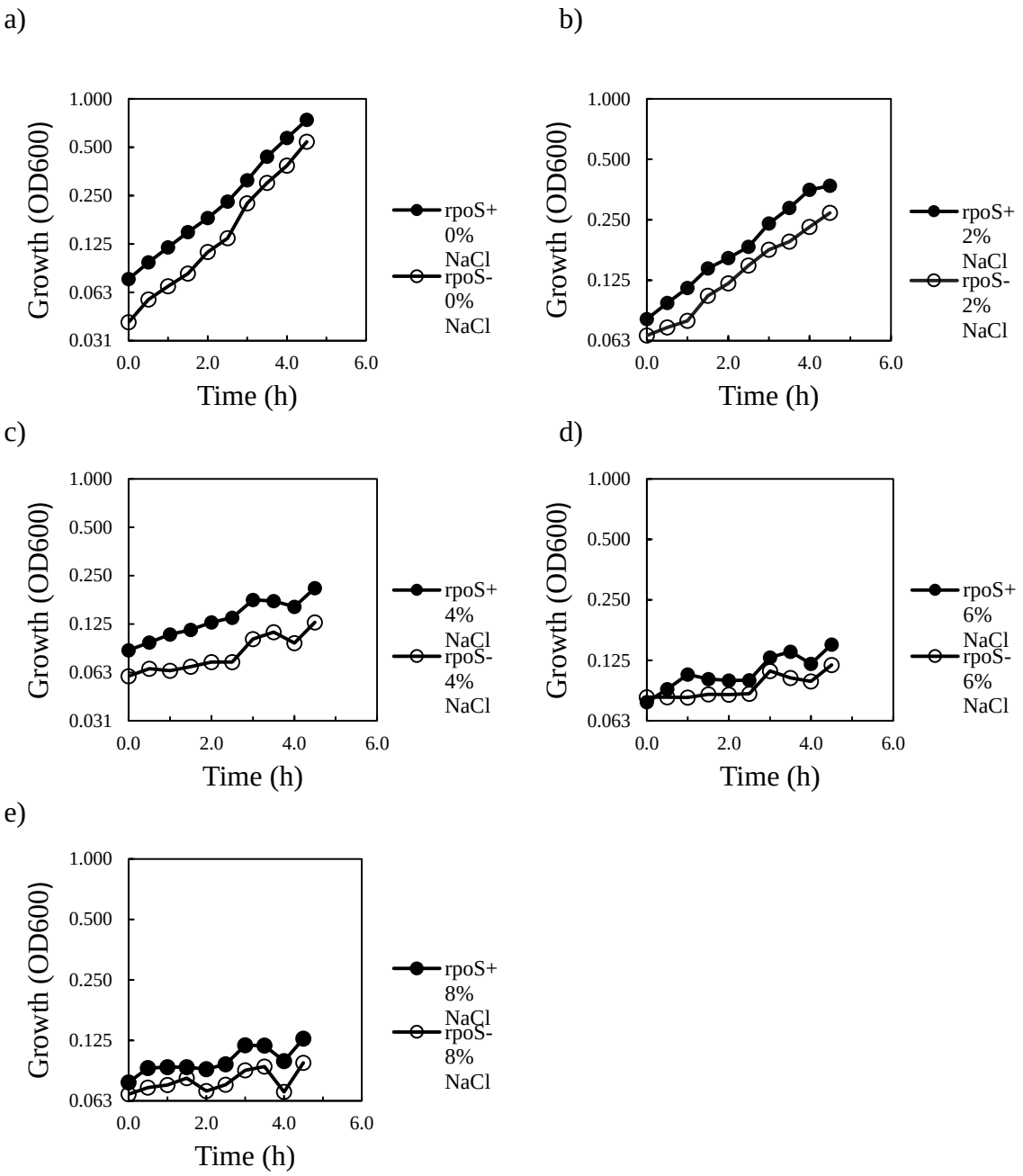


Figure 8. Effect of RpoS on growth in high salt.

E. coli MG1655 (*rpoS*$^{+}$) and HS2210 (*rpoS*$^{-}$) were grown to exponential phase and assayed for growth in M9 media with NaCl by absorbance. Growth curves show absorbance at 600 nm with time in a) 0%, b) 2%, c) 4%, d) 6%, and e) 8% NaCl. *E. coli* cultures were grown and assayed in triplicate.

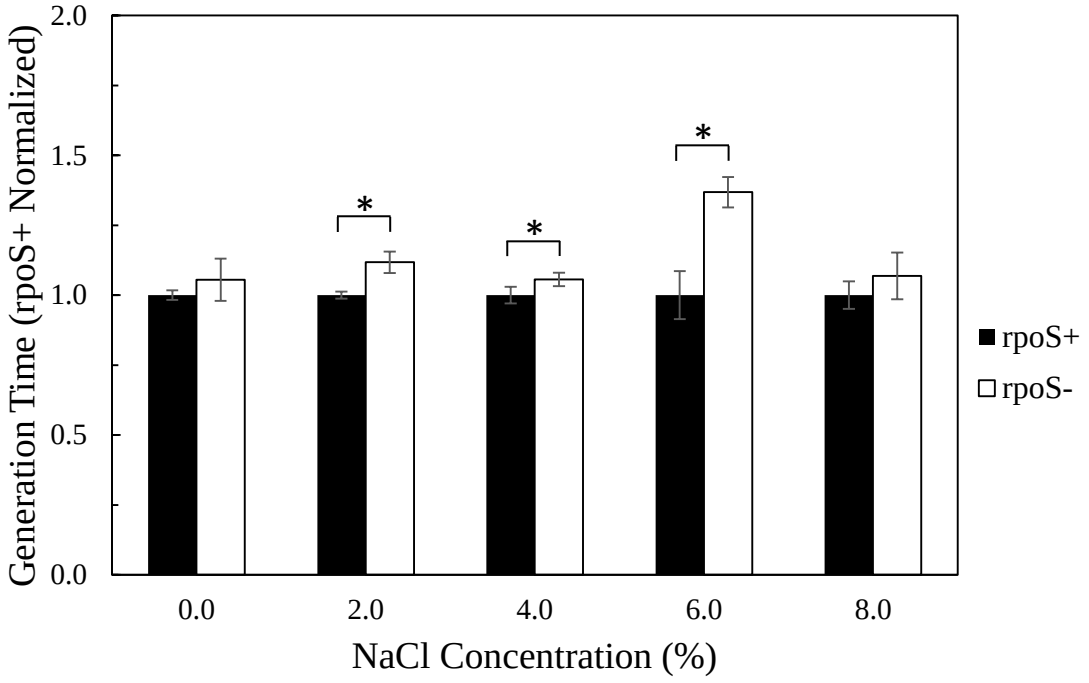


Figure 9. Effect of RpoS on generation times in high salt.

E. coli MG1655 (*rpoS*$^+$) and HS2210 (*rpoS*$^-$) were grown in triplicate to exponential phase and assayed for growth in M9 media with NaCl by absorbance. Generation times were calculated by averaging the slope of the linear range of $\log_2$-transformed absorbance values with time from each replicate. Values are normalized with MG1655 (*rpoS*$^+$). Differences in generation time between *rpoS*$^+$ and *rpoS*$^-$ *E. coli* are significant in 2, 4, and 6% NaCl ($p < 0.05$, unpaired samples t-test).

1.5.4 Viability of RpoS mutants in high salt

rpoS$^-$ *E. coli* strains were plated on 6% NaCl to select for gain of function mutations. Colonies were observed at a frequency of 6.5×10^{-7}- 4.1×10^{-6} per cell. The blue colony phenotype occurred at a rate of $5.9 \times 10^{-8} – 1.5 \times 10^{-7}$ per cell (3-9% of total colonies) while the white colony phenotype was observed at a rate of $5.9 \times 10^{-7} – 4.0 \times 10^{-6}$ per cell (91-97% of total colonies) (Table 3). RpoS reversion was confirmed among blue colonies by

sequence analysis, while RpoS null mutations were maintained among white colonies (Figure 10).

Table 3. Growth of RpoS mutants in high salt.

rpoS^{-} *E. coli* derivatives were plated on LB X-gal added with 6% NaCl. Colonies displaying either a blue or white phenotype were counted to determine frequency. *E. coli* cultures were grown and assayed in triplicate.

Strain	Blue Phenotype		White Phenotype		Total
	Frequency	% Colonies	Frequency	% Colonies	Frequency
SS13	5.9E-08 +/- 4.3E-08	9.1	5.9E-07 +/- 3.0E-07	90.9	6.5E-07 +/- 3.5E-07
SS35	1.5E-07 +/- 1.6E-08	4.5	3.3E-06 +/- 6.5E-07	95.5	3.4E-06 +/- 6.4E-07
SS53	1.3E-07 +/- 1.2E-07	3.3	4.0E-06 +/- 1.6E-06	96.7	4.1E-06 +/- 1.7E-06

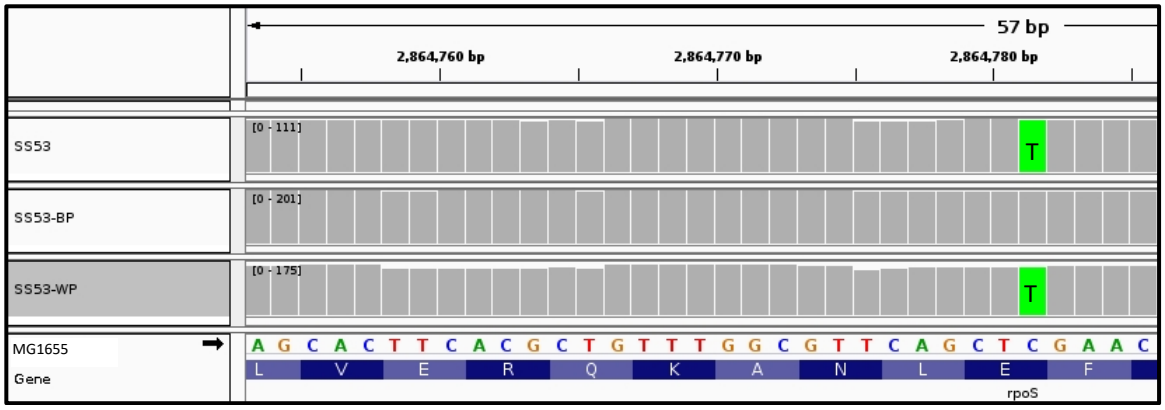


Figure 10. Reversion of RpoS in high salt.

E. coli SS53 (*rpoS*-) was plated on LB X-gal with 6% NaCl. Colonies displaying either a blue or white phenotype were selected for Illumina© sequencing. The *rpoS* E265* null mutation (green) was lost in the blue phenotype colony (SS53-BP) and maintained in the white phenotype colony (SS53-WP). This image was obtained from Broad Institute© Integrative Genomics Viewer.

1.5.5 Identification of second-site salt-resistant mutations

White colony isolates from *rpoS*$^{-}$ *E. coli* strains (20 from SS53, 5 from SS13, and 5 from SS35) were analyzed for second-site mutations at alternative gene loci which may confer salt resistance. 27 of 30 isolates had sufficient mapping coverage for mutation analysis. A total of 16 mutations were identified among 15 of the 27 remaining isolates with an average coverage (i.e. read depth) of 50.63, frequency of 97.50%, and phred-scaled quality score of 35.77 (Table 4). Nucleotide conflicts at the corresponding position in the reference genome, if present, did not exceed a frequency of 0.87%.

15 mutations were identified within the coding regions of 11 genes involved in adaptation to osmotic stress (e.g. cell wall synthesis and biofilm formation). A single intergenic mutation occurred within an operator upstream of *metB*, a gene involved in methionine biosynthesis and upregulated in hyperosmotic environments (Kocharunchitt et al., 2014).

Mutations within *wcaL* (biofilm formation; (Danese et al., 2000)) were identified in multiple isolates while those within *mraY* (cell wall synthesis; (Bouhss et al., 2004)) were found in isolates derived from different *rpoS* null mutants (SS13 and SS53). 3 of the 16 mutations resulted in the generation of a stop codon within the corresponding genes *ihfA* (chromosome structure; (Freundlich et al., 1992)), *aroK* (chorismate biosynthesis; (Lobner-Olesen et al., 1992)), and *rfaH* (biofilm formation; (Beloin et al., 2006)). Additionally, 3 of the 16 mutations resulted in a frameshift within the genes *wcaL* (biofilm formation; (Danese et al., 2000)), *ihfB* (chromosome structure; (Freundlich et al., 1992)), and *lpxM* (biofilm formation; (Theilacker et al., 2009)).

RpoS null mutations were maintained in all white colony isolates with sufficient mapping coverage for analysis. Loss of catalase activity was confirmed among all white colony isolates indicating loss of RpoS function.

Table 4. Second-site candidate mutations for salt resistance.

White colony isolates from *rpoS*$^{-}$ *E. coli* strains were analyzed for second-site mutations at alternative gene loci which confer salt resistance in the absence of RpoS function.

Strain	*rpoS*/RpoS mutation	Second-site mutation	Type	Gene affected	Region	Amino acid change	Relevant function	Source
SS13	L234*	97080G>A	SNV	*mraY*	CDS	R360H	Translocase; cell wall peptidoglycan synthesis; maintains defined cell shape	(Nanninga, 1998) (Bouhss et al., 2004)
SS35	95_96insTTAGTAGA	2116349C>T	SNV	*wcaL*	CDS	G260S	Glycosyltransferase; colanic acid biosynthesis; upregulated during biofilm formation	(Wang et al., 2012) (Danese et al., 2000)
SS35	95_96insTTAGTAGA	2116887delC	Deletion	*wcaL*	CDS	L80fs	Glycosyltransferase; colanic acid biosynthesis; upregulated during biofilm formation	(Wang et al., 2012) (Danese et al., 2000)
SS35	95_96insTTAGTAGA	2192114T>C	SNV	*yehD*	CDS	I36V	Uncharacterized fimbrial-like protein; could contribute to surface adhesion in environmental niches	(Korea et al., 2010)
SS35	95_96insTTAGTAGA	2318399A>T	SNV	*rcsC*	CDS	L492Q	Sensor histidine kinase; regulation of colanic acid capsule synthesis genes during osmotic stress	(Sledjeski et al., 1996)
SS35	95_96insTTAGTAGA	4128590C>T	SNV	*metB*	Operator	-	Cystathionine gamma-synthase; methionine biosynthesis; upregulated during osmotic stress	(Holbrook et al., 1990) (Kocharunchitt et al., 2014)
SS53	E265*	96069C>A	SNV	*mraY*	CDS	T23K	Translocase; cell wall peptidoglycan synthesis; maintains defined cell shape	(Nanninga, 1998) (Bouhss et al., 2004)

Strain	*rpoS*/RpoS mutation	Second-site mutation	Type	Gene affected	Region	Amino acid change	Relevant function	Source
SS53	E265*	964075delG	Deletion	*ihfB*	CDS	G83fs	Integration host factor subunit; involved in chromosome structure	(Freundlich et al., 1992)
SS53	E265*	1626593T>G	SNV	*dcp*	CDS	Q263P	Peptidyl-dipeptidase; hydrolyzes unblocked C-terminal dipeptides from oligopeptides	(Yaron, 1976)
SS53	E265*	1795293A>T	SNV	*ihfA*	CDS	L87*	Integration host factor subunit; involved in chromosome structure	(Freundlich et al., 1992)
SS53	E265*	1939582delT	Deletion	*lpxM*	CDS	G204fs	Myristoyltransferase; glycolipid biosynthesis; involved in biofilm formation	(Clementz et al., 1997) (Theilacker et al., 2009)
SS53	E265*	3518627C>A	SNV	*aroK*	CDS	E146*	Shikimate kinase; chorismate biosynthesis	(Lobner-Olesen et al., 1992)
SS53	E265*	3743911C>A	SNV	*yiaL*	CDS	A57E	Uncharacterized; belongs to TabA/YhcH/YiaL family which influences biofilm formation and fimbriae	(Kim et al., 2009)
SS53	E265*	4024408C>T	SNV	*rfaH*	CDS	R138H	Transcription antitermination protein; represses biofilm formation	(Beloin et al., 2006)
SS53	E265*	4024801A>T	SNV	*rfaH*	CDS	L7Q	Transcription antitermination protein; represses biofilm formation	(Beloin et al., 2006)
SS53	E265*	4024817G>A	SNV	*rfaH*	CDS	Q2*	Transcription antitermination protein; represses biofilm formation	(Beloin et al., 2006)

SNV = single nucleotide variant; CDS = coding sequence; IGR = intergenic region; fs = frameshift

1.6. DISCUSSION

1.6.1 Summary and rationale

This study investigated selection for the restoration of RpoS function in *E. coli* under natural conditions (i.e. high salt). *rpoS*$^{+}$ and *rpoS*$^{-}$ *E.coli* strains were first assayed for growth in 6% NaCl to confirm the role of RpoS in adaptation to osmotic stress. The growth assay identified a 37% increase in generation time for the *rpoS*$^{-}$ strain relative to *rpoS*$^{+}$ in 6% NaCl and a selective pressure for RpoS function in high osmolarity environments. This difference is compounded during exponential growth on solid media where single cells may undergo more than 20 cell division cycles and allows for isolation of larger salt resistant colonies. *rpoS*$^{-}$ *E.coli* strains were then plated on 6% NaCl to select for gain of function mutations. Colonies were observed at a frequency of 10^{-7}-10^{-6} per cell. 3-9% displayed a blue phenotype indicative of *rpoS* reversion and 91-97% displayed a white phenotype indicative of non-functional RpoS and suggestive of second-site mutations at alternative gene loci which allow for growth in high salt. Sequence analysis confirmed *rpoS* reversion among blue colonies and provides evidence for an RpoS mutational switch among *E. coli* at the population level. This model predicts the loss of RpoS function in poor carbon environments and subsequent restoration of function during environmental stress (Ferenci, 2003). It is therefore an important consideration for the interaction of *E. coli* with naturally-occurring *rpoS*$^{-}$ populations in the environment (Chiang et al., 2011).

1.6.2 RpoS reversion frequency

While only a limited fraction of salt resistant mutants contained RpoS activity, the observed RpoS reversion frequency (i.e. 10^{-8} – 10^{-7} per cell) is orders of magnitude greater than expected by spontaneous reversion of the RpoS null mutations used in this study (i.e. 10^{-10} per cell, based on a spontaneous mutation rate of 10^{-3} per genome per generation in wild-type *E. coli* (Lee et al., 2012) and a 10^{-7} probability of a given mutation occurring in a single stop codon within a 4.6 Mbp genome). This reversion may occur at the site of the substitution or insertion mutation to restore the original sequence of the parent *E. coli* strain (true reversion) or at adjacent sites in *rpoS* which correct for the nonsense or frameshift mutation (second-site reversion).

Reversion frequency of the RpoS null strain SS35 containing a frameshift mutation (i.e. 10^{-7} per cell) was comparable with strains SS13 and SS53 containing nonsense mutations in RpoS (i.e. 10^{-8} – 10^{-7} per cell). Insertion/deletion mutations which may correct this frameshift have been observed at about $1/10^{th}$ the rate of base pair substitutions which may rescue nonsense mutations in wild-type *E. coli* (Lee et al., 2012), suggesting that RpoS reversion in SS35 should be an order of magnitude less frequent than in SS13 and SS53. This discrepancy may be due to the larger number of nucleotide sites available for restoration of the reading frame through additional insertion/deletion mutations upstream of the 8 base pair insertion in SS35. It should be noted that additional downstream mutations would not correct for the two stop codons present within the 8 base pair insertion.

1.6.3 Second-site mutation & resistance

A large fraction of colonies observed during salt viability assays displayed an *osmY*$^{-}$ phenotype indicative of non-functional RpoS. Confirmation of *rpoS* null mutations among white colony isolates indicated that alternative mechanisms may contribute to osmotic stress resistance apart from RpoS. These mechanisms may occur alongside RpoS regulation, where mutations in alternative regulators influence the expression of proteins within the RpoS regulon. The stress proteins UspA and UspB belong to both the cAMP-CRP and RpoS regulons and show decreased expression in Δ*cya* relative to Δ*rpoS E. coli* strains, suggesting that cAMP-CRP may influence stress resistance in conjunction with RpoS (Dong et al., 2008; Franchini et al., 2015). Other mutations may occur within RpoS regulon proteins to affect expression, stability, or degradation and contribute to stress resistance in the absence of RpoS. *E. coli* isolates displaying a white phenotype were therefore sequenced to identify prospective mutations which may confer salt resistance through these methods.

Sequence analysis of white colony isolates identified 15 suppression mutations within the coding regions of 11 genes involved in adaptation to osmotic stress- the majority of which were involved in bacterial cell wall synthesis or biofilm formation. The gene *mraY*, for instance, codes for a translocase involved in cell wall peptidoglycan synthesis and maintaining a defined cell shape (Nanninga, 1998; Bouhss et al., 2004). *mraY* mutations my affect cell structure in hyperosmotic conditions and were observed in isolates derived from distinct *rpoS* null mutants (SS13 and SS53). The genes *wcaL*, *lpxM*, and *rfaH* respectively code for a glycosyltransferase involved in colonic acid biosynthesis (Wang et

al., 2002), a myristoyltransferase involved in glycolipid biosynthesis (Clementz et al., 1997), and a transcription antitermination protein (Beloin et al., 2006). Mutations within *wcaL* and *lpxM*, genes upregulated during biofilm formation, may influence survival during environmental stress (Danese et al., 2000; Theilacker et al., 2009). 1 of 3 identified mutations in *rfaH*, shown to repress biofilm formation in *E. coli*, contained a null mutation which generated a stop codon at the second amino acid position (Beloin et al., 2006). A single intergenic mutation occurred within an operator upstream of *metB*, a gene involved in methionine biosynthesis and upregulated in hyperosmotic environments (Kocharunchitt et al., 2014). Methionine is essential for the initiation and elongation of proteins and synthesis of polyamines, purines, and pyridamines, which suggests that increased methionine levels may be important for growth of *E. coli* in high osmolarity environments (Ron et al., 1990; Gur et al., 2002). Multiple mutations were observed among separate isolates within the genes *mraY* (cell wall synthesis; (Nanninga, 1998)), *rfaH* (biofilm formation; (Beloin et al., 2006)), and *wcaL* (biofilm formation; (Danese et al., 2000)), providing strong evidence that these genes contribute to salt resistance in *E. coli* and that this analysis was exhaustive in its identification of second-site candidates. These candidate mutations may serve as targets of future research concerning the adaptation of *E. coli* under osmotic stress in the absence of RpoS.

1.6.4 Limitations of *E. coli* MG1655 as a reference sequence

E.coli GC4468-derived *rpoS* null mutants were aligned with an annotated *E. coli* MG1655 reference genome to produce consensus sequences for each parent strain in

second-site mutation analyses. Genes present in these null mutants and absent in *E. coli* MG1655 were not considered in the identification of second-site mutations and may confer salt resistance in the absence of RpoS activity. The exclusion of these genes may explain the fraction of white colony isolates which did not contain second-site mutations.following sequence analysis.

1.6.5 Limitations of *osmY* as a reporter of RpoS activity

The RpoS-dependent *osmY* gene used as a reporter of RpoS activity in salt resistant mutants encodes a periplasmic protein and is upregulated in hyperosmotic environments (Yim et al., 1992). Salt resistance conferred by *osmY* introduces a bias in the accurate reporting of RpoS restoration among revertants. High salt conditions may select for mutations which upregulate *osmY* in the absence of RpoS and result in false-positive mutants which display a blue phenotype without RpoS activity. Future studies in RpoS reversion should use RpoS regulon reporter genes which do not confer a selective advantage for resistance to the RpoS-inducing stress.

1.6.6 Implications of the RpoS mutational switch & future directions

An RpoS mutational switch influences the adaptation of natural *E. coli* populations in high stress conditions along with *E. coli* persistence and pathogenicity through the loss and gain of RpoS function under antagonistic environments. Selection for the loss of RpoS function has been observed in both poor nutrient conditions (Notley-McRobb et al., 2002; Chen et al., 2004) and during extended incubation (King et al., 2004). Both of these

conditions are likely encountered in the natural environment, where *E. coli* has remained culturable for over 68 days without significant changes in cell count (Davies et al., 1995). Osmotic stress selecting for the restoration of RpoS function may similarly occur in nature (e.g. during the dessication of urine (Putnam, 1971)). These contrasting environments establish a mutational equilibrium among natural *E. coli* populations, a proportion of which contains null mutations in the *rpoS* gene (Chiang et al., 2011). RpoS function is essential for the entry of *E. coli* into the viable but non-culturable (VBNC) state following exposure to particular environmental stresses (Boaretti et al., 2003). O-island virulence factors within enterohaemorrhagic *E. coli* are similarly regulated by RpoS and contribute to microbial pathogenesis (e.g. the LEE island elements *ler*, *cesF*, and *Z5139*) (Dong et al., 2009). An RpoS mutational equilibrium influences both *E. coli* persistence and virulence through antagonistic pleiotrophic regulation (e.g. the selection for loss of RpoS function in limiting nutrients contrasts the positive selection for RpoS-mediated persistence or virulence when exposed to additional environmental stress).

The RpoS sigma factor regulates gene expression following UV radiation (Sammartano et al., 1986), acid exposure (Small et al., 1994), heat shock (Hengge-Aronis et al., 1991), and oxidative in addition to osmotic stress (Sammartano et al., 1986). Future work may investigate positive selection for the gain of RpoS function and establishment of an RpoS mutational switch in the presence of these alternative stressors. Significant findings would explain the prevalence of *rpoS*$^{-}$ *E. coli* in the natural environment.

CHAPTER 2:

Data analysis & microbial profiling of the Niagara Region

2.1 ABSTRACT

The monitoring of waterborne pathogens in recreational beaches is of considerable importance for public health and safety. Labor-intensive plate count methods are currently employed by water surveillance authorities to quantify fecal indicator bacteria (FIB) and estimate the degree of fecal contamination of a given water area. While this technique is effective in identifying coliform bacteria, plate counts are delayed by colony growth and many taxa which harbor more resilient pathogenic microorganisms remain undetected. In light of these limitations, the application of next-generation sequencing (NGS) technology can potentially augment traditional monitoring methods and may provide added value to the sampling regimens implemented by conservation authorities and municipal officials. To investigate the use of NGS in identifying microbial taxa within the Niagara Region of Ontario, Canada, we collected water samples from Lake Erie, Lake Ontario, and nearby areas and examined the metagenome of microbial communities with high-throughput 16S rRNA gene sequencing. A QIIME (Quantitative Insights Into Microbial Ecology) analysis of sequence data identified significant differences in relative microbial abundance with respect to sample metadata (e.g. location and subtype), significant correlations between relative abundance and quantitative parameters (e.g. *Escherichia coli* counts and fecal DNA markers), and detected pathogen-containing taxa at a relative abundance of 0.1-1.5%. Further analysis identified indicator species abundant in and characteristic of particular sample groups (e.g. location or subtype). These results show that sequence-based analyses can be used in conjunction

with traditional identification methods to profile the metagenomic community of environmental samples and predict water quality.

2.2. INTRODUCTION

This literature review will introduce the emerging field of metagenomics and discuss the sampling initiatives for water quality monitoring in the Niagara Region. It will focus on methodologies for DNA sequencing and analysis of microbial communities in the natural environment. Finally, it will review recent metagenomics studies on microbial diversity in the Great Lakes.

2.2.1 Metagenomics overview

Microbial communities exhibit broad diversity in a host of natural environments. Early advances in the elucidation of this diversity were made in 1970s by Carl Woese, who suggested that sequence variation in ribosomal RNA (rRNA) genes provided insight into the evolutionary history of microorganisms (Pace et al., 2012). Woese's discovery of the kingdom Archaea and contribution to the universal "tree of life" laid the foundation for future developments in microbial diversity.

Methodologies for characterizing microbial communities have evolved from simple plate counting, which allowed only identification of culturable microbes, to structural (e.g. phospholipid-derived fatty acid analyses), biochemical (e.g. sole-carbon-source utilization assays), and early molecular-based techniques (e.g. DNA microarray hybridization) (Theron et al., 2000). Subsequent development of sequenced-based analyses resulted in the emergence and popularization of metagenomics – the collection, purification, and study of mixed genetic material from communities of organisms (Touchon et al., 2009). A general

shift from more traditional methods of characterization to metagenomic approaches has allowed for the robust identification of both culturable and non-culturable diversity.

2.2.2 Niagara Region sampling

2.2.2.1 Projects

The Niagara sample collection spans 6 projects in coordination with several water conservation authorities and municipal officials. The details of each project are included in Table 5.

Table 5. Niagara Region project summary.

	Project	Partner	Agency	Municipality	Number of Samples
1	SO/CSO rain event and selected watershed monitoring	Mark Green	St. Catharine's Environmental Services	St. Catharine's	~200
2	Diurnal microbial monitoring	Glen Hudgin	Niagara Region Public Health	Niagara Region	~200
3	Watershed survey	Joshua Diamond	Niagara Peninsula Conservation Authority	Niagara Region	~220
4	Queens Royal/Niagara River Bacteroidales monitoring	Tom Edge	Environment Canada	Niagara Region	~200
5	DNA purification technology testing	Wonsik Kim	Norgen Biotek	Niagara Region	~50
6	Source tracking and well monitoring	Trevor Imhoff	Wainfleet	Wainfleet	TBD

2.2.2.2 Procedures

Within the Niagara region of Southern Ontario, lake and pore water samples are typically collected from 5 sites along a transect spanning the length of the sampled beach. Lake samples are obtained at a depth of approximately 1 meter and 15-30 cm beneath the water surface, while pore samples are collected by displacing approximately 0.3 meters of beach sediment and sampling the resultant pool of water (Pike, 2008). Source water samples may additionally be obtained from sites designated sources of pollution by the Niagara Conservation Authority. Creek water, stormwater outfall, and combined sewer overflow are collected as general grab samples.

2.2.2.3 E. coli & water quality monitoring

The pervasiveness of *Escherichia coli* within animal fecal matter (Tenaillon et al., 2010) and pathogenicity observed among select serotypes (Dong et al., 2009) has led to the bacterium's widespread use as an indicator of water quality in the natural environment. Early investigations of bacterial prevalence in U.S. beaches identified correlations between *E. coli* densities and gastrointestinal illness in swimmers (Dufour, 1984). On the basis of this data, the U.S. Environmental Protection Agency (EPA) established a 30-day geometric mean and maximum count of 126 and 235 *E. coli* colony forming units (CFU) per 100mL of freshwater as respective standards for water quality (USEPA, 2005). Similar guidelines were later published by the Public Health Agency of Canada, specifying a maximum acceptable count of 200 *E. coli* CFU/100ml (Canada). The adopted provincial guideline for

the recreational use of beaches within Ontario is currently 100 CFU/100ml- a geometric mean of samples taken from several sites along the beach of interest (PHO, 2013).

2.2.3 DNA sequencing

2.2.3.1 Whole-genome vs. targeted sequencing

Both whole-genome and targeted sequencing methodologies may be used to elucidate microbial diversity through metagenomic analysis. A whole-genome shotgun approach involves the random sequencing of DNA fragments from a mixed sample and is essential for the reconstruction of individual genomes from microbial communities. Functional metagenomic studies which associate gene or protein complements with particular functions in the environment are also dependent on the indiscriminate sequencing of genetic material. While whole-genome sequencing provides detailed genomic information from mixed populations, it is generally associated with significant financial cost and intensive computational analysis. More targeted sequencing approaches circumvent these shortcomings and have gained widespread use in identifying taxonomic distributions among microbial communities. The 16S ribosomal RNA gene, which codes for an RNA component of the prokaryotic ribosome and contains species-specific sequence regions, has become a standard for the classification of microbial species (Kolbert et al., 1999). 16S rDNA sequencing requires universal primers capable of annealing to conserved 16S regions and amplifying hypervariable sites associated with speciation. Though targeted sequencing approaches provide a much more comprehensive profile of microbial diversity,

consideration must be given to primer specificity and the effects of sequencing biases on species inclusion.

2.2.3.2 Sequencing depth

There exist inherent trade-offs between the depth of sequencing (i.e. sequence coverage at a particular nucleotide position) and quantity of samples included in a metagenomics design. Deep sequencing of few samples allows for the identification of rare taxa or genes within microbial communities, and is essential for reconstructing individual genomes from mixed populations. While offering distinct advantages for a more focused characterization of microbial diversity, this approach is limited in its generation of statistically significant comparisons between samples and determination of biotic or abiotic community-structuring factors. In contrast, a design incorporating shallow sequencing of many samples more clearly identifies spatial and temporal community dynamics (i.e. changes in microbial populations across space or with time). A shallow sequencing approach may also be used to provide direction for deeper sequencing analyses without significant financial or computational requirements (Knight et al., 2012).

2.2.4 Diversity analysis

2.2.4.1 Pipeline design

The establishment of a robust metagenomics pipeline is essential for the effective and reproducible characterization of microbial communities. This procedure begins with the quality filtration and formatting of sequence data (i.e. removal of cloning vector or

adapter sequences) and is followed by the alignment sequences to a reference database of operational taxonomic units (OTUs). In a closed-reference OTU approach, sequences which do not align to the reference database are excluded from downstream processing. Conversely, the use of open-reference OTU picking allows for the clustering of sequences with no reference match and de novo generation of OTUs. An entirely de novo-based approach for OTU generation may be taken when a reference database is unavailable, though suffers from a lack of parallelization (i.e. the simultaneous execution of multiple calculations) and increase computational needs.

To generate meaningful comparisons between samples, OTU results must be rarified to a singular depth in OTU counts/sample (an OTU count refers to the number of non-unique OTUs called for a given sample). Since all samples with less than the chosen depth are excluded from downstream analyses, compromises are typically made between the number of samples and depth of diversity analyzed.

Several analysis platforms have been designed for the characterization of microbial diversity. QIIME or Quantitative Insights into Microbial Diversity refers to an open-source software package designed to analyze sequence data from microbial communities (Caporaso et al., 2010). The QIIME package accepts raw sequence output and performs a collection of metagenomic functions, including OTU picking, taxonomic assignment, construction of phylogenetic trees, and the visualization of publication-quality graphics. Other platforms, including mothur (Schloss et al., 2009), EstimateS (Colwell et al., 1994), and MG-RAST (Meyer et al., 2008) perform similar analyses of microbial communities.

2.2.4.2 Characterizing alpha & beta diversity

The analysis of both alpha diversity within and beta diversity between metagenomic samples provides insight into the role of spatial or temporal factors in shaping microbial communities.

Alpha rarefaction techniques identify species richness (i.e. a count of species within a community) by generating the number of unique OTUs with increasing sampling depth. Rarefaction curves typically show rapid initial growth as most common species are reported at lower sampling depths and plateau as species richness becomes saturated. Rarefaction analyses may be used to determine the minimum sampling depth required to accurately characterize community diversity and reduce both financial and computational costs.

Rank abundance curves are generated to depict both species richness, a count of the number of species, and evenness, a measure of variation in species abundance, within microbial communities. Species are ranked on abundance and graphed on abundance vs. species rank. Steep rank abundance curves denote low species evenness as high ranking species have much greater abundance than low ranking species. Conversely, shallow rank abundance curves denote high species evenness as most species are present in similar abundance.

Beta diversity analyses make statistical comparisons between samples to elucidate differences in microbial communities. Phylogenetic tests (P tests) are conducted by generating a phylogenetic tree from a collection of samples and determining the number of changes needed to produce the existing topology using an objective standard (e.g. parsimony). UniFrac, an alternative and widely used metric for comparing diversity in

microbial studies, generates distance values between samples based on the fraction of total branch lengths unique to any sample in a combined phylogenetic tree (Lozupone et al., 2005). A qualitative or unweighted UniFrac analysis assigns equal representation to all branch lengths and is ideal for elucidating factors that affect the presence or absence of taxa. In contrast, a quantitative or weighted analysis accounts for the relative abundance of taxa in samples and is used to identify factors which affect relative community composition. Statistical analyses are conducted by Monte Carlo method, which randomly permutes tip assignments and identifies how often simulations have more extreme values than the observed data.

Principal coordinates analyses (PCoA) utilize a matrix of distance values to generate a coordinate matrix which minimizes the strain loss function (Buja et al., 2008). The coordinate matrix clusters samples by similarity and can be visualized in multiple dimensions (i.e. three-dimensional PCoA plots). The percentage associated with each principal coordinate axis indicates axis contribution to variation.

Alpha and beta diversity analyses enable the comparison of microbial communities within and between an assortment of sample parameters (e.g. location, subtype, or rain events).

2.2.5 Microbial diversity in the Great Lakes

The application of metagenomics analyses to characterize microbial diversity in the Great Lakes and surrounding bodies of water is a relatively new area of interest. A previous study examined Cyanobacterial communities in Great Lakes microbial mats and identified

metabolic versatility and low species diversity in low oxygen environments (Voorhies et al., 2012). More recent investigations have focussed on the distributions of viral communities in ballast water (Kim et al., 2015) along with the spatiotemporal dynamics of virus occurrence in Lake Ontario and Lake Erie (Mohiuddin et al., 2015). This lack of current research suggests a need for robust characterization of microbial community dynamics in the Great Lakes, with particular focus on bacterial distributions.

2.3 PROJECT OUTLINE

2.3.1 Objectives

1. Development of an automated workflow for characterizing microbial diversity among metagenomic samples. 16S sequence data must be filtered for quality, annotated with appropriate metadata, and subjected to diversity analyses.
2. Analysis of microbial diversity within Niagara collection across qualitative (e.g. sample location and subtype) and quantitative (e.g. *E. coli* CFU counts, Bacteroidales DNA markers) parameters.
    a. Characterization of pathogen-containing genera (i.e. genera known to harbour water pathogens).
    b. Characterization of indicator OTUs for sample groups of interest (i.e. those unique to or abundant in particular sample groups relative to others).

2.3.2 Significance

This project investigates the use of next-generation sequencing technology to augment traditional water quality monitoring methods and add value to sampling regimens employed by conservation authorities. Incorporation of 16S sequence data into sampling protocols addresses current limitations in monitoring techniques, including the detection of more resilient non-coliform bacteria. The 16S workflow developed during this project may also be implemented in future microbial studies (e.g. mouse microbiome dynamics).

2.4 METHODS

2.4.1 Sample collection

Water samples were collected from Lake Erie, Lake Ontario, and nearby areas (i.e. creeks, SOs, and CSOs) within the Niagara Region by several sampling agencies (see Table 5). Beach water samples were obtained at chest or knee-depth while pore samples were collected along shorelines (sampling protocol depicted in Figure 11). Creek water, SO, and CSO were collected as general grab samples.

Figure 11. Beach sampling protocol.

Water samples were obtained at chest or knee-depth 15-30 cm beneath the water surface (left). Pore samples were collected by displacing 0.3 m of beach sediment and sampling the resultant water pool (right).

2.4.2 Sample processing & sequencing

Water samples were processed for DNA extraction using Norgan Biotek© Bacterial Genomic DNA Isolation Kits. 16S rDNA sequences were amplified with primers flanking

the V3-V4 region using a polymerase chain reaction (PCR) protocol (Klindworth et al., 2013). The amplification schematic for this reaction is illustrated in Figure 12. 16S DNA was prepared with a Nextera XT Library Preparation Kit and sequenced with Illumina© MiSeq paired-end technology producing 2x300 base pair (bp) reads at 100-fold coverage.

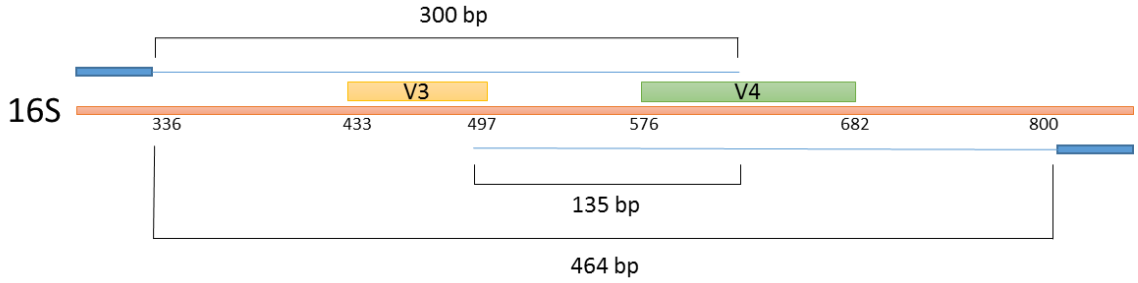


Figure 12. 16S rDNA amplification schematic.

Forward and reverse primers flank the V3-V4 hypervariable region. Illumina© MiSeq paired-end sequencing at 2x300 bp reads produces a 464 bp amplicon and 135 bp overlap.

2.4.3 Data organization

Sample and sequencing metadata were compiled into a single spreadsheet and supplemented with taxonomic distributions of the top 10 most abundant phyla and pathogen-containing genera using QIIME (Caporaso et al., 2010). Metadata categories are listed in Table 6.

Table 6. Niagara collection summary data.

Sample Metadata		Sequencing	QIIME Data	
Sample ID		Index	OTU Counts/Sample	
Sample Name		Yield (Mb)	Top 10 Phyla - Relative Abundance	Proteobacteria
Location		% PF		Bacteroidetes
Sublocation		# Reads		Actinobacteria
Type		% of raw clusters		Cyanobacteria
Subtype		% Perfect Index		Firmicutes
Temperature (°C)		% One Mismatch		Verrucomicrobia
Turbidity		% of >= Q30		Chloroflexi
pH		Mean Quality		Planctomycetes
Rain				Acidobacteria
Date				Nitrospirae
Time				Other Phyla
Sampling Agency			Pathogen-containing Genera - Relative Abundance	*Clostridium*
Sampling Coordinator				*Mycobacterium*
Comments				*Legionella*
DNA Stats	DNA (ng/uL)			*Vibrio*
	260/280			*Yersinia*
Qubit DNA Conc (ng/uL)				*Leptospira*
DC Counts (CFU/100mL)				*Salmonella*
Endpoint PCR	BAC32			*Campylobacter*
	HF183			*Escherichia*
	CF128			
	GULL2			
	DG37			
qPCR (CFU/mL)	GenBac			
	Hum			
	Cow			
	Gull			
	DG37			
NLET Analysis	CAFFEINE			
	CRBMZPN			
	CODEINE			
	COTININE			
	ACTMNPHN			
	ACSLFM			
	SUCRLOSE			

2.4.4 Diversity analysis

Illumina© FASTQ input was processed by Perl script and analyzed using QIIME (Caporaso et al., 2010). A workflow for the analysis is depicted in Figure 13. The shell script fastx-toolkit_greengenes_install.sh installs both FASTX-Toolkit for quality trimming and the Greengenes 16S reference database for taxonomic classification (DeSantis et al., 2006). The script qiime_workflow.pl accepts a) CPU thread number, b) reads to sample from each file, and metadata categories for c) core diversity and d) subset analyses as input.

FASTQ read files were first decompressed and trimmed for quality using FASTQ Quality Trimmer (quality threshold = 25). Forward and reverse read files were then converted to FASTA format, combined, annotated with QIIME identifiers, and submitted for closed-reference OTU picking (minimum query alignment = 50%; sequence similarity threshold = 97%) using the Greengenes 16S database (OTU identity = 97%). A BIOM (biological observation matrix) summary table was generated from the resulting OTU table and provides OTU count/sample statistics.

A sampling depth of 5230 OTU counts/sample was chosen during OTU table rarefaction for standardization of core, supplementary, and subset analyses (see Figure 6 for depth justification). Core diversity analyses (i.e. alpha and beta) were then conducted using specified metadata (i.e. sample location and subtype) and parameters found in the parameters.txt text file, followed by supplementary analyses (e.g. rank-abundance curves, heatmaps, and bipartite networks). Taxonomy tables containing relative abundance

distributions were generated from the kingdom to species level. Subset analyses for sample locations and subtypes were subsequently conducted.

A log file was generated during the execution of qiime_workflow.pl to catalogue each process. Instructions for diversity analysis are included within a readme text file. The readme file, shell and Perl scripts, and parameter file have been appended (File S7-S10).

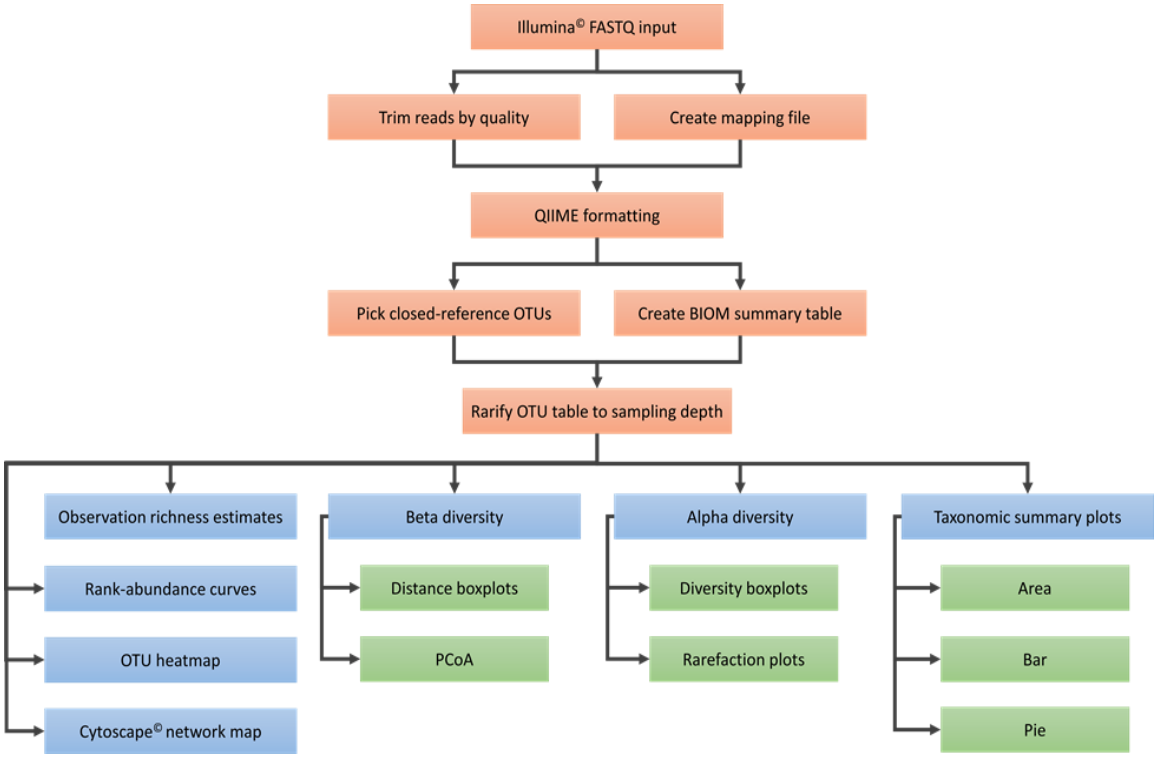


Figure 13. Diversity analysis workflow.

Water samples were collected from regions surrounding the Great Lakes. Metagenomic DNA was extracted, sequenced with Illumina© MiSeq technology, and characterized using QIIME.

2.4.5 Species richness analysis

Species richness curves were generated by subsampling OTU tables from 10-50,000 OTU counts/sample using QIIME and fitting a saturation growth-rate model to the data using CurveExpert©. Species richness was estimated using the limit of the curve as OTU counts/sample approach infinity.

2.4.6 Indicator OTU analysis

Indicator OTUs (i.e. those abundant in and representative of particular sample groups) were identified using the indicspecies R package (Dufrene et al., 1997; De Caceres et al., 2009). The package evaluates the strength (i.e. indicator value or IndVal) of relationships between species occurrence/abundance and sample groups by assessing the mean abundance of a species in a given sample group relative to all groups and the frequency of occurrence of that species among sites in the sample group. Statistical inferences for these relationships are determined using permutations tests.

2.5. RESULTS

2.5.1 Illumina sequencing and QIIME OTU fetching

16S DNA was sequenced with Illumina© MiSeq paired-end technology in 8 runs and analyzed using QIIME. Runs averaged 117,626-445,280 reads/sample with a Phred quality score of 15.63-29.80. OTUs were fetched at an average of 5,156-201,150 counts/sample. Samples comprising run 4 were resequenced in runs 5 and 7 due to low read quality and taxonomic information. Runs 4 and 5 were excluded from subsequent diversity analyses. Summary statistics for sequencing runs are included in Table 7. Samples from runs 1-3 and 6-8 are ranked by OTU count in Figure 14.

Table 7. Illumina run summary.

847 samples from the Niagara Region were sequenced with Illumina© MiSeq technology (2x300 bp reads) in 8 runs and analyzed using QIIME.

Run	Number of samples	Number of reads/sample		PhiX sequencing control (reads)	Phred quality score/sample		OTU counts (QIIME w/ FASTQ Quality Trimmer)/sample	
		Average	SD		Average	SD	Average	SD
1	96	445280	270912	5699070	29.07	2.56	169945	108102
2	96	440415	221353	8856936	25.42	4.17	142477	71159
3	96	414833	137953	9059452	28.7	2.15	201150	71286
4	192	244907	113657	6387810	19.89	2.33	42640	18088
5	192	251236	242935	6409364	15.63	0.58	5156	2369
6	192	208795	140489	5504976	27.53	0.47	110631	77550
7	192	117626	50153	10445866	28.65	0.31	71809	31960
8	192	191604	148889	11098154	29.80	0.33	106375	85192

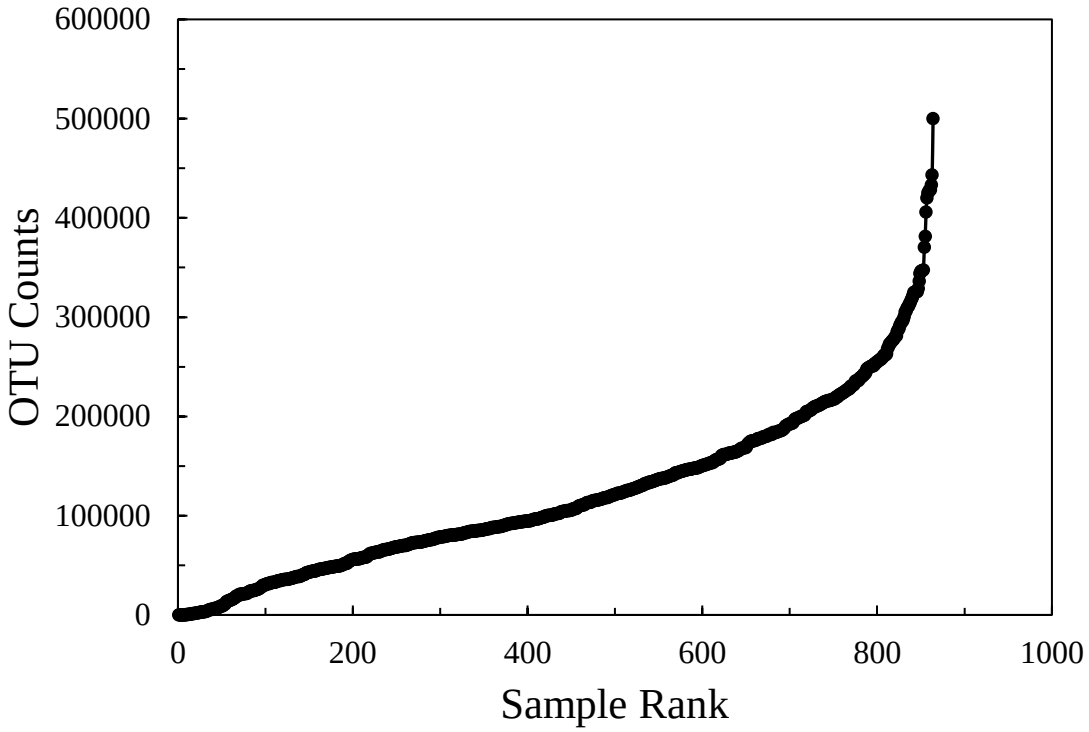


Figure 14. Rank-OTU count curve for Niagara collection.

847 samples from the Niagara Region were sequenced with Illumina© MiSeq technology (2x300 bp reads) and analyzed using QIIME.

2.5.2 Determination of species richness and minimum sampling depth

Figure 15 depicts the average number of species for run 3 (96 samples) following OTU rarefaction from 10-50,000 OTU counts/sample. Species richness plateaus at approximately 105 species/sample with increasing OTU sampling depth.

To determine the minimum sampling depth required to characterize microbial diversity at the species level, relative standard errors were calculated for each species abundance and averaged across samples following OTU rarefaction from 10-50,000 OTU counts/sample (Figure 16). This average relative standard error plateaus at approximately

5,000 OTU counts/sample. A sampling depth of 5,000 OTU counts/sample is therefore sufficient to produce the expected variation in abundance at the species level.

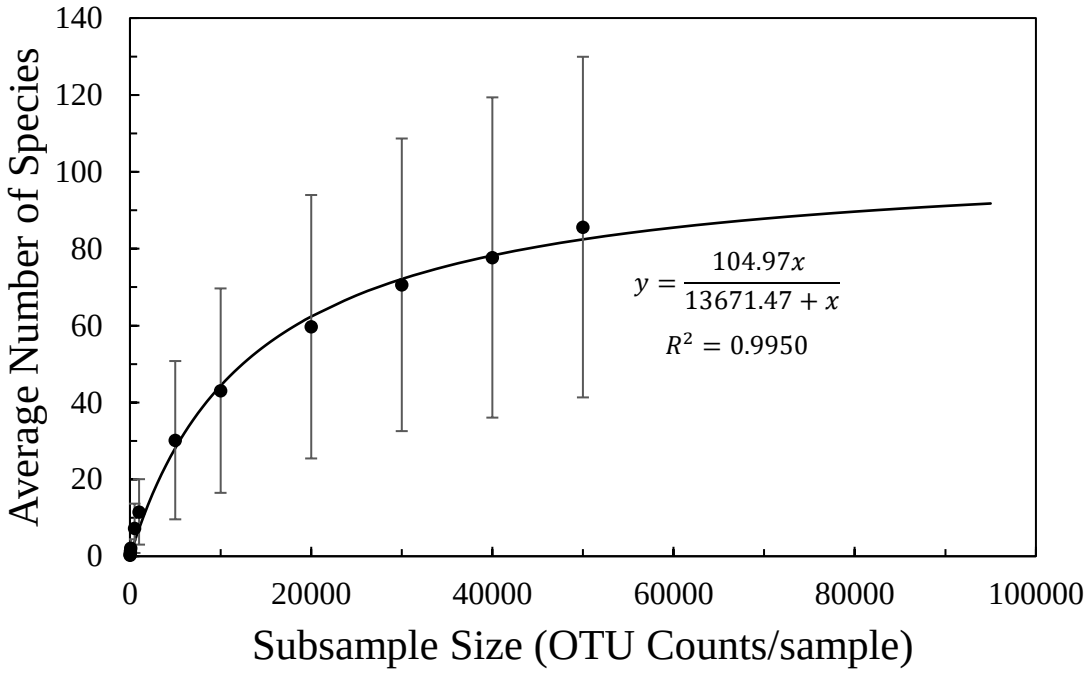


Figure 15. Species richness following OTU rarefaction.

OTU tables from Run 3 (96 samples) were subsampled from 10-50,000 OTU counts/sample and assessed for species richness.

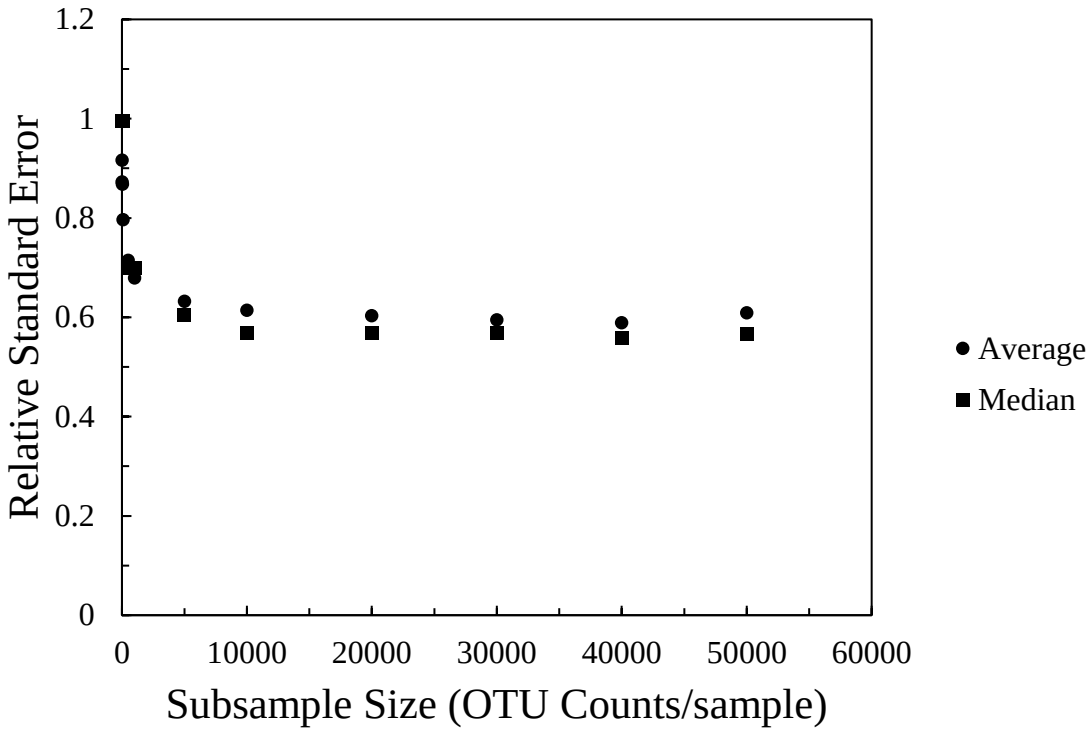


Figure 16. Species abundance variation following OTU rarefaction.

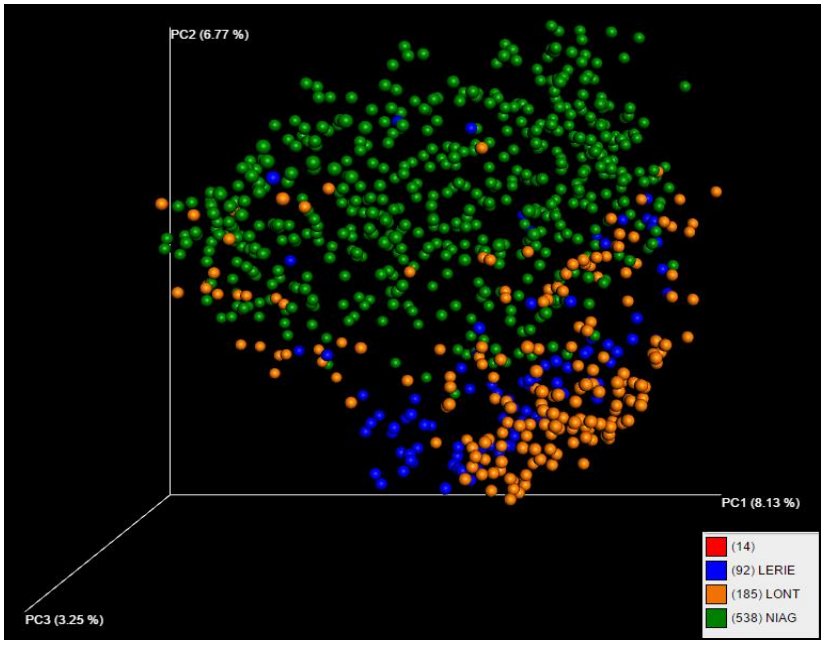
OTU tables from Run 3 (96 samples) were subsampled from 10-50,000 OTU counts/sample. Relative standard errors were calculated for each species abundance and averaged for each OTU subset.

2.5.3 Characterization of sample collection by location, subtype, and rain incidence

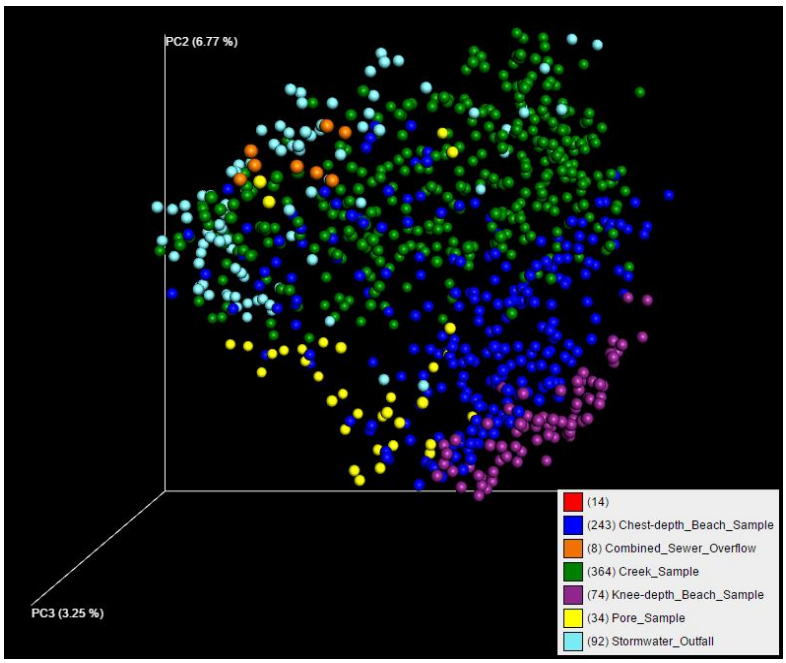
An unweighted principal components analysis (PCoA) was conducted on sample taxonomic distributions. Figure 17 depicts samples according to a) location, b) subtype, and c) rain events during collection. Demarcation of samples in Figure 17a and 17b suggest that both location and subtype influence microbial composition. The demarcation seen among sample locations is likely attributable to subtype, as samples of certain subtypes (e.g. creek and combined sewer overflow) were only collected from particular locations (i.e. the Niagara region). Lack of demarcation in Figure 17c suggests that the occurrence

of rain during sample collection has little influence on microbial communities. Additional diversity analyses were conducted to investigate these relationships.

a)



b)



c)

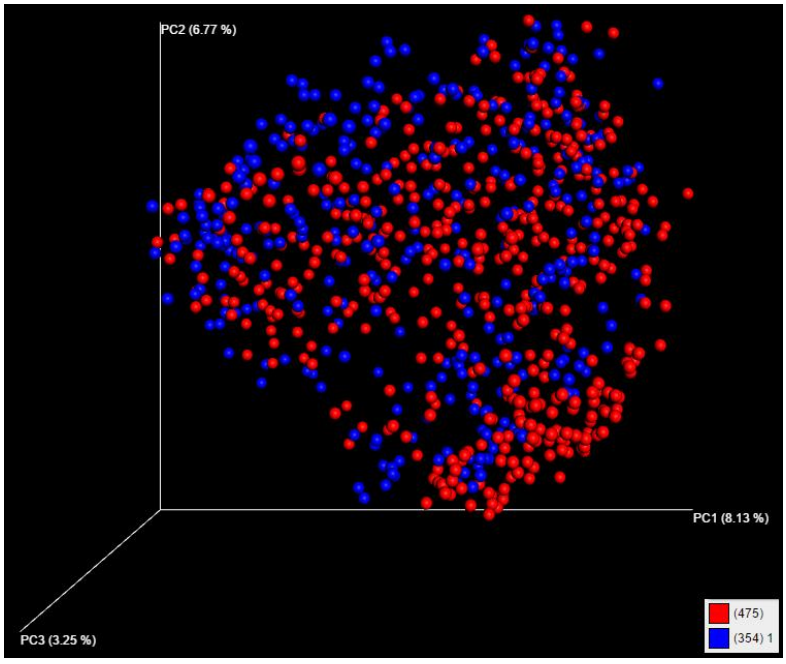


Figure 17. Microbial diversity across sample location, subtype, and rain incidence.

847 samples from the Niagara Region were sequenced with Illumina© MiSeq technology (2x300 bp reads) and analyzed using QIIME. Unweighted PCoA plots depicting a) sample location, b) sample subtype, and c) rain events during collection are shown.

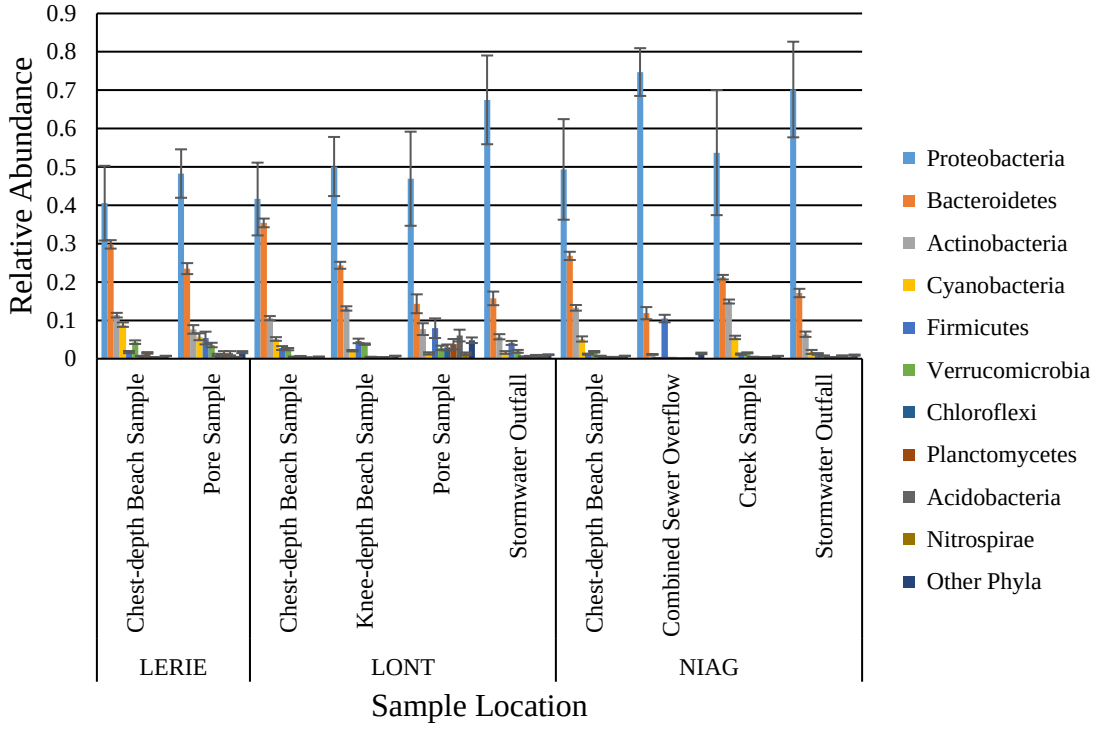
2.5.4 Relationships between subtype and microbial communities

Figure 18 depicts taxonomic distributions of the top 10 most abundant phyla across sample location and subtype. A significant increase in the relative abundance of phyla Chloroflexi, Plantomycetes, Acidobacteria, and Nitrospirae was observed in pore samples compared to chest or knee-depth water samples from Lake Ontario ($p < 0.01$, Welch's unpaired t-test) (Figure 18b).

Figure 19 depicts taxonomic distributions of Proteobacteria classes across sample location and subtype. Combined sewer overflow samples in the Niagara Region contained a smaller relative abundance of Alpha- and Betaproteobacteria and larger relative abundance of Epislon- and Gammaproteobacteria compared to beach, creek, or stormwater outfall samples ($p < 0.01$, Welch's unpaired t-test) (Figure 19b).

Figure 20 depicts taxonomic distributions of pathogen-containing genera across sample location and subtype. A significant increase in the relative abundance of genera *Vibrio* and *Yersinia* was observed in stormwater outfall samples compared to pore or water samples from Lake Ontario ($p < 0.01$, Welch's unpaired t-test) (Figure 20b). Both pore and stormwater outfall samples from Lake Ontario contained a larger relative abundance of the genus *Legionella* compared to water samples ($p < 0.01$, Welch's unpaired t-test) (Figure 20b).

a)



b)

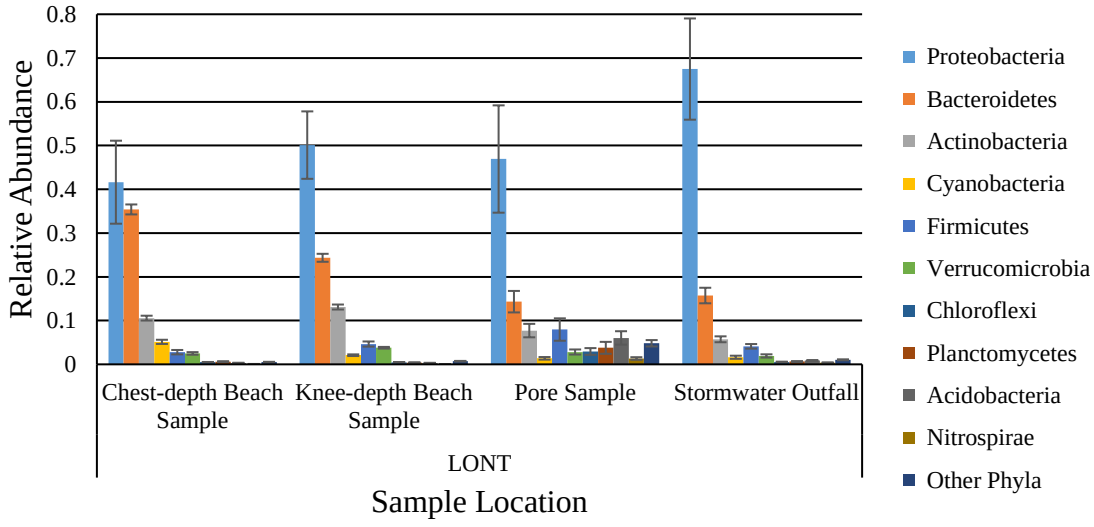
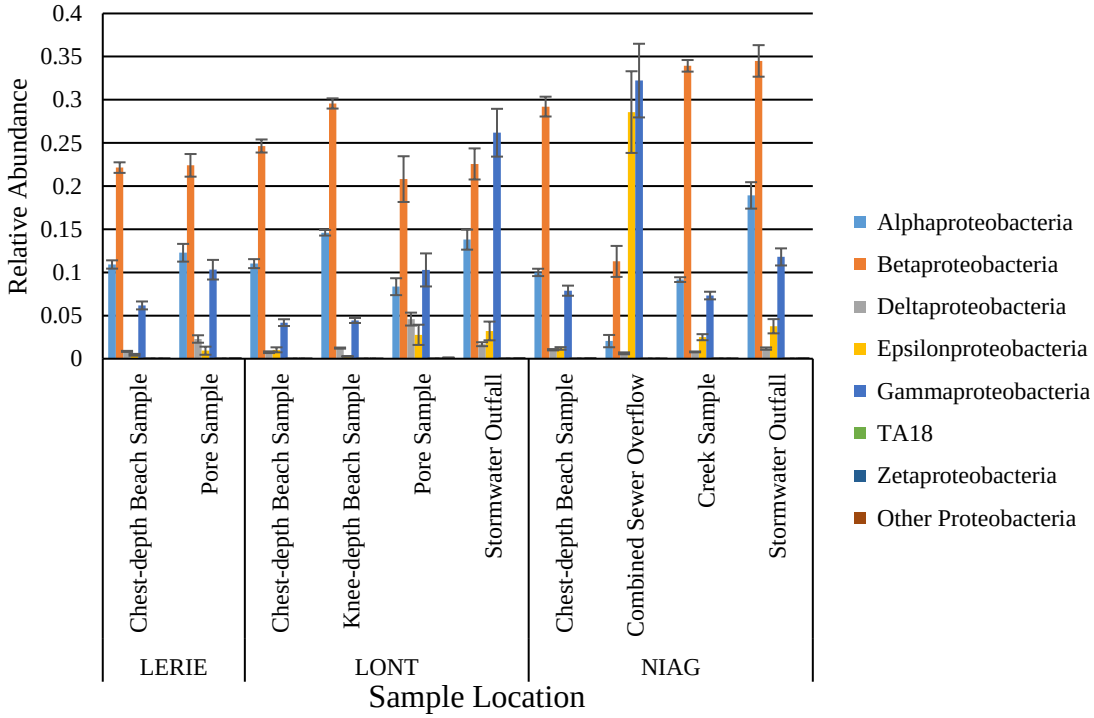


Figure 18. Phylum distribution across sample location and subtype.

a) Relative abundance of the top 10 most abundant phyla is depicted across sample location and subtype. Abundance distributions for samples collected from b) Lake Ontario are included.

a)



b)

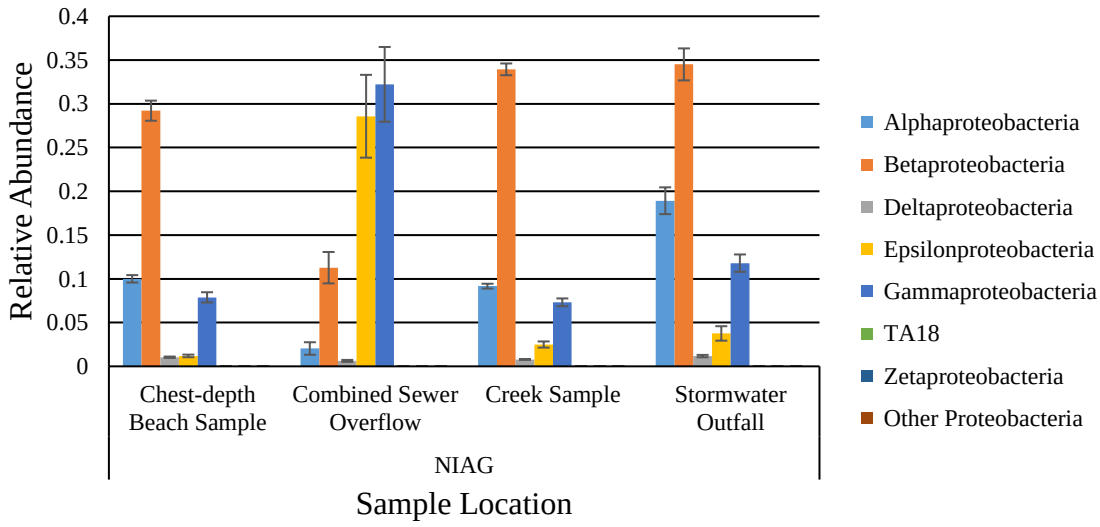
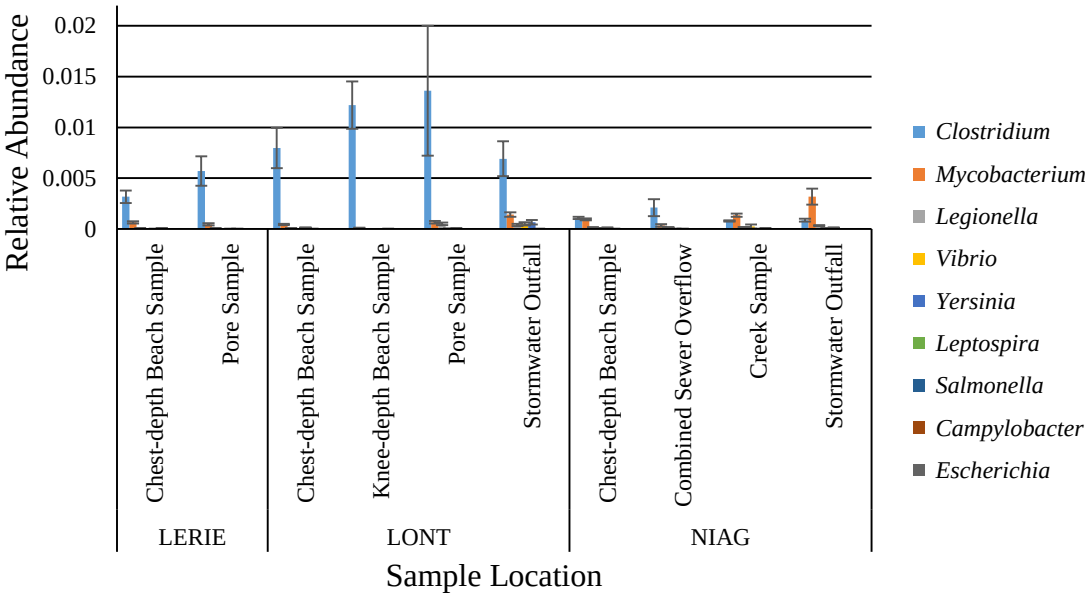


Figure 19. Proteobacteria distribution across sample location and subtype.

a) Relative abundance of Proteobacteria classes is depicted across sample location and subtype. Abundance distributions for samples collected from b) nearby areas within the Niagara Region are included.

a)



b)

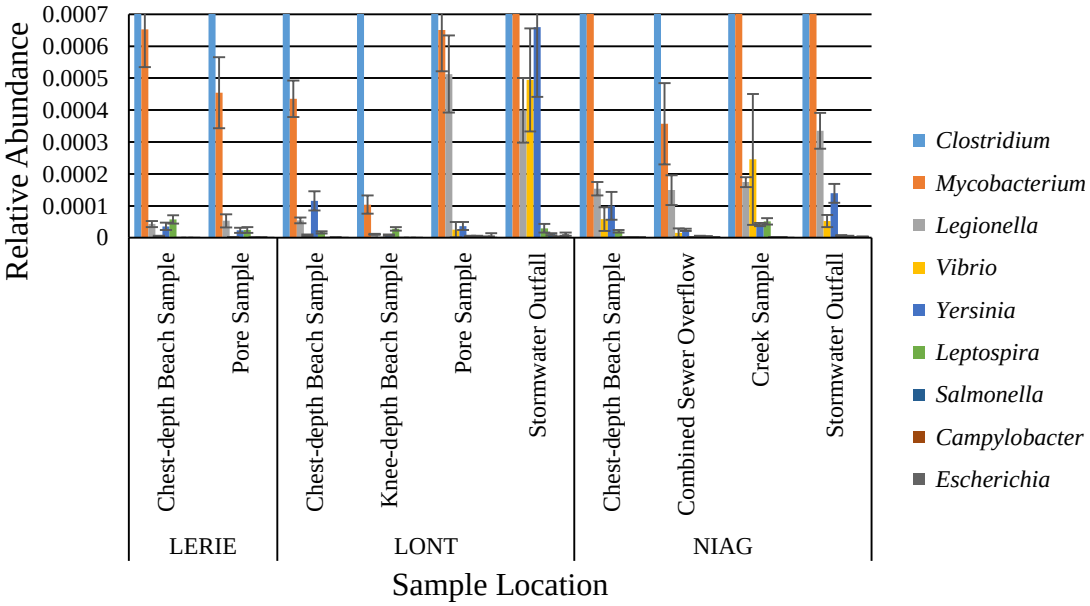


Figure 20. Pathogen-containing genus distribution across sample location and subtype.

a) Relative abundance of genera known to harbour water pathogens is depicted across sample location and subtype. Distributions for b) less abundant pathogen-containing genera are included.

2.5.5 Relationships between quantitative metadata and microbial communities

A regression analysis was conducted between taxa of interest (i.e. the top 10 most abundant phyla, Cyanobacteria classes, and pathogen-containing genera) and quantitative metadata (i.e. sample turbidity, *E. coli* counts, and Bacteroidales markers). Pearson correlation coefficients and Student's T-test results for each comparison are listed in Table 8. The analysis yielded weak but significant correlations among several comparisons, including positive correlations between the class Gammaproteobacteria and genera *Campylobacter* and *E. coli* plate counts, as well as between the phylum Firmicutes and all 3 Bacteroidales DNA markers ($p < 0.01$, Student's T-test). Figure 21 depicts the significant correlation between Gammaproteobacteria relative abundance and *E. coli* CFU counts ($p < 0.01$, Student's T-test).

Table 8. Correlations between quantitative metadata and microbial diversity.

Summary statistics are provided for the top 10 most abundant phyla, Cyanobacteria, and pathogen-containing genera with sample turbidity, *E. coli* CFU counts, and qPCR Bacteroidales markers. Correlation coefficients (R) are coloured by magnitude. Significant correlations ($p < 0.05$) are highlighted in green.

		qPCR (CFU/mL)																			
		Turbidity				*E. coli* Counts				GenBac				Hum				Gull			
		R	R^2	t	p	R	R^2	t	p	R	R^2	t	p	R	R^2	t	p	R	R^2	t	p
Top 10 Phyla	Alphaproteobacteria	-0.15	0.02	2.65	0.01	-0.17	0.03	2.63	0.01	-0.24	0.06	4.70	0.00	-0.18	0.03	2.19	0.03	-0.18	0.03	2.19	0.03
	Betaproteobacteria	-0.04	0.00	0.66	0.51	-0.20	0.04	3.11	0.00	-0.01	0.00	0.23	0.81	-0.18	0.03	2.23	0.03	-0.18	0.03	2.23	0.03
	Deltaproteobacteria	0.25	0.06	4.41	0.00	-0.06	0.00	0.91	0.36	-0.03	0.00	0.51	0.61	-0.05	0.00	0.65	0.52	-0.05	0.00	0.65	0.52
	Epsilonproteobacteria	0.19	0.04	3.24	0.00	0.46	0.21	7.72	0.00	0.23	0.05	4.50	0.00	0.33	0.11	4.33	0.00	0.33	0.11	4.33	0.00
	Gammaproteobacteria	0.10	0.01	1.63	0.10	0.39	0.16	6.44	0.00	-0.03	0.00	0.54	0.59	0.24	0.06	3.04	0.00	0.24	0.06	3.04	0.00
	TA18	-0.06	0.00	0.99	0.32	-0.11	0.01	1.70	0.09	-0.05	0.00	0.88	0.38	-0.03	0.00	0.35	0.73	-0.03	0.00	0.35	0.73
	Zetaproteobacteria	-0.03	0.00	0.58	0.56	-0.03	0.00	0.52	0.60	-0.03	0.00	0.65	0.52	-0.03	0.00	0.38	0.71	-0.03	0.00	0.38	0.71
	Other Proteobacteria	0.05	0.00	0.79	0.43	-0.06	0.00	0.88	0.38	-0.04	0.00	0.81	0.42	-0.04	0.00	0.46	0.65	-0.04	0.00	0.46	0.65
	Bacteroidetes	-0.14	0.02	2.44	0.02	-0.18	0.03	2.72	0.01	0.01	0.00	0.17	0.86	-0.11	0.01	1.35	0.18	-0.11	0.01	1.35	0.18
	Actinobacteria	-0.07	0.00	1.11	0.27	-0.19	0.04	2.94	0.00	-0.22	0.05	4.26	0.00	-0.16	0.03	1.99	0.05	-0.16	0.03	1.99	0.05
	Cyanobacteria	-0.08	0.01	1.33	0.19	-0.14	0.02	2.18	0.03	-0.12	0.02	2.34	0.02	-0.09	0.01	1.18	0.24	-0.09	0.01	1.18	0.24
	Firmicutes	0.27	0.07	4.79	0.00	0.29	0.08	4.55	0.00	0.61	0.37	14.54	0.00	0.54	0.29	7.84	0.00	0.54	0.29	7.84	0.00
	Verrucomicrobia	0.06	0.00	1.01	0.31	-0.14	0.02	2.16	0.03	-0.05	0.00	1.01	0.31	-0.08	0.01	0.99	0.32	-0.08	0.01	0.99	0.32
	Chloroflexi	0.01	0.00	0.25	0.81	-0.11	0.01	1.65	0.10	-0.09	0.01	1.73	0.08	-0.09	0.01	1.14	0.26	-0.09	0.01	1.14	0.26
	Planctomycetes	0.00	0.00	0.01	0.99	-0.10	0.01	1.48	0.14	-0.04	0.00	0.74	0.46	-0.07	0.00	0.84	0.40	-0.07	0.00	0.84	0.40
	Acidobacteria	0.17	0.03	2.91	0.00	-0.07	0.00	1.05	0.30	-0.05	0.00	0.88	0.38	-0.03	0.00	0.31	0.75	-0.03	0.00	0.31	0.75
	Nitrospirae	0.10	0.01	1.73	0.08	-0.07	0.01	1.10	0.27	-0.07	0.01	1.38	0.17	-0.04	0.00	0.51	0.61	-0.04	0.00	0.51	0.61
	Other Phyla	0.38	0.15	7.04	0.00	0.10	0.01	1.51	0.13	0.08	0.01	1.47	0.14	-0.01	0.00	0.11	0.91	-0.01	0.00	0.11	0.91
Cyanoacteria	Chloroplast	-0.05	0.00	0.79	0.43	-0.13	0.02	1.93	0.05	-0.12	0.01	2.24	0.03	-0.03	0.00	0.38	0.70	-0.07	0.01	0.92	0.36
	Synechococcophycideae	-0.10	0.01	1.76	0.08	-0.11	0.01	1.69	0.09	-0.06	0.00	1.22	0.22	-0.19	0.04	2.50	0.01	-0.09	0.01	1.17	0.24
	Oscillatoriophycideae	-0.04	0.00	0.69	0.49	-0.02	0.00	0.37	0.71	-0.02	0.00	0.30	0.77	-0.05	0.00	0.60	0.55	-0.04	0.00	0.49	0.63
	4C0d-2	-0.02	0.00	0.31	0.75	-0.02	0.00	0.32	0.75	0.03	0.00	0.58	0.57	0.21	0.04	2.73	0.01	0.01	0.00	0.14	0.89
	Nostocophycideae	-0.08	0.01	1.41	0.16	-0.02	0.00	0.37	0.71	-0.02	0.00	0.42	0.67	0.04	0.00	0.49	0.63	-0.03	0.00	0.39	0.70
	ML635J-21	0.16	0.02	2.67	0.01	-0.07	0.01	1.07	0.29	-0.04	0.00	0.85	0.40	0.00	0.00	0.05	0.96	0.00	0.00	0.05	0.96
	Gloeobacterophycideae	-0.03	0.00	0.59	0.56	-0.01	0.00	0.16	0.87	-0.02	0.00	0.32	0.75	-0.02	0.00	0.30	0.76	-0.02	0.00	0.23	0.82
	Other Cyanobacteria	0.07	0.01	1.25	0.21	-0.04	0.00	0.53	0.60	0.05	0.00	0.94	0.35	-0.06	0.00	0.73	0.47	0.00	0.00	0.01	0.99
Pathogen-Containing Genera	*Clostridium*	0.09	0.01	1.55	0.12	-0.01	0.00	0.18	0.86	0.66	0.43	16.39	0.00	0.04	0.00	0.54	0.59	0.02	0.00	0.27	0.79
	Mycobacterium	-0.09	0.01	1.56	0.12	-0.07	0.01	1.10	0.27	-0.10	0.01	1.84	0.07	-0.06	0.00	0.75	0.46	-0.03	0.00	0.42	0.68
	Legionella	-0.01	0.00	0.11	0.92	-0.01	0.00	0.16	0.87	-0.07	0.01	1.37	0.17	0.12	0.01	1.50	0.14	0.00	0.00	0.04	0.97
	Vibrio	-0.09	0.01	1.61	0.11	-0.02	0.00	0.27	0.79	0.02	0.00	0.46	0.65	0.54	0.29	8.18	0.00	-0.03	0.00	0.36	0.72
	Yersinia	-0.12	0.01	2.05	0.04	-0.04	0.00	0.57	0.57	-0.06	0.00	1.04	0.30	-0.06	0.00	0.82	0.41	-0.03	0.00	0.41	0.68
	Leptospira	0.06	0.00	1.02	0.31	-0.08	0.01	1.19	0.24	-0.03	0.00	0.56	0.58	0.00	0.00	0.02	0.98	-0.03	0.00	0.37	0.71
	Salmonella	-0.02	0.00	0.36	0.72	0.04	0.00	0.63	0.53	-0.06	0.00	1.19	0.23	-0.03	0.00	0.41	0.68	0.06	0.00	0.77	0.44
	Campylobacter	0.10	0.01	1.65	0.10	0.17	0.03	2.66	0.01	-0.04	0.00	0.73	0.46	0.04	0.00	0.49	0.62	0.30	0.09	3.95	0.00
	Escherichia	-0.04	0.00	0.70	0.48	-0.01	0.00	0.16	0.88	-0.01	0.00	0.25	0.80	0.02	0.00	0.25	0.80	0.02	0.00	0.29	0.78

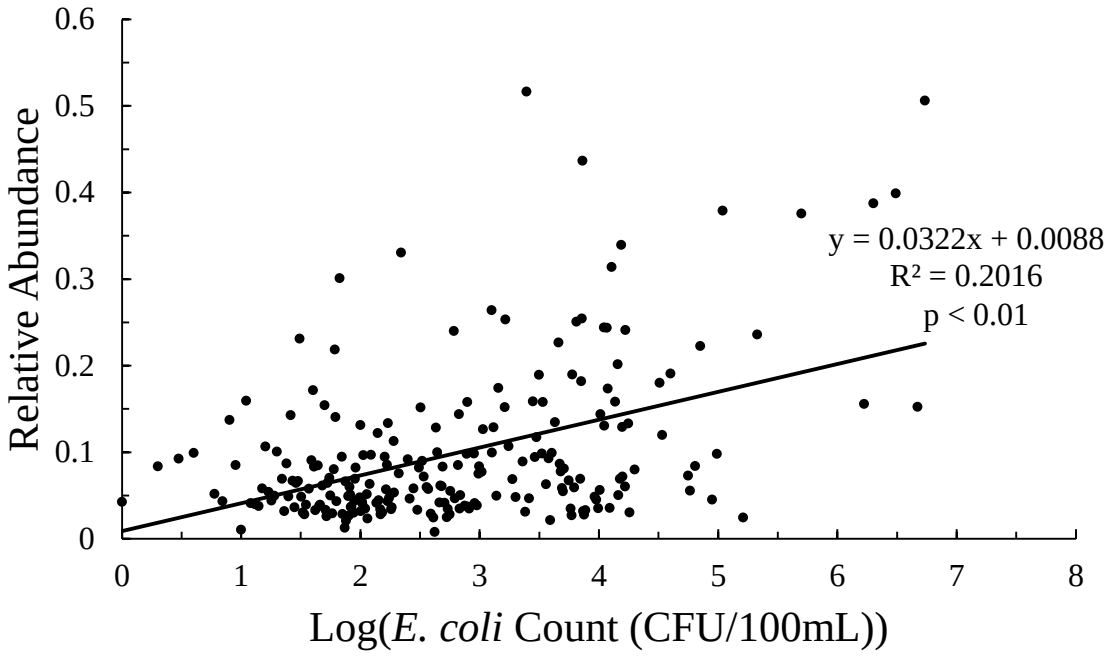


Figure 21. Gammaproteobacteria abundance across *E. coli* counts.

Relative abundance of Gammaproteobacteria is depicted across *E. coli* counts in CFU/100mL.

2.5.6 Diurnal changes in microbial communities

Figures 22-25 depict diurnal changes in the top 10 most abundant phyla (22), Proteobacteria (23) and Cyanobacteria (24) classes, and pathogen-containing genera (25) at Lakeside Beach in St. Catharines, Ontario. At the phylum and Proteobacteria level, the relative abundances of taxa were relatively stable over 8 hours of sampling (Figures 22 and 23). Within the phylum Cyanobacteria, both Chloroplast and Synechococcophycideae decreased in relative abundance over 8 hours (Figure 24a), while the class Oscillatoriophycideae peaked at 3:00 PM and decreased by 8:00 PM (Figure 24b). Additionally, the pathogen-containing genera *Clostridium* (Figure 25a) and

Mycobacterium (Figure 25b) both increased in relative abundance over the course of sampling.

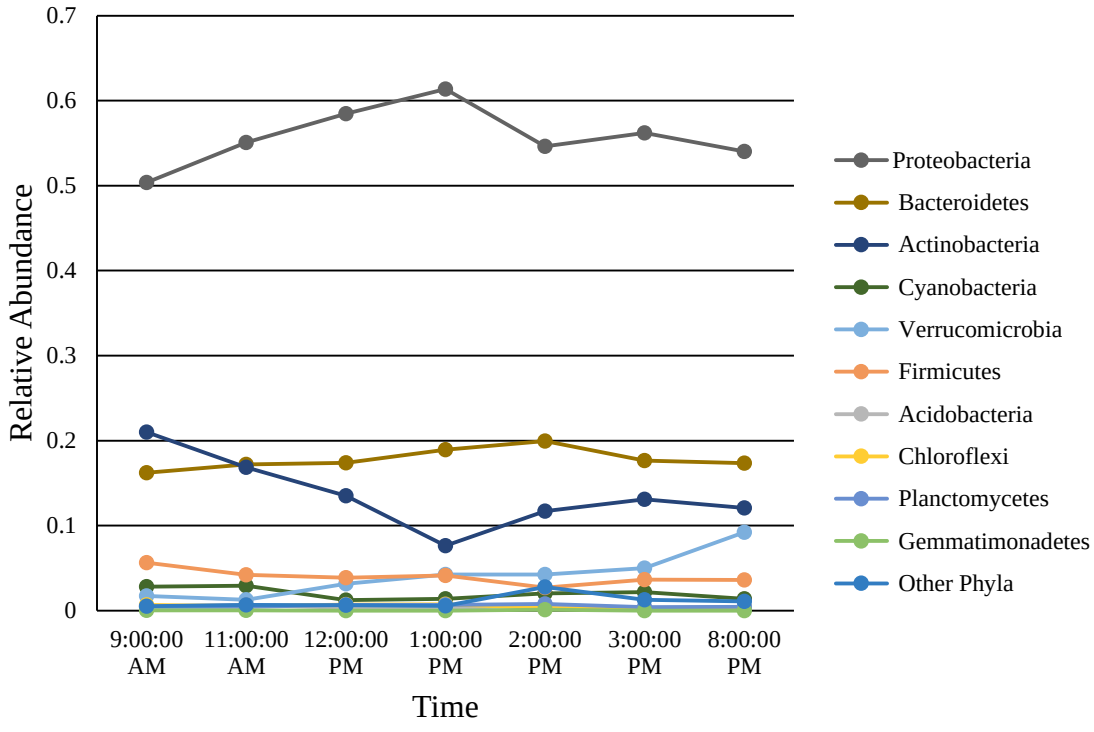


Figure 22. Diurnal changes in Lake Ontario phyla.

Water samples were collected from Lakeside Beach in St. Catharines, Ontario on August 13, 2014.

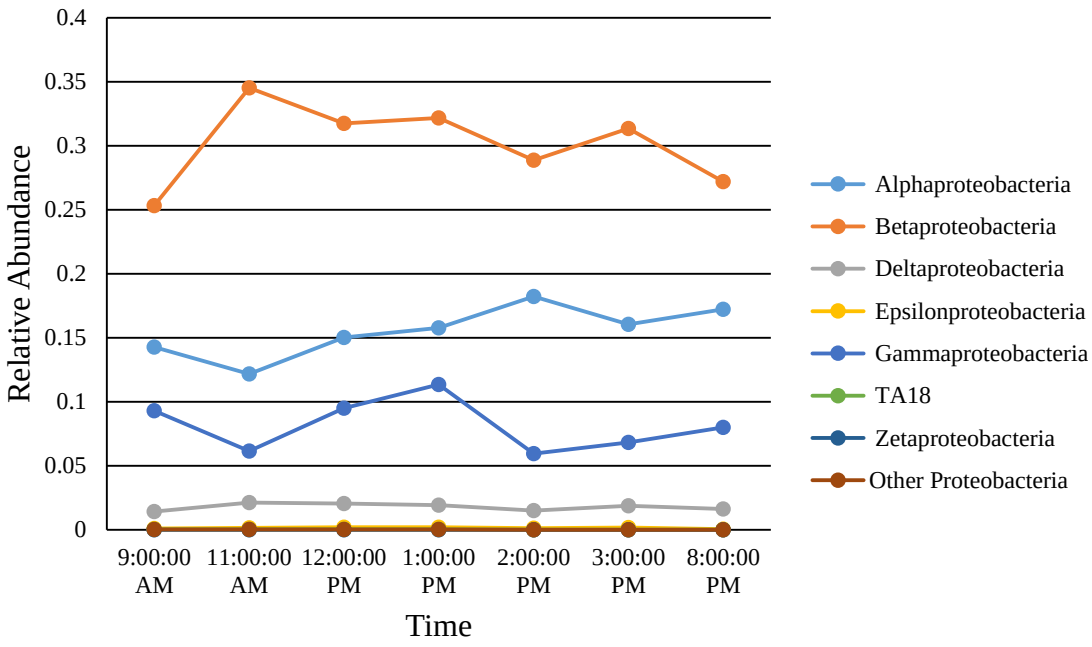
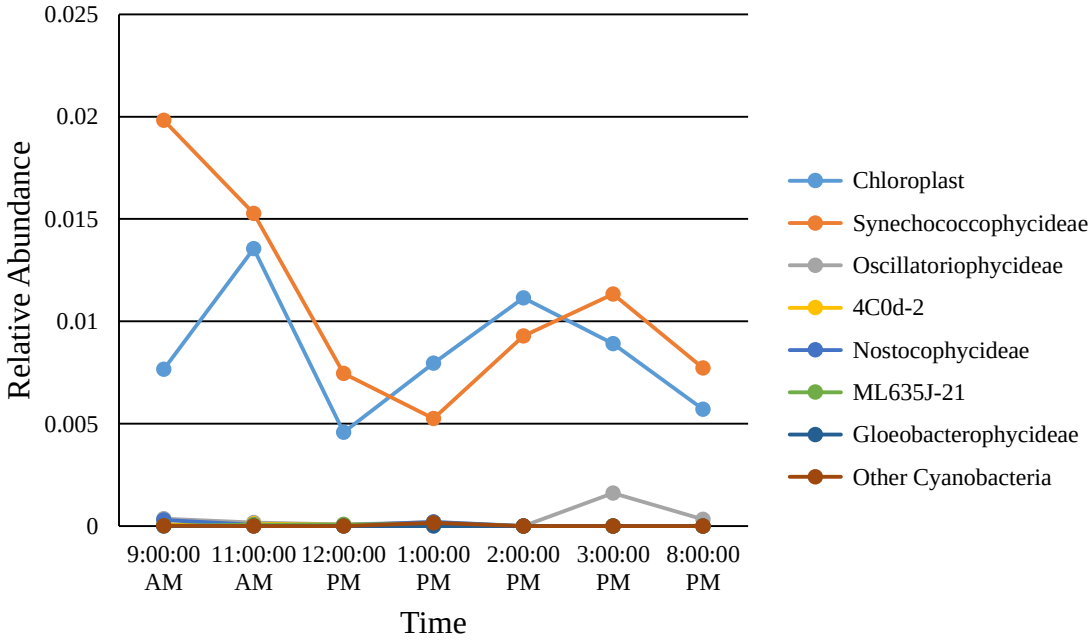


Figure 23. Diurnal changes in Lake Ontario Proteobacteria.

Water samples were collected from Lakeside Beach in St. Catharines, Ontario on August 13, 2014.

a)



b)

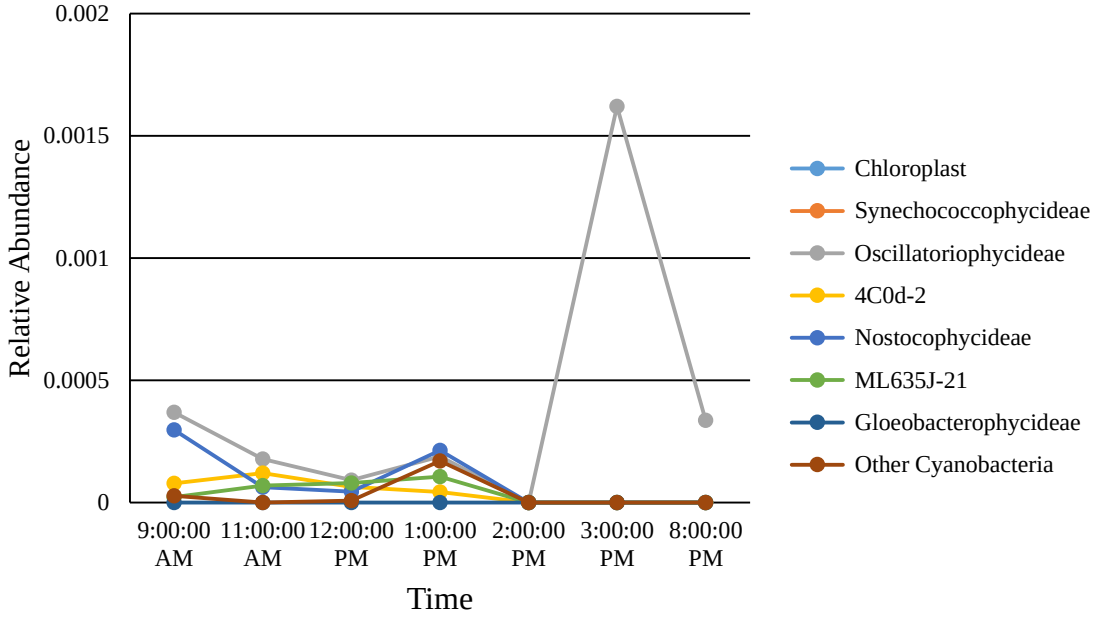
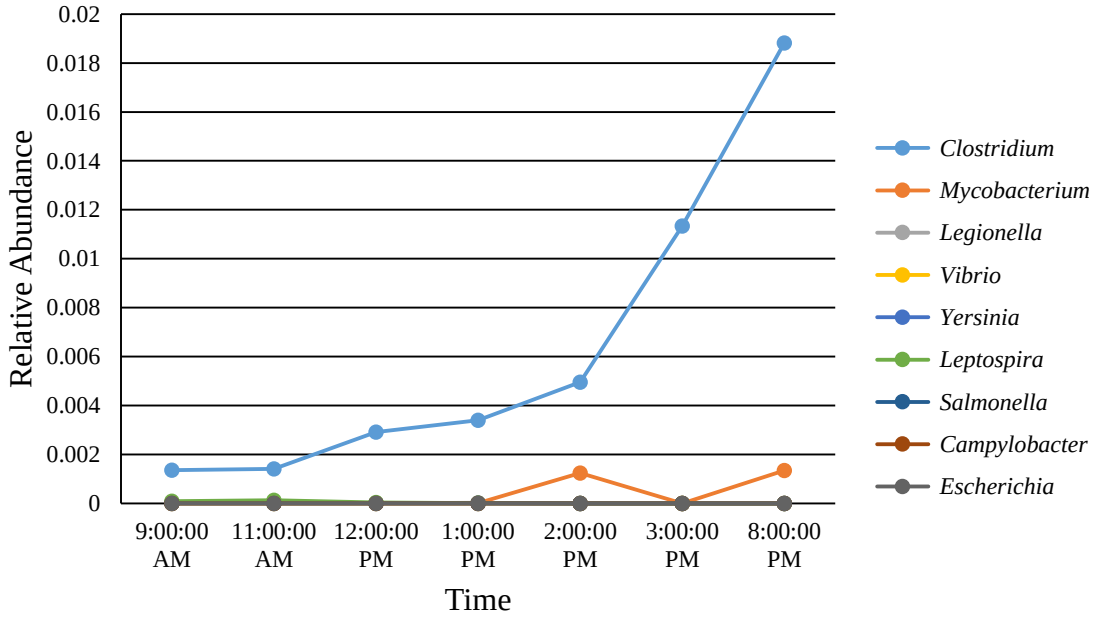


Figure 24. Diurnal changes in Lake Ontario Cyanobacteria.

a) Water samples were collected from Lakeside Beach in St. Catharines, Ontario on August 13, 2014. Distributions for b) less abundant classes are included.

a)



b)

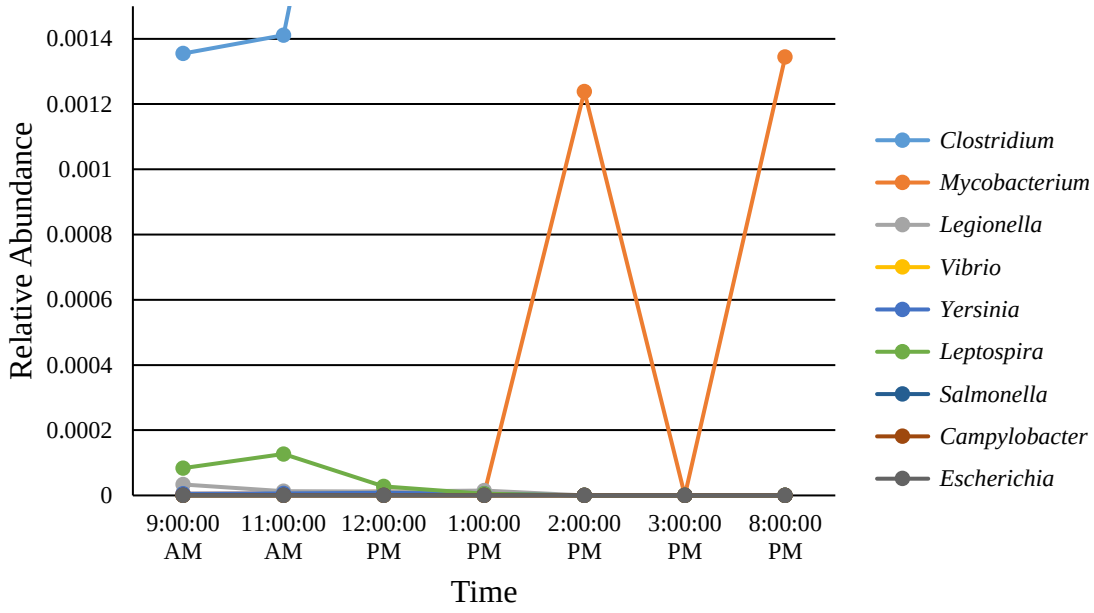


Figure 25. Diurnal changes in Lake Ontario pathogen-containing genera.

a) Water samples were collected from Lakeside Beach in St. Catharines, Ontario on August 13, 2014. Distributions for b) less abundant genera are included.

2.5.7 Microbial changes following rain events

Figures 26-29 depict changes in the top 10 most abundant phyla (26), Proteobacteria (27) and Cyanobacteria (28) classes, and pathogen-containing genera (29) at Golden Boulevard in St. Catharines, Ontario over the course of a rain event. At the phylum level, both Proteobacteria and Cyanobacteria increased while Bacteroidetes decreased in relative abundance over 10 hours of sampling (Figure 26). This increase was accounted for by Alphaproteobacteria (Figure 27) and the classes 4C0d-2 (Figure 28a) and Chloroplast (Figure 28b) in the phylum Cyanobacteria. Additionally, the pathogen-containing genera

Mycobacterium (Figure 29a) and *Clostridium* (Figure 29b) both peaked in relative abundance after 10 hours of sampling.

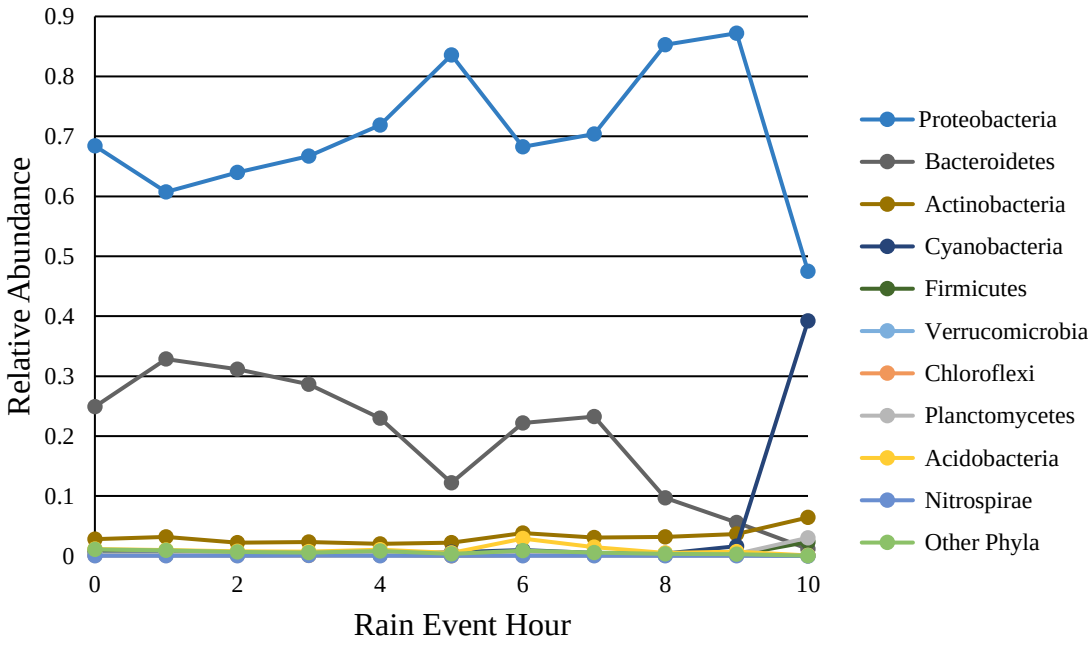


Figure 26. Temporal changes in stormwater outfall phyla during rain event.

Stormwater outfall from Golden Boulevard in St. Catharines, Ontario was autosampled every hour during a rain even on September 6, 2014.

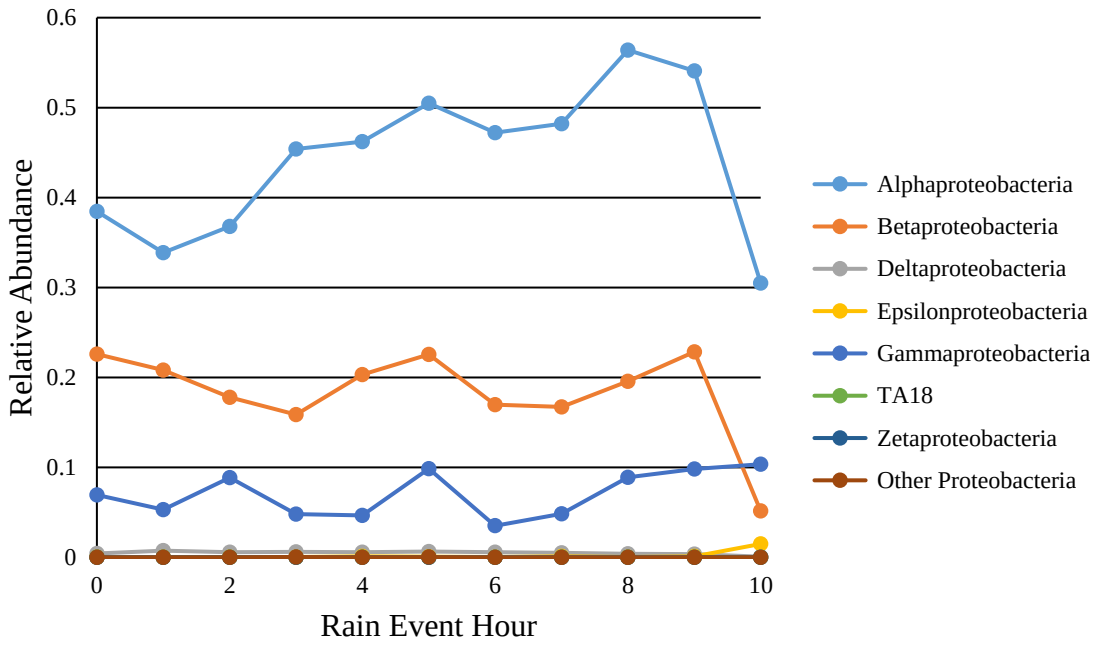
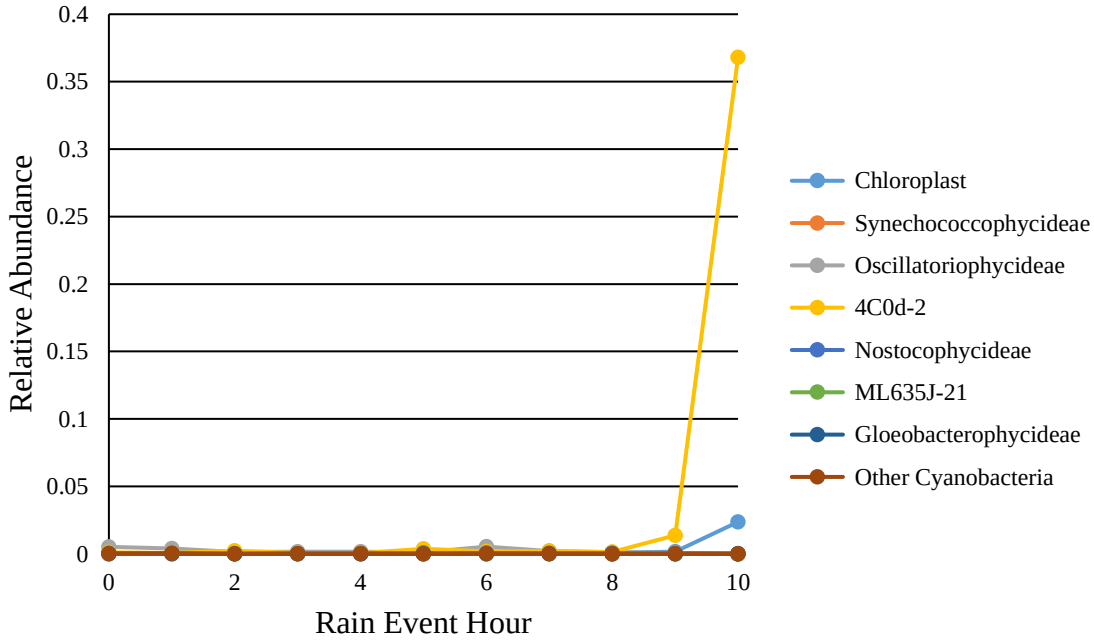


Figure 27. Temporal changes in stormwater outfall Proteobacteria during rain event.

Stormwater outfall from Golden Boulevard in St. Catharines, Ontario was autosampled every hour during a rain even on September 6, 2014.

a)



b)

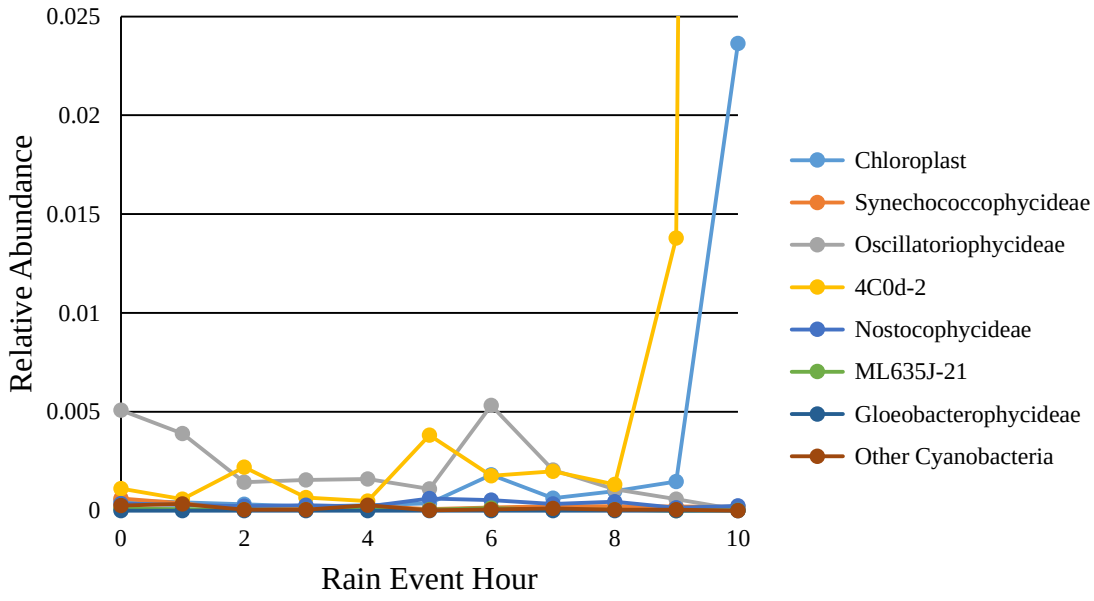
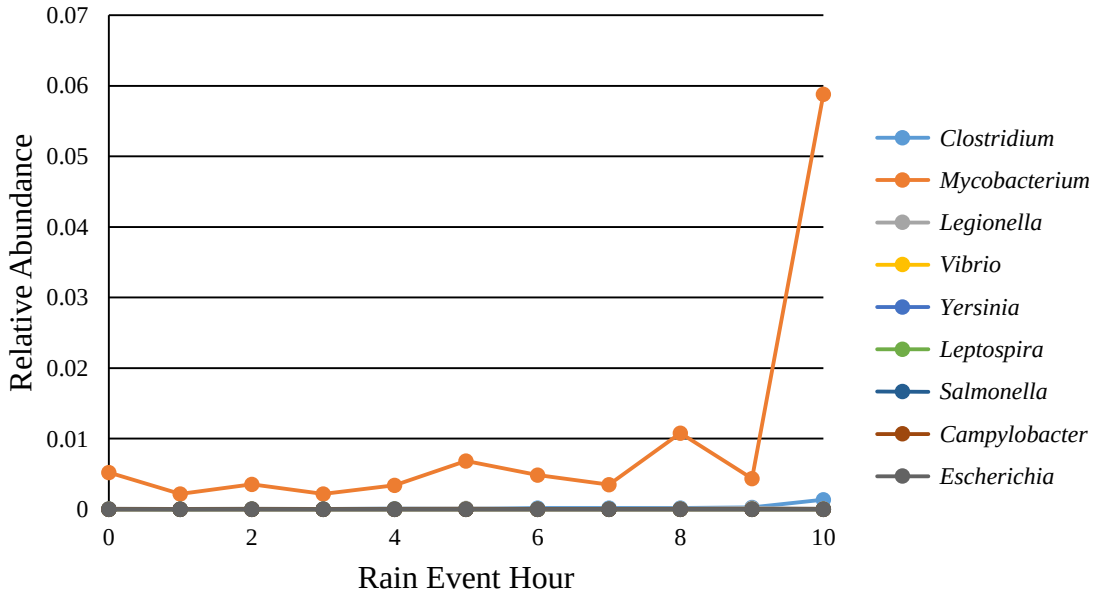


Figure 28. Temporal changes in stormwater outfall Cyanobacteria during rain event.

a) Stormwater outfall from Golden Boulevard in St. Catharines, Ontario was autosampled every hour during a rain even on September 6, 2014. Distributions for b) less abundant classes are included.

a)



b)

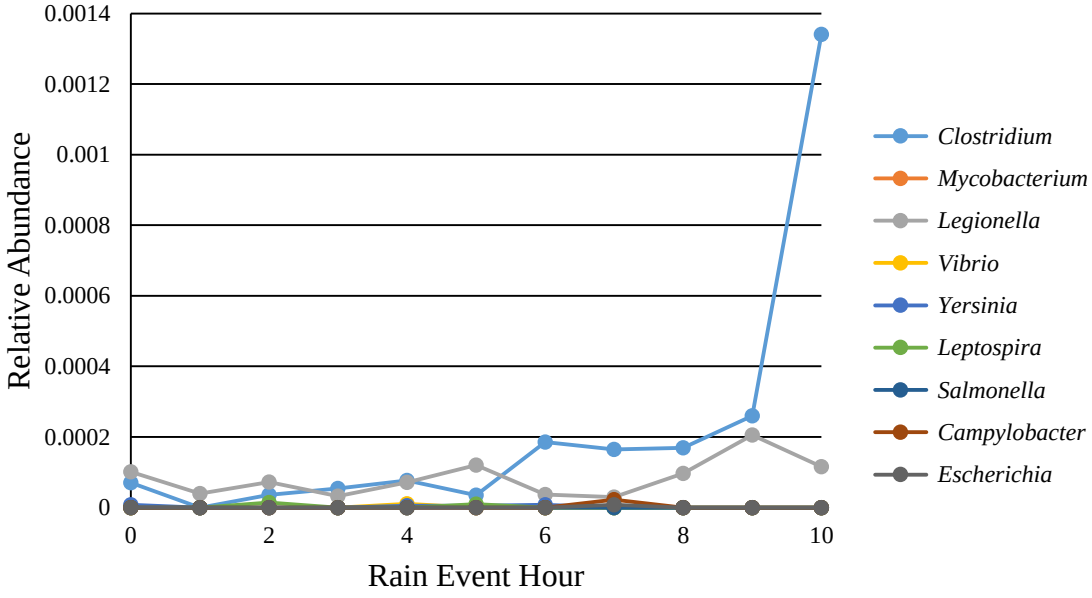


Figure 29. Temporal changes in stormwater outfall pathogen-containing genera during rain event.

a) Stormwater outfall from Golden Boulevard in St. Catharines, Ontario was autosampled every hour during a rain even on September 6, 2014. Distributions for b) less abundant genera are included.

2.5.8 Indicator OTU identification by sample location and subtype

Indicator OTUs were generated using the indicspecies R package and ranked according to decreasing p-value and increasing indicator value (i.e. IndVal). Tables x and y respectively list the top 5 indicator OTUs for each sample location and subtype.

Water samples collected from Lake Erie were characterized by 100 indicator OTUs, including the Cyanobacteria genera *Paulinella* and *Snowella* and the thermophilic genera *Caldilinea*. Those collected from Lake Ontario were characterized by 21 OTUs, including species from the orders Sphingomonadales, Methylococcales, and Methylophilales. 105 indicator OTUs were identified among water samples from the surrounding region, including the Cyanobacteria genera *Pseudanabaena*.

Combined sewer overflow and stormwater outfall samples were respectively characterized by 289 and 98 indicator OTUs, including the human commensal genera *Arcobacter* and *Leptotrichia* in CSO and the Actinobacteria families Geodermatophilaceae and Nocardiaceae in SO. 169 and 9 indicator OTUs were respectively identified among lake pore and water samples, including members of the thermophilic family Thermodesulfovibrionaceae and the soil bacteria phylum Gemmatimonadetes in pore samples and the cattle bacterium and opportunistic human pathogen *Prevotella melaninogenica* in water samples.

Table 9. Top 5 indicator OTUs for sample locations.

Indicator OTUs were identified for each location using the indicspecies R package and ranked according to decreasing p-value and increasing indicator value (IndVal; max = 1).

Location	Number of indicator OTUs	Top 5 indicator OTUs		
		OTU	IndVal	p-value
Lake Erie	100	k__Bacteria.p__Cyanobacteria.c__Synechococcophycideae.o__Synechococcales.f__Synechococcaceae.g__Paulinella.s__	0.864	0.001
		k__Bacteria.p__Cyanobacteria.c__Oscillatoriophycideae.o__Chroococcales.f__Gomphosphaeriaceae.g__.s__	0.798	0.001
		k__Bacteria.p__Cyanobacteria.c__Oscillatoriophycideae.o__Chroococcales.f__Cyanobacteriaceae.g__Cyanobacterium.s__	0.797	0.001
		k__Bacteria.p__Cyanobacteria.c__Oscillatoriophycideae.o__Chroococcales.f__Gomphosphaeriaceae.g__Snowella.s__	0.795	0.001
		k__Bacteria.p__Chloroflexi.c__Anaerolineae.o__Caldilineales.f__Caldilineaceae.g__Caldilinea.s__	0.773	0.001
Lake Ontario	21	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Sphingomonadales.f__Sphingomonadaceae.g__Novosphingobium.s__capsulatum	0.685	0.001
		k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Methylococcales.f__Crenotrichaceae.g__Crenothrix.s__polyspora	0.666	0.001
		k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Methylophilales.f__Methylophilaceae.g__Methylovorus.s__glucosotrophus	0.627	0.001
		k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Microbacteriaceae.g__Agrococcus.s__	0.555	0.001
		k__Bacteria.p__OP3.c__koll11.o__GIF10.f__.g__.s__	0.548	0.001

Location	Number of indicator OTUs	Top 5 indicator OTUs		
		OTU	IndVal	p-value
Niagara Region	105	k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae.g__PSB.M.3.s__	0.944	0.001
		k__Bacteria.p__Bacteroidetes.c__.Saprospirae..o__.Saprospirales..f__Saprospiraceae.g__Haliscomenobacter.s__	0.942	0.001
		k__Bacteria.p__Planctomycetes.c__Phycisphaerae.o__Phycisphaerales.f__.g__.s__	0.935	0.001
		k__Bacteria.p__Planctomycetes.c__Planctomycetia.o__Gemmatales.f__Isosphaeraceae.g__.s__	0.927	0.001
		k__Bacteria.p__Cyanobacteria.c__Synechococcophycideae.o__Pseudanabaenales.f__Pseudanabaenaceae.g__Pseudanabaena.s__	0.92	0.001

Table 10. Top 5 indicator OTUs for sample subtypes.

Indicator OTUs were identified for each subtype using the indicspecies R package and ranked according to decreasing p-value and increasing indicator value (IndVal; max = 1).

Subtype	Number of indicator OTUs	Top 5 indicator OTUs		
		OTU	IndVal	p-value
Chest-depth beach sample	1	k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Bacillaceae.g__Bacillus.s__thermoamylovorans	0.375	0.027
Combined sewer overflow	289	k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae.g__Chelonobacter.s__	0.999	0.001
		k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Neisseriales.f__Neisseriaceae.g__Microvirgula.s__	0.979	0.001
		k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria.o__Campylobacterales.f__Campylobacteraceae.g__Arcobacter.s__cryaerophilus	0.976	0.001
		k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Actinomycetaceae.g__.s__	0.971	0.001
		k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Leptotrichiaceae.g__Leptotrichia.s__	0.968	0.001
Creek sample	5	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodospirillales.f__Rhodospirillaceae.g__Rhodospirillum.s__	0.358	0.023
		k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__RF16.g__.s__	0.345	0.023
		k__Bacteria.p__Chlorobi.c__Chlorobia.o__Chlorobiales.f__Chlorobiaceae.g__Chlorobaculum.s__	0.311	0.034
		k__Bacteria.p__Cyanobacteria.c__Oscillatoriophycideae.o__Chroococcales.f__Gomphosphaeriaceae.g__Woronichinia.s__naegeliana	0.261	0.049

Subtype	Number of indicator OTUs	Top 5 indicator OTUs		
		OTU	IndVal	p-value 8
		k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Dermabacteraceae.g__Helcobacillus.s__massiliensis	0.259	0.048
Knee-depth beach sample	8	k__Bacteria.p__Verrucomicrobia.c__.Methylacidiphilae..o__Methylacidiphilales.f__LD19.g__.s__	0.933	0.001
		k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__.Tissierellaceae..g__Sedimentibacter.s__	0.75	0.001
		k__Bacteria.p__Armatimonadetes.c__Armatimonadia.o__Armatimonadales.f__Armatimonadaceae.g__.s__	0.638	0.002
		k__Bacteria.p__Cyanobacteria.c__4C0d.2.o__SM2F09.f__.g__.s__	0.517	0.008
		k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Prevotellaceae.g__Prevotella.s__melaninogenica	0.388	0.002
Pore sample	169	k__Bacteria.p__Nitrospirae.c__Nitrospira.o__Nitrospirales.f__.Thermodesulfovibrionaceae..g__GOUTA19.s__	0.861	0.001
		k__Bacteria.p__Actinobacteria.c__Acidimicrobiia.o__Acidimicrobiales.f__koll13.g__.s__	0.835	0.001
		k__Bacteria.p__Gemmatimonadetes.c__Gemmatimonadetes.o__.f__.g__.s__	0.821	0.001
		k__Bacteria.p__Nitrospirae.c__Nitrospira.o__Nitrospirales.f__.Thermodesulfovibrionaceae..g__HB118.s__	0.819	0.001
		k__Bacteria.p__Chlorobi.c__BSV26.o__PK329.f__.g__.s__	0.818	0.001
Stormwater outfall	98	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodospirillales.f__Rhodospirillaceae.g__Skermanella.s__	0.777	0.004
		k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Geodermatophilaceae.g__.s__	0.768	0.001
		k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Nocardiaceae.g__.s__	0.746	0.001
		k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Caulobacterales.f__Caulobacteraceae.g__Asticcacaulis.s__	0.739	0.002
		k__Bacteria.p__Chloroflexi.c__SAR202.o__.f__.g__.s__	0.682	0.001

2.6. DISCUSSION

2.6.1 Summary and rationale

This study investigated the use of next-generation sequencing technology to characterize microbial diversity in the Niagara Region and augment traditional water quality monitoring methods. Water samples were collected from Lake Erie, Lake Ontario, and nearby areas and examined with 16S rRNA gene sequencing. An automated workflow was first developed in Perl for microbial community analyses using QIIME. The OTU table generated from the sample collection was then analyzed across qualitative and quantitative parameters. Differences in microbial communities between sample locations and subtypes were investigated on the basis of clear demarcation of sample groups in a principal coordinates analysis. Significant correlations were identified between microbial diversity and quantitative parameters (i.e. *E. coli* count and fecal DNA marker data from water quality monitoring initiatives). Genera known to harbour water pathogens (e.g. *Clostridium*, *Escherichia*, and *Mycobacterium* among others) were observed at a relative abundance of 0.1-1.5%. Further analyses identified indicator OTUs (i.e. those abundant in and characteristic of particular sample groups) within sample locations and subtypes. These results indicate that sequence-based analyses can be used in combination with traditional identification methods to profile microbial diversity and predict water quality.

2.6.2 Phylum and Proteobacteria distributions across sample location and subtype

Microbial distributions of the top 10 most abundant phyla and Proteobacteria classes were compared across sample locations and subtypes. Notably, pore samples were

enriched in Acidobacteria relative to water samples from Lake Ontario. The phylum Acidobacteria is characteristic of soil environments and should therefore occupy a larger niche within the microbial communities of samples collected from Lake shorelines (Quaiser et al., 2003). Similarly, combined sewer overflow samples contained a larger relative abundance of Epislon- and Gammaproteobacteria than beach, creek, or stormwater outfall samples in the Niagara Region. Epsilon- and Gammaproteobacteria typically inhabit the digestive tracts of animals as either symbionts or pathogens, suggesting contamination of the Niagara watershed with animal fecal pollution.

2.6.3 Identification and implications of pathogen-containing genera

The relative abundance of genera known to harbour microbial water pathogens was also examined across sample locations and subtypes. The greatest proportion of pathogen-containing genera was found among Lake Ontario samples (0.9-1.5%) relative to Lake Erie and surrounding areas (0.1-0.7%). The genus *Clostridium* predominated within fractions from Lake Ontario and Lake Erie, while the majority of samples from the Niagara watershed contained a larger relative abundance of *Mycobacterium*.

Whether or not water pathogens are present in pathogen-containing genera identified by metagenomic analyses is an important concern for water quality monitoring and public health. For instance, the *Escherichia coli* species typically forms a probiotic, mutualistic relationship with its host and only a small proportion of strains are considered pathogenic (Altenhoefer et al., 2004). The specific detection of pathogenic species is limited to the resolution of the sequencing analysis. In the case of 16S rRNA sequencing,

many OTUs fetched during alignment resolve at the level of genera or higher taxonomic classifications. 16S analyses present a challenge in adequately characterizing the abundance of pathogenic organisms and may be addressed with whole-genome sequencing methods which allow for more robust characterization.

2.6.4 Correlation of taxonomic data with *E. coli* plate counts

A regression analysis was conducted between taxa of interest and quantitative parameters (i.e. sample turbidity, *E. coli* plate counts, and Bacteroidales markers) to identify correlations between metagenomic data and traditional sampling methods. A significant correlation was identified between *E. coli* counts and the class Gammaproteobacteria, of which the genus *Escherichia* is a member. No significant correlation was found, however, with the *Escherichia* genus alone. These results suggest a limited resolution for *Escherichia coli* at the genus level. The detection of *E. coli* 16S sequences at higher taxonomic levels would support the correlation between Gammaproteobacteria and *E. coli* plate counts.

2.6.5 Temporal changes in microbial diversity

Microbial community dynamics were observed diurnally over 8 hours of sampling and over 10 hours during the course of a rain event. While phylum and Proteobacteria distributions were relatively stable over the 8 hour period, a decrease in the relative abundance of select Cyanobacteria and increase in *Clostridium* and *Mycobacterium* was observed in the afternoon. By contrast, phylum distributions shifted during the rain event

while select Proteobacteria, Cyanobacteria, and pathogen-containing genera increased in relative abundance.

The genus *Clostridium* is a fecal indicator prevalent in the wet sand of public beaches (Heaney et al., 2012). A progressive increase in the relative abundance of *Clostridium* over an 11 hour diurnal sampling period (i.e. 9:00am – 8:00pm) suggests that recreational beach activity may be responsible for stirring beach sand and introducing *Clostridium* species into the aquatic environment.

Several datasets showed a large variation in relative abundance during diurnal monitoring and would therefore benefit from additional sampling sites and an extended sampling duration. It is also possible that changes in *Clostridium* and *Mycobacterium* abundance during the rain event were confounded by diurnal behaviour observed at the same time of day. An analysis of microbial dynamics during rain events at a different time or controlling for diurnal behaviour would reduce confounding effects.

2.6.6 Identification of indicator OTUs

A large collection of OTUs abundant in and indicative of particular sample groups (i.e. location and subtype) were identified for the Niagara Region. Notable indicator taxa include several Cyanobacteria genera characteristic of Lake Erie samples and human commensal genera characteristic of combined sewer overflow (CSO). OTUs of this nature may be used to predict algal bloom or CSO pollution sources in future microbial monitoring studies.

2.6.7 Quantitative metagenomics analysis and future directions

The incorporation of quantitative DNA measurements into microbial diversity analyses may provide a more accurate picture of microbial dynamics in the Niagara Region. The use of a biological spiking control (i.e. an organism not expected to be present in the sample) allows for quantification of DNA present in the sample and generation of absolute values for species abundance. For instance, a 2010 study spiked fungal community samples with control spores and observed that read abundance is generally quantitative within species but suffers from sequence biases between species (Amend et al.). Stool samples were similarly spiked with *S. oneidensis* cells in a recent quantitative metagenomics analysis of the human microbiome (Jones et al., 2015). In contrast, metagenomics studies have also refuted the validity of DNA quantification in analyzing microbial diversity. A 2011 study which spiked microbial community samples with *S. oneidensis* cells observed large variation in control abundance for both forward and reverse 16S primers (Zhou et al.). Similar spiking of zooplankton community samples with indicator species larvae produced an order of magnitude variation in control abundance among different communities (Sun et al., 2015).

Future work may investigate the application and validity of quantitative DNA measurements in characterizing microbial diversity in the Niagara Region. Considerations for generating absolute abundance values from measurements of DNA concentration include the presence of eukaryotic DNA in water samples as well as bias during PCR amplification and sequencing.

REFERENCES

Altenhoefer, A., Oswald, S., Sonnenborn, U., Enders, C., Schulze, J., Hacker, J. and Oelschlaeger, T. A. (2004). The probiotic Escherichia coli strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens. *FEMS Immunol Med Microbiol* **40**(3): 223-229.

Amend, A. S., Seifert, K. A. and Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol Ecol* **19**(24): 5555-5565.

Beloin, C., Michaelis, K., Lindner, K., Landini, P., Hacker, J., Ghigo, J. M. and Dobrindt, U. (2006). The Transcriptional Antiterminator RfaH Represses Biofilm Formation in Escherichia coli. *J Bacteriol* **188**(4): 1316-1331.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., . . . Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**: Unit 19.10.11-21.

Boaretti, M., Lleo, M. M., Bonato, B., Signoretto, C. and Canepari, P. (2003). Involvement of rpoS in the survival of Escherichia coli in the viable but non-culturable state. *Environ Microbiol* **5**(10): 986-996.

Bougdour, A., Wickner, S. and Gottesman, S. (2006). Modulating RssB activity: IraP, a novel regulator of sigma(S) stability in Escherichia coli. *Genes Dev* **20**(7): 884-897.

Bouhss, A., Crouvoisier, M., Blanot, D. and Mengin-Lecreulx, D. (2004). Purification and Characterization of the Bacterial MraY Translocase Catalyzing the First Membrane Step of Peptidoglycan Biosynthesis. *Journal of Biological Chemistry* **279**(29): 29974-29980.

Brittnacher, M. J., Fong, C., Hayden, H. S., Jacobs, M. A., Radey, M. and Rohmer, L. (2011). PGAT: a multistrain analysis resource for microbial genomes. *Bioinformatics* **27**(17): 2429-2430.

Brown, L. and Elliott, T. (1997). Mutations that increase expression of the rpoS gene and decrease its dependence on hfq function in Salmonella typhimurium. *Journal of Bacteriology* **179**(3): 656-662.

Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H. and Chen, L. (2008). Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics* **17**(2): 444-472.

Canada, H. a. W. Guidlines for Canadian Recreational Water Quality. Ottawa, Canadian Government Publishing Centre**:** p. 14.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**(5): 335-336.

Carlioz, A. and Touati, D. (1986). Isolation of superoxide dismutase mutants in Escherichia coli: is superoxide dismutase necessary for aerobic life? *Embo j* **5**(3): 623-630.

Chase-Topping, M. E., McKendrick, I. J., Pearce, M. C., MacDonald, P., Matthews, L., Halliday, J., . . . Woolhouse, M. E. (2007). Risk factors for the presence of high-level shedders of Escherichia coli O157 on Scottish farms. *J Clin Microbiol* **45**(5): 1594-1603.

Chen, G., Patten, C. L. and Schellhorn, H. E. (2004). Positive selection for loss of RpoS function in Escherichia coli. *Mutat Res* **554**(1-2): 193-203.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E., Wetter, T. and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* **14**(6): 1147-1159.

Chiang, S. M., Dong, T., Edge, T. A. and Schellhorn, H. E. (2011). Phenotypic diversity caused by differential RpoS activity among environmental Escherichia coli isolates. *Appl Environ Microbiol* **77**(22): 7915-7923.

Chiang, S. M. and Schellhorn, H. E. (2010). Evolution of the RpoS regulon: origin of RpoS and the conservation of RpoS-dependent regulation in bacteria. *J Mol Evol* **70**(6): 557-571.

Clementz, T., Zhou, Z. and Raetz, C. R. (1997). Function of the Escherichia coli msbB gene, a multicopy suppressor of htrB knockouts, in the acylation of lipid A. Acylation by MsbB follows laurate incorporation by HtrB. *J Biol Chem* **272**(16): 10353-10360.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci* **345**(1311): 101-118.

Danese, P. N., Pratt, L. A. and Kolter, R. (2000). Exopolysaccharide production is required for development of Escherichia coli K-12 biofilm architecture. *J Bacteriol* **182**(12): 3593-3596.

Davies, C. M., Long, J. A., Donald, M. and Ashbolt, N. J. (1995). Survival of fecal microorganisms in marine and freshwater sediments. *Appl Environ Microbiol* **61**(5): 1888-1896.

De Caceres, M. and Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**(12): 3566-3574.

Demple, B., Halbrook, J., and Linn, S. (1983). *Escherichia coli xth* mutants are hypersentitive to hydrogen peroxide. *J Bacteriol.* **153**(2): 1079-1082.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**(7): 5069-5072.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007). SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res* **17**(11): 1697-1706.

Dong, T., Chiang, S. M., Joyce, C., Yu, R. and Schellhorn, H. E. (2009). Polymorphism and selection of rpoS in pathogenic Escherichia coli. *BMC Microbiol* **9**: 118.

Dong, T., Joyce, C., and Schellhorn, H.E. (2008). The role of RpoS in bacterial adaptation. Bacterial physiology - a molecular approach. W. M. El-Sharoud. Berlin, Germany, Springer**:** 313-339.

Dong, T., Kirchhof, M. G. and Schellhorn, H. E. (2008). RpoS regulation of gene expression during exponential growth of Escherichia coli K12. *Mol Genet Genomics* **279**(3): 267-277.

Dufour, A. P. (1984). Health Effects Criteria for Fresh Recreational Waters, U.S. Environmental Protection Agency

Dufrene, M. and Legendre, P. (1997). Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach. *Ecological Monographs* **67**(3): 345-366.

Edberg, S. C., Rice, E. W., Karlin, R. J. and Allen, M. J. (2000). Escherichia coli: the best biological drinking water indicator for public health protection. *Symp Ser Soc Appl Microbiol*(29): 106s-116s.

Ferenci, T. (2003). What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli? *Trends Microbiol* **11**(10): 457-461.

Fish, J. T. and Pettibone, G. W. (1995). Influence of freshwater sediment on the survival of Escherichia coli and Salmonella sp. as measured by three methods of enumeration. *Lett Appl Microbiol* **20**(5): 277-281.

Franchini, A. G., Ihssen, J. and Egli, T. (2015). Effect of Global Regulators RpoS and Cyclic-AMP/CRP on the Catabolome and Transcriptome of <italic>Escherichia coli</italic> K12 during Carbon- and Energy-Limited Growth. *PLoS ONE* **10**(7): e0133793.

Freundlich, M., Ramani, N., Mathew, E., Sirko, A. and Tsui, P. (1992). The role of integration host factor In gene expression in Escherichia coli. *Molecular Microbiology* **6**(18): 2557-2563.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., . . . Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**(10): 1451-1455.

Goecks, J., Nekrutenko, A. and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86.

Gur, E., Biran, D., Gazit, E. and Ron, E. Z. (2002). In vivo aggregation of a single enzyme limits growth of Escherichia coli at elevated temperatures. *Mol Microbiol* **46**(5): 1391-1397.

Heaney, C. D., Sams, E., Dufour, A. P., Brenner, K. P., Haugland, R. A., Chern, E., . . . Wade, T. J. (2012). Fecal indicators in sand, sand contact, and risk of enteric illness among beachgoers. *Epidemiology* **23**(1): 95-106.

Hengge-Aronis, R., Klein, W., Lange, R., Rimmele, M. and Boos, W. (1991). Trehalose synthesis genes are controlled by the putative sigma factor encoded by rpoS and are involved in stationary-phase thermotolerance in Escherichia coli. *J Bacteriol* **173**(24): 7918-7924.

Holbrook, E. L., Greene, R. C. and Krueger, J. H. (1990). Purification and properties of cystathionine gamma-synthase from overproducing strains of Escherichia coli. *Biochemistry* **29**(2): 435-442.

Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol* **54**: 499-518.

Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N., . . . Venter, J. C. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A* **112**(45): 14024-14029.

Jung, J. U., Gutierrez, C., Martin, F., Ardourel, M. and Villarejo, M. (1990). Transcription of osmB, a gene encoding an Escherichia coli lipoprotein, is regulated by dual signals. Osmotic stress and stationary phase. *J Biol Chem* **265**(18): 10574-10581.

Kabir, M. S., Sagara, T., Oshima, T., Kawagoe, Y., Mori, H., Tsunedomi, R. and Yamada, M. (2004). Effects of mutations in the rpoS gene on cell viability and global gene expression under nitrogen starvation in Escherichia coli. *Microbiology* **150**(Pt 8): 2543-2553.

Kapanidis, A. N., Margeat, E., Laurence, T. A., Doose, S., Ho, S. O., Mukhopadhyay, J., . . . Weiss, S. (2005). Retention of transcription initiation factor sigma70 in transcription elongation: single-molecule analysis. *Mol Cell* **20**(3): 347-356.

Kim, Y., Aw, T. G., Teal, T. K. and Rose, J. B. (2015). Metagenomic Investigation of Viral Communities in Ballast Water. *Environ Sci Technol* **49**(14): 8396-8407.

Kim, Y., Wang, X., Ma, Q., Zhang, X. S. and Wood, T. K. (2009). Toxin-antitoxin systems in Escherichia coli influence biofilm formation through YjgK (TabA) and fimbriae. *J Bacteriol* **191**(4): 1258-1267.

King, T., Ishihama, A., Kori, A. and Ferenci, T. (2004). A regulatory trade-off as a source of strain variation in the species Escherichia coli. *J Bacteriol* **186**(17): 5614-5620.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glockner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**(1): e1.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., . . . Gilbert, J. A. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nat Biotech* **30**(6): 513-520.

Kocharunchitt, C., King, T., Gobius, K., Bowman, J. P. and Ross, T. (2014). Global Genome Response of <italic>Escherichia coli</italic> O157:H7 Sakai during Dynamic Changes in Growth Kinetics Induced by an Abrupt Downshift in Water Activity. *PLoS ONE* **9**(3): e90422.

Kolbert, C. P. and Persing, D. H. (1999). Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr Opin Microbiol* **2**(3): 299-305.

Korea, C. G., Badouraly, R., Prevost, M. C., Ghigo, J. M. and Beloin, C. (2010). Escherichia coli K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. *Environ Microbiol* **12**(7): 1957-1977.

Lange, R. and Hengge-Aronis, R. (1991). Identification of a central regulator of stationary-phase gene expression in Escherichia coli. *Mol Microbiol* **5**(1): 49-59.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**(4): 357-359.

Lederberg, J., and Tatum, E.L. (1946). Gene Recombination in *Escherichia coli Nature* **158**: 558.

Lee, H., Popodi, E., Tang, H. and Foster, P. L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A* **109**(41): E2774-2783.

Lennox, E. S. (1955). Transduction of linked genetic characters of the host by bacteriophage P1. *Virology* **1**(2): 190-206.

Lobner-Olesen, A. and Marinus, M. G. (1992). Identification of the gene (aroK) encoding shikimic acid kinase I of Escherichia coli. *J Bacteriol* **174**(2): 525-529.

Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**(12): 8228-8235.

Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T. and Gottesman, S. (1998). DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proceedings of the National Academy of Sciences* **95**(21): 12462-12467.

Majdalani, N., Hernandez, D. and Gottesman, S. (2002). Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol* **46**(3): 813-826.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., . . . Edwards, R. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Mika, F., and Hengge, R. (2005). A two-component phophotransfer network involving ArcB, ArcA, and RssB coordinates synthesis and proteolysis of sigmaS (RpoS) in E. *coli*. *Genes Dev.* **19**(22): 2770-2781.

Miller, J. H. (1992). A short course in bacterial genetics: a laboratory manual and handbook for Escherichia coli and related bacteria. NY, Cold Spring Harbor.

Mohiuddin, M. and Schellhorn, H. E. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**.

Nanninga, N. (1998). Morphogenesis of Escherichia coli. *Microbiol Mol Biol Rev* **62**(1): 110-129.

Notley-McRobb, L., King, T. and Ferenci, T. (2002). rpoS mutations and loss of general stress resistance in Escherichia coli populations as a consequence of conflict between competing stress responses. *J Bacteriol* **184**(3): 806-811.

Pace, N. R., Sapp, J. and Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc Natl Acad Sci U S A* **109**(4): 1011-1018.

Patten, C. L., Kirchhof, M. G., Schertzberg, M. R., Morton, R. A. and Schellhorn, H. E. (2004). Microarray analysis of RpoS-mediated gene expression in Escherichia coli K-12. *Mol Genet Genomics* **272**(5): 580-591.

Peng, Y., Soper, T. J. and Woodson, S. A. (2013). Positional Effects of AAN Motifs in rpoS Regulation by sRNAs and Hfq. *J Mol Biol*.

Pevzner, P. (2000). Mapping Assembly. Computational Molecular Biology: An Algorithmic Approach, MIT Press**:** 314.

Phadtare, S. and Inouye, M. (2001). Role of CspC and CspE in regulation of expression of RpoS and UspA, the stress response proteins in Escherichia coli. *J Bacteriol* **183**(4): 1205-1214.

PHO (2013). Public Health Inspector's guide to the principles and practices of environmental microbilogy.

Pike, K. (2008). Towards Safe Harbours 2008 - Hamilton Harbour Beaches. Hamilton, BARC Monitoring Committee.

Putnam, D. F. (1971). Composition and concentrative properties of human urine, National Aeronautics and Space Administration; [distributed by National Technical Information Service, Springfield, Va.].

Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J. and Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol Microbiol* **50**(2): 563-575.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., . . . Ravel, J. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* **190**(20): 6881-6893.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotech* **29**(1): 24-26.

Rockabrand, D., Livers, K., Austin, T., Kaiser, R., Jensen, D., Burgess, R. and Blum, P. (1998). Roles of DnaK and RpoS in starvation-induced thermotolerance of Escherichia coli. *J Bacteriol* **180**(4): 846-854.

Ron, E. Z., Alajem, S., Biran, D. and Grossman, N. (1990). Adaptation of Escherichia coli to elevated temperatures: the metA gene product is a heat shock protein. *Antonie Van Leeuwenhoek* **58**(3): 169-174.

Rothberg, J. M. and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nat Biotechnol* **26**(10): 1117-1124.

Sammartano, L. J., Tuveson, R. W. and Davenport, R. (1986). Control of sensitivity to inactivation by H2O2 and broad-spectrum near-UV radiation by the Escherichia coli katF locus. *J Bacteriol* **168**(1): 13-21.

Schellhorn, H. E., Audia, J. P., Wei, L. I. and Chang, L. (1998). Identification of conserved, RpoS-dependent stationary-phase genes of Escherichia coli. *J Bacteriol* **180**(23): 6283-6291.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**(23): 7537-7541.

Sledjeski, D. D. and Gottesman, S. (1996). Osmotic shock induction of capsule synthesis in Escherichia coli K-12. *J Bacteriol* **178**(4): 1204-1206.

Small, P., Blankenhorn, D., Welty, D., Zinser, E. and Slonczewski, J. L. (1994). Acid and base resistance in Escherichia coli and Shigella flexneri: role of rpoS and growth pH. *J Bacteriol* **176**(6): 1729-1737.

Sun, C., Zhao, Y., Li, H., Dong, Y., MacIsaac, H. J. and Zhan, A. (2015). Unreliable quantitation of species abundance based on high-throughput sequencing data of zooplankton communities. *Aquatic Biology* **24**(1): 9-15.

Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* **8**(3): 207-217.

Theilacker, C., Sanchez-Carballo, P., Toma, I., Fabretti, F., Sava, I., Kropec, A., . . . Huebner, J. (2009). Glycolipids are involved in biofilm accumulation and prolonged bacteraemia in Enterococcus faecalis. *Molecular Microbiology* **71**(4): 1055-1069.

Theron, J. and Cloete, T. E. (2000). Molecular techniques for determining microbial diversity and community structure in natural environments. *Crit Rev Microbiol* **26**(1): 37-57.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**(2): 178-192.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., . . . Denamur, E. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5**(1): e1000344.

Tramonti, A., Visca, P., De Canio, M., Falconi, M. and De Biase, D. (2002). Functional characterization and regulation of gadX, a gene encoding an AraC/XylS-like transcriptional activator of the Escherichia coli glutamic acid decarboxylase system. *J Bacteriol* **184**(10): 2603-2613.

USEPA (2005). State of the Great Lakes. *State of the Great Lakes Ecosystem Conference*.

Voorhies, A. A., Biddanda, B. A., Kendall, S. T., Jain, S., Marcus, D. N., Nold, S. C., . . . Dick, G. J. (2012). Cyanobacterial life at low O(2): community genomics and function reveal metabolic versatility and extremely low diversity in a Great Lakes sinkhole mat. *Geobiology* **10**(3): 250-267.

Wang, K., Liu, E., Song, S., Wang, X., Zhu, Y., Ye, J. and Zhang, H. (2012). Characterization of Edwardsiella tarda rpoN: roles in sigma(70) family regulation, growth, stress adaption and virulence toward fish. *Arch Microbiol* **194**(6): 493-504.

Wang, L., Huskic, S., Cisterne, A., Rothemund, D. and Reeves, P. R. (2002). The O-Antigen Gene Cluster of Escherichia coli O55:H7 and Identification of a New UDP-GlcNAc C4 Epimerase Gene. *J Bacteriol* **184**(10): 2620-2625.

Yaron, A. (1976). [50] Dipeptidyl carboxypeptidase from Escherichia coli. Methods in Enzymology, Academic Press. **Volume 45:** 599-610.

Yim, H. H. and Villarejo, M. (1992). osmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in Escherichia coli. *J Bacteriol* **174**(11): 3637-3644.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., . . . Yang, Y. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**(8): 1303-1313.

Zhou, Y., Gottesman, S., Hoskins, J. R., Maurizi, M. R. and Wickner, S. (2001). The RssB response regulator directly targets sigma(S) for degradation by ClpXP. *Genes Dev* **15**(5): 627-637.

APPENDIX A: SUPPLEMENTARY FILES, TABLES, AND FIGURES

A.1 Chapter 1 – supplementary information

File S1. readme.txt.

```
Setting up Galaxy:

Run "setup.sh".

        $ sh setup.sh


Running Galaxy:

Run "start.sh".

        $ sh start.sh


Logging in:

User -> login

        email address: admin@galaxy.com
        password: password


Contig alignment - lastZ:

Fasta file uploaded for use as reference must contain "gi" as sequence identifier.
For example:

>gi
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAA
AAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAG
GTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACAT
CCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGC
GTACAGGAAACACAGAAAAAAG


NCBI BLAST:

Run "fetch_taxonomy.sh".

```

```
$ sh fetch_taxonomy.sh
```

RNA analysis - Tophat for Illumina and Cufflinks:

Both reference sequence (FASTA format) and reference annotation (GFF format) used in alignment must have the same identifying information. For convenience, the reference sequence identifer can be edited to match the header (first column) of the reference annotation. For example:

Reference sequence:

```
>NC_000913.2
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAA
AAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAG
GTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACAT
CCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGC
GTACAGGAAACACAGAAAAAAG
```

Reference annotation:

```
##gff-version 3
#!gff-spec-version 1.20
#!processor NCBI annotwriter
##sequence-region NC_000913.2 1 4639675
##species
http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145
NC_000913.2 RefSeqregion 1 4639675 . + .
        ID=id0;Dbxref=taxon:511145;Is_circular=true;gbkey=Src;genome=chrom
osome;mol_type=genomic DNA;old-name=Escherichia coli K12;strain=K-
12;substrain=MG1655
NC_000913.2 RefSeqgene 190 255 . + .
        ID=gene0;Name=thrL;Dbxref=EcoGene:EG11277,GeneID:944742;gbkey
=Gene;gene=thrL;gene_synonym=ECK0001,JW4367;locus_tag=b0001
NC_000913.2 RefSeqCDS 190 255 . + 0
        ID=cds0;Name=NP_414542.1;Parent=gene0;Dbxref=ASAP:ABE-
0000006,UniProtKB%2FSwiss-
Prot:P0AD86,Genbank:NP_414542.1,EcoGene:EG11277,GeneID:944742;gbke
```

=CDS;product=thr operon leader peptide;protein_id=NP_414542.1;transl_table=11

**If using Cuffcompare, duplicate gene entries (same ID) are NOT permitted within the reference annotation (GFF) file. These must be manually edited prior to Cufflinks analysis. Duplicate gene entries can be identified using the "GFF Duplicate Identification" workflow in Galaxy.

**If using the 'E. coli MG1655' Transcript Assembly and Comparison workflows, the Database/Build of Tophat accepted hits output files must be set to 'E. coli (NC_000913.2) (E_coli_MG1655_NC_000913.2)' in Attributes.

SNP/Indel Calling:

The SnpEff program may output several errors during calling (e.g. ERROR_OUT_OF_CHROMOSOME_RANGE; WARNING_TRANSCRIPT_NO_START_CODON). These can be ignored as data may still be accessed.

Visualization:

Files submitted for visualization must have Database/Build set for the associated reference sequence. In History:

Edit attributes -> Database/Build -> Select reference -> Save

**For viewing contig alignment files against 'E. coli MG1655', use 'E. coli MG1655 (gi)'.
**For viewing RNA transcript alignment files against 'E. coli MG1655', use 'E. coli MG1655 (NC_000913.2)'.

File S2. setup.sh.

```
#!/bin/sh

user=$(whoami)
sudo chown -R $user *
chgrp -R $user *

cd /usr/local/bin
sudo rm velveth velvetg
```

```
cd -
cd shed_tools/toolshed.g2.bx.psu.edu/repos/edward-kirton/velvet_toolsuite/4afe13ac23b6/velvet_toolsuite/velvet
sudo cp velveth velvetg /usr/local/bin

cd -
cd visualization
sudo cp bedGraphToBigWig wigToBigWig bedtools /usr/local/bin
exit
```

File S3. start.sh.

```
#!/bin/sh

sudo chmod 777 galaxy-dist/tool-data/shared/jars/FastQC/fastqc /usr/local/bin/bedGraphToBigWig /usr/local/bin/wigToBigWig

cd galaxy-dist
sudo stop galaxy
sudo sh run.sh &
sleep 25 ; chromium-browser http://localhost:8080/

exit
```

File S4. fetch_taxonomy.sh.

```
#!/bin/sh

cd galaxy-dist/scripts/taxonomy/
sudo sh processTaxonomy.sh
user=$(whoami)
sudo chown -R $user *
chgrp -R $user *
cp -r * ../../tool-data/taxonomy/

exit
```

File S5. pangenome_annotation.pl.

```
#!/usr/bin/perl

open($ffn1, "<", "@ARGV[0]") ;
```

```
open($output, ">>", "output.ffn") ;
$found1 = 0 ;
$found2 = 0 ;
while (<$ffn1>) {
        if ($_ =~ m/>(.*)\s\|/) {
                $id = $1 ;
                $found1 = 1 ;
                }
        if($found1 == 1) {
                open($tsv, "<", "@ARGV[1]") ;
                while (<$tsv>) {
                        @tsv = split(/\t/) ;
                        if (@tsv[3] =~ $id) {
                                $gene = @tsv[5] ;
                                open($ffn2, "<", "@ARGV[0]") ;
                                while (<$ffn2>) {
                                        if ($_ =~ /$id\s\|/) {
                                                s/$id/$gene/ ;
                                                print $output $_ ;
                                                $found2 = 1 ;
                                                next ;
                                                }
                                        if ($found2 == 1) {
                                                if ($_ =~ m/>(.*)\s\|/ || eof) {
                                                        $found1 = 0 ;
                                                        $found2 = 0 ;
                                                        }
                                                else{
                                                        print $output $_ ;
                                                        }
                                                }
                                        }
                                close $ffn2 ;
                                }
                        }
                }
                close $tsv ;
close $output ;
close $ffn1 ;
```

File S6. second-site_annotation.pl.

```
#!/usr/bin/perl
```

```
$in1 = @ARGV[0];
`mkdir ../output`;
my @reads1 = `ls`;
foreach (@reads1){
	chomp;
	if($_ =~ m/^.*variants.txt$/){
		$in2 = $_;
		open($output, ">>", "../output/$in2\_output.txt");
		open($variants, "<", "$in2");
		$found1 = 0;
		while(<$variants>){
			$_ =~ s/\r\n//;
			if($found1 == 0){
				$header1 = $_;
				open($conflicts, "<", "$in1");
				while(<$conflicts>){
					$header2 = $_;
					print $output "$header1\t$header2";
					last;
					}
				close $conflicts;
				$found1 = 1;
				}
			else{
				$found2 = 0;
				$line1 = $_;
				@line1 = split("\t", $_);
				$ref = @line1[0];
				open($conflicts, "<", "$in1");
				while(<$conflicts>){
					$line2 = $_;
					@line2 = split("\t", $_);
					if($ref eq @line2[0]){
						print $output "$line1\t$line2";
						$found2 = 1;
						}
					}
				close $conflicts;
				if($found2 == 0){
					print $output "$line1\tno conflict\n";
					}
				}
			}
```

```
45      close $variants;
46                              close $output;
47                              }
48              }
```

A.2 Chapter 2 – supplementary information

File S7. readme.txt.

```
Setting up script in QIIME virtual box:

        cd workflow/
        sh fastx-toolkit_greengenes_install.sh


Setting up read files:

Place zipped FASTQ read files in '16S_data' directory.


Setting up mapping.txt file:

Edit mapping file with appropriate sample metadata.

The 'mapping.txt' file in the 'QIIME' directory contains sample ID's and supplementary information in tab-delimited format. Please note:

- SampleID's must correspond to the sample name for each FASTQ file (e.g. '4' for the '4_S4_L001_R1_001.fastq' forward and '4_S4_L001_R2_001.fastq' reverse read files)
- BarcodeSequence and LinkerPrimerSequence categories can be left blank (3 tab's from SampleID to InputFileName)
- InputFileName is included in the format 'SampleID'_combined.fna (e.g. '4_combined.fna')
- Category headers must not contain any special characters (e.g. '/')
- Any metadata categories that are chosen for analysis must not contain white spaces (e.g. use 'Lake_Ontario' instead of 'Lake Ontario')

See 'mapping.txt' in QIIME/ for sample categories.


Analyzing samples with QIIME:

        cd QIIME/
        qiime
        perl ../qiime_workflow.pl

You will be required to input the following information:

- The number of threads (CPU cores) available for processing (e.g. '4')
```

- The number of reads to sample from each file (enter 'ALL' for all reads) (e.g. '100000')
- A comma-delimited list of metadata categories on which to conduct core diversity analyses (e.g. 'Location,Subtype,Rain')
- A comma-delimited list of metadata categories on which to conduct subset analyses (e.g. 'Location,Subtype,Rain')
- Towards the end of the workflow, an **even** value for sampling depth (e.g. '1000'). You will need to determine this value from the file 'otus/table_summary.txt' for diversity analyses. Many of the analyses that follow require that there are an equal number of sequences in each sample, so you need to review the Counts/sample detail and decide what depth you'd like. Any samples that don't have at least that many sequences will not be included in the analyses, so this is always a trade-off between the number of sequences you throw away and the number of samples you throw away. For some perspective on this, see Kuczynski 2010.

Viewing QIIME output:

- HTML output can be viewed by accessing the .html file within the 'core_diversity_analyses' directory
- Taxonomy tables should be opened with a spreadsheet processor (e.g. Microsoft Excel)
- Network files within the "cytoscape_bipartite_network' directory can be visualized with the program Cytoscape (downloaded separately)
- A summary of the workflow can be accessed with 'QIIME/log.txt'

File S8. fastx-toolkit_greengenes_install.sh.

```
#!/bin/sh

sudo apt-get --assume-yes install fastx-toolkit

wget
ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_8_otus.tar.gz;
tar -xzvf gg_13_8_otus.tar.gz;
\rm -r gg_13_8_otus.tar.gz;
```

File S9. qiime_workflow.pl.

```
#!/usr/bin/perl

##Request user input
```

```
print "Please consult system specifications and enter the number of threads available for processing:\n";
$threads = <STDIN>;
chomp $threads;

print "Please enter the number of reads to sample from each file:\n";
$fastq_depth = <STDIN>;
chomp $fastq_depth;
$fastq_lines = 4*$fastq_depth;

print "Please enter the mapping.txt metadata categories to analyze in comma-delimited format (e.g. Location,Subtype,Rain):\n";
$metadata = <STDIN>;
chomp $metadata;

print "Please enter the mapping.txt metadata categories to use for subset analyses in comma-delimited format (e.g. Location,Subtype,Rain):\n";
$subset1 = <STDIN>;
chomp $subset1;

print "Executing analyses...\n";

##Create log file

open($log, ">", "log.txt");
select $log;

##Trim read files by quality; convert to fasta; reverse-compliment reverse reads; combine forward and reverse reads for each sample

print scalar localtime();
print " - Filtering read files by quality; formatting read files for QIIME...\n";

`mkdir reads`;
chdir "../16S_data";

my @folder = `ls`;
foreach (@folder){
	chomp;
	$folder = $_;
	chdir "$folder";
	my @zip = `ls`;
	foreach (@zip){
```

```
47
48                      chomp;
49                      if($_ =~ m/.*\.fastq\.gz$/){
50                              `mv $_ ../$_`;
51                              `gunzip ../$_`;
52                              }
53                      }
54              chdir "..";
55              `rm -r $folder`;
56              }
57
58      my @reads = `ls`;
59      foreach (@reads){
60              chomp;
61              ($sample1 = $_) =~ s/^(.+)\_.+\_.+\_.+\_.+$/$1/;
62              ($sample2 = $sample1) =~ s/\_/\./g;
63              ($reads_base = $_) =~ s/(.)\.[^.]+$/$1/;
64              chdir "../QIIME/reads";
65              if($fastq_depth eq "ALL"){
66                      `fastq_quality_trimmer      -t      25      -i      ../../16S_data/$_      -o
67      $reads_base\_trimmed\_final.fastq`;
68                      }
69              else{
70                      `fastq_quality_trimmer      -t      25      -i      ../../16S_data/$_      -o
71      $reads_base\_trimmed\_1.fastq`;
72                      open($fastq, "<", "$reads_base\_trimmed\_1.fastq");
73                      $fastq_count = () = <$fastq>;
74                      if($fastq_count > $fastq_lines){
75                              open($fastq, "<", "$reads_base\_trimmed\_1.fastq");
76                              open($fastq_short,                                              ">>",
77      "$reads_base\_trimmed\_final.fastq");
78                              $count = 0;
79                              while(<$fastq>){
80                                      unless($count == $fastq_lines){
81                                              print $fastq_short $_;
82                                              $count++;
83                                              }
84                                      }
85                              close $fastq;
86                              close $fastq_short;
87                              }
88                      else{
89                              print "\t $_ - Insufficient data\n";
```

```
                `cp $reads_base\_trimmed\_1.fastq $reads_base\_trimmed\_final.fastq`;
                }
        }
    `convert_fastaqual_fastq.py -f $reads_base\_trimmed\_final.fastq -F -c fastq_to_fastaqual`;
    if($reads_base =~ m/.2\_001+$/){
        `adjust_seq_orientation.py -i $reads_base\_trimmed\_final.fna`;
        $reads = "$reads_base\_trimmed\_final\_rc.fna";
        if(-e "combined/$sample2\_combined.fna") {
            `cat $reads >> combined/$sample2\_combined.fna`;
            }
        else{
            `mkdir -p combined`;
            `cp $reads combined/$sample2\_combined.fna`;
            }
        }
    else{
        $reads = "$reads_base\_trimmed\_final.fna";
        if(-e "combined/$sample2\_combined.fna") {
            `cat $reads >> combined/$sample2\_combined.fna`;
            }
        else{
            `mkdir -p combined`;
            `cp $reads combined/$sample2\_combined.fna`;
            }
        }
    chdir "../../16S_data";
    }

print scalar localtime();
print " - Read files formatted\n";

##Add qiime labels to fasta files; pick closed reference OTUs; generate BIOM summary table

print scalar localtime();
print " - Picking closed reference OTUs...\n";

chdir "../QIIME/reads";
`add_qiime_labels.py -m ../mapping.txt -i combined -c InputFileName -n 1`;
chdir "..";
```

```
`pick_closed_reference_otus.py -a -O $threads -i reads/combined_seqs.fna -r
../gg_13_8_otus/rep_set/97_otus.fasta -t
../gg_13_8_otus/taxonomy/97_otu_taxonomy.txt -o otus`;
`biom summarize-table -i otus/otu_table.biom -o otus/table_summary.txt`;

print scalar localtime();
print " - OTUs picked\n";

#The key piece of information you need to pull from this output is the depth of
sequencing that should be used in diversity analyses. Many of the analyses that
follow require that there are an equal number of sequences in each sample, so you
need to review the Counts/sample detail and decide what depth you'd like. Any
samples that don't have at least that many sequences will not be included in the
analyses, so this is always a trade-off between the number of sequences you throw
away and the number of samples you throw away. For some perspective on this,
see Kuczynski 2010.

select STDOUT;
print "Please consult otus/table_summary.txt and enter value for sampling depth
(greatest even value which excludes minimal samples):\n" ;
$depth2 = <STDIN>;
chomp $depth2;

##Conduct core diversity analyses

select $log;
print scalar localtime();
print " - Conducting core diversity analyses...\n";

`core_diversity_analyses.py -a -O $threads -o core_diversity_analyses -i
otus/otu_table.biom -m mapping.txt -e $depth2 -t
../gg_13_8_otus/trees/97_otus.tree -c $metadata -p parameters.txt`;

print scalar localtime();
print " - Analyses completed\n";

print scalar localtime();
print " - Conducting supplementary diversity analyses...\n";

##Estimate observation richness

`estimate_observation_richness.py -i otus/otu_table.biom -o
observation_richness_estimates`;
```

```
##Make rank-abundance graph

`mkdir rank_abundance_graph`;
chdir "rank_abundance_graph";
`plot_rank_abundance_graph.py -i ../otus/otu_table.biom -s "*" -o graph`;
chdir "..";

##Build weighted UPGMA-cluster tree (for otu heatmap)

`upgma_cluster.py -i
core_diversity_analyses/bdiv_even$depth2/weighted_unifrac_dm.txt -o
weighted_UPGMA_cluster_tree`;

##Make otu heatmap

`mkdir heatmap`;
chdir "heatmap";
`make_otu_heatmap.py -i ../otus/otu_table.biom -o heatmap --sample_tree
../weighted_UPGMA_cluster_tree`;
chdir "..";
`mv weighted_UPGMA_cluster_tree heatmap/weighted_UPGMA_cluster_tree`;

##Make bipartite network (for Cytoscape network mapping)

`make_bipartite_network.py -i otus/otu_table.biom -m mapping.txt -o
cytoscape_bipartite_network -k taxonomy --md_fields 'k,p,c,o,f,g,s'`;

##Make taxonomy tables

`mkdir taxonomy_tables`;
`mkdir taxonomy_tables/data`;
chdir "taxonomy_tables/data";

#Summarize sample taxonomy

mkdir "taxa_plots";
`sort_otu_table.py -i ../../otus/otu_table.biom -o taxa_plots/otu_table_sorted.biom`;
`summarize_taxa.py -i taxa_plots/otu_table_sorted.biom -o taxa_plots/ --level
1,2,3,4,5,6,7`;

#Summarize category taxonomies

@category = split(",", $metadata);
```

```
218                     foreach(@category){
219                             mkdir "taxa_plots_$_";
220                             `collapse_samples.py -b table_mc$depth4.biom -m
221     ../../mapping.txt --output_biom_fp taxa_plots_$_/$_\_otu_table.biom --
222     output_mapping_fp taxa_plots_$_/$_\_mapping.txt --collapse_fields
223     BarcodeSequence,$_`;
224                             `sort_otu_table.py -i taxa_plots_$_/$_\_otu_table.biom -o
225     taxa_plots_$_/$_\_otu_table_sorted.biom`;
226                             `summarize_taxa.py -i
227     taxa_plots_$_/$_\_otu_table_sorted.biom -o taxa_plots_$_/ --level 1,2,3,4,5,6,7`;
228                             }
229
230     #Rename taxonomy tables
231
232     my @diversity = `ls`;
233     foreach(@diversity){
234             chomp;
235             if($_ =~ m/taxa\_plots.*$/){
236                     chdir "$_";
237                     if($_ =~ m/taxa\_plots.+$/){
238                             ($parameter = $_) =~ s/taxa\_plots\_(.+)$/$1/;
239                             chdir "../..";
240                             `mkdir $parameter`;
241                             chdir "data/$_";
242                             my @tables = `ls`;
243                             foreach(@tables){
244                                     chomp;
245                                     if($_ =~ m/.*sorted\_L1\.txt/){
246                                             `cp $_ ../../$parameter/"Kingdom.txt"`;
247
248                                             }
249                                     if($_ =~ m/.*sorted\_L2\.txt/){
250                                             `cp $_ ../../$parameter/"Phylum.txt"`;
251
252                                             }
253                                     if($_ =~ m/.*sorted\_L3\.txt/){
254                                             `cp $_ ../../$parameter/"Class.txt"`;
255
256                                             }
257                                     if($_ =~ m/.*sorted\_L4\.txt/){
258                                             `cp $_ ../../$parameter/"Order.txt"`;
259
260                                             }
```

```
261                             if($_ =~ m/.*sorted\_L5\.txt/){
262                                     `cp $_ ../../$parameter/"Family.txt"`;
263
264                                     }
265                             if($_ =~ m/.*sorted\_L6\.txt/){
266                                     `cp $_ ../../$parameter/"Genus.txt"`;
267
268                                     }
269                             if($_ =~ m/.*sorted\_L7\.txt/){
270                                     `cp $_ ../../$parameter/"Species.txt"`;
271
272                                     }
273                             }
274                     chdir "..";
275                     }
276             else{
277                     $parameter = "Sample";
278                     chdir "../..";
279                     `mkdir $parameter`;
280                     chdir "data/$_";
281                     my @tables = `ls`;
282                     foreach(@tables){
283                             chomp;
284                             if($_ =~ m/.*sorted\_L1\.txt/){
285                                     `cp $_ ../../$parameter/"Kingdom.txt"`;
286
287                                     }
288                             if($_ =~ m/.*sorted\_L2\.txt/){
289                                     `cp $_ ../../$parameter/"Phylum.txt"`;
290
291                                     }
292                             if($_ =~ m/.*sorted\_L3\.txt/){
293                                     `cp $_ ../../$parameter/"Class.txt"`;
294
295                                     }
296                             if($_ =~ m/.*sorted\_L4\.txt/){
297                                     `cp $_ ../../$parameter/"Order.txt"`;
298
299                                     }
300                             if($_ =~ m/.*sorted\_L5\.txt/){
301                                     `cp $_ ../../$parameter/"Family.txt"`;
302
303                                     }
```

```
304                                     if($_ =~ m/.*sorted\_L6\.txt/){
305                                             `cp $_ ../../$parameter/"Genus.txt"`;
306
307                                             }
308                                     if($_ =~ m/.*sorted\_L7\.txt/){
309                                             `cp $_ ../../$parameter/"Species.txt"`;
310
311                                             }
312                                     }
313                             chdir "..";
314                             }
315                     }
316             }
317     chdir "../..";
318
319     print scalar localtime();
320     print " - Analyses completed\n";
321
322     ##Conduct sample subset analyses
323
324     print scalar localtime();
325     print " - Conducting sample subset analyses...\n";
326
327     `mkdir subset_analyses`;
328     chdir "subset_analyses";
329     @subset1 = split(",", $subset1);
330
331     #Identify unique values within each metadata category (e.g. "LONT" & "LERIE"
332     within "Location")
333
334     foreach(@subset1){
335             chomp;
336             $found1 = 0;
337             $count1a = 0;
338             $subset2 = $_;
339             `mkdir $subset2`;
340             chdir "$subset2";
341             open($mapping, "<", "../../mapping.txt");
342             while(<$mapping>){
343                     @line = split("\t", $_);
344                     if($found1 == 1) {
345                                     push (@subset3, @line[$count1b]);
346                                     }
```

```
347                     else{
348                             foreach(@line){
349                                     chomp;
350                                     $count1a++;
351                                     if($_ =~ $subset2){
352                                             $found1 = 1;
353                                             $count1b = $count1a - 1;
354                                             last;
355                                             }
356                                     }
357                             }
358                     }
359             close $mapping;
360             my %seen;
361             my @unique = grep { not $seen{$_} ++ } @subset3;
362             undef @subset3;
363
364             foreach(@unique){
365                     chomp;
366                     $unique = $_;
367                     `mkdir $unique`;
368                     chdir "$unique";
369                     `cp ../../../parameters.txt .`;
370
371     #Select samples of interest from OTU table for each unique metadata value (e.g.
372     samples with location "LONT")
373
374                     `mkdir otus`;
375                     `filter_samples_from_otu_table.py -i ../../../otus/otu_table.biom -o
376     otus/otu_table.biom -m ../../../mapping.txt --output_mapping_fp mapping.txt -s
377     $subset2\:$unique -n $depth2`;
378
379     #Remove metadata category of interest from $metadata (e.g. remove "Location"
380     from category analysis for location subsets)
381
382                     @metadata2 = split(",", $metadata);
383                     foreach(@metadata2){
384                             chomp;
385                             unless($_ eq $subset2){
386                                     push(@metadata3,$_);
387                                     }
388                             }
389
```

```
#Remove metadata categories if not present in filtered mapping file (i.e. QIIME will filter all categories containing identical values for each sample from the mapping file; these will be excluded from subset analysis)

		open($mapping, "<", "mapping.txt");
			while(<$mapping>){
				@line2 = split("\t", $_);
				last;
				}
			close $mapping;
			foreach(@metadata3){
				chomp;
				$metadata3 = $_;
				if(grep{$_ eq $metadata3} @line2){
					push(@metadata4,$_);
					}
				}
		$metadata4 = join(',',@metadata4);
		undef @metadata3;
		undef @metadata4;

#Suppress beta diversity analyses for subsets with 3 or fewer samples

		$count2a = 0;
		open($mapping, "<", "mapping.txt");
			while(<$mapping>){
				$count2a++;
				}
		close $mapping;
		$count2b = $count2a - 1;
		if($count2b <= 3){
			$suppress = "--suppress_beta_diversity";
			}
		else{
			undef $suppress;
			}

#Conduct analyses

		if(!$metadata4){
			`core_diversity_analyses.py -a -O $threads -o core_diversity_analyses -i otus/otu_table.biom -m mapping.txt -e $depth2 -t ../../../../gg_13_8_otus/trees/97_otus.tree -p parameters.txt $suppress`;
```

```
433                         }
434                 else{
435                         `core_diversity_analyses.py -a -O $threads -o
436 core_diversity_analyses -i otus/otu_table.biom -m mapping.txt -e $depth2 -t
437 ../../../../gg_13_8_otus/trees/97_otus.tree -c $metadata4 -p parameters.txt
438 $suppress`;
439                         }
440                 #`estimate_observation_richness.py -i otus/otu_table.biom -o
441 observation_richness_estimates`;
442                 `mkdir rank_abundance_graph`;
443                 chdir "rank_abundance_graph";
444                 `plot_rank_abundance_graph.py -i ../otus/otu_table.biom -s "*" -o
445 graph`;
446                 chdir "..";
447                 unless($count2b <= 3){
448                         `upgma_cluster.py -i
449 core_diversity_analyses/bdiv_even$depth2/weighted_unifrac_dm.txt -o
450 weighted_UPGMA_cluster_tree`;
451                         `mkdir heatmap`;
452                         chdir "heatmap";
453                         `make_otu_heatmap.py -i ../otus/otu_table.biom -o heatmap
454 --sample_tree ../weighted_UPGMA_cluster_tree`;
455                         chdir "..";
456                         `mv weighted_UPGMA_cluster_tree
457 heatmap/weighted_UPGMA_cluster_tree`;
458                         }
459                 `make_bipartite_network.py -i otus/otu_table.biom -m mapping.txt
460 -o cytoscape_bipartite_network -k taxonomy --md_fields 'k,p,c,o,f,g,s'`;
461                 `mkdir taxonomy_tables`;
462                 chdir "core_diversity_analyses";
463                 my @diversity = `ls`;
464                 foreach(@diversity){
465                         chomp;
466                         if($_ =~ m/taxa\_plots.*$/){
467                                 chdir "$_";
468                                 if($_ =~ m/taxa\_plots.+$/){
469                                         ($parameter = $_) =~
470 s/taxa\_plots\_(.+)$/$1/;
471                                         chdir "../../taxonomy_tables";
472                                         `mkdir $parameter`;
473                                         chdir "../core_diversity_analyses/$_";
474                                         my @tables = `ls`;
475                                         foreach(@tables){
```

```
476                                                     chomp;
477                                                     if($_ =~ m/.*sorted\_L1\.txt/){
478                                                             `cp                             $_
479     ../../taxonomy_tables/$parameter/"Kingdom.txt"`;
480                                                             }
481                                                     if($_ =~ m/.*sorted\_L2\.txt/){
482                                                             `cp                             $_
483     ../../taxonomy_tables/$parameter/"Phylum.txt"`;
484                                                             }
485                                                     if($_ =~ m/.*sorted\_L3\.txt/){
486                                                             `cp                             $_
487     ../../taxonomy_tables/$parameter/"Class.txt"`;
488                                                             }
489                                                     if($_ =~ m/.*sorted\_L4\.txt/){
490                                                             `cp                             $_
491     ../../taxonomy_tables/$parameter/"Order.txt"`;
492                                                             }
493                                                     if($_ =~ m/.*sorted\_L5\.txt/){
494                                                             `cp                             $_
495     ../../taxonomy_tables/$parameter/"Family.txt"`;
496                                                             }
497                                                     if($_ =~ m/.*sorted\_L6\.txt/){
498                                                             `cp                             $_
499     ../../taxonomy_tables/$parameter/"Genus.txt"`;
500                                                             }
501                                                     if($_ =~ m/.*sorted\_L7\.txt/){
502                                                             `cp                             $_
503     ../../taxonomy_tables/$parameter/"Species.txt"`;
504                                                             }
505                                                     }
506                                             chdir "..";
507                                             }
508                                     else{
509                                             $parameter = "Sample";
510                                             chdir "../../taxonomy_tables";
511                                             `mkdir $parameter`;
512                                             chdir "../core_diversity_analyses/$_";
513                                             my @tables = `ls`;
514                                             foreach(@tables){
515                                                     chomp;
516                                                     if($_ =~ m/.*sorted\_L1\.txt/){
517                                                             `cp                             $_
518     ../../taxonomy_tables/$parameter/"Kingdom.txt"`;
```

```
519                                                     }
520                                                 if($_ =~ m/.*sorted\_L2\.txt/){
521                                                     `cp                         $_
522     ../../taxonomy_tables/$parameter/"Phylum.txt"`;
523                                                     }
524                                                 if($_ =~ m/.*sorted\_L3\.txt/){
525                                                     `cp                         $_
526     ../../taxonomy_tables/$parameter/"Class.txt"`;
527                                                     }
528                                                 if($_ =~ m/.*sorted\_L4\.txt/){
529                                                     `cp                         $_
530     ../../taxonomy_tables/$parameter/"Order.txt"`;
531                                                     }
532                                                 if($_ =~ m/.*sorted\_L5\.txt/){
533                                                     `cp                         $_
534     ../../taxonomy_tables/$parameter/"Family.txt"`;
535                                                     }
536                                                 if($_ =~ m/.*sorted\_L6\.txt/){
537                                                     `cp                         $_
538     ../../taxonomy_tables/$parameter/"Genus.txt"`;
539                                                     }
540                                                 if($_ =~ m/.*sorted\_L7\.txt/)z
541                                                 }
542                                             chdir "..";
543                                             }
544                                         }
545                                     }
546                                 chdir "../..";
547                                 }
548                             chdir "..";
549                             }
550
551     print scalar localtime();
552     print " - Analyses completed\n";
553
554     close $log;
555
556     print "Analyses completed\n";
```

File S10. parameters.txt.

```
1    summarize_taxa:level 1,2,3,4,5
2    plot_taxa_summary:chart_type        area,bar,pie
```