Genomics as a Lens into the Population Structure and Evolutionary Dynamics of Freshwater Microbes

by

Sarah L. R. Stevens

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2019

Date of final oral examination: 01/16/2019

The dissertation is approved by the following members of the Final Oral Committee:

Katherine D. McMahon, Professor, Bacteriology
Cameron R. Currie, Professor, Bacteriology
Emily Stanley, Professor, Zoology
Garret Suen, Associate Professor, Bacteriology
Rex R. Malmstrom, Research Scientist, DOE Joint Genome Institute

ACKNOWLEDGMENTS

I would first like to thank my advisor Dr. Katherine McMahon. I can't imagine a better role model than Trina. She has been an amazing and supportive mentor throughout my PhD. She encouraged me to follow my interests in both my research and professional development. I'd also like to thank Dr. Rex Malmstrom who as been my secondary advisor for his support. We started working together on my first chapter and his influence over the rest of my work has been immeasurable. I'd also like to thank the rest of my committee for your support and feedback throughout.

I'm also very grateful for all the support of my friends and family. Thank you to my husband, Chad, your constant support kept me going and helped me believe I could do it. Even when I was super stressed you were there with food and hugs. Thank you to my parents, your emphasis on the importance of education got me started on this path early. Your continued support helped me make it this far. Thank you to my in-laws for your enthusiasm for this venture. A special thanks to Aunt Allison, who helped Madison feel like home. Thanks to Cristina, my shoulder to cry on and statistics guru. When I wore you down to being my friend, I didn't know the depth of friendship I would get in return. Thanks to Meng, who listened to my stress and encouraged me with your outside perspective. Thanks to Robin, who has been with me from my first days in the lab and is the best proofreader of anyone I know. Thanks to Elizabeth, first mentee then friend. Sometimes you don't realize how much you know until you teach it to someone else.

My final thank you goes to the Carpentries community. It has been a blessing of its own to be a part of such a wonderful group of people who are helping others.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Microorganisms are important players in the ongoing biogeochemical cycling of the environment. While studies of microbial genomes have begun to increase our understanding of these nutrient transformations, much remains to be learned about how the microbial populations performing these functions are changing through time in response to evolutionary forces. In the three studies presented in this thesis, I used single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) as references to recruit reads from time series metagenomes taken from two freshwater lakes. This allowed me to track the diversity and abundance of 'sequence-discrete' populations of microbes across the whole genome through three or more years.

In the first study, I observed how each of a phylogenetically diverse group of populations was changing in Trout Bog, a dystrophic lake. I observed a genome-wide sweep in the a population of Chlorobium over the course of the time series. I also found evidence that some of the other populations I tracked had experienced gene-specific sweeps prior to the start of the time series. This suggests that co-occurring populations may be controlled by different evolutionary forces.

In the second chapter, I studied populations of two very common and abundant freshwater taxa. I saw that these two taxa contained populations with very different structures. These structural differences may be explained by a recent diversification among the populations represented in one group. The abundance patterns of this same group suggest that there is still significant niche overlap between these populations.

In the final chapter, I investigated how the patterns of gene abundance and diversity underlie the whole population's abundance and diversity. I observed that all of the six *Polynucleobacter* populations recovered were fairly persistent through time, though one was considerably more abundant. None of the six populations recovered were dominated by a single strain though there is a trend between abundance and SNV homogeneity. Additionally, I characterized the core and accessory genome based using metagenomics through time and differences in evolutionary signatures between these genes. This work provides a framework for breaking down across the levels of biological organization.

1 GENOME-WIDE SELECTIVE SWEEPS AND GENE-SPECIFIC SWEEPS IN NATURAL BACTERIAL POPULATIONS

Matthew L. Bendall*[1], Sarah L.R. Stevens*[2], Leong-Keat Chan[1], Stephanie Malfatti[1], Patrick Schwientek[1], Julien Tremblay[1], Wendy Schackwitz[1], Joel Martin[1], Amrita Pati[1], Brian Bushnell[1], Jeff Froula[1], Dongwan Kang[1], Susannah G. Tringe[1], Stefan Bertilsson[3], Mary A. Moran[4], Ashley Shade[5], Ryan J. Newton[6], Katherine D. McMahon[2,7] and Rex R. Malmstrom[1]

[1]*DOE Joint Genome Institute, Walnut Creek*; [2]*Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA*; [3]*Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden*; [4]*Department of Marine Sciences, University of Georgia, Athens, GA, USA*; [5]*Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA*; [6]*School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA*; [7]*Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA*

* These authors contributed equally to this work.

MLB, SLRS, KDM, SB, MM, and RM conceived the research. LC, SM, PS, JT, WS, JM, AP, BB, JF, DK, SGT, AS, RJN, and RM conducted experiments and generated the data. MLB, SLRS, LC, and RRM analyzed the data. MLB, SLRS, KDM, and RRM wrote the manuscript.

Publication: Bendall, Matthew L, Sarah LR Stevens, Leong-Keat Chan, …, Katherine D McMahon, and Rex R Malmstrom. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal* 10(7):1589–1601.

1.1 Abstract

Multiple models describe the formation and evolution of distinct microbial phylogenetic groups. These evolutionary models make different predictions regarding how adaptive alleles spread through populations and how genetic diversity is maintained. Processes predicted by competing evolutionary models, for example, genome-wide selective sweeps vs gene-specific sweeps, could be captured in natural populations using time-series metagenomics if the approach were applied over a sufficiently long time frame. Direct observations of either process would help resolve how distinct microbial groups evolve. Here, from a 9-year metagenomic study of a freshwater lake (2005–2013), we explore changes in single-nucleotide polymorphism (SNP) frequencies and patterns of gene gain and loss in 30 bacterial populations. SNP analyses revealed substantial genetic heterogeneity within these populations, although the degree of heterogeneity varied by >1000-fold among populations. SNP allele frequencies also changed dramatically over time within some populations. Interestingly, nearly all SNP variants were slowly purged over several years from one population of green sulfur bacteria, while at the same time multiple genes either swept through or were lost from this population. These patterns were consistent with a genome-wide selective sweep in progress, a process predicted by the 'ecotype model' of speciation but not previously observed in nature. In contrast, other populations contained large, SNP-free genomic regions that appear to have swept independently through the populations prior to the study without purging diversity elsewhere in the genome. Evidence for both genome-wide and gene-specific sweeps suggests that different models of bacterial speciation may apply to different populations coexisting in the same environment.

1.2 Introduction

Microbial communities are composed of genetically and ecologically distinct groups. Multiple evolutionary models have been proposed to explain the formation of distinct groups,

and these models often assume a different balance between the forces of recombination and selection. The 'ecotype model' is perhaps the most prominent, and it assumes recombination within ecologically coherent populations is low enough that if a population member gains an advantageous trait, then that member will likely take over the population before the trait can spread to other members via recombination (Cohan, 2001; Cohan and Perry, 2007). As a result, genetic heterogeneity is purged from the population, that is, the population experiences a genome-wide selective sweep. In this model, distinct phylogenetic groups form after ecologically divergent populations undergo a series of genome-wide sweeps (Cohan, 2001; Cohan and Perry, 2007). Support for the ecotype model, however, is largely based on theoretical simulations (Cohan, 1994; Majewski and Cohan, 1999), and thus far genome-wide sweeps have not been observed in natural populations (Cordero and Polz, 2014; Shapiro and Polz, 2014b). In fact, recent comparative genomic analyses support an alternate model where recombination rates are high, and advantageous genes are exchanged among population members without initiating genome-wide sweeps (Whitaker et al., 2005; Fraser et al., 2007; Cadillo-Quiroz et al., 2012; Shapiro et al., 2012). Direct, time-resolved observations of either genes or genomes sweeping through natural populations would help to determine which mechanisms drive diversification in microbial assemblages.

Genetic diversification can be observed directly by sequencing bacterial populations at various time points throughout their evolutionary history (Barrick et al., 2009; Maharjan et al., 2012; Herron and Doebeli, 2013). In long-term evolutionary studies of Escherichia coli cultures, for example, DNA sequencing has revealed numerous single-nucleotide polymorphisms (SNPs) appearing spontaneously and, in some cases, becoming fixed, over thousands of generations (Barrick and Lenski, 2009; Barrick et al., 2009; Lee et al., 2012). Exploring genetic changes within natural populations is the next step in understanding how bacteria evolve and diverge into distinct groups. Investigating natural communities will, for example, provide a more complete picture of how genome composition is impacted by natural processes, such as horizontal gene transfer, the direct uptake of free DNA and

interactions with viruses—processes that are not typically addressed in laboratory-based studies (Barrick et al., 2009; Maharjan et al., 2012; Herron and Doebeli, 2013). This approach will also expand our view to include new microbial groups whose rates of growth, mutation and recombination may differ substantially from isolates grown in the laboratory.

Time-series metagenomics has the potential to identify genetically and ecologically distinct groups within natural microbial communities and reveal the mechanisms leading to their diversification. For example, de novo assembly of metagenomic data can generate reference genomes of uncultivated microbes (Tyson et al., 2004; Iverson et al., 2012; Wrighton et al., 2012; Albertsen et al., 2013; Sharon et al., 2013), while recruitment of metagenomic reads to reference genomes can reveal genetic heterogeneity within discrete populations (Konstantinidis and DeLong, 2008; Caro-Quintero and Konstantinidis, 2012). Metagenomics can also provide insights into the evolutionary processes within natural communities by uncovering evidence for genome recombination among microbes and providing direct measurements of nucleotide substitution rates (Tyson et al., 2004; Allen et al., 2007; Simmons et al., 2008; Denef and Banfield, 2012). Repeated metagenomic sampling of an environment, if applied over a sufficiently long time frame, could also capture other evolutionary patterns such as genome-wide selective sweeps, a process that has not been directly observed in natural populations to date (Cordero and Polz, 2014; Shapiro and Polz, 2014b).

Here we use metagenomics to explore the genome dynamics and diversification processes of freshwater bacterial groups over a 9-year period. As part of this study, we perform shotgun sequencing of a freshwater lake microbial community sampled at 63 time points from 2005 to 2013 and reconstruct 30 genomes from a variety of bacterial groups. To better understand the ecological and evolutionary processes at work within natural communities, we analyze these genomes, and the populations they represent, for changes in gene content and SNP-level heterogeneity over the 9-year period.

1.3 Materials and Methods

DNA sampling and sequencing

Trout Bog Lake is located in Wisconsin, USA and surrounded by boreal forests and a sphagnum mat that supply large amounts of terrestrially derived organic matter to the lake. Surface area is ~11000m2, a maximum depth of 9m and a mean pH of 5.1. Depth integrated water samples were collected from the hypolimnion layer at 63 different time points during ice-free periods from 2005 to 2013 and from the epilimnion layer at 45 time points from 2007 to 2009 (Supplementary Table A.2S4) and filtered on 0.2 micrometer pore-size polyethersulfone Supor filters (Pall Corp., Port Washington, NY, USA) prior to storage at 80°C. DNA was later purified from these filters using the FastDNA Kit (MP Biomedicals, Burlingame, CA, USA).

DNA sequencing was performed at the Department of Energy Joint Genome Institute (Walnut Creek, CA, USA). Four libraries (two from each layer of water column) were amplified following the standard Illumina TruSeq (Illumina, San Diego, CA, USA) protocol and sequenced on the Illumina GA IIx platform (Illumina), while all other libraries remained unamplified and were sequenced on the HiSeq 2500 platform (Illumina). Paired-end sequences of 2 150bp were generated for all libraries. Libraries from samples collected between 2007 and 2009 were generated simultaneously in a 96-well plate, and samples from different years were pooled together for sequencing. Samples collected in 2005, 2012 and 2013 were also processed simultaneously in a 96-well plate prior to pooling and sequencing. Sequence reads were merged with the FLASH v1.0.3 (Magoc and Salzberg, 2011) with a mismatch value of $\leqslant 0.25$ and a minimum of 10 overlapping bases from paired sequences, resulting in merged read lengths of 150–290bp. Metagenomic sequence reads are publicly available on the JGI Genome Portal (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism= TroutBog-metagenomicdata).

Merged reads from all samples collected between 2007 and 2009 were pooled by layer into two combined assemblies using SOAPdenovo (Luo et al., 2012) with k-mer sizes of 107, 111, 115, 119, 123 and 127 (Supplementary Table A.2S5). Contigs from SOAPdenovo assemblies were combined into a final assembly using Minimus (Sommer et al., 2007). Samples from 2005, 2012 and 2013 were sequenced at a later date so that changes in SNP allele frequencies and patterns of gene gain/loss could be followed over a longer time period (see below), and these sequences were not included in the combined assembly.

Binning metagenomic contigs into genomes

Contigs $\geqslant$ 2.5kbp were organized into genomes based on tetranucleotide sequence composition and overall contig coverage patterns using the binning tool MetaBat (Kang et al., 2015). Coverage levels at 45 time points collected between 2007 and 2009 were determined from metagenomic reads mapping with $\geqslant$ 95% sequence identity using the Burrows–Wheeler aligner (BWA)-backtrack alignment algorithm with n=0.05 (Li and Durbin, 2009). To minimize the chance of incorrectly binning contigs from different organisms, MetaBat was run with 'very specific' settings. Genome bins with $\geqslant$ 10-fold coverage in $\geqslant$ 3 years of the time-series study were then manually curated to ensure all contigs shared similar abundance patterns (Supplementary Figure A.1S2). Contig coverage levels in curated genome bins had an average correlation coefficient of 0.995, with the median bin coverage.

Gene prediction and annotation

Gene prediction and annotation for metagenomic reconstructions was performed using the DOE Joint Genome Institute's Integrated Microbial Genome database tool (Markowitz et al., 2012). Genome completeness was estimated using the two methods published previously based on the fraction of broadly shared genes recovered in each genome (Rinke et al., 2013; Parks et al., 2015) Supplementary Table A.2S6. Accession numbers for publically available genomes deposited in IMG are listed in Supplementary Table A.2S7.

Phylogenetic analysis and average sequence identities

Genomes were classified based on the taxonomic assignments from a subset of 37 conserved marker genes, mostly ribosomal proteins, extracted from the reconstructed genomes using PhyloSift (Darling et al., 2014). Marker genes with cumulative probability masses <0.80 were removed. Genomes were assigned to the finest taxonomic scale for which all marker genes agreed, ranging the phylum level for some genomes down to genus level for others. TM7-1225 was initially only classified to the domain Bacteria using this approach, but the population was assigned to the TM7 phylum through phylogenetic analysis of marker genes from previously published TM7 genomes. Marker genes in other TM7 genomes were identified and concatenated using Phylosift, and a maximum likelihood tree was generated using RAxML with the Dayhoff substitution model Supplementary Figure A.1S6 (Stamatakis, 2014). Bootstraps were generated with 100 replicates using RAxML's rapid bootstrap function.

Identifying sequence-discrete populations

Metagenomic reads were mapped to the reconstructed genomes using BBmap (https://sourceforge.net/projects/bbmap/), with minimum alignment identity cutoff of 0.60. BBmap was selected for this particular mapping step owing to ease in mapping with low-percent identity reads. The genome location and percentage of identity for each mapped read was extracted from the alignments, and the fraction of reads mapping with 60–100% nucleotide identity to each genome was determined for all time points. A large drop in coverage around 95% identity was observed for all genomes (Figure 1.1). This coverage discontinuity was used to identify the boundary of 'sequence-discrete' populations, although the vast majority of reads mapping with high identity (>95%) actually mapped with 99% identity.

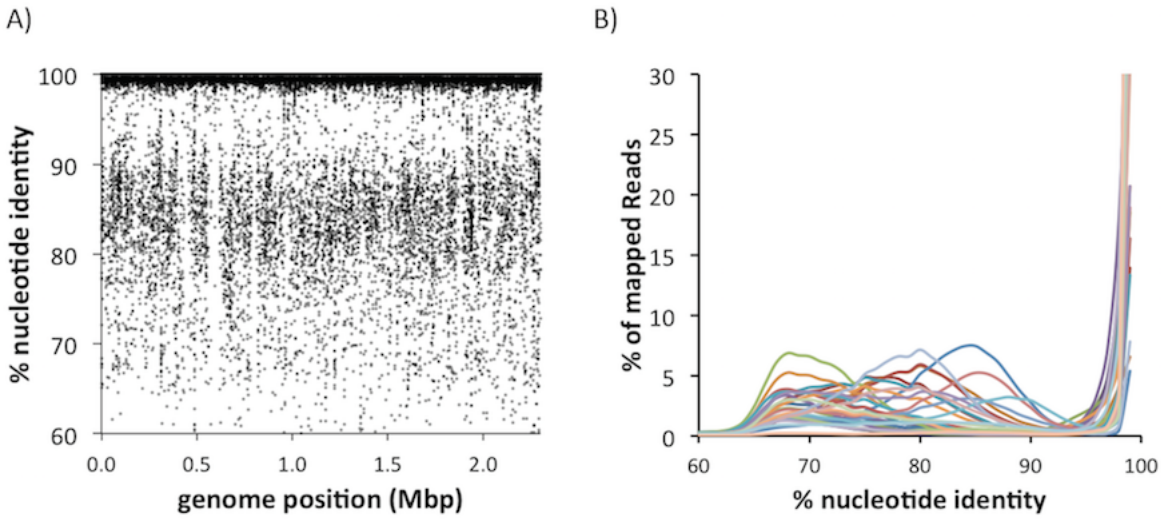

Figure 1.1: 'Sequence-discrete' populations revealed by metagenomic read mapping. (a) An example recruitment plot of 50000 shotgun reads mapping across the Chlorobium-111 genome at various nucleotide identity levels. Each dot represents a read. (b) Summary of reads mapping at each percentage of nucleotide identity level for all genomes. Each line represents a different genome. A distinct lack of coverage around 95% identity was observed in all genomes. The y axis (percentage of mapped reads) of panel (b) was truncated at 30% to illustrate this coverage discontinuity.

SNP identification and analysis

SNPs were discovered by first mapping reads with 95% nucleotide sequence identity from each time point to the reference genomes using BWA. The vast majority of recruited reads actually mapped with 99% identity. As many individual time points lacked sufficient coverage for confident SNP discovery, we combined the alignments from samples collected in the same year to ensure at least 10-fold coverage per time period. Each of these yearly time periods were treated as a sample, and variant positions were identified using the multi-sample genotype likelihood model implemented in the GATK UnifiedGenotyper tool v.2.7-2 (McKenna et al., 2010; DePristo et al., 2011). The tool was run in 'DISCOVERY' mode, which did not require known variants as input, and ploidy was set to 1. To ensure only high-confidence SNPs were examined, an initial filter was applied to remove SNP loci with multiple alternate alleles, low quality scores (Q<30, 99.9%) or low genotype quality in one

or more samples (Q<30, 99.9%) (Supplementary Table A.2S8). We then removed a small fraction of outlier SNPs with unusually high or low coverage, that is, >1.5 interquartile ranges below the first quartile or above the third quartile. These SNPs do not necessarily represent all single-nucleotide variation in the populations because the references genomes were not complete. Some rare SNPs might also be overlooked despite high sequencing coverage.

The reconstructed genomes were temporal composites assembled from reads collected from 2007 to 2009, and ultimately only a single allele at each SNP locus was selected by the assembly algorithm, thus referring to the assembled allele as the 'reference' was somewhat arbitrary. For consistency, the 'reference' allele was chosen to be the majority allele observed at the final time period. This choice simplified figure construction and had no impact on patterns of gain and loss of diversity. Allele frequencies were calculated based on the number of reads observed with the reference or alternate allele.

Gene gain and loss over time

To identify genes whose relative abundance in the population changed significantly over the course of this study, we compared gene coverage between the first and last year with $$10-fold coverage using the Metastats software (Paulson et al., 2011). Coverage was determined as the number of metagenomic reads mapping with 95% sequence identity to each gene at each time point. Gene coverage was normalized by gene length, and spurious short gene annotations (<450bp) were excluded from the analysis. Gene frequency was estimated as the coverage of each gene divided by the median coverage of all other genes in the genome. A frequency of 1 implies each cell in the population encoded one copy of the gene. Genes were considered to be gained or lost from a population if the gene frequency changed by a magnitude of >0.4 copies per cell with a false discovery rate of $\leqslant 0.01$ using the Metastats test.

Identifying putative sites of historical gene-specific sweeps

Potential sites of gene-specific sweeps were identified as regions with unusually low numbers of SNPs relative to the rest of the genome. The probability that region of any size would contain no SNPs was modeled as a Poisson distribution that assumed SNPs were distributed uniformly and occurred with an average rate equal to the total number of SNPs divided by genome size. The chance of finding a SNP-free region of any size in a genome was then determined as the Poisson probability multiplied by the genome size minus the region size. In a 1-Mbp genome, for example, the Poisson probability of a 1-Kbp region lacking SNPs would be multiplied by 999000, that is, the number of unique 1-Kbp regions found in a 1-Mbp genome. Genome regions with anomalously low numbers of SNPs were identified, with a significance cutoff of $P<0.0001$.

1.4 Results and Discussion

Genome assembly from metagenomic data

Bacterial genomes were reconstructed from a combined assembly of metagenomic sequences collected at several time points. Contigs generated from this combined assembly were organized into genome bins based on tetranucleotide sequence composition and differences in contig coverage levels throughout the time series. The unique temporal abundance pattern of each genome bin (Supplementary Figure A.1S1), and the tight synchronization of contig coverage within bins (Supplementary Figure A.1S2), allowed us to confidently distinguish closely related genomes based on coverage differences (Albertsen et al., 2013; Sharon et al., 2013). We then focused our analyses on 30 reconstructed genomes that had 10-fold sequence coverage in at least three different years from 2005 to 2013 (Table 1.1). These genomes belonged to 13 classes distributed among 6 phyla; some could only be classified to the phylum level while others were classified to the genus level based

Genome name	Environment	Genome size (bp)	Contigs	Genes	Genome comp. (a/b)
Actinobacterium-149	Epilimnion	764,032	95	917	64/74
Nitrosomonadales-439	Epilimnion	996,711	125	1094	67/69
Polynucleobacter-567	Epilimnion	1,660,228	93	1777	72/62
Rickettsia-755	Epilimnion	1,013,290	136	1149	98/100
Betaproteobacteria-788	Epilimnion	990,006	133	1125	57/52
Methylophilaceae-913	Epilimnion	942,700	85	1111	76/99
Opitutae-1301	Epilimnion	2,036,179	101	1943	95/100
Opitutae-1800	Epilimnion	2,186,907	124	1,998	90/100
Actinobacterium-2057	Epilimnion	971,617	97	1063	74/58
Chlorobium-111	Hypolimnion	2,314,202	74	2319	92/100
Polynucleobacter-238	Hypolimnion	1,314,366	121	1475	66/52
Holophagales-254	Hypolimnion	2,981,798	188	2862	80/56
Desulfocapsa-433	Hypolimnion	3,073,408	152	2864	77/66
Methylotenera-545	Hypolimnion	1,431,993	51	1439	90/82
Actinobacterium-680	Hypolimnion	1,257,796	81	1353	73/60
Polynucleobacter-941	Hypolimnion	1,496,525	68	1581	57/54
TM7-1225	Hypolimnion	915,278	14	993	63/90
Methylobacter-1380	Hypolimnion	2,299,825	136	2072	68/59
Methylotenera-1381	Hypolimnion	1,077,715	49	1131	50/46
Sulfurimonas-1998	Hypolimnion	2,301,184	60	2383	98/100
Methylobacter-2062	Hypolimnion	3,124,798	188	2919	94/89
Bacteroidales-2086	Hypolimnion	3,680,027	151	2965	72/59
Actinobacterium-2152	Hypolimnion	845,311	113	980	61/64
Opitutae-2519	Hypolimnion	1,808,963	100	1654	72/88
Methylophilaceae-2902	Hypolimnion	1,002,927	75	1180	62/65
Desulfobulbus-2922	Hypolimnion	3,798,404	58	3387	93/92
Actinobacterium-3180	Hypolimnion	1,149,636	85	1251	67/54
Gallionella-3415	Hypolimnion	2,657,023	54	2637	97/95
Chlorobium-3520	Hypolimnion	2,156,671	83	2242	89/100
Acidomicrobium-3765	Hypolimnion	1,315,659	42	1392	76/94

Table 1.1: Genomes reconstructed from metagenomic-combined assembly

Genome completeness estimated using the approaches of Parks et al. (2015) (a) and Rinke et al. (2013) (b)

on availability of related reference genomes (Supplementary Figure A.1S3). Estimates of genome completeness ranged from ~50 to 100% (Table 1.1).

Genetic heterogeneity in natural populations

The recovered genomes were assembled from sequences collected at several time points and do not reflect the exact genetic make up of any single cell, as is the case with all metagenomic constructs (Tyson et al., 2004; Simmons et al., 2008; Denef and Banfield, 2012). Instead, they are composites that represent populations of cells with high sequence similarity. These populations were visualized by recruiting metagenomic reads at various sequence identity levels to each composite reference genome (Konstantinidis and DeLong, 2008; Caro-Quintero and Konstantinidis, 2012). In every case, metagenomic recruitment revealed 'sequence-discrete' populations whose reads typically mapped with $\geqslant$ 99% nucleotide identity to reference genomes and closely related populations whose reads mapped with <90% identity (Figure 1.1). A large drop in coverage around 95% sequence identity was observed in all genomes (Figure 1.1b). This is a common feature in metagenomic recruitment plots, and it marks the boundary between these operationally defined sequence-discrete populations and other closely related sympatric populations (Tyson et al., 2004; Konstantinidis and DeLong, 2008; Caro-Quintero et al., 2011; Oh et al., 2011; Caro-Quintero and Konstantinidis, 2012). The terms 'population' and 'sequence-discrete population' are used interchangeably for the remainder of this manuscript.

Sequence-discrete populations are not clonal but instead are composed of highly similar, co-occurring genotypes that contain some degree of genetic diversity (Caro-Quintero et al., 2011; Caro-Quintero and Konstantinidis, 2012). Previous studies suggest that levels of intra-population diversity are lower than those among strains of the same named species (Konstantinidis and Tiedje, 2005; Konstantinidis and DeLong, 2008; Caro-Quintero et al., 2011; Caro-Quintero and Konstantinidis, 2012). This implies members of sequence-discrete populations may have highly similar, if not identical, ecological roles (Caro-Quintero and Konstantinidis, 2012), although the ecological coherence of these populations has not been demonstrated.

We examined intra-population diversity by identifying SNPs within sequence-discrete

populations (Tyson et al., 2004; Hunt et al., 2008). By recruiting highly similar reads from all time points, the vast majority of which mapped with 99% nucleotide identity, we found numerous SNPs in each population, ranging from 8501 SNPs in Holophagales-254 to only 3 SNPs in TM7-1225 (Table 1.2). Most populations had >1800 SNPs per Mbp, but four populations had <50 SNPs per Mbp, including the nearly clonal TM7-1225 population. Although abundant populations had higher coverage levels and thus more power to detect rare SNPs, coverage depths alone could not account for the large differences in SNP counts among populations—up to three orders of magnitude in some cases (Figure 1.2; Supplementary Table A.2S1). For example, Methylotenera-1381 had eightfold more SNPs per Mbp than its close relative Methylotenera-545 even though Methylotenera-545 had higher metagenomic coverage. This suggests that intra-population diversity levels varied dramatically between phylogenetic groups, including closely related populations belonging to the same genus (Supplementary Figure A.2S3).

Large differences in diversity among populations could result from a number of processes. For example, populations with fewer SNPs might have immigrated to the lake more recently and had less time to diversify (that is, founder effect) or may have lower mutation/substitution rates or could have more recently experienced a purge of diversity than populations with higher SNP counts. Indeed, the extraordinarily low number of SNPs in TM7-1225 suggests that this population is either quite new to the ecosystem or it experienced a periodic selective event that essentially produced a clonal population shortly before the start of this study (Table 1.2).

Most SNPs within the sequence-discrete populations did not result in amino-acid substitutions (Table 1.2). Instead, SNPs were typically silent or located in intergenic regions. Nonsense mutations generating premature stop codons were found in several populations, indicating some genotypes within these populations encoded nonfunctional genes, although these mutations typically accounted for <0.1% of SNPs (Table 1.2). The small proportion of nonsynonymous SNPs might indicate that purifying selection was driving

Genome name	Total SNPs	SNPs / Mbp	Syn SNPs (a)	Miss (b)	Non (c)	Inter (d)
Actinobacterium-149	3514	4599	2914	460	3	136
Nitrosomonadales-439	1772	1753	1378	275	3	91
Polynucleobacter-567	4571	2753	3627	710	3	231
Rickettsia-755	45	44	18	11	2	14
Betaproteobacteria-788	6244	6188	5039	851	3	231
Methylophilaceae-913	3003	3186	2223	656	1	123
Opitutae-1301	6437	3161	5257	893	4	283
Opitutae-1800	3839	1743	2924	663	1	223
Actinobacterium-2057	2238	2182	1659	377	0	84
Chlorobium-111	3111	1344	1498	1127	22	464
Polynucleobacter-238	6451	4908	3418	738	1	2291
Holophagales-254	8501	2851	5605	2004	10	881
Desulfocapsa-433	4995	1625	3187	1037	1	770
Methylotenera-545	279	195	132	120	1	26
Actinobacterium-680	297	236	189	47	1	60
Polynucleobacter-941	4269	2853	2971	971	4	323
TM7-1225	3	3	0	1	0	2
Methylobacter-1380	1381	600	951	197	1	232
Methylotenera-1381	1779	1651	1153	434	2	190
Sulfurimonas-1998	279	121	154	95	1	29
Methylobacter-2062	6660	2131	3908	1515	14	1223
Bacteroidales-2086	4256	1157	2389	1231	12	623
Actinobacterium-2152	4209	4979	3400	597	3	209
Opitutae-2519	8036	4442	6254	1246	4	531
Methylophilaceae-2902	2943	2934	2115	712	2	113
Desulfobulbus-2922	145	38	43	70	3	29
Actinobacterium-3180	2111	1836	1551	318	2	240
Gallionella-3415	69	26	35	23	0	11
Chlorobium-3520	4146	1922	2317	1180	11	637
Acidomicrobium-3765	2126	1616	1505	477	0	143

Table 1.2: Summary of single-nucleotide polymorphisms (SNPs)

Synonymous SNPs (a), Nonsynonymous SNPs Missense (b), Nonsynonymous SNPs Nonsense (c), Intergenic (d)

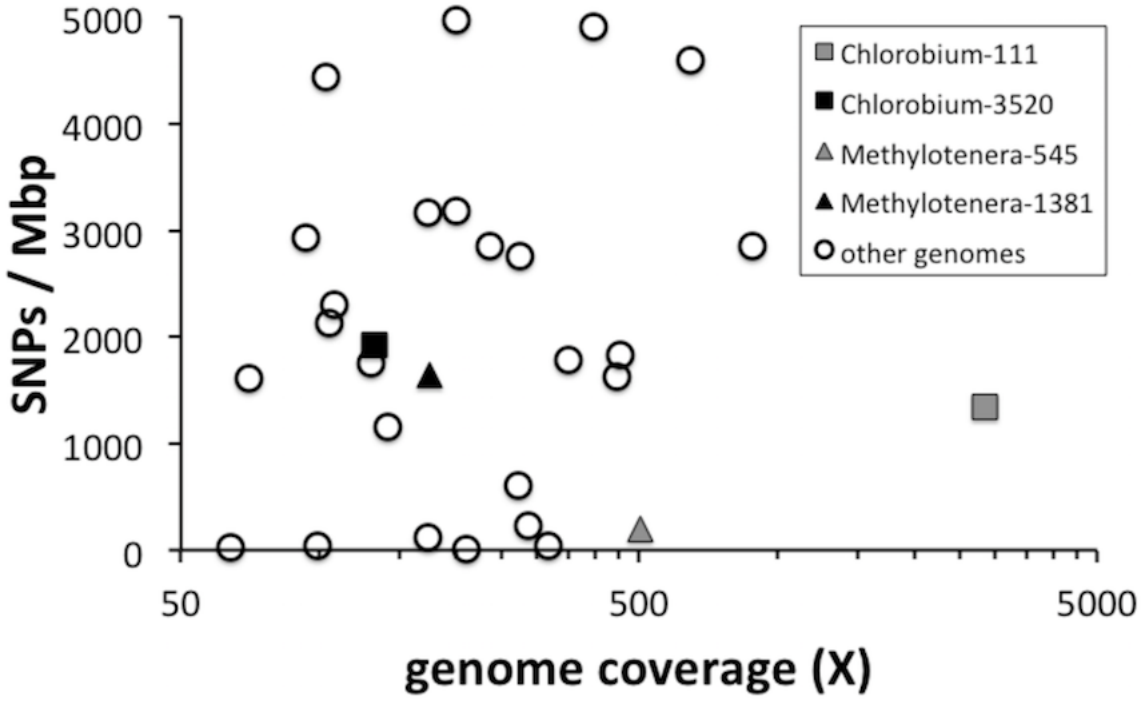

Figure 1.2: Differences in SNP-level heterogeneity among coexisting populations. The number of SNPs found in each sequence-discrete population, normalized to genome size (SNPs per Mbp), varied by three orders of magnitude among populations with similar coverage levels. Although the power to identify low-frequency SNPs increases with greater genome coverage, populations with many SNPs were not necessarily sequenced deeper than those with few SNPs. Two pairs of closely related populations are highlighted to illustrate this point.

mutation accumulation in most populations we surveyed (Simmons et al., 2008). The preponderance of synonymous mutations also suggests that most genetic variation within these sequence-discrete populations might be neutral, thus allowing many highly similar genotypes to coexist without outcompeting each other.

Purges of diversity in natural populations

Next we asked whether the degree of genetic heterogeneity within each population, as revealed by the proportions of SNP variants in the metagenomic reads, changed over the 9-year study period. SNP allele frequencies varied over time in all populations, although the fraction of total SNPs dominated by a single allele remained relatively low in most years

(for example, Actinobacterium-2152, Figures 1.3a and c; Supplementary Figure A.1S4). This suggests that the overall level of genetic heterogeneity in most populations did not change dramatically. However, in a few populations SNP allele frequencies did shift considerably and many SNP loci were dominated by a single allele (Figures 1.3b and d; Supplementary Figure A.1S4), indicating large changes in the relative abundance of different genotypes within these sequence-discrete populations. For example, Bacteroidales-2086 was composed of many genotypes with comparable abundances in 2007, 2008 and 2012—based on the more even distribution of SNP allele frequencies in these years—whereas large shifts in allele frequencies throughout the genome suggests that one genotype, or perhaps a few, dominated the population in 2005, 2009 and 2013 (Figures 1.3b and d). Diversity levels also shifted substantially from year to year within Methylobacter-1380, Methylotenera-1381 and Sulfurimonas-1998 (Supplementary Figure A.1S4).

The most dramatic change in allele frequencies was observed in the Chlorobium-111 population, which initially displayed a high degree of SNP-level heterogeneity, but slowly lost most of this diversity over the course of the study. That is, the frequency of alternate alleles in the population was close to zero at nearly all SNP sites by 2013 (Figure 1.4a; Supplementary Figure A.1S4). These SNP sites were not localized to specific genomic regions (Supplementary Figure A.1S5). This pattern did not result from differences in coverage (Supplementary Figure A.1S1) or differences in library creation and sequencing steps (see Methods and Materials section). Nor was it the result of inter-population dynamics where a different sequence-discrete population displaced the Chlorobium-111 population; this process would appear as a drop in coverage in Chlorobium-111, not a change in SNP allele frequencies. The simultaneous trend towards fixation at nearly all SNP sites, which were spread throughout the genome, indicates a steady and substantial loss of genetic heterogeneity within the population.

In addition to SNP dynamics, our time series also revealed patterns of gene gain and loss within the Chlorobium-111 population. The relative abundance of eight genes slowly

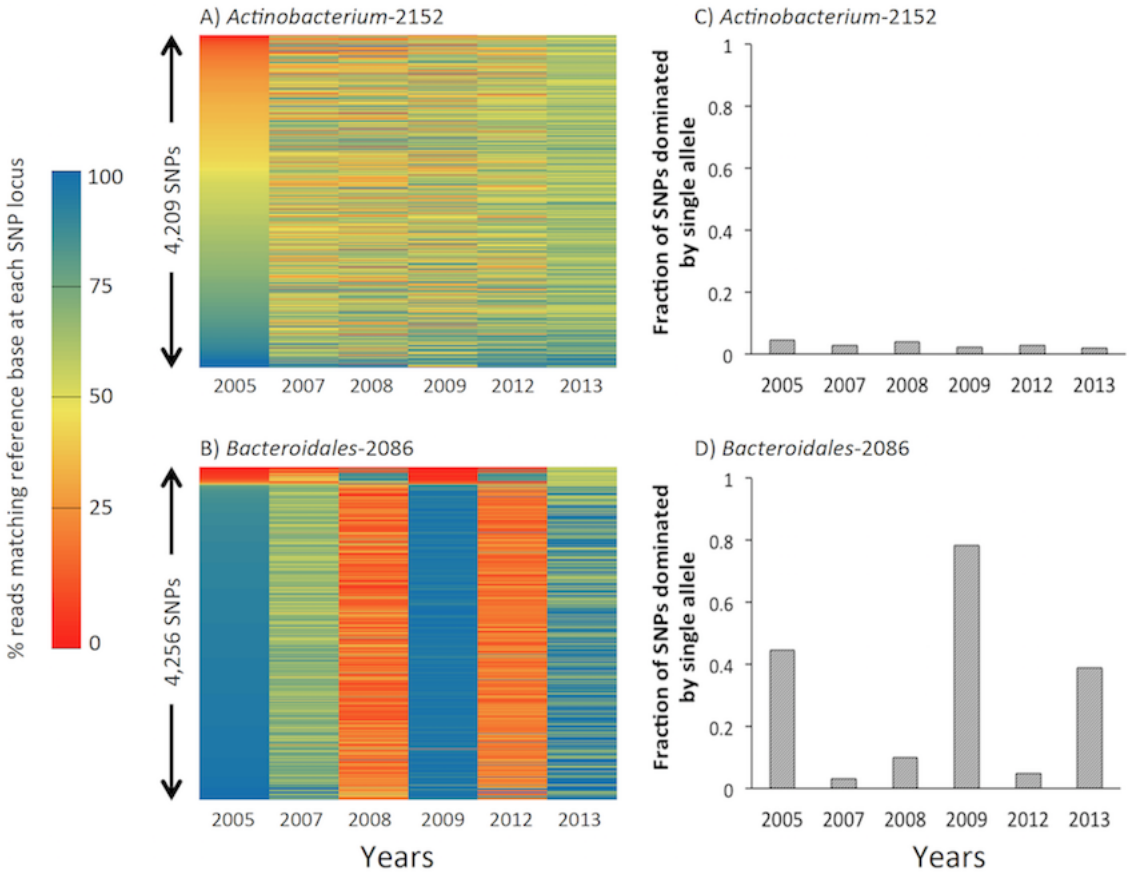

Figure 1.3: Temporal dynamics of SNP allele frequencies within different populations. (a, b) Two examples of populations with different SNP dynamics. SNPs are arrayed along the y axis, with each row representing one SNP locus. SNP color indicates allele frequency, that is, the percentage of metagenomic reads supporting the reference allele during each time period. SNPs dominated by a single allele appear either as red (few reads matching reference base) or blue (most reads matching reference base). SNPs are arranged in ascending order along the y axis based on allele frequency in 2005. (c, d) Fraction of SNPs dominated by single allele ($\geqslant$ 95% frequency) in each year. Broad patterns of allele frequencies were determined by combining sequence data for each year.

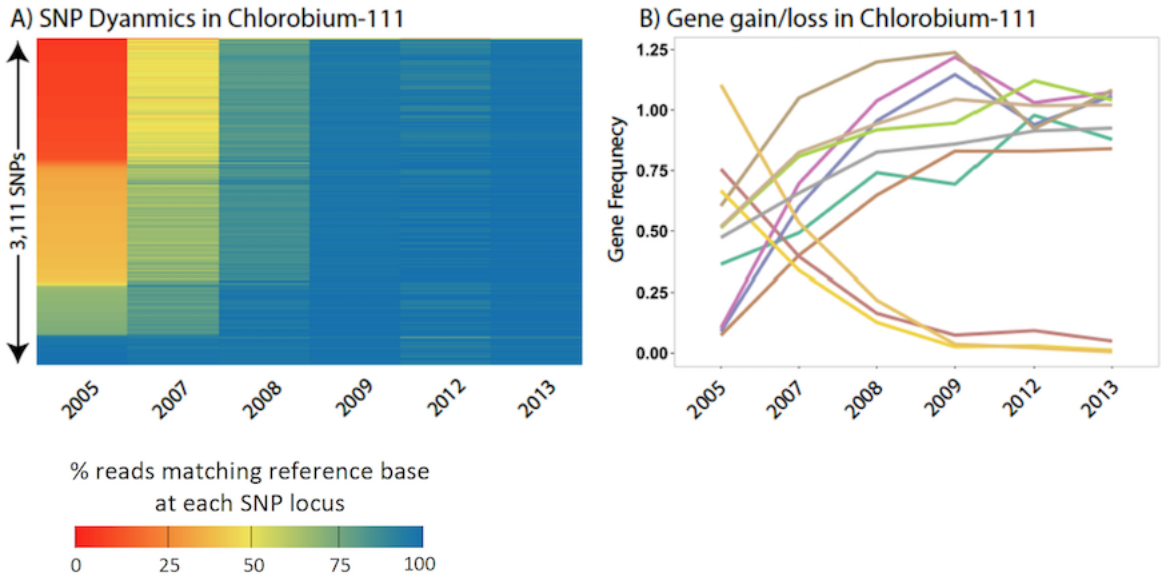

Figure 1.4: Temporal trends in SNP allele frequencies and gene content in a natural Chlorobium population. (a) SNPs are arrayed along the y axis, with each row representing one SNP locus. SNP color indicates allele frequency, that is, the percentage of metagenomic reads supporting the reference allele during each year. (b) Relative abundance of genes gained or lost from Chlorobium-111. A gene frequency of 1 equates to single copy per cell. Gene annotations and locus IDs are listed in Supplementary Table A.2S2. Broad patterns of allele frequencies and gene abundances were determined by combining sequence data for each year.

increased until they were encoded by nearly every cell in the population (Figure 1.4b; Supplementary Table A.2S2). Two of the genes were adjacent while the others were scattered throughout the genome. These dynamics, when viewed alongside the simultaneous genome-wide purge of SNPs, suggests that these genes were acquired horizontally in one genotype at some point prior to this study and increased in abundance as the genotype (or its descendant lineage) took over the population. Simultaneously, three genes slowly decreased until <10% of cells in the Chlorobium-111 population encoded them in 2013, indicating that the newly dominant lineage lacked these genes (Figure 1.4b).

The dramatic loss of SNP-level heterogeneity and the patterns of gene gain and loss in the Chlorobium-111 population were consistent with a genome-wide selective sweep in progress, a process predicted by the ecotype model for bacterial diversification (Cohan,

2001; Cohan and Perry, 2007). In this model, genetic diversity accumulates within ecologically coherent populations and is periodically lost when one member of a population outcompetes all others after gaining an advantageous trait through mutation or horizontal gene transfer (Cohan and Perry, 2007). I n such an event, diversity would be purged at all loci in the population as the less fit members of the population were replaced. If this process were captured in a metagenomic time-series study, then we would expect nearly all SNPs in the population to trend toward fixation, while at the same time some genes would sweep through or be swept from the population—the same patterns we observed in Chlorobium-111 (Figure 1.4). In this scenario, we would also expect the vast majority of SNP variants to be neutral, at least with regards to the selective pressure driving the sweep, and their dynamics would merely trace the process of selection based on their genomic linkage to some advantageous trait in the winning lineage. That is, the SNPs in Chlorobium-111 did not arise de novo during this study, and it is not clear which alleles, if any, were specifically selected based on a fitness advantage they provided; most SNPs were simply 'genomic hitchhikers'. Similarly, it is not clear if the genes we observed sweeping through the population provided an advantage, or if they, much like the neutral SNPs, merely traced the putative sweep based on their linkage to other unidentified alleles that improved fitness. It was not obvious from functional annotations, when available, how the gain or loss of these genes might have provided an advantage (Supplementary Table A.2S2).

The predicted result of genome-wide sweeps and the ecotype model is the formation of sequence clusters that represent ecologically distinct groups (Cohan and Perry, 2007). The existence of such sequence clusters in other systems has been taken as evidence for the ecotype model, but to our knowledge this study provides the first direct observations of a natural population appearing to undergo a genome-wide sweep (Cordero and Polz, 2014; Shapiro and Polz, 2014b). Of course, Chlorobium-111 was not completely clonal by 2013, indicating that the sweep was not yet complete or the population was experiencing a

'soft sweep' where selection favored a few genotypes from a large and diverse population. In this scenario, the persistent genotypes would have acquired an advantageous allele independently or via intra-population recombination prior to selection (Messer and Petrov, 2013). Thus a selective sweep would not purge sequence differences among genotypes encoding the advantageous allele. As the time between trait acquisition and selection increases, periodic selection is more likely to produce some form of soft sweep in natural populations rather than a theoretical 'hard sweep' (Messer and Petrov, 2013). In addition, even though populations were sequenced deeply over 9 years, it is possible that diversity could be maintained below detection limits and reappear on longer time scales. Although acknowledging this caveat, we believe the patterns observed in Chlorobium-111 and the discovery of four populations with <50 SNPs per Mbp, including the nearly clonal TM7-1225 population (Table 1.2), suggest tshat genome-wide sweeps are occurring in natural populations.

Based on the observed patterns, the Chlorobium-111 population appears to follow a different model of bacterial diversification than some other microbes. For example, through comparative genomic analysis of closely related Vibrio cyclitrophicus isolates, Shapiro et al. (2012) found that divergence between ecologically distinct groups was likely driven by gene-specific sweeps followed by preferential recombination within micro-niche-adapted populations and not by genome-wide sweeps. High recombination rates also appear to prevent periodic selection and to preserve genome-wide diversity in populations of Sulfolobus islandicus and Synechococcus dwelling in hot springs (Whitaker et al., 2005; Cadillo-Quiroz et al., 2012; Rosen et al., 2015). Conversely, although we could not measure recombination with only a single reconstructed genome representing each population, it appears that intra-population recombination rates were too low to prevent a massive and long-term purge of diversity within Chlorobium-111.

Preservation of intra-population diversity

Models invoking either genome-wide or gene-specific sweeps are not mutually exclusive (Doolittle, 2012), and it is possible both mechanisms shape the genetic diversity of microbial populations. For example, genome-wide sweeps may occur in groups with lower recombination rates, whereas gene-specific sweeps occur in other groups with inherently high recombination rates, for example, Helicobacter pylori (Falush et al., 2001) and presumably V. cyclitrophicus, S. islandicus and Synechcococcus (Whitaker et al., 2005; Shapiro et al., 2012; Rosen et al., 2015). Twenty-nine out of the 30 populations analyzed did not undergo genome-wide sweeps during the course of our study, suggesting either that periodic selection events are rare and that these populations did not experience strong selective pressures during the course of our study or that other mechanisms preserved diversity within these populations.

To determine whether recombination preserved diversity in some of the populations, we next searched for genes sweeping through populations, as was seen in Chlorobium-111, but without a corresponding genome-wide purge of SNPs. However, we did not find clear evidence of gene-specific sweeps in any of the populations during the course of this study. Gene-specific sweeps could have been missed if the genes were not part of the assembled genomes, but we might have expected to capture a gene sweep in at least 1 of the other 29 populations if such sweeps were common. Gene-specific sweeps could also have been missed if the sweeping genes only differed by a few nucleotides from homologs already found in the populations. In fact, there were examples in some populations where a few adjacent SNPs trended toward fixation while genome-wide diversity was maintained, a pattern not only consistent with a gene variant sweeping independently through a population but also consistent with a shift in the relative abundance of different genotypes—the latter process occurred in all populations (Supplementary Figure A.1S4). If populations did not experience gene-specific sweeps during the course of the study, then perhaps diversity was preserved through other mechanisms such as 'kill the winner' interactions where

viruses suppress rapidly growing genotypes within a population (Thingstad, 1998, 2000; Rodriguez-Brito et al., 2010). Interestingly, such top–down pressures were not sufficient to prevent the steady and massive loss of diversity that occurred within the Chlorobium-111 population over several years.

Although gene-specific sweeps were not directly observed during the course of the time series, SNP recruitment patterns indicate that large genome regions may have swept independently through some populations prior to the study period. For example, Polynucleobacter-238 had 6451 SNPs located throughout the genome except for in a statistically anomalous 21kbp region that lacked SNPs entirely (P<0.0001; Supplementary Figure A.1S5; Supplementary Table A.2S3). Large SNP-free regions of 41kbp, 9–25kbp, 22–23kbp, 11kbp and 12kbp were also found in Methylobacter-2062, Holophagales-254, Opitutae-1800, Opitutae-1301 and Methylophilaceae-913, respectively (P<0.0001; Supplementary Table A.2S3). If a genome region swept independently through a population, then this region would appear as an island of localized homogeneity within a heterogeneous genomic background (Guttman and Dykhuizen, 1994) —the same pattern observed in these six populations.

Large, SNP-free regions could also arise according to the 'adapt globally, act locally' model where a generally advantageous allele is shared between closely related ecotypes and triggers independent genome-wide sweeps in each (Majewski and Cohan, 1999). The six sequence-discrete populations were each clearly composed of many different genotypes based on the large range of SNP allele frequencies observed during the same time period—SNP allele frequencies would be similar at all loci if each population was composed of only two genotypes. Thus, for the 'adapt globally, act locally' model to apply, each sequence-discrete population would have to be composed of several coexisting ecotypes with inter-ecotype recombination rates sufficient for the allele to spread among all ecotypes but with intra-ecotype recombination rates too low to prevent genome-wide sweeps. Definitively distinguishing between this model and a single recombining population that experienced a gene-specific sweep may not be possible with our data, although the latter would seem

to be the more parsimonious explanation.

Gene annotations provide little insight into why the particular regions might have swept independently (Supplementary Table A.2S3), but the presence of these large SNP-free regions indicates that diversity within some populations may be maintained through frequent recombination. In addition, the evidence for gene-specific sweeps suggests that some populations in the lake might evolve following the model proposed for V. cyclitrophicus and S. islandicus where recombination rates are high, genes sweep independently and sequence divergence results from barriers to recombination between microniche-adapted populations (Whitaker et al., 2005; Fraser et al., 2009; Cadillo-Quiroz et al., 2012; Shapiro et al., 2012). Thus it appears that different evolutionary models might apply to different populations coexisting in the same environment.

Sequence-discrete populations and theoretical ecotypes

According to the 'stable model', an ecotype is a population of closely related genotypes whose members are ecologically similar and can coexist until one member/lineage gains a selective advantage and takes over the population by outcompeting all others (Cohan, 2001; Cohan and Perry, 2007). The model also assumes that periodic selection in one ecotype is independent from selection in other closely related, co-occurring ecotypes (Cohan, 2001; Cohan and Perry, 2007). However, the existence of these theoretically defined ecotypes has not been clearly demonstrated previously. The term 'ecotype' has been applied to various microbial groups, for example, clades of Prochlorococcus adapted to different light, temperature and mixing regimes (Moore and Chisholm, 1999; Rocap et al., 2003; Johnson, 2006; Malmstrom et al., 2010), but here and elsewhere the term follows the broader historical designation for subgroups within a species adapted to different environments and does not necessarily fit the more formal definition predicted by the ecotype evolutionary model and its variations (Turesson, 1922; Clausen et al., 1940; Coleman and Chisholm, 2007).

The sequence-discrete populations in this study, which were defined based on patterns

in metagenomic read recruitment, appear to match the description of theoretical ecotypes in some ways. For example, populations were composed of many closely related genotypes that were able to coexist at similar abundance levels for years. In some populations, a single genotype (or lineage of genotypes) was able to displace the other population members, implying that they all shared the same ecological niche (Figures 1.3b and 1.44, Supplementary Figure A.1S4). Furthermore, timing and magnitude of diversity purges differed between sympatric populations (that is, Chlorobium-111 vs Chlorobium-3520), suggesting that closely related sequence-discrete populations could undergo sweeps independently (Supplementary Figure A.1S4). The Chlorobium populations were separated in sequence space by the coverage discontinuity around 95% nucleotide sequence identity—for example, metagenomic reads mapping with $\geqslant$ 99% sequence identity to Chlorobium-111 also mapped with ~70–90% similarity to Chlorobium-3520, and vice versa—indicating that these populations could not be more similar and still remain sequence discrete (Figure 1.1). Thus closely related populations on either side of the coverage discontinuity appear to be ecologically distinct and behave in some ways similar to the theoretically predicted ecotypes.

If sequence-discrete populations behave similar to ecotypes in general, then coverage discontinuities in metagenomic read recruitment could be used to define ecotype boundaries. Ecotypes are expected to form distinct sequence clusters at the furthest tips of phylogenetic trees constructed from marker genes (Cohan, 2001; Cohan and Perry, 2007), but it remains unclear what level of sequence similarity, if any, demarcates an ecotype. In fact, any cutoff is likely to vary depending on the marker gene or the phylogenetic group in question, whereas the boundaries of sequence-discrete populations are determined empirically through read recruitment. For reference, the common marker genes recA and rpoB (Eisen, 1995; Dahllof et al., 2000; Walsh, 2004) both displayed 97% amino-acid sequence identity between the sympatric Chlorobium populations, while the other 1594 shared genes had an average amino-acid identity of 84%. Additional evidence of ecological

coherence within sequence-discrete populations will clarify the connections between these operationally defined populations and theoretical ecotypes.

1.5 Conclusions

In this study, we examined ecological and evolutionary patterns within natural bacterial communities through direct, time-resolved observations. From a metagenomic time-series study, we identified tractable populations that were genetically and ecologically distinct. We also observed substantial genetic heterogeneity within these populations, although the degree of heterogeneity varied by orders of magnitude between closely related, co-occurring populations. The purge of genetic heterogeneity from one of these populations, identified by changes in SNP allele frequencies, suggests that natural populations can experience genome-wide sweeps, a process not previously observed in situ (Cordero and Polz, 2014; Shapiro and Polz, 2014b). In other populations, evidence of historical gene-specific sweeps was uncovered, indicating that diversity within co-occurring populations may be controlled by different mechanisms and explained by different evolutionary models (Whitaker et al., 2005; Fraser et al., 2009; Cadillo-Quiroz et al., 2012; Shapiro et al., 2012).

These observations raise a variety of questions, such as: Are certain mechanisms of speciation (for example, genome-wide vs gene-specific sweeps) more common in certain environments or microbial groups? Do multiple mechanisms act on the same groups? How long does it take for genes or genomes to sweep through populations? At what rates do natural populations accumulate mutations? How does dispersal of highly similar genotypes impact population boundaries? We believe metagenomic time-series studies of different microbial groups inhabiting different environments will help answer these questions.

1.6 Acknowledgments

We thank JF Cheng, T Woyke, C Rinke, T Glavina del Rio, M Huntemann, N Ivanova, B Oyserman, B Foster and B Crary for their assistance with data analyses. We also thank J Shapiro and R Stepanauskus for their comments on an early draft of the manuscript. Work conducted by the US Department of Energy Joint Genome Institute was supported by the DOE Office of Science (DE-AC02-05CH11231). KDM acknowledges funding from the United States National Science Foundation Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-0822700), an INSPIRE award (DEB- 1344254) and a CAREER award (CBET-0738309). This material is based upon work supported by the National Institute of Food and Agriculture, United States Department of Agriculture, under ID number WIS01516 (to KDM).

2 CONTRASTING PATTERNS OF GENOME-LEVEL DIVERSITY ACROSS DISTINCT CO-OCCURRING BACTERIAL POPULATIONS

Sarahi L. Garcia[1,2]*, Sarah L.R. Stevens[1]*, Benjamin Crary[3], Manuel Martinez-Garcia[4], Ramunas Stepanauskas[5], Tanja Woyke[6], Susannah G. Tringe[6], Siv G.E. Andersson[7], Stefan Bertilsson[2], Rex R. Malmstrom[6] and Katherine D. McMahon[1,3]

[1]*Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA*; [2]*Department of Ecology and Genetics, Limnology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden*; [3]*Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA*; [4]*Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain*; [5]*Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA*; [6]*DOE Joint Genome Institute, Walnut Creek, CA, USA*; [7]*Department of Molecular Evolution, Uppsala University, Uppsala*

* These authors contributed equally to this work.

SLG, SLRS, RM, SB, and KDM conceived the research. RM, MMG, TW, and SGT conducted experiments and generated the data. SLG, SLRS, and KDM analyzed the data. SLG, SLRS, and BC prepared the figures. SLG, SLRS, RM, SB, and KDM wrote the manuscript. All authors participated in revision of the manuscript.

Publication: Garcia, Sarahi L., Sarah L. R. Stevens, Benjamin Crary, Manuel Martinez-Garcia, Ramunas Stepanauskas, TanjaWoyke, Susannah G. Tringe, Siv G. E. Andersson, Stefan Bertilsson, Rex R. Malmstrom, and Katherine D. McMahon. 2018. Contrasting patterns of genomelevel diversity across distinct co-occurring bacterial populations. *The ISME Journal* 12(3): 742–755.

2.1 Abstract

To understand the forces driving differentiation and diversification in wild bacterial populations, we must be able to delineate and track ecologically relevant units through space and time. Mapping metagenomic sequences to reference genomes derived from the same environment can reveal genetic heterogeneity within populations, and in some cases, be used to identify boundaries between genetically similar, but ecologically distinct, populations. Here we examine population-level heterogeneity within abundant and ubiquitous freshwater bacterial groups such as the acI Actinobacteria and LD12 Alphaproteobacteria (the freshwater sister clade to the marine SAR11) using 33 single-cell genomes and a 5-year metagenomic time series. The single-cell genomes grouped into 15 monophyletic clusters (termed "tribes") that share at least 97.9% 16S rRNA identity. Distinct populations were identified within most tribes based on the patterns of metagenomic read recruitments to single-cell genomes representing these tribes. Genetically distinct populations within tribes of the acI Actinobacterial lineage living in the same lake had different seasonal abundance patterns, suggesting these populations were also ecologically distinct. In contrast, sympatric LD12 populations were less genetically differentiated. This suggests that within one lake, some freshwater lineages harbor genetically discrete (but still closely related) and ecologically distinct populations, while other lineages are composed of less differentiated populations with overlapping niches. Our results point at an interplay of evolutionary and ecological forces acting on these communities that can be observed in real time.

2.2 Introduction

Bacteria represent a significant biomass component in almost all ecosystems and drive most biogeochemical cycles on Earth. Yet, we know little about the population structure of bacteria in natural ecosystems and have yet to find and define the boundaries for ecological populations. Cohesive temporal dynamics and associations inferred from distribution

patterns have been documented for many habitats and these observations are consistent with the notion of such populations as locally coexisting members of a species (Shapiro and Polz, 2014b). The most compelling cases are from collections of closely related isolates (Shapiro and Polz, 2014b; Hanage et al., 2005; Luo et al., 2011), but cultured species represent only a very small portion of the bacteria populating the Earth (Hug et al., 2016; Amann et al., 1995; Kaeberlein, 2002), and thus we still know little about the most abundant lineages. Therefore, it is critical to study microorganisms in their natural environments (Little et al., 2008), in order to test if and how their population-level heterogeneity differs from the established models based on isolates. The advent of culture-independent approaches, such as single-cell genomics and metagenomics, provides an opportunity for gaining new insights about genome-level diversity at the population level for organisms that are currently difficult or impossible to culture.

The delineation of ecologically differentiated lineages within complex microbial communities remains controversial because direct evidence for such differentiation is usually sparse (Hunt et al., 2008). Additionally, the appropriate level of phylogenetic resolution defining ecologically equivalent groups has not yet been established and likely varies across different groups (Fuhrman et al., 2015). Past explorations for defining such groups have used genome-wide average nucleotide identity (gANI) across shared regions of isolate genome sequences (Konstantinidis and Tiedje, 2005; Varghese et al., 2015). These studies have found that gANI greater than 94–96% unites past classical species definitions and separates known sequenced strains into consistent and distinct groups. Genetically distinct populations have been identified in microbial communities using metagenomics by mapping reads against reference genomes and noting a coverage gap at 90–95% identity (Kashtan et al., 2014; Konstantinidis and DeLong, 2008; Bendall et al., 2016; Oh et al., 2011; Caro-Quintero and Konstantinidis, 2012). Reads mapping with identities above the coverage discontinuity have been defined as originating from a "sequence-discrete population" (SDP) of genetically nearly identical cells that are distinct from other cells whose sequences

map with identities below the coverage discontinuity (Bendall et al., 2016). For the remainder of the manuscript, we will use the terms "population" and "sequence-discrete population" interchangeably.

We used a combination of time series metagenomics and single-cell genomics to define genetic diversification within ubiquitous and abundant freshwater lineages such as acI and tribe LD12. The term "tribe" was previously coined to delineate these groups using 16S rRNA gene sequences, where tribes are defined by monophyly and >97.9% within-clade 16S rRNA gene sequence identity (Newton et al., 2007, 2011). Freshwater microbial ecology researchers generally discuss and track these tribes as coherent units that are ecologically distinct from one another. A primary motivation for the present study was the challenge of moving beyond 16S rRNA sequence identity to delineate ecologically relevant taxonomic units given observed patterns of population-level heterogeneity, using shared genomic content. This study includes 33 single amplified genomes (SAGs) representing 15 phylogenetically coherent groups (i.e., freshwater "tribes").

The SAGs in this study originated from four lakes geographically isolated from one another and represent a rich source of reference genomes that can be used to recruit metagenomic reads in order to study population-level heterogeneity and dynamics through time in naturally assembled communities. Two of the lineages featured in the present study are the abundant and ubiquitous freshwater Actinobacteria acI and Alphaproteobactera alfV containing the freshwater SAR11 sister clade, tribe LD12. Members of these lineages are intriguing in their own right, as they represent groups of free-living ultramicrobacteria that dominate many freshwater ecosystems (Ghai et al., 2014; Glockner et al., 2000; Heinrich et al., 2013; Rösel et al., 2012; Salcher et al., 2010, 2011; Warnecke et al., 2005; Zwart et al., 2002). They differ markedly with respect to within-lineage diversity: LD12 is the sole tribe defined within the freshwater alfV lineage, while the acI lineage is comprised of 13 tribes (Newton et al., 2011). The acI and alfV are not easy to cultivate in monocultures (Kang et al., 2017) (though see ref. Henson et al. (2018b), published after this work but

prior to this thesis as Henson et al. (2018a)) and share a large number of genomic and cellular traits. First, both lineages have genomes with GC content values lower than 40% and estimated sizes of about 1.5Mb or less (Kang et al., 2017; Garcia et al., 2013; Ghylin et al., 2014; Zaremba-Niedzwiedzka et al., 2013; Eiler et al., 2016). These genome characteristics are all the more striking since most cultivated species in the Alphaproteobacteria and Actinobacteria have GC-rich genomes up to 10Mb in size. Second, both lineages have evolved by massive gene loss (Zaremba-Niedzwiedzka et al., 2013). Third, the fraction of gained genes is only about 10% of the lost genes. Fourth, both groups of bacteria have small cell volumes (Heinrich et al., 2013; Salcher et al., 2011). However, acI and alfV seem to employ different substrate niche specialization. While acI is thought to primarily use polyamines, oligopeptides, and carbohydrates, alfV specializes in carboxylic acids and lipids (Ghylin et al., 2014; Eiler et al., 2016; Salcher et al., 2013).

By combining genome information from 21 previously published (Ghylin et al., 2014; Zaremba-Niedzwiedzka et al., 2013) and 12 new SAGs from different freshwater lineages and an extensive 5-year time series of lake metagenomes (94 samples), we investigated the population-level heterogeneity of such ubiquitous freshwater bacteria for the first time. Our results confirm the existence of coherent sequence-discrete populations within these ubiquitous freshwater bacterial groups in natural communities and we could trace the abundance and gANI of these populations over monthly to seasonal time scales. Our work demonstrates the power of combining time series metagenomics and single-cell genomics for studying bacterial diversification and for describing ecologically meaningful population-level heterogeneity within communities inhabiting natural ecosystems.

2.3 Results

The SAG collection represents multiple clades within cosmopolitan freshwater lineages

We analyzed 33 SAGs from four different freshwater lakes. Twenty-one of these SAGs were previously analyzed for their genomic features and phylogenetic relationships (Garcia et al., 2013; Ghylin et al., 2014; Zaremba-Niedzwiedzka et al., 2013; Eiler et al., 2016). The 33 SAGs had total assembly sizes between 0.33 and 2.42Mbp and were organized into 8–103 contigs with GC contents between 29.1 and 51.7% (Table 2.3). Estimated genome completeness, calculated using two different methods, ranged between 30 and 99%. Throughout the paper, we will use mostly the shorter name version to facilitate reading, for example, M14 in place of AAA027-M14.

Genome name	Genome OID in IMG/MER	Phylum/ Class Abbr.	Tribe	NCBI Taxon ID	Lake Abbr.	Collection date (M/D/Y)	Assembly size (Mb)	Est. Comp. (dark matter) (%)	Est. Genome Comp. (checkm) (%)	Number of contigs	GC content (%)	Citation
AAA278-O22	2236661007	Actino	acI-A1	932044	Da	09/18/09	1.14	96.72	74.4	43	47.6	Ghylin et al. (2014)
AAA027-M14	2236661003	Actino	acI-A1	932041	Me	12/5/09	0.82	31.18	43.1	22	47.3	Ghylin et al. (2014)
IMCC25003	2602042019	Actino	acI-A1		So	13-Jun	1.35	NA	NA	1	49.1	Kang et al. (2017)
IMCC26103	2602042020	Actino	acI-A4		So	14-Apr	1.46	NA	NA	1	47.0	Kang et al. (2017)
AAA028-I14	2619618809	Actino	acI-A6	938457	Me	12/5/09	0.78	35.97	39.66	54	45.2	This paper
AAA044-N04	2236661005	Actino	acI-A7	932043	Da	04/28/09	1.29	80.74	79.59	23	45.6	Ghylin et al. (2014)
AAA041-L13	2519899769	Actino	acI-A7	932042	Da	04/28/09	1.38	63.95	74.14	103	44.2	This paper
AAA024-D14	2264265190	Actino	acI-A7	932039	Sp	05/28/09	0.78	72.74	48.4	82	45.4	Ghylin et al. (2014)
AAA023-J06	2236661001	Actino	acI-A7	932038	Sp	05/28/09	0.70	37.57	34.48	98	45.1	Ghylin et al. (2014)
IMCC19121	2602042021	Actino	acI-A7		So	11-Oct	1.51	NA	NA	1	45.5	Kang et al. (2017)
AB141-P03	2236876028	Actino	acI-B1	1053690	St	05/25/10	0.66	43.96	45.98	66	40.8	Ghylin et al. (2014)
AAA278-I18	2236661006	Actino	acI-B1	938557	Da	09/18/09	0.94	73.54	63.73	54	41.4	Ghylin et al. (2014)
AAA028-A23	2236661004	Actino	acI-B1	932036	Me	12/5/09	0.83	89.53	57.56	64	41.5	Ghylin et al. (2014)
AAA027-L06	2505679121	Actino	acI-B1	913338	Me	12/5/09	1.16	100.00	76.59	75	41.7	Garcia et al. (2013)
AAA027-J17	2236661002	Actino	acI-B1	932040	Me	12/5/09	0.97	87.93	65.26	81	42.1	Ghylin et al. (2014)
AAA023-D18	2236661009	Actino	acI-B1	932037	Sp	05/28/09	0.75	64.75	44.22	67	39.6	Ghylin et al. (2014)
AAA044-D11	2619618811	Actino	acI-B4	938518	Da	04/28/09	1.15	92.73	66.18	30	44.2	This paper
AAA027-D23	2524023172	Actino	acSTL-A1	938429	Me	12/5/09	0.94	44.76	44.01	18	48	This paper
AAA028-N15	2619618810	Actino	acTH1-A1	938467	Me	12/5/09	0.83	42.37	45.98	19	38	This paper
AAA027-G08	2619618806	Bacter	bacI-A1	938698	Me	12/5/09	1.32	63.95	59.36	36	35.5	This paper
AAA027-K21	2619618803	Beta	betIII-A1	938785	Me	12/5/09	1.38	67.95	42.1	21	51.5	This paper
AAA027-N21	2619618807	Bacter	Flavo-A2	938709	Me	12/5/09	2.21	100	92.44	36	33.1	This paper
AAA487-M09	2236347068	Alpha	LD12	938672	Da	09/18/09	0.63	60.75	53.15	97	29.1	Zaremba-Niedzwiedzka et al. (2013)
AAA280-P20	2236876029	Alpha	LD12	938665	Da	09/18/09	0.72	82.33	65.06	65	29.6	Zaremba-Niedzwiedzka et al. (2013)

Genome name	Genome OID in IMG/MER	Phylum/ Class Abbr.	Tribe	NCBI Taxon ID	Lake Abbr.	Collection date (M/D/Y)	Assembly size (Mb)	Est. Comp. (dark matter) (%)	Est. Genome Comp. (checkm) (%)	Number of contigs	GC content (%)	Citation
AAA280-B11	2236876032	Alpha	LD12	938663	Da	09/18/09	0.67	44.76	51.24	47	29.8	Zaremba-Niedzwiedzka et al. (2013)
AAA028-D10	2236347069	Alpha	LD12	938641	Me	12/5/09	0.93	75.14	81.64	57	29.6	Zaremba-Niedzwiedzka et al. (2013)
AAA028-C07	2236661008	Alpha	LD12	938639	Me	12/5/09	0.85	75.94	74.12	32	29.5	Zaremba-Niedzwiedzka et al. (2013)
AAA027-L15	2236876031	Alpha	LD12	938633	Me	12/5/09	0.72	53.56	68.68	56	29.4	Zaremba-Niedzwiedzka et al. (2013)
AAA027-J10	2236876030	Alpha	LD12	938631	Me	12/5/09	0.79	68.75	69.56	82	29.8	Zaremba-Niedzwiedzka et al. (2013)
AAA027-C06	2264265094	Alpha	LD12	938624	Me	12/5/09	0.78	83.13	82.29	90	29.6	Zaremba-Niedzwiedzka et al. (2013)
AAA024-N17	2236876027	Alpha	LD12	938623	Sp	05/28/09	0.33	19.18	30.19	45	30.1	Zaremba-Niedzwiedzka et al. (2013)
AAA023-L09	2236661000	Alpha	LD12	938615	Sp	05/28/09	0.77	58.35	68.1	76	29.4	Zaremba-Niedzwiedzka et al. (2013)
AAA028-K02	2619618804	Beta	LD28	938797	Me	12/5/09	0.56	31.18	34.48	8	37.5	This paper
AAA027-I06	2619618802	Beta	Lhab-A1	938781	Me	12/5/09	1.52	50.36	39.38	79	50.9	This paper
AAA027-I19	2619618805	Verruco	Opiput-aceae	939126	Me	12/5/09	2.42	55.96	54.58	63	51.7	This paper
AAA027-C02	2619618801	Beta	PnecC	938772	Me	12/5/09	1.27	58.35	61.93	49	43.7	This paper

Metadata for the 33 SAGs and genomes from ref. Kang et al. (2017) The Genome OID is the object identifier for the genome record in the Joint Genome Institute's IMG/MER Database. Estimated genome completeness was calculated using CheckM as described in the main text and (Parks et al., 2015). NA not applicable. Phylum/Class Abbrevations - Actino : Actinobacteria, Bacter : Bacteroidetes, Beta : Betaproteobacteria, Alpha : Alphaproteobacteria, Verruco : Verrucomicrobia. Lake Abbrevations - Da : Damariscotta, Me : Mendota, So : Soyang, Sp : Sparkling, St : Stechlin.

The 33 SAGs in the study represent 15 different previously defined freshwater "tribes" (that are each monophyletic and defined by >97.9% within-clade 16S rRNA gene sequence identity, measured across the nearly full-length 16S rRNA gene) (Newton et al., 2007, 2011). Ten tribes are represented by only one SAG each, while four tribes (LD12, acI-A1, acI-A7, and acI-B1) have more than one SAG representative in our data set. In addition to their classification based on 16S rRNA genes, the nine SAGs that were the only representatives of their lineage were classified using protein coding marker genes and PhyloSift (Darling et al., 2014) (Table A.4S1). To illustrate phylogenetic and taxonomic placement of the LD12 and acI SAGs, we used the PhyloPhlAn pipeline (Segata et al., 2013) to generate a multi-gene tree (Fig. 2.1a, b). The tree topology was consistent with previous phylogentic reconstructions for LD12 (Zaremba-Niedzwiedzka et al., 2013) and acI (Newton et al., 2007; Ghylin et al., 2014). The tree supported the 16S rRNA gene-based tribe designations but did not reveal a clear biogeographic pattern, in agreement with previous analyses, i.e., members of the same tribes were found in different lakes (Zaremba-Niedzwiedzka et al., 2013). However, our SAG collection was not designed to explore biogeography and much deeper sampling of each population would be needed to address this question rigorously.

Genome-wide nucleotide identity is consistent with phylogeny

Although multi-locus phylogenies supported the 16S rRNA gene-based phylogeny, we wondered whether gANI could similarly be used to demarcate one tribe from another. To this end, we determined the pairwise gANI for genomes in the set of four tribes that each contained more than one SAG representative. This general approach has been proposed as a way to compare genome pairs using a single metric that robustly reflects phylogenetic and taxonomic groupings obtained using other polyphasic methods (Konstantinidis and Tiedje, 2005; Varghese et al., 2015). We asked whether all genome pairs from the same tribe shared a consistent minimum gANI. Most SAGs shared gANI of at least 78% and alignment fractions greater than 40% with other members of the same tribe (Fig. 2.1c and Table A.4S2).

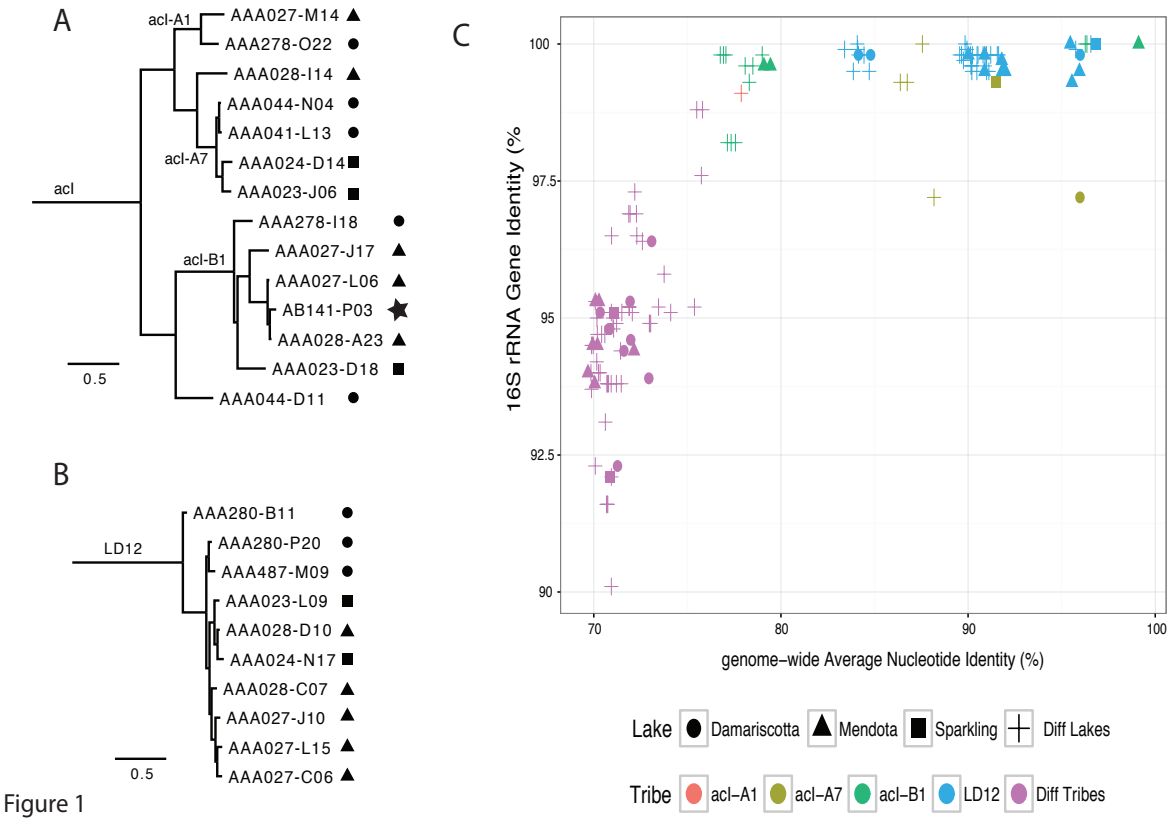

Figure 1

Figure 2.1: Phylogenetic and sequence identity relationships between SAGs. a Phylogenetic tree of acI SAGs based on conserved single copy genes selected by PhyloPhlAn. Amino acid sequences from 400 genes were aligned. The tree topology is consistent with 16S rRNA gene-based phylogenies (Ghylin et al., 2014). SAGs L06 and A23 are part of the same sequence-discrete population (SDP) as defined in the text and further based on data shown in Fig. 2. b Phylogenetic tree of LD12 SAGs based on conserved single copy genes selected by PhyloPhlAn, representing 400 genes. The tree topology was consistent with prior work that provided evidence for finer-scale groups within the LD12 tribe (Zaremba-Niedzwiedzka et al., 2013). c Genome-wide nucleotide identity (gANI) vs. 16S rRNA gene identity for pairs of SAGs. Alignment fractions for homologous genomic regions and 16S rRNA genes are given in Table A.4S2. Shapes indicate the lake the tribe is from, if same, otherwise different lake is indicated. Colors indicate the tribe a pair is from, if same, otherwise different tribe is indicated. The arrow denotes the L06–A23 pair.

Most pairs from the same tribe that were also recovered from the same lake shared at least 84% gANI, but some pairs were much more similar (gANI above 95%). gANIs between pairs belonging to different tribes but still within the same lineage were markedly lower and typically below 74% (e.g., acI-A1 vs. acI-B1) (Fig. 2.1c and Table A.4S2).

Although gANI is a useful univariate metric for comparing genome pairs, it masks the differences in sequence similarity of individual genes or genome regions that arise due to varying rates of divergence across loci. This variation can be visualized by plotting the frequency distribution of nucleotide identities calculated using a sliding window across the genome (Konstantinidis and Tiedje, 2005). We asked whether different homologous genomic regions from two SAGs would have markedly different nucleotide identities even if they were from the same tribe. We used the most complete SAGs from the acI-B1 and LD12 tribes as reference genomes and calculated nucleotide identity using a sliding window with other SAGs from the same respective tribe and visualized the results as a frequency distribution (Fig. 2.2 and Fig. A.3S1). The acI-B1 SAGs featuring the highest gANI (L06 and A23) were both from Lake Mendota and shared nucleotide identity consistently greater than 95% with a peak at 99–100%, suggesting they belong to the same SDP. The acI-B1 SAG P03 recovered from a lake in Germany had a frequency distribution with a peak more near 97% and a distinctly different shape. Other acI-B1 SAGs shared genomic regions with primarily 80–85% nucleotide identity. This was even true for J17, which was also collected from Lake Mendota and shared an average gANI of 79% with L06/A23 (Table A.4S2), suggesting that cells belonging to the same tribe (acI-B1) and living in the same environment can have substantial genetic differences. The LD12 SAGs, which all belonged to the same tribe, also displayed three distinct patterns, with one peak near 85%, several near 91%, and two near 97%. Lake origin did not appear to explain these differences. That is, some LD12 cells from Lake Mendota were more similar to LD12 cells from Sparkling Lake than to other LD12 cells from Lake Mendota.

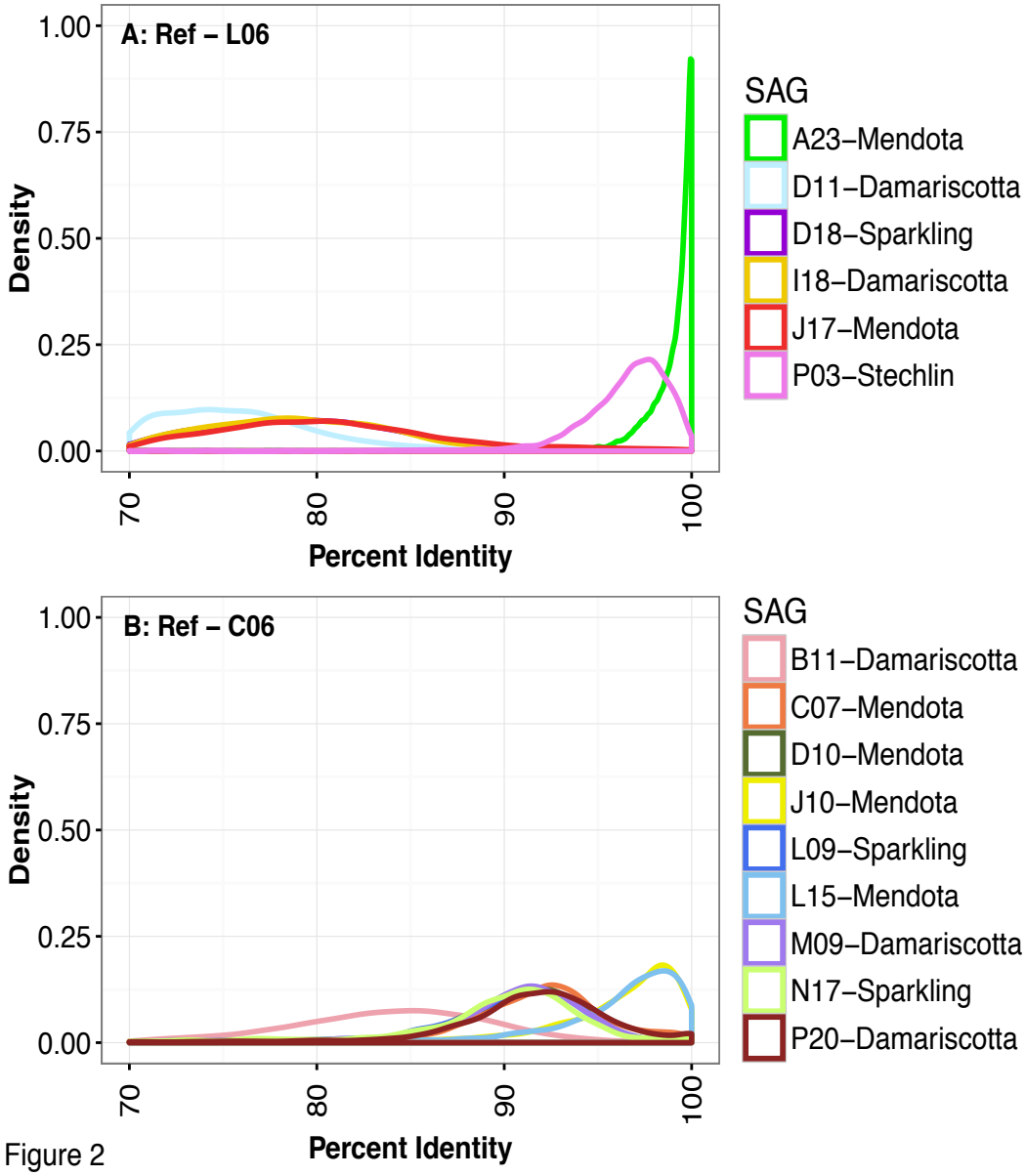

Figure 2.2: Nucleotide identity density plots for SAG vs. SAG genome-wide comparison using a sliding window. Results are shown for two reference SAGs representing the most complete genomes from the most thoroughly sampled tribes. All SAG pairs were from the same tribe. Nucleotide identity was calculated with blastn using 301bp fragments that overlapped by 150bp. a acI-B1 SAGs and other selected acI SAGs vs. L06. Note that the purple line (D18) is hidden underneath the orange (I18) and red (J17) lines. b Selected LD12 SAGs vs. C06. Note the dark blue line (L09) is hidden under the light green (N17) line. Group designations match those shown in Fig. 1b, as proposed previously (Zaremba-Niedzwiedzka et al., 2013). An expanded multi-panel version of the same data is shown in Fig. A.3S1, for clarity.

Diversity of wild populations inferred using SAGs

The variety of patterns observed in Fig. 2 indicated substantial within-tribe variability even among cells recovered from the same lake. This made us wonder if tribes were composed of genetically and ecologically distinct populations coexisting in the same environment. SAGs can serve as relevant reference points to study the diversity of abundant populations sampled using shotgun metagenomics by recruiting metagenomic reads and examining the extent of nucleotide identity for each aligned read (Stepanauskas, 2012). The results can also be used to identify sequence-discrete populations whose boundaries are revealed by recruitment patterns and specifically the dramatic drop in coverage observed around 95% sequence identity (Konstantinidis and DeLong, 2008; Bendall et al., 2016; Caro-Quintero and Konstantinidis, 2012). We asked whether such SDPs could be identified using metagenomic reads from Lake Mendota, WI, USA, by mapping them to the 33 SAGs, 19 of which were collected from this lake.

Each of the SAGs was first used to recruit reads from a single metagenomic data set collected from Lake Mendota on 29 April 2009 (Fig. A.3S2). This time point was chosen because it was the sample collected closest to the date on which the single cells were collected (12 May 2009). Frequency distribution plots of the same data (Fig. 2.3 and Fig. A.3S3) revealed patterns that were similar to those obtained with SAG pairs (Fig. 2.2). The five acI-SAGs from Lake Mendota (J17, L06, A23, M14, and I14) recruited more reads than the acI-SAGs from other lakes, with many reads recruiting at nucleotide identity greater than 97.5% (Fig. 2.3). All of the acI-SAGs also recruited many reads at 60–90% identity (Fig. 2.3), creating the characteristic bimodal distribution observed in previous work Caro-Quintero and Konstantinidis (2012). Based on these results, we hereafter consider reads sharing >97.5% nucleotide identity as coming from the same, operationally defined population (i.e., SDP) as the reference SAG. Thus, the acI lineage in Lake Mendota on 29 April 2009 was composed of multiple SDPs. Interestingly, the acI-B1 tribe in Lake Mendota, a subset of the acI lineage, appeared to be composed of at least two coexisting

and genetically distinct populations, one represented by SAG J17 and the other by SAGs A23 and L06, consistent with the pairwise gANI observed using only the SAGs (Fig. 2.2).

To determine if we recovered representative SAGs from all acI populations in Lake Mendota, we next performed recruitments competitively, allowing each read to only map to the SAG with the greatest percent identity (Fig. A.3S4). Since the patterns in Fig. 2.3 were generated by non-competitive recruiting, some reads mapping with 100% identity to one SAG might for example also have mapped with 60–90% identity to SAGs from different SDPs. Under competitive recruiting conditions, the resulting frequency distributions changed and the fraction of reads recruiting with 60–90% identity to each acI SAG dropped dramatically (Fig. A.3S4). However, a secondary peak around 80% identity still remained in most cases, and it is possible these reads originated from cells belonging to other acI populations lacking a representative SAG.

LD12 SAGs collected from Lake Mendota (C06, J10, L15, C07, and D10) also had a distinctive peak of recruited reads at >97.5% sequence identity (Fig. 2.3), although the overall shape of the recruitment patterns differed dramatically from those of the acI lineage. For example, LD12 SAGs had a secondary recruitment peak at ~92% identity, whereas the acI SAGs had secondary peaks at ~75% with non-competitive recruiting (Fig. A.3S4). This suggests the SDPs within the LD12 tribe were more similar genetically than populations comprising the acI-B1 tribe. In fact, the populations were sufficiently similar that the hallmark coverage discontinuity below 97% similarity was not particularly pronounced (Fig. 2.3). Under competitive recruiting conditions, the LD12 recruitment distribution plots had remarkably different shapes (Fig. A.3S4B, D), as compared to the uncompetitive recruiting conditions (Fig. 2.3), and each SAG had only a single peak at >97.5% identity. This suggests the majority of LD12 cells in Lake Mendota belong to SDPs represented by the SAGs in our collection.

All but one (I06) of the other freshwater SAGs in this study that were collected from Lake Mendota generated the distinctive read recruitment frequency peak above 97.5% identity

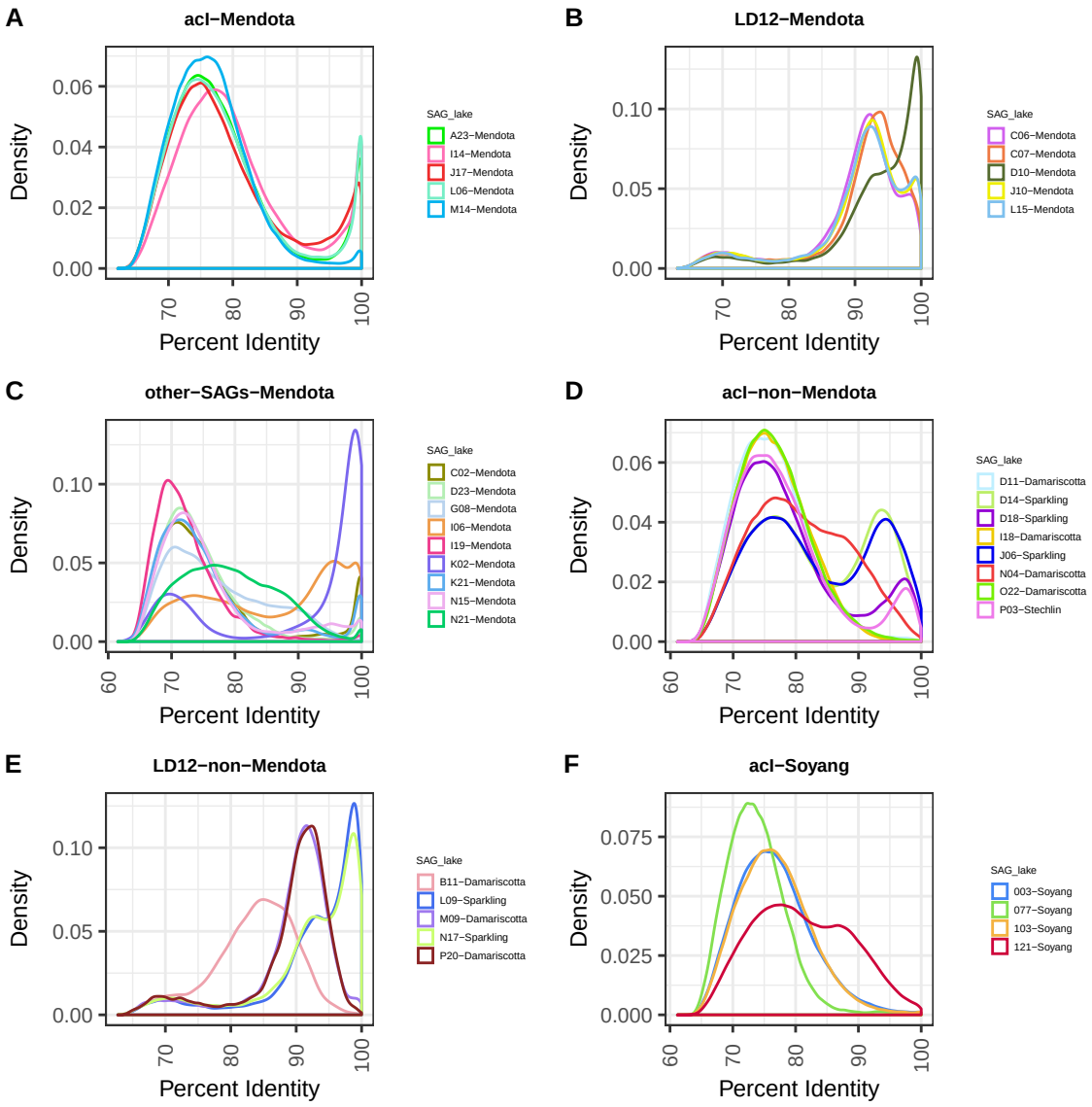

Figure 2.3: Mapping metagenomic reads from Lake Mendota to SAGs and four genomes from Lake Soyang. (Kang et al., 2017). The x-axis represents nucleotide identity of the recruited reads. The metagenome sample was collected from Lake Mendota on 29 April 2009. Reads were only counted if they aligned over a minimum of 200bp. Recruitments were not competitive, meaning that each read could recruit to multiple SAGs. Analogous competitive recruitments that required each read to recruit to only one SAG are presented in Fig. A.3S4. The non-competitive recruitment showed the close relationship of the LD12 populations that is not visible in the competitive recruitment. An expanded multi-panel version of the same data is presented in Fig. A.3S3 for clarity. Each panel represents a different subset of the SAGs: *a* acI from Mendota, *b* acI not from Mendota, *c* LD12 from Mendota (group members demarcated in legend), *d* LD12 not from Mendota (group members demarcated in legend), *e* other freshwater groups from Mendota, f genomes from Lake Soyang, Korea. Regarding the other freshwater groups from Mendota, since each of these SAGs represent just one tribe, it is not appropriate to infer any general conclusions for these populations or tribes, but we present them here to show the intriguing diversity of recruitment patterns. We finally underscore the need to more deeply sample individual population members using SAGs, to better capture and describe the range of variation in population heterogeneity.

(Fig. 2.3) that was observed for acI (Fig. 2.3). A negligible number of reads recruited to the SAGs collected from other lakes under the competitive recruiting conditions (data not shown).

Four complete acI genomes recovered from Lake Soyang in Korea were recently published, and we included these in our recruitment analysis (Fig. 2.3). Three of the SAGs exhibited recruitment frequency distributions analogous to those obtained using acI SAGs from Sparkling Lake and Damariscotta Lake (Fig. 2.3), with very few reads mapping above 90% ANI. The distribution from one SAG (IMCC19121) was remarkably similar to that obtained from SAG N04, which was recovered from Damariscotta Lake in Maine. Both IMCC19121 and N04 are members of the acI-A7 tribe and share 89.8% ANI with each other.

Are sequence-discrete populations within a tribe ecologically discrete too?

Results from a single metagenome sample suggested that individual tribes were composed of multiple genetically distinct populations that could be delineated and tracked using metagenomic read recruitment. Next, we hypothesized that these populations might also be ecologically distinct and fill different realized niches. If so, we might expect these populations to display different temporal abundance patterns. We followed changes in population abundance through time by recruiting reads from a 5-year metagenomic time series applying a nucleotide identity cutoff of 97.5%, using only those SAGs derived from Lake Mendota. SAGs from the LD12 tribe recruited more reads than all of the acI SAGs summed together, on almost all sample dates (Fig. A.4S5).

Using the relative number of reads recruited as a proxy for abundance, we found the J17 population, which belonged to the acI-B1 tribe, to be the most abundant acI population in almost every sample (Figs. 2.4a and 2.5a). The abundance of the J17 population was poorly correlated over time with the other acI-B1 population represented by L06 (maximum Spearman rank correlation=0.256), indicating each population had a different temporal

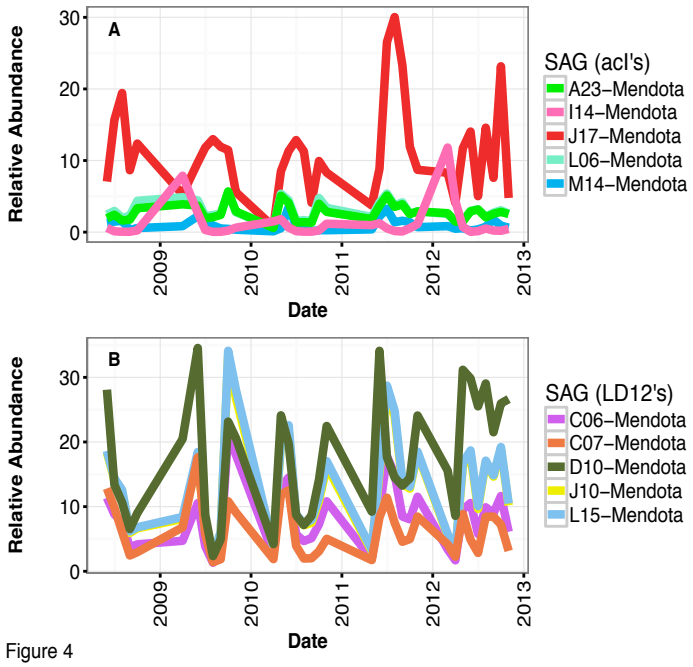

Figure 2.4: Sequence-discrete population abundance in Lake Mendota over time, as measured by the relative number of reads recruited to each SAG using blastn. All SAGs and samples are from Lake Mendota. Timepoints are pooled by month. Filtering criteria: 97.5% ANI and 200bp alignment length. Recruitment was done using the most strict definition of competitive described in the methods, meaning any read that matched equally well to more than one SAG was not counted at all. Colors for each SAG are the same as in Figs. 2 and 3. Relative abundance was calculated by normalizing the number of basepairs that recruited to each SAG by dividing by the genome size and the pooled metagenome size. The normalized number was then multiplied by the average pooled metagenome size. a Relative abundance for each acI-B1 SAG. b Relative abundance for each LD12 SAG. Membership in the groups defined in Fig. 1b and by ref. (Zaremba-Niedzwiedzka et al., 2013) are denoted in the legend

abundance pattern.

In contrast to the acI-B1 tribe, the populations comprising the LD12 tribe had highly similar abundance patterns (Fig. 2.4b and Fig. A.4S6). The abundances of J10, L15, and C06 populations were strongly correlated (Spearman rank correlation=0.996–0.999) (Fig. A.3S8 and Table A.4S5) and tended to peak both in Spring and Fall (Fig. A.3S6). The D10 population was the most abundant in the data set but its abundance was not as strongly correlated to the other LD12 populations (Spearman rank correlation=0.705–0.725) (Fig. A.3S8 and Table A.4S5). The C07 population was the least abundant but was also

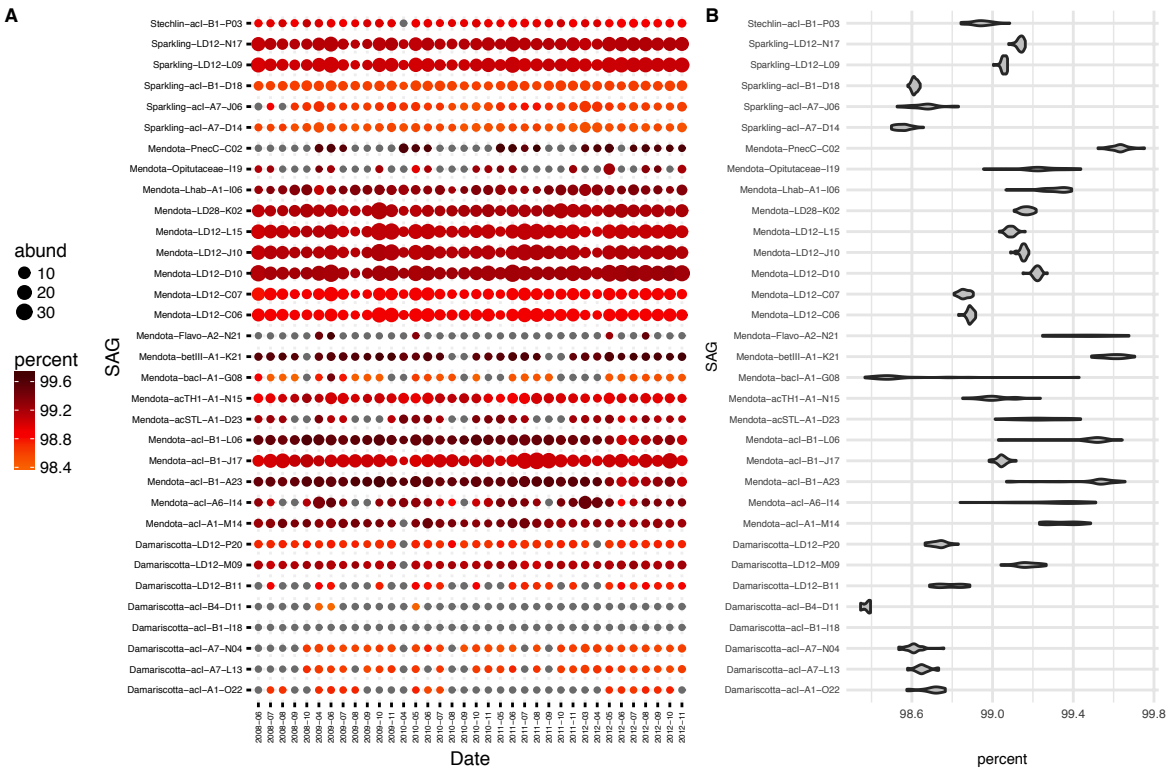

Figure 2.5: Abundance and ANI for SAGs through timeseries. *a* Metagenomic read recruitment using the SAGs from Lake Mendota. SAGs are in rows with bubbles representing all metagenomes from a particular month recruited against SAG. Filtering criteria: 97.5% ANI and 200bp alignment length. Color scale indicates the ANI of the recruited metagenome reads. Bubble size represents the average coverage per base in the reference SAG divided by the size of the metagenome, multiplied by the average size of all metagenomes (1.34 Gigabases). Gray bubbles indicate that fewer than 200 reads recruited to the SAG in that month. Note that the resulting values do not represent a true measure of absolute abundance, but allow for quantitative comparison of month-to-month variation in population-level abundance. Recruitments were performed competitively, meaning that each read was counted for only one SAG, unless the read hit two SAGs equally well in which case it was counted for both SAGs. *b* Variation in ANI for each SAG, across all 30 metagenomes from throughout the five years. Variation was not calculated for a SAG unless at least 10 months recruited more than 200 hits each. The data underlying these plots can be found in Table A.4S6

correlated to both the J10-L15-C06 populations and the D10 population (Spearman rank correlation=0.850–0.873).

Does the genetic diversity of populations change over time?

We also examined the extent to which within-population diversity varied through time by quantifying changes in population-wide ANI, i.e., the average identity of all reads mapping with at least 97.5% identity (Fig. 2.5b). For this purpose, we only recruited reads to SAGs recovered from Lake Mendota. More abundant populations (such as LD12 and acI-B1 J17) generally had lower population-wide ANI variance through time compared to some less abundant populations (such as acSTL-A1-D23 and acI-A6-I14). For example, the SAG bacI-A1 G08 population had relatively high population-wide ANI in June 2009, around the time when the sample was collected for SAG library collection, but had markedly lower ANI on all other dates. One interesting exception to this observation was a significantly lower ANI for the relatively abundant acI-B1 L06–A23 population in 2012, as compared to 2007–2011 (Mann–Whitney U-test; p=1.4e-06).

2.4 Discussion

Comparative genomics can reveal the diversity and structure of bacterial populations. This approach is particularly powerful when applied using single cells recovered from environmental samples (SAGs) and shotgun metagenomes from the same or similar ecosystems. Here we used a combination of 33 SAGs and 94 metagenomes collected over 5 years to ask the following questions: (1) How well do individual SAGs represent the population-level diversity found in natural communities? (2) Do common freshwater bacterial groups have similar patterns of population-level diversity? and (3) How stable is population abundance and diversity through time? We used the answers to these questions to gain insight into the population-level diversity and ecology of the cosmopolitan and abundant freshwater

bacteria, alfV-LD12 (Alphaproteobacteria) and acI (Actinobacteria).

Sequence-discrete populations could be delineated in the Lake Mendota metagenome using our 33 SAGs as references, as has previously been demonstrated in other lakes using genomes assembled from metagenomes (Bendall et al., 2016; Caro-Quintero and Konstantinidis, 2012). We interpret the occurrence of these populations in the context of previously defined phylogenetically coherent and ostensibly ecologically distinct "tribes" composed of cells with >97.9% 16S rRNA identity (Newton et al., 2011). We conclude that the freshwater tribes can contain multiple sequence-discrete populations. The converse is, of course, not true: sequence-discrete populations can never represent multiple tribes because tribes are by definition more distantly related to one another than genomes sharing a minimum of 97.5% gANI.

Pairwise gANI analysis of SAGs and metagenomic read recruitment indicated that cells belonging to the same tribe but inhabiting different lakes were usually genetically distinct. For example, SAGs collected from other lakes generally recruited very few reads from Lake Mendota at ANI >97.5% while many recruited a substantial number of reads in the 89–92% range (Fig. 2.3). However, there were two prominent exceptions: LD12 N17 and L09, both of which are from Sparkling Lake. N17 and L09 share 97% gANI with Mendota SAG D10, which is substantially higher than the average (88%) and median (90%) within-tribe gANI (Table A.4S2). These SAGs also recruited roughly the same number of reads with >97.5% identity as did the LD12 SAGs from Lake Mendota, though around 17% (L09) and 23% (N17) of the base pairs in the genomes did not recruit any reads. This implies that some gene content was present in the Sparkling Lake populations but missing in Lake Mendota. However, 10% of the base pairs in the D10 genome also did not recruit any reads, even though it was from Lake Mendota. We examined the phylogenetic distribution of low-coverage contigs and did not discern any evidence of contamination. This rare genome content could represent flexible or low frequency genes in the population, or contamination in the SAG preparation (Blainey, 2013). However, it could also represent

systematic coverage bias, a phenomenon that we are not able to rule out with the data at hand.

In Lake Mendota, acI cells are organized into genetically discrete populations, but the forces creating this organization remain unknown. The consistent lack of coverage around 90–97% identity in recruitment plots indicates Lake Mendota lacks acI genotypes sharing this degree of sequence similarity with our SAGs, or at least that these putative genotypes were consistently at much lower abundances than their close relatives over the 5 years surveyed. The P03 SAG from Stechlin Lake shares gANI of 96% with acI-B1 SAGs from Mendota, indicating that genotypes within this locally excluded sequence space do exist, at least as long as they are from different environments. We infer the persistence of the coverage discontinuity between populations to be less a factor of dispersal limitation and more likely the result of competitive exclusion and barriers to recombination within Mendota populations. Additional SAG and metagenomic studies are necessary to determine if similar coverage discontinuities are observed in other phylogenetic groups and in different environments. However, we do note that others have observed similar population-level diversity in other lakes (Bendall et al., 2016; Caro-Quintero and Konstantinidis, 2012) and marine ecosystems (Konstantinidis and DeLong, 2008).

We know that both acI tribes and LD12 vary in abundance over seasonal and annual time scales, based on previous work using 16S rRNA gene sequencing, quantitative PCR, and FISH (Heinrich et al., 2013; Salcher et al., 2011; Allgaier and Grossart, 2006; Eiler et al., 2012). Here we used our SAGs to track such populations at monthly intervals over 5 years (Fig. 2.4 and Fig. A.3S5). The results confirmed prior work that showed acI tribes and LD12 are among the most abundant non-cyanobacterial groups in Lake Mendota (Newton et al., 2011), but also revealed dynamics at unprecedentedly high phylogenetic resolution. Based on our extensive comparison of how SAGs recruited relative to one another, we are confident that our metagenomic recruitment filters allowed us to delineate discrete populations that would not be possible to resolve using more traditional and widely used

methods (e.g., 16S rRNA gene sequencing or FISH). However, we do note that our acI SAG collection to date does not seem to fully capture the full diversity of acI populations in the lake, as evidenced by the residual peak of reads matching our SAGs at ~70–80% ANI, even under competitive recruiting conditions. For example, we roughly estimate that our acI SAGs captured only 12% of the resident acI metagenome on 29 April 2009, as compared to 50% of the LD12 metagenome (Table A.4S3). Thus, we cannot completely rule out the possibility that we missed strong correlations among other acI populations that we could not detect.

However, the most striking finding of our study was that metagenomic recruitments to LD12 SAGs yielded dramatically different patterns compared to the acI lineage. We discovered that LD12 populations were not as strongly genetically separated as acI populations; pairwise gANIs between SAGs were higher and recruitment plots showed secondary peaks between 90 and 95% identity (Fig. 2.3), the same range where coverage of acI SAGs was at a minimum (Fig. 2.3). Under a competitive recruitment analysis, wherein each read is counted only once and attributed to the best match SAG, the secondary peaks disappear (Fig. A.3S4), indicating the LD12 SAGs represent highly similar, but still genetically discrete, populations. Temporal abundance patterns of these LD12 populations were strongly correlated over 5 years, whereas acI populations showed much lower correlation within tribes (Fig. A.3S8). This suggests that the acI-B1 populations are ecologically distinct (i.e., occupying temporally discrete niches) while LD12 populations are less differentiated with respect to niche dimensions, leading to co-occurrence and synchronization of temporal abundance patterns. LD12 is a particularly fascinating group because it is also a subclade of the broader SAR11 clade, with hypothesized ancient transition from marine to freshwater (Logares et al., 2010) followed by specialization through gene flux and mutation, with comparatively low recombination rates (Zaremba-Niedzwiedzka et al., 2013). Over time, low recombination rates and relatively low selection levels should lead to large genetic divergence among coexisting populations. Thus, we propose that LD12 populations are

simply at earlier stages of differentiation as compared to acI populations, although we cannot exclude that something fundamental about their lifestyle is "holding" the populations together genetically and ecologically. This is particularly interesting in light of recent reports of unusually high recombination rates in LD12 (Zaremba-Niedzwiedzka et al., 2013), pointing to the need to further investigate contrasting population structures and what these structures mean for the ecophysiology of the organisms. We do note that it is also possible that the highly correlated LD12 populations are each occupying unidentified distinct niches that are unrelated to the temporal correlation, allowing these slightly genetically differentiated populations to co-occur while being ecologically distinct. In any case, the lack of coherence among acI-B1 populations challenges our concept of tribes as ecologically coherent units and suggests that freshwater microbial ecologists re-examine conventions for tracking these units through space and time. Taken together, these observations suggest fundamental differences in evolutionary history and/or lifestyles among these abundant and ubiquitous freshwater bacteria.

The metagenomic recruitments allowed us to also examine the extent to which diversity varied within and among populations as well as how diversity changed over time. We calculated the population-wide ANI for reads that recruited only above 97.5% and found the resulting value was remarkably stable through time for most of the abundant populations (Fig. 2.5b). This was particularly true for the LD12 populations. However, one striking contrast was the acI-B1 population represented by L06/A23, which had consistent population-wide ANI of 99.3% during 2008–2011, but 99.0% during 2012 (Mann–Whitney U-test p=1.4e06). Similar shifts were observed previously in sequence-discrete populations inhabiting Trout Bog Lake, indicating this could be a common phenomenon among freshwater clades (Bendall et al., 2016). Unlike the genome-wide selective sweep observed in one Chlorobium population from Trout Bog Lake, the distribution of single nucleotide polymorphisms within the L06/A23 population before and after 2012 exhibited no clear pattern of gene- or genome-wide sweep (data not shown). That is, it seems that the increase

in population-wide gANI resulted in a change in the relative abundance of individual genotypes, rather than a single new genotype overtaking the population. It is difficult or impossible to separate genotypes within sequence-discrete populations using short-read shotgun sequencing, so further work using long-read technologies will be needed to link SNPs in populations to individual genomes. This kind of approach will likely be required to tease apart the paths leading to diversification within and among populations.

2.5 Methods

Single amplified genomes (SAGs)

Water samples (1ml) were collected from the upper 0.5–1m of each of four lakes (Mendota, Sparkling, Damariscotta, and Stechlin) and cryopreserved, as previously described (Garcia et al., 2013; Martinez-Garcia et al., 2012). These lakes were originally selected because they represent different freshwater trophic status (eutrophic, oligotrophic, mesoeutrophic, and oligotrophic, respectively) and geographic regions (Wisconsin and Maine, USA, and Germany). Bacterial SAGs were generated by fluorescence-activated cell sorting (FACS) and multiple displacement amplification (MDA), and identified by PCR-sequencing of their 16S rRNA genes at the Bigelow Laboratory Single Cell Genomics Center (SCGC; http://scgc.bigelow.org). Thirty-two SAGs from lakes Mendota, Sparkling, and Damariscotta were selected for sequencing based on the previously sequenced 16S rRNA gene as well as the kinetics of the MDA reactions (Martinez-Garcia et al., 2012). The one SAG from Lake Stechlin was selected from a separate library because its 16S rRNA gene was 100% identical to an acI-B1 SAG previously analyzed (AAA027-L06) (Garcia et al., 2013). In the present study, we analyze 21 previously published and 12 new SAGs. All 33 SAGs were analyzed (Table 2.3) after genome sequencing, assembly, contamination removal, and annotation as previously described (Ghylin et al., 2014). Estimation of completeness was done using CheckM (Parks et al., 2015) and the gene markers from a recent study examining a large

collection of draft environmental genomes (Rinke et al., 2013).

Tree construction, average amino acid and average nucleotide identity (AAI, ANI)

A phylogenomic analysis was conducted using PhyloPhlAn (Segata et al., 2013). ANI was calculated by using the method described in ref. Konstantinidis and Tiedje (2005) with fragment size of 1000, minimum alignment length of 700bp, percent identity of 70, and e value of 0.001. AAI was calculated by averaging the identity of the reciprocal best hits from the BLASTP searches of the predicted proteins for each pair of genomes. 16S rRNA gene similarity for each pair was calculated using the overlapping region in an alignment created using a multiple alignment (default options) in Geneious Version R6 (Kearse et al., 2012). Additional classifications were carried out using PhyloSift version 1.0.1, which examines 37 conserved single copy marker genes and places them into a phylogenetic reference tree (Darling et al., 2014).

SAG-to-SAG recruitments

SAG pairs from the same tribe were used to examine the frequency distribution of nucleotide identities across homologous regions of the two genomes. In order to create a sliding window for comparison, the contigs of all SAGs were shredded into 301bp fragments with 150bp overlap. Two SAGs were selected as reference genomes: L06 as the most complete from the tribe acI-B1 and C06 as the most complete LD12. The contigs of each of the two selected SAGs were used as a reference for recruiting from the shredded SAGs using Blast 2.2.28 (Camacho et al., 2009). Ribosomal RNAs were masked from the SAGs prior to performing blast.

Five-year time series metagenome data: sampling, sequencing, and recruitments

Samples were collected from Lake Mendota over the course of 5 years, as previously described (Kara et al., 2013; Shade et al., 2007). Lake Mendota, Madison, Wisconsin, (N 43°06, W 89°24) is one of the most well-studied lakes in the world, and is a long-term ecological research site affiliated with the Center for Limnology at the University of Wisconsin Madison (Carpenter et al., 2006). It is dimictic and eutrophic with an average depth of 12.8m, maximum depth of 25.3m, and total surface area of 3938ha. Depth-integrated water samples were collected from 0 to 12m of the epilimnion (upper mixed layer) at 94 different time points during ice-free periods from summer 2008 to summer 2012, and filtered onto 0.2micrometer pore-size polyethersulfone filters (Supor, Pall) prior to storage at 80°C. DNA was later purified from these filters using the FastDNA kit (MP Biomedicals). DNA sequencing was performed at the Department of Energy Joint Genome Institute using standard protocols (Walnut Creek, CA, USA). DNA from the 94 samples was used to generate libraries that were sequenced on the Illumina HiSeq 2000 platform. Paired-end sequences of 2150bp were generated for all libraries. Adapter sequences, low-quality reads (i.e., 80% of bases had quality scores <20), and reads dominated by short repeats of 3bp were removed. The remaining high-quality reads were merged with the fast length adjustment of short reads (Magoc and Salzberg, 2011) with a mismatch value of 0.25 and a minimum of 10 overlapping bases from paired sequences, resulting in merged read lengths of 150–290bp (Table A.4S4). Metagenomes were pooled by month to reduce the time series data to 30 observations and increase coverage. Original records can be found as a group in JGI's Genome Portal: http://genome.jgi.doe.gov/Mendota_metaG.

All contigs from each of the 33 SAGs were used as a reference to recruit reads from the Mendota metagenomes using blastn. Metagenome reads that recruited to the SAGs were filtered and only alignments 200bp or longer were considered. An additional filter requiring an alignment percent identity of at least 97.5% was applied when analyzing

the metagenome time series. Ribosomal RNAs were masked from the SAGs prior to performing the recruitments. Relative abundance was calculated by normalizing the number of basepairs that recruited to each SAG by the genome and pooled metagenome size and multiplying all by the average pooled metagenome size. When appropriate to the research question, recruitment was conducted "competitively," meaning that if a read recruited to more than one SAG it was only counted for the best hit SAG. In this case, if a read recruited equally well to both SAGs, it was counted for both. In some cases, we applied an even stricter definition of "competitive" and did not count any read that recruited equally well to more than one SAG. For Fig. 2.3, recruitment was conducted "non-competitively," meaning that reads could be counted for multiple SAGs as long as the hits met the filtering criteria. We note that this a commonly used approach developed by other researchers (Konstantinidis and Tiedje, 2005; Konstantinidis and DeLong, 2008). The figure and table legends contain the information necessary to discern which kind of recruitment criteria were applied for that specific analysis.

Statistics, visualization, and reproducible methods

Data sets were analyzed and results were visualized using custom scripts written in R (R Core Team, 2014) and python. Pipeline and scripts for analysis can be found at https://github.com/McMahonLab/blast2ani.

2.6 Acknowledgments

We thank Dr Todd Miller and Sara Yeo for collecting the original water samples used to retrieve single cells from Lake Mendota and Sparkling Lake. We thank the Joint Genome Institute for supporting this work through the Community Science Program, performing the bioinformatics, and providing technical support. We thank Moritz Buck for informatics and statistical support. The work conducted by the U.S. Department of Energy Joint Genome

Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02–05CH11231. This research was performed using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. KDM acknowledges funding from the United States National Science Foundation (NSF) Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-0822700), an INSPIRE award (DEB-1344254), and the Swedish Wenner-Gren Foundation. RS acknowledges funding from NSF (DEB-0841933, EF-0633142, and OCE-821374). SB acknowledges funding from the Swedish Research Council. SLG thanks and acknowledges the JSMC for funding. MMG acknowledges funding from Ministry of Economy and Competitiveness (CGL2013–405064-R and SAF2013-49267-EXP).

3 POPULATION AND GENE DYNAMICS OF POLYNUCLEOBACTER POPULATIONS THROUGH A METAGENOMIC TIME-SERIES

Sarah L.R. Stevens[1], Cristina M. Herren[2], Jeff Froula[3],
Rob Egan[3], Torben Nielsen[3], Rex R Malmstrom[3] and Katherine D McMahon[1,4]

[1]*Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA*; [3]*Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, USA*; [3]*DOE Joint Genome Institute, Walnut Creek*; [4]*Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA*

SLRS, CMH, RM, and KDM conceived the research. SLRS, CMH, RM, and KDM analyzed the data. SLRS prepared the figures. SLRS and KDM wrote the manuscript.

In Prep for Submission

3.1 Abstract

While we can identify and track genetically discrete populations through time, we know little about how genes within these populations change in response to fluctuating conditions in natural ecosystems. To understand how the diversity and frequency of genes within a population changes through time, we recovered metagenome-assembled genomes (MAGs) from 6 distinct *Polynucleobacter* populations and used them as references for a 45-sample metagenomic time series spanning 3 years in a bog lake. This allowed us to track the diversity and abundance of individual sequence-discrete populations and the frequency of genes within the population. At the population level, we found that while all of the *Polynucleobacter* populations were detected in each year sampled, one population was considerably more abundant throughout the time series. We also found that none of the populations were dominated by a single or small number of strains through time, based on the fraction of single nucleotide variants (SNVs) dominated by a single allele, termed SNV homogeneity. Based on gene frequency, we characterized the core and accessory genes for each of the populations and investigated the relationship between SNV density and gene frequency. By tracking 6 *Polynucleobacter* populations over 3 years and 45 time points, we found that one population was consistently dominant, each population consistently contained many strains, high frequency genes had higher SNV density, and most of the genes had a low percentage of non-synonymous SNVs.

3.2 Introduction

Microbial species concepts are actively debated, but some definitions have successfully explained observed data. As part of understanding what makes up a single species, microbiologists have been exploring the pan-genome of different taxa for over a decade (Tettelin et al., 2005). However, only recently have researchers started to incorporate metagenomic information into their understanding of the 'metapangenome' by characterizing genes that

are core and accessory to a taxon (Delmont and Eren, 2018). To improve our understanding of the pan-genome of individual, genetically separate populations, we extended the 'meta-pangenome' using time series analysis. We used shotgun metagenomics to determine the core and accessory genes of discrete populations and to track the abundance and diversity of bacterial populations from a freshwater lake through a three-year, 45-sample time series.

In this work we operationally define populations as the genetically separate or "sequence-discrete" populations previously observed (Konstantinidis and DeLong, 2008; Caro-Quintero et al., 2011; Caro-Quintero and Konstantinidis, 2012). Therefore, we will use "population" and "sequence-discrete population" interchangeably. These populations can be observed in metagenomic samples by mapping reads to reference genomes. They are differentiated from the next most closely related population by a coverage discontinuity, where read coverage drops between populations due to a lack of individuals with intermediate genotypes. This discontinuity is generally found at approximately 95% nucleotide identity. Our past work has used this population definition to look at how the genome-wide diversity of freshwater microbial populations changes through time (Bendall et al., 2016). We also used this definition to examine the different population structures and dynamics of two highly abundant, cosmopolitan groups of freshwater bacteria, LD12 and acI (Garcia et al., 2018). In both of these previous works, we focused primarily on population abundance and diversity metrics. In this work, we begin by studying the genome-wide abundance and diversity within sequence-discrete populations. We then zoom in further to examine the variation in gene frequency and characterize the core and accessory genes within *Polynucleobacter* populations.

We chose to focus on *Polynucleobacter* because members of this genus have a cosmopolitan distribution (Zwart et al., 2002; Wu and Hahn, 2006a; Newton et al., 2011) and are often the most abundant freshwater "lineage" found in dystrophic lakes, including in our study system, Trout Bog Lake (Linz et al., 2017). *Polynucleobacter asymbioticus* is the most well studied free-living species from this genus. From a comparative genomics analysis, *Polynu-*

cleobacter asymbioticus has a streamlined genome and is thought to have a passive lifestyle (Hahn et al., 2012), high rates of within-population recombination, and large functional diversity which can be shared between members of the *Polynucleobacter* genus by horizontal gene transfer (HGT) of genomic islands (Hoetzinger and Hahn, 2017; Hoetzinger et al., 2017). *Polynucleobacter* populations seem to be strongly influenced by the phytoplankton community as has been seen in Lake Mondsee (Austria) and in Trout Bog Lake (Wisconsin, U.S.A.) (Wu and Hahn, 2006b; Paver et al., 2015).

For this study, we identified and tracked six *Polynucleobacter* populations though a 45-sample, three-year time series (2007-2009) in Trout Bog Lake (Wisconsin, U.S.A.), a dystrophic, bog lake in northern Wisconsin. First, we used genome-wide average nucleotide identity (gANI) to find a high matching set (HMS) of metagenome-assembled genomes (MAGs), which all belong to the same population. Next, we used the representative set of genes from each HMS to perform a phylogenomic analysis of these populations and place them in the context of each other and other known *Polynucleobacter*. Additionally, we clustered genes from all six HMSs to find the *Polynucleobacter* single copy conserved genes (Pnec-SCCGs) shared between them. Four closely related single-cell amplified genomes (SAGs) from Trout Bog which belonged to the same population as one HMS allowed us to interrogate the gene recovery between these different methods. In order to track the changes in single nucleotide variant (SNV) frequency through time, we determined the fraction of SNVs dominated by a single allele (SNV homogeneity) and tracked this for each of the 6 populations through the time series. Finally, we characterized the frequencies for all the genes in each population and determined the relationship between SNV density and gene frequency.

3.3 Materials and Methods

Sample Collection, Extraction, and Sequencing

Trout Bog is a dystrophic lake located in Wisconsin, USA. It is surrounded by boreal forests and a Sphagnum mat which leaches terrestrially-derived organic matter into the lake. The surface area of Trout Bog is ~11000m^2, a maximum depth of 9m and a mean pH of 5.1. As described in Bendall et al. (2016), samples were collected from a 1 meter, depth integrated portion of the hypolimnion layer at 45 different time points during ice-free periods from 2006 to 2008. Samples were filtered on 0.2 micrometers pore-size polyethersulfone Supor filters (Pall Corp., Port Washington, NY, USA) and subsequently stored at 80°C. DNA was later purified from these filters using the FastDNA Kit (MP Biomedicals, Burlingame, CA, USA). DNA sequencing was performed at the Department of Energy Joint Genome Institute (Walnut Creek, CA, USA) as described in Bendall et al. (2016).

Assembly

For assembly of each sample, reads were filtered using the default settings of BBtools trimming to a minimum quality of Q17 (Bushnell, 2018). Then BFC was used for error correction, trimming reads containing singleton k-mers, using k-mer size 21 and expected genome size 10 gigabases (Li, 2015). Finally, the reads from each sample were assembled separately using the assembly only option of metaSPAdes v.3.10.1 and the following k-mer sizes: 21,33,55,77,99,127 (Nurk et al., 2017). The resulting assembly statistics (JGI library ID, collection date, IMG genome ID, number of contigs, size, N50, L50, GC content) can be found in Table 3.1.

Library ID	Date	IMG Genome ID	of contigs	size (bp)	N50	L50	GC content
IHXI	2007-05-28	3300021113	28546	17931747	3860	770	0.458

IHWF	2007-06-07	3300020684	73048	49921885	10174	904	0.474
IHUS	2007-06-13	3300020680	64194	41364591	9747	766	0.454
IHWI	2007-06-27	3300020730	201551	130505942	28234	766	0.471
IHWN	2007-07-02	3300020706	133328	85831074	18008	758	0.478
IHWG	2007-07-12	3300020702	126318	79125120	18743	713	0.472
IHWH	2007-07-25	3300020735	236622	158939914	30109	842	0.467
IHUZ	2007-07-31	3300020726	186317	126876739	24227	870	0.474
IHSA	2007-08-09	3300020699	119892	76017719	17662	733	0.473
IHWA	2007-08-20	3300020713	155074	96472500	23815	695	0.479
IHWY	2007-08-27	3300020690	97669	59027541	15728	661	0.490
IHWB	2007-09-10	3300020692	101120	61011128	14993	668	0.480
IHWU	2007-09-17	3300020700	127984	75507841	20559	622	0.484
IHXG	2007-10-01	3300020693	100570	66856621	18639	722	0.491
IHXO	2007-10-16	3300020703	134746	76999047	22932	581	0.459
IHXN	2007-11-05	3300020708	148865	85622164	27442	610	0.489
IHXW	2007-11-14	3300020682	62510	49769621	4722	1327	0.451
IHPY	2008-05-22	3300020697	99484	83406058	10917	1244	0.458
IHSB	2008-05-29	3300020687	73059	61544036	7956	1216	0.448
IHWS	2008-06-13	3300020679	43795	31359678	5090	926	0.447
IHWX	2008-07-01	3300020709	128561	99511045	14692	1015	0.462
IHWW	2008-07-08	3300020683	65294	50543708	7094	1129	0.459
IHXP	2008-07-15	3300020734	209672	153665877	30731	870	0.461
IHXT	2008-07-22	3300020721	162805	118711634	21994	866	0.460
IHXH	2008-07-29	3300020711	137030	93131989	18097	866	0.442
IHXA	2008-08-05	3300020707	126191	87797378	19314	781	0.455
IHXF	2008-08-12	3300020688	80409	57125852	12219	842	0.448

IHXU	2008-08-19	3300020691	87054	63037302	12235	885	0.448
IHXS	2008-08-25	3300020698	114615	76825112	12767	846	0.445
IHXX	2008-09-09	3300020701	117002	84540196	16382	883	0.454
IHPO	2008-09-20	3300021116	55282	33786504	7431	669	0.439
IHWP	2008-10-04	3300020724	184296	123385235	32506	732	0.457
IHPN	2008-10-23	3300020722	174776	113763045	34649	710	0.478
IHWT	2009-05-29	3300020727	174471	129099922	28106	906	0.444
IHTZ	2009-06-03	3300021135	263988	192126459	26872	986	0.472
IHXY	2009-06-15	3300020717	152419	111462832	17558	1052	0.484
IHUN	2009-06-23	3300020723	159179	135276742	17633	1258	0.485
IHUO	2009-06-29	3300020704	114345	94544317	13565	1139	0.490
IHUP	2009-07-07	3300020729	185169	137756086	29063	925	0.481
IHUA	2009-07-13	3300020720	166764	118262174	18603	939	0.497
IHUB	2009-07-21	3300020681	65254	49622392	6172	1190	0.500
IHUC	2009-07-27	3300020715	143314	107414156	19341	939	0.491
IHUH	2009-08-03	3300020712	142247	95972496	18765	855	0.488
IHUI	2009-08-11	3300020685	76220	49153280	13427	680	0.474
IHUF	2009-08-18	3300020719	177480	113005853	25132	745	0.482

Table 3.1: Assembly Statistics for Each Metagenomic Timepoint Sequenced

Read Preprocessing, Mapping, and Binning

Prior to mapping, reads were merged using the bbmerge tool from BBtools using default settings with the addition of the qtrim2 option (Bushnell, 2018). Next, reads were trimmed to Q17 using the default settings of bbqc (Bushnell, 2018). Reads from all samples in the time series were mapped to each assembly for binning with differential coverage using

bbmap from BBtools with a minimum identity threshold of 0.95. Bins were recovered from each individual sample assembly using the default settings of Metabat (Kang et al., 2015).

Filtering, Classifying, and Dereplicating Medium-High Quality Bins

Bin quality was assessed using CheckM v. 1.0.11, default settings with lineage specific option (Parks et al., 2015). Only bins with greater than 50% completeness and less than 10 percent redundancy/contamination, were analyzed further (Supplementary Tables A.1, A.2). These thresholds are consistent with Genomic Standards Consortium's (GSC) Minimum Information about a Metagenome-Assembled Genome (MIMAG) standards for 'medium quality' genomes (Bowers et al., 2017). Assembly statistics (HMS, number of contigs, size, GC content, N50, L50) for these genomes can be found in Supplementary Table A.1.

The medium quality MAGs were classified using a custom classification method. Open reading frames and annotations were predicted using IMG's annotation pipeline (Huntemann et al., 2015). For classification, the annotated genes were subset to only those which had RPSBLAST hits to phylogenetically conserved COGs (COG0016, COG0049, COG0051, COG0052, COG0072, COG0080, COG0081, COG0087, COG0088, COG0090, COG0091, COG0092, COG0093 , COG0094, COG0096, COG0097, COG0098, COG0099, COG0100, COG0102, COG0103, COG0164, COG0181, COG0184, COG0185, COG0186, COG0197, COG0198, COG0200, COG0244, COG0255, COG0532, COG3590) (Camacho et al., 2009; Darling et al., 2014; Galperin et al., 2015) . In order for a MAG to be classified at each taxonomic rank three criteria had to be met for the phylogenetically conserved genes listed above: at least 3 of the genes were required to have the same classification at that rank, at least 70% of the genes were required to have the same classification at that rank, and the genes were required to have an amino acid identity match to the IMG database greater than the threshold for that taxonomic rank (Kingdom:.20, Phylum:.45, Class:.49, Order:.53, Family:.61, Genus:.70, Species:.90, Taxon_name:.97). The taxonomic rank thresholds were

chosen based on the analysis AAI thresholds determined in Luo et al. (2014)(Supplementary Table A.3).

MAGs were identified as belonging to the *Polynucleobacter* genus if they were classified or belonged to the same High Matching Set (HMS) as MAGs classified as *Polynucleobacter*. There were six *Polynucleobacter* HMSs which contained 76 medium to high quality MAGs (Supplementary Tables A.1, A.2, A.3). To find which *Polynucleobacter* genomes were part of the same (HMS), genome-wide average nucleotide identity(gANI) and alignment fraction(AF) was calculated between every pair of genomes which were classified to the same phylum (Varghese et al., 2015). MAGs were considered to belong to the same HMS if they had an AF greater than 0.50 and an gANI of greater than 95% (Tables 3.2, A.1).

A representative gene for each cluster was chosen in order to 'deduplicate' the genes in an HMS prior to the abundance, gene frequency, and SNV analyses. Prior to clustering, genes and contigs which could not be confidently assigned to a particular HMS were removed using a Blast analysis (Camacho et al., 2009). These included any gene which had greater than 95% nucleotide identity over a region that was similar to the length of a metagenomic read, 150 basepair. To find the representative gene set, the nucleotide sequences for all genes predicted in an HMS were compared against each other using Blast 2.2.31+ (Camacho et al., 2009). Results with query coverage greater than 70% with identity greater than 95% were then clustered based on their normalized bit score using MCL v.14-137 (van Dongen, 2000; Enright et al., 2002) with an inflation score of 1.1. Representative genes were chosen from the 'best' MAG (most complete with least redundancy) which contained a gene represented in the cluster (Table 3.2).

Phylogenetic tree construction

Reference genomes were chosen from IMG because they were classified as belonging to the genus *Polynucleobacter* or were isolated from freshwater and members of the Burkholderiales order (Huntemann et al., 2015). Only one reference genome was chosen for each genus

HMS	# of Bins	Total Genes in HMS	RepGenes in HMS	Average Relative Abundance	Average Depth of Coverage
HMS10	1	1514	1030	0.447	5.064
HMS18	4	5850	1507	0.669	7.601
HMS19	34	61820	2248	5.707	65.227
HMS23	16	30741	2068	1.546	17.716
HMS28	13	32427	2871	0.611	6.932
HMS3	8	13539	2130	0.453	5.459

Table 3.2: HMS stats

outside *Polynucleobacter*. A multilocus alignment was constructed by concatenating the alignments of a set of 31 phylogenetically conserved genes. These genes were the same COGs used for the classification step above. These genes were found in references and MAGs by identifying the corresponding COGs which matched those gene annotations (Galperin et al., 2015). Each COG set was then aligned using the default settings of MAFFT v7.407 (Katoh et al., 2002; Katoh and Standley, 2013). COGs were concatenated using catfasta2phyml and then trimmed using the automated settings of TrimAI on the CIPRES Science Gateway V. 3.3 (Nylander, 2018; Capella-Gutiérrez et al., 2009; Miller et al., 2010). The final phylogenetic tree (Figure 3.3) was also made using RAxML on CIPRES (Stamatakis, 2014; Miller et al., 2010). The phylogenetic tree was visualized, edited for readability, and rooted on midpoint using iTOL (Letunic and Bork, 2016).

Clustering Homologous Genes among and within HMSs

In order to find the shared, orthologous genes between all six HMSs, Blastp (Blast 2.2.31+) was run on all the pairwise combinations of amino acid sequences for all HMSs (Camacho et al., 2009). The blast results were then filtered, keeping on those with > 70% query coverage and > 60% identity, and clustered using MCL v.14-137 by their normalized bit score, with an inflation value of 1.1 (van Dongen, 2000; Enright et al., 2002). A custom script was used to find clusters with a single gene from each of the six HMSs. These genes

are termed the *Polynucleobacter* Single Copy Core Genes (Pnec-SCCGs).

To compare homologous genes for HMS19 and the SAGs from Trout Bog belonging to the same sequence-discrete population, clustering was repeated using the amino acid sequences with thresholds of > 70% query coverage and > 90% amino acid identity.

Calculating Coverage, Abundance, and Gene Frequency

We calculated coverage (reads) per gene in each bin from the same mapping described in mapping section above. Coverage for each gene in each sample was calculated by taking the average coverage calculated by bedtools for each basepair (Quinlan and Hall, 2010). The coverage values for each gene were then normalized by the number of reads per million reads in each metagenome to estimate abundance (Reads per Million, RPM) of each gene. As a proxy for the abundance of each population, we calculated the average abundance of all the representative genes in an HMS. To find within sequence-discrete population gene frequency, we normalized each gene abundance value by the average gene abundance for that HMS in that sample. Thus, these values can be interpreted as relatively the portion of cells in the population which had that gene in their genome.

Calling SNVs in HMSs Representative Genes

Due to its lower sensitivity to minimum coverage, VarScan was chosen for calling single nucleotide variants (SNVs) in the reads mapped to the genomes (Zojer et al., 2017; Koboldt et al., 2012). First samtools was run to generate a mpileup file (Li et al., 2009a; Li, 2011). VarScan was run for each timepoint and assembly with the default settings. SNVs were then parsed into their respective MAGs. Outlier SNVs were removed if their average summed coverage across all samples was more than 3 standard deviations away from the average. Additionally, SNVs were only considered if they had a coverage value $\geqslant 8$ and a p-value < 0.05 in at least 2 of the 45 timepoints. The resulting number of SNVs for each HMS can be found in Table 3.4.

Similar to Bendall et al. (2016), using custom scripts we calculated the fraction of SNVs dominated by a single allele, here called SNV homogeneity, for each timepoint. For this analysis we considered SNVs to be dominated by a single allele if they were > 90% one variant. SNV homogeneity was only calculated for samples where the average coverage for the HMS was > 10.

Figure Construction and Computational Analysis

Unless otherwise stated all analysis was done using a combination of custom bash, python, R scripts (RStudio Team, 2016; R Core Team, 2018; Python Core Team, 2018; Free Software Foundation, 2018). Figures were all constructed using R and Rstudio (RStudio Team, 2016; R Core Team, 2018). The following libraries and modules were used for python analysis: pandas, ipython, glob, argparse, numpy, Bio.
The following libraries were used for R analysis and figure creation: ggplot2, dplyr, tidyr, reshape2, cowplot, RColorBrewer, forcats, magrittr, moments, segmented, grid, UpSetR, plotly, heatmaply, dendextend.

3.4 Results

Overview of MAGs and their High Matching Sets (HMSs)

In order to track population abundance and gene dynamics for members of the *Polynucleobacter* genus through our Trout Bog time-series, we assembled and binned 6 high matching sets (HMSs) containing 76 medium to high quality MAGs (Supplementary Tables A.1, A.2). These HMSs are distinct from one another and would be defined as different species using < 95% nucleotide identity across their shared genomic content as a cutoff (Konstantinidis and Tiedje, 2005; Kim et al., 2014; Varghese et al., 2015). The number of MAGs in an HMS ranged from 1 - 34 and more abundant HMSs tended to have more MAGs. The MAGs ranged from 1.18 - 2.57 MB in size and had from 73 to 312 contigs with estimated completeness of 55.6%

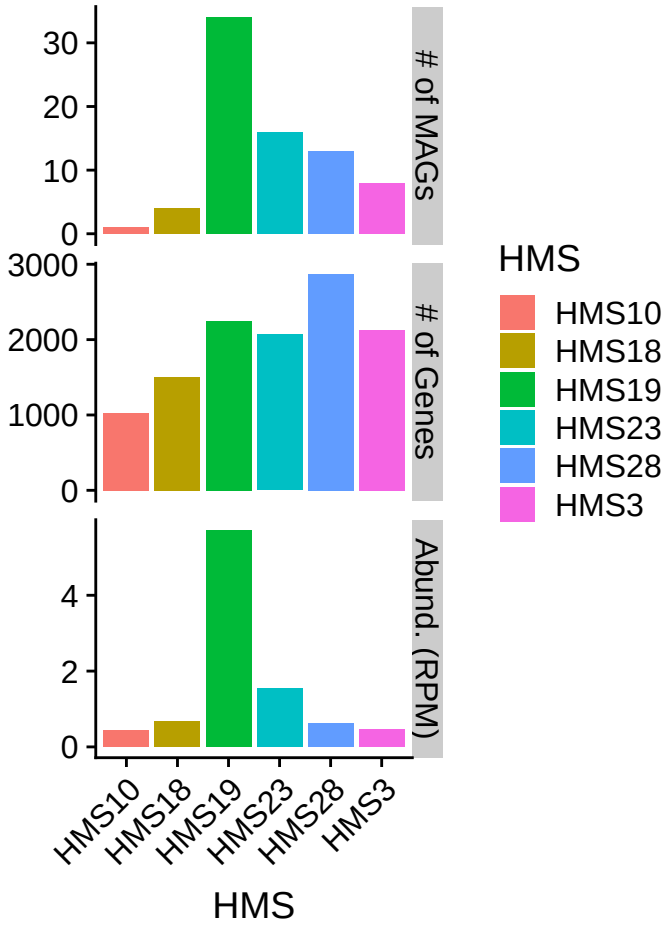

Figure 3.1: HMS statistics. Top: the number of MAGs which clustered into that HMS based on gANI/AF similarity. Middle: the number of non-redundant genes for that HMS after clustering the homologous genes and choosing a representative. Bottom: the average abundance of each HMS (average coverage of genes in HMS normalized by the number of reads in each metagenome).

- 98.8% and redundancy of 0% - 7.77%, as calculated by checkM (Parks et al., 2015). The genes from MAGs in the same HMS were then clustered to find the non-redundant set of representative genes (Figure 3.1). The pairwise genome-wide average nucleotide identities (gANI) between the HMSs ranged from 74.87 - 78.57 with alignment fractions (AF) ranging from 0.31 - 0.70 3.4.

Of the recovered *Polynucleobacter*, HMS19 was the most abundant throughout the metagenomic time series (Figure 3.2). HMS23 had relatively lower abundance in most of 2007 and 2009. It then bloomed to a much higher abundance in 2008, perhaps starting in

GENOME1	GENOME2	ANI(1->2)	ANI(2->1)	AF(1->2)	AF(2->1)
HMS10	HMS19	76.92	76.95	0.72	0.31
HMS18	HMS10	76.05	75.99	0.33	0.53
HMS18	HMS19	75.74	75.80	0.71	0.49
HMS18	HMS23	75.47	75.45	0.71	0.52
HMS18	HMS28	75.10	75.10	0.70	0.37
HMS18	HMS3	74.90	74.87	0.69	0.50
HMS23	HMS10	78.57	78.53	0.35	0.76
HMS23	HMS19	77.49	77.50	0.73	0.69
HMS23	HMS28	76.62	76.61	0.65	0.47
HMS23	HMS3	76.38	76.37	0.65	0.64
HMS28	HMS10	76.29	76.27	0.21	0.63
HMS28	HMS19	77.21	77.20	0.47	0.63
HMS28	HMS3	76.76	76.74	0.50	0.70
HMS3	HMS10	75.79	75.76	0.29	0.64

Table 3.3: Genome-wide average nucleotide identity (gANI/ANI) and alignment fraction (AF) between all HMSs

November 2007, which made it the second most abundant HMS over the whole time series. This corresponded with a higher average abundance for all the recovered *Polynucleobacter* HMSs in 2008. HMS18, HMS28, HMS3, and HMS10 were generally much less abundant than HMS19 over the whole time series. HMS3 became somewhat more abundant in 2009, making it the second most abundant HMS from that year.

A phylogenetic tree was constructed to understand how these HMSs were related to each other and the known diversity of *Polynucleobacter* (Figure 3.3). The most abundant HMS, HMS19, grouped very closely with 4 single-cell amplified genomes (SAGs) which were also collected from Trout Bog hypolimnion and two MAGs from mixed cultures (Garcia et al., 2018). One of the mixed culture MAGS was from the epilimnion of Trout Bog and the other was from the similarly dystrophic Lake Grosse Fuchskuhle (Brandenburg, Germany). The 4 SAGs collected from Trout Bog's hypolimnion each had an average nucleotide identity of ~98% with an alignment fraction (AF) of > 0.79 to HMS19 (Varghese et al., 2015). The most deeply branching HMS was HMS18, which also grouped with two mixed culture MAGs from Trout Bog, with the closest one having 97.8 gANI (0.82 AF) to

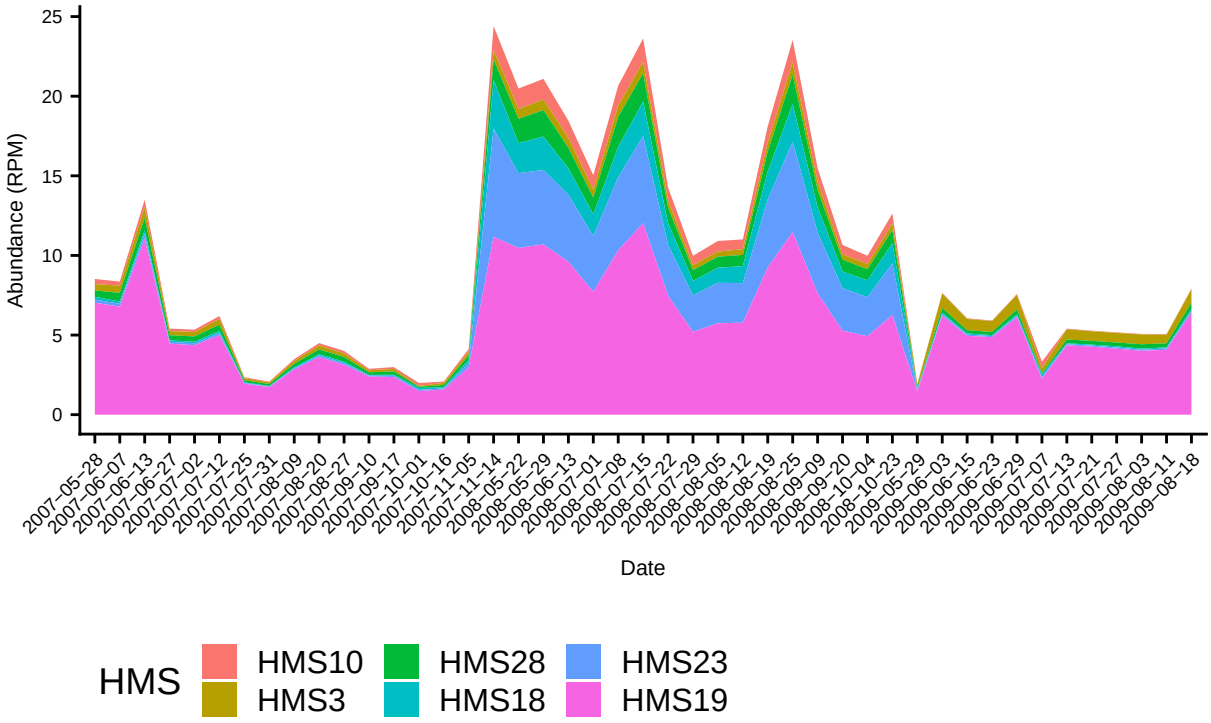

Figure 3.2: HMS Abundance Through Time. Abundance of each HMS in Trout Bog throughout the metagenomic time series. Area colored indicates the abundance for that HMS. Abundance was calculated by normalizing the average coverage values across all genes in the HMS by the number of reads in each metagenome.

HMS18. HMS28 grouped most closely with *Polynucleobacter sphagniphilus* MWH-Weng1-1 which was also isolated from another similar lake (Hahn et al., 2017). HMS28 would not be considered part of the same species by most definitions since its gANI is only 76.9 (0.50 AF) to MWH-Weng1-1. HMS3 grouped most closely with a SAG also taken from Trout Bog (99% gANI, 0.86 AF). It also grouped with a isolate genome which was recently characterized as a new species of PnecC, *Polynucleobacter meluiroseus* AP-Melu-1000-B4, though it shared only 76 gANI (0.67 AF) (Pitt et al., 2018). HMS10 grouped most closely with *Polynucleobacter campilacus* MWH-Feld-100(78 gANI, 0.71 AF to HMS10), and basal to the clade which contained most of the *necessarius* and *asymbioticus* genomes. HMS23 would be considered part of the *Polynucleobacter necessarius asymbioticus* species under many definitions since it had ~97 gANI(~0.86 AF) to both QLW-P1DATA-2 and MWH-Tro-8-2-5GR (Hoetzinger and

Figure 3.3: Phylogenetic tree of *Polynucleobacter* HMSs. Tree was constructed using the concatenated alignment of 31 amino acid sequences (mostly ribosomal proteins). Midpoint rooted tree showing the six *Polynucleobacter* HMSs in the context of known *Polynucleobacter* reference genomes and other genomes from the Burkholderiales order. HMSs are colored in orange.

Hahn, 2017; Hoetzinger et al., 2017).

By clustering the genes predicted for each HMS together, we found many homologous genes shared among the six *Polynucleobacter* HMSs (Figure 3.4). Between 66.5 - 90.1% genes found in each HMS had a homolog in at least one other HMS. There were 526 gene clusters

found in every HMS. Of those, 427 were single copy in each HMS and were subsequently termed the Polynucleobacter Single Copy Core Genes (Pnec-SCCGs). The number of shared gene clusters may be underestimated due to incompleteness in the MAGs comprising the HMSs. Each HMS also had unique genes recovered only in that HMS. HMS28 had the largest number of genes recovered and also had the most unique genes. However, the number unique genes for each HMS did not always reflect the total number of genes, as HMS3 had slightly more singletons than HMS19. It is possible that some genes identified as unique are in fact shared among the HMSs due to incomplete genome reconstruction, however the unrecovered regions of the genome may also harbor other unique genes.

Gene Content Differences between SAGs and HMS19 Representative Genes

The SAGs which are in the same sequence-discrete population as HMS19 provide an opportunity to learn more about the limitations and strengths of SAGs and MAGs. The relationships between SAGs and MAGs have been investigated in several lineages of bacteria in the Baltic Sea (Alneberg et al., 2018)., however with our dataset we can use the time series aspect to investigate if the genes missing in the MAGs have lower gene frequencies. After clustering the HMS19 and SAG genes together based on homology, 83.1% of the gene clusters identified had a gene representative from HMS19 (Supplementary Figures A.1 and A.3). The SAG genes missing from HMS19 had significantly lower coverage than those which were found in the MAG (Figure 3.5). SAGs MCM14TBH076, MCM14TBH017, and MCM14TBH079 had large differences ($p < 0.001$ for all) in the average coverages of the genes with homologs in HMS19 vs those without. SAG MCM14TBH064 had the smallest but still significant difference in average coverages between the two groups of genes ($p = 0.008$). There were three genes found in SAGs MCM14TBH076 and MCM14TBH017 which had much higher coverage than the rest, greater than 350 reads. These genes were all found to be transposases, which may account for their high coverage.

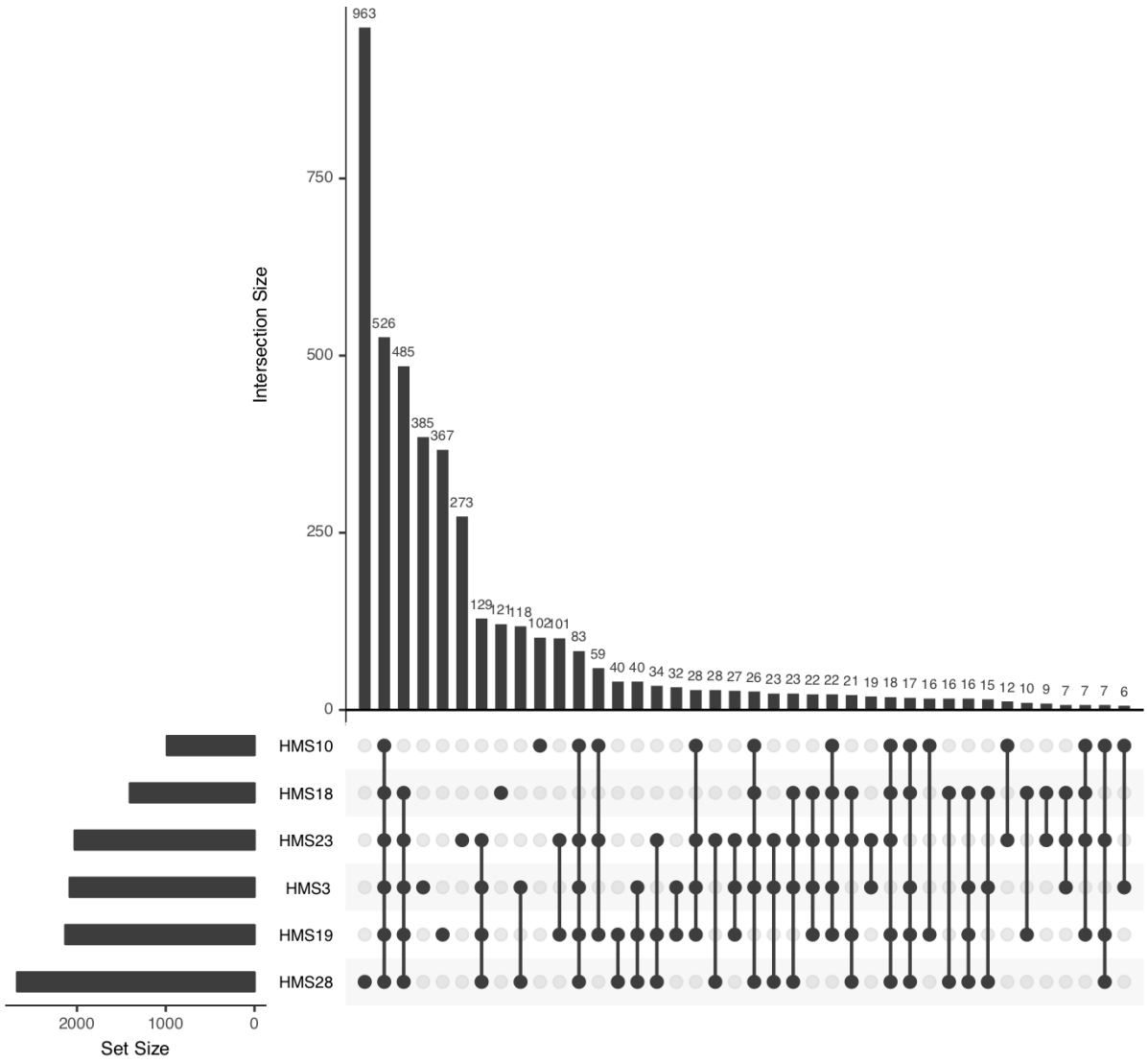

Figure 3.4: Grouped Histogram of Homologous Genes Between HMSs. Homologous genes were clustered by their normalized bit scores resulting from a pairwise comparison of all amino acid sequences with blastp(Camacho et al., 2009). Only homologous genes with query coverage >70 and amino acid identity >60 were clustered. The set size histogram on the bottom left shows the number of genes in each HMS. The upper histogram shows the number of genes in the group highlighted by the black dots on the x axis. Groups are ordered by their size and only the largest 40 groups are shown.

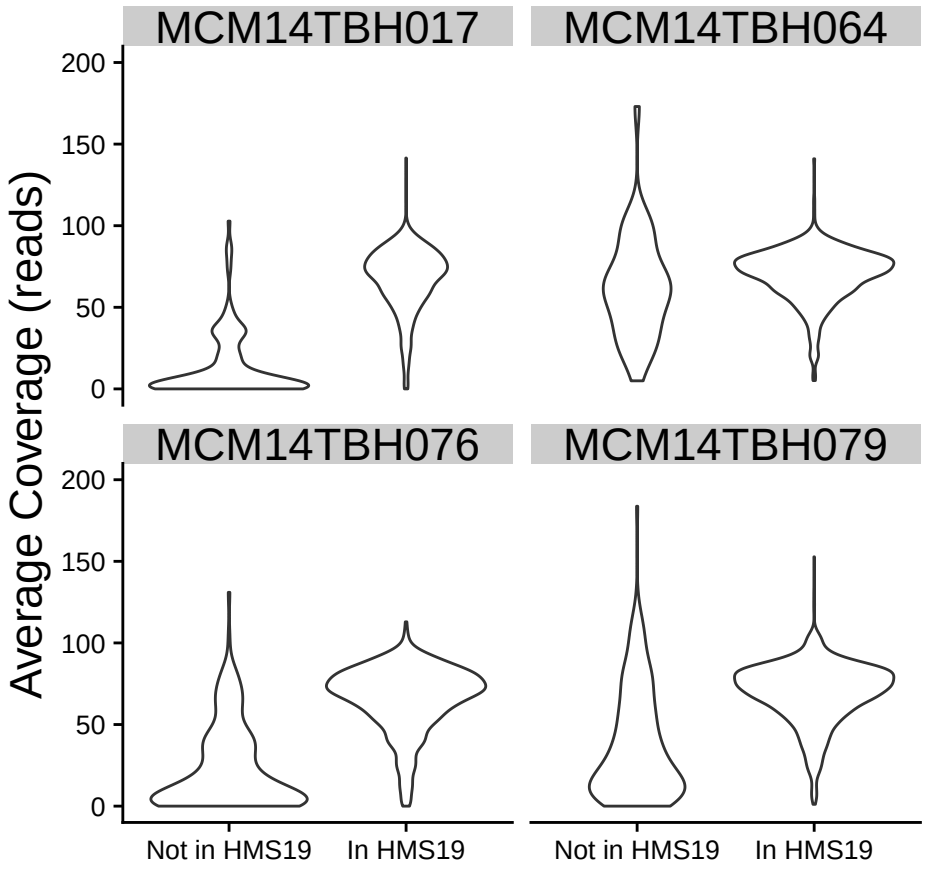

Figure 3.5: Average Coverages for Genes in HMS19 and SAGs. Violin plots (i.e. distributions) of average metagenomic read coverage for the genes in each SAG across the time series, split into the genes which were assembled in HMS19 and those which were not. Two transposases with very high coverage (>350 reads) were removed from MCM14TBH076 and 1 was removed from MCM14TBH017. Y-axis labels are displayed on the right of each plot.

Population Diversity Analysis

To understand how these changes in abundance may affect the diversity within each population we identified the single nucleotide variants (SNVs) in the reads mapped to each reference HMS. Low frequency SNVs are less likely to be detected by our method because we only included SNVs with an abundance higher than 8 in at least 2 time points. As such, we likely detect rarer SNVs only in the most abundant population, HMS19. HMS18 and HMS19 had an order of magnitude more SNVs detected than the other four HMSs

hms	Total SNVs	Synonymous	Nonsynonymous	Percent Synonymous
HMS10	2696	2099	597	77.9
HMS18	23392	19733	3659	84.4
HMS19	37508	30535	6973	81.4
HMS23	6362	5304	1058	83.4
HMS28	4060	3087	973	76.0
HMS3	6769	5132	1637	75.8

Table 3.4: SNV counts for each HMS

(Table 3.4, Figure 3.6). The number of SNVs detected in HMS18, which had relatively low abundance, suggests that HMS18 may truly have a higher SNV density than the other low abundance populations. The percentage of synonymous SNVs, those which don't result in an amino acid change, in each population was relatively consistent, ranging from 75.8 - 84.4% (Table 3.4). While all of the HMSs SNVs were mostly heterogenous, i.e. not dominated by a single allele, HMS19 had the most homogeneous SNVs (Figures 3.7 and A.2). HMS19 had a single timepoint with 26.8% SNV homogeneity and an average of 9.80% SNV homogeneity across all timepoints.

The abundance of HMS19 had a significant linear relationship with homogeneity (R^2 = 0.437, $p < 0.001$) (Figure 3.8). While not a clonal sweep since the values always remain below 30%, this indicated that a subset of strains are blooming when there is a rise in abundance.

Gene Frequency Analysis

For each HMS we performed a gene frequency analysis to characterize the core and flexible genome. We investigated the relationship between gene frequencies for the whole HMS to the genes we had determined to be Pnec-SCCGs for each sample and found that the SCCG average frequency followed the average gene frequency closely though the timeseries (Supplementary Figures A.4, A.5, A.6, A.7, A.8, A.9). Based on this relationship, we used the 95% confidence interval (CI) of the Pnec-SCCG average frequency to determine gene

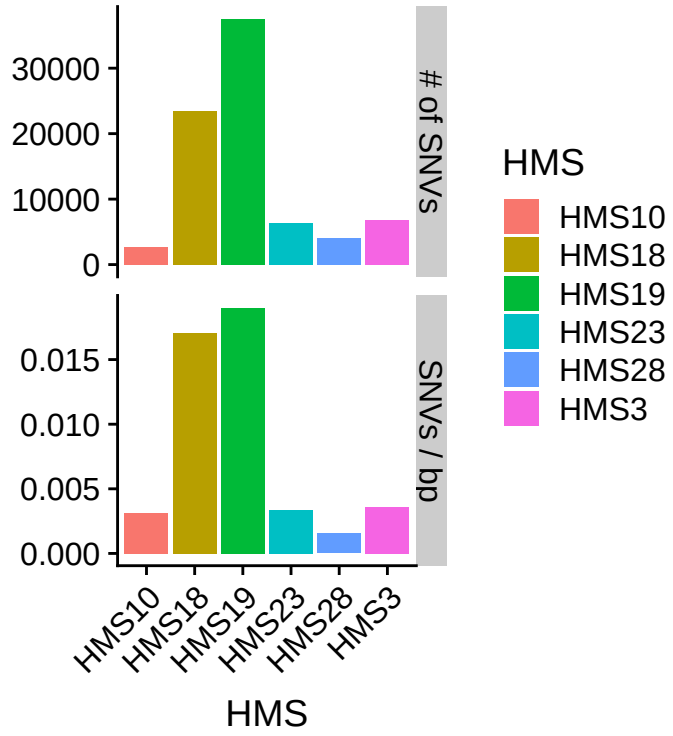

Figure 3.6: HMS SNV statistics. Top: the number of SNVs in the representative genes for each HMS. Bottom: the number of SNVs per basepair for each HMS. Y-axis labels are displayed on the right of each plot.

frequency groups (Figure 3.9 and Supplementary Figures A.10, A.11, A.12, A.13, A.14). We considered genes within this interval to be 'high' frequency since their frequency suggests they are typically present in every cell in the sequence-discrete population. Genes with average frequencies above the 95% CI were labeled as 'multicopy' since their gene frequency suggests that they may be present in more than one copy in each cell. Similarly, genes with below the 95% CI were labeled as 'low' frequency genes. We expected that it may be difficult to recover the low frequency genes from the low coverage genomes. While our frequency groups label 'low' frequency genes for each HMS, only the average gene frequency histograms for HMS19 and HMS23 have a secondary peak in the low frequency range (Figure 3.9 and Supplementary Figures A.10, A.11, A.12, 3.9, A.13, A.14).

To verify these genes frequency groups, we clustered the genes based on their gene frequencies throughout the time series (Figure 3.10 and Supplementary Figures A.10, A.11,

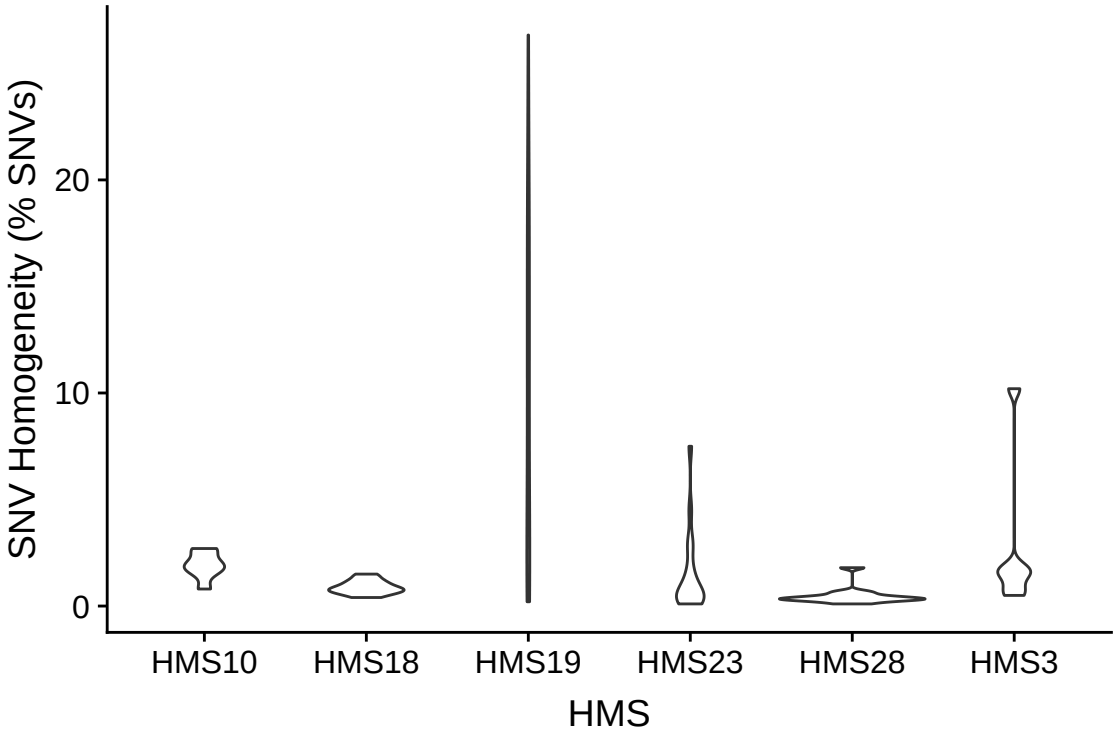

Figure 3.7: HMS SNV Homogeneity. Distributions of each HMS's SNV homogeneity across all samples. A SNV is considered homogeneous when it is >90% one variant. SNV homogeneity is defined as the percentage of homogeneous SNVs in an HMS. Homogeneity was not calculated for samples where average gene coverage for the HMS was below 10.

A.12, A.13, A.14). The genes in each HMS did typically group with other genes from their same frequency group, especially in the more abundant populations (HMS19 and HMS23). When clustering was done on each sample based on its gene frequency pattern, typically samples in the same year grouped together and 2007 and 2008 grouped together. These groups differ from the abundance patterns for all *Polynucleobacter* HMSs, where all populations were more abundant in 2008 and less abundant in 2007 and 2009.

Relationship between Gene Frequency and Selection Signatures (SNV Density and Non-Synonymous Percentage)

We hypothesized that there might be a relationship between gene frequency and selection which might be observed by differing the SNV density. We examined the relationship be-

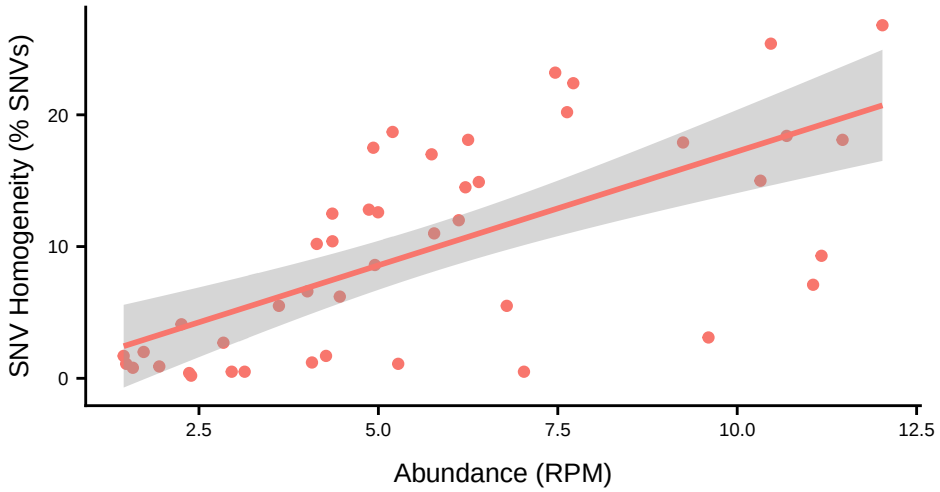

Figure 3.8: Abundance verses SNV Homogeneity. Each point represents the abundance and SNV homogeneity for HMS19 at a single time point. A SNV is considered homogeneous when it is >90% one variant. SNV homogeneity is defined as the percentage of homogeneous SNVs in an HMS. Grey shading indicates the 95% confidence interval for the linear model line.

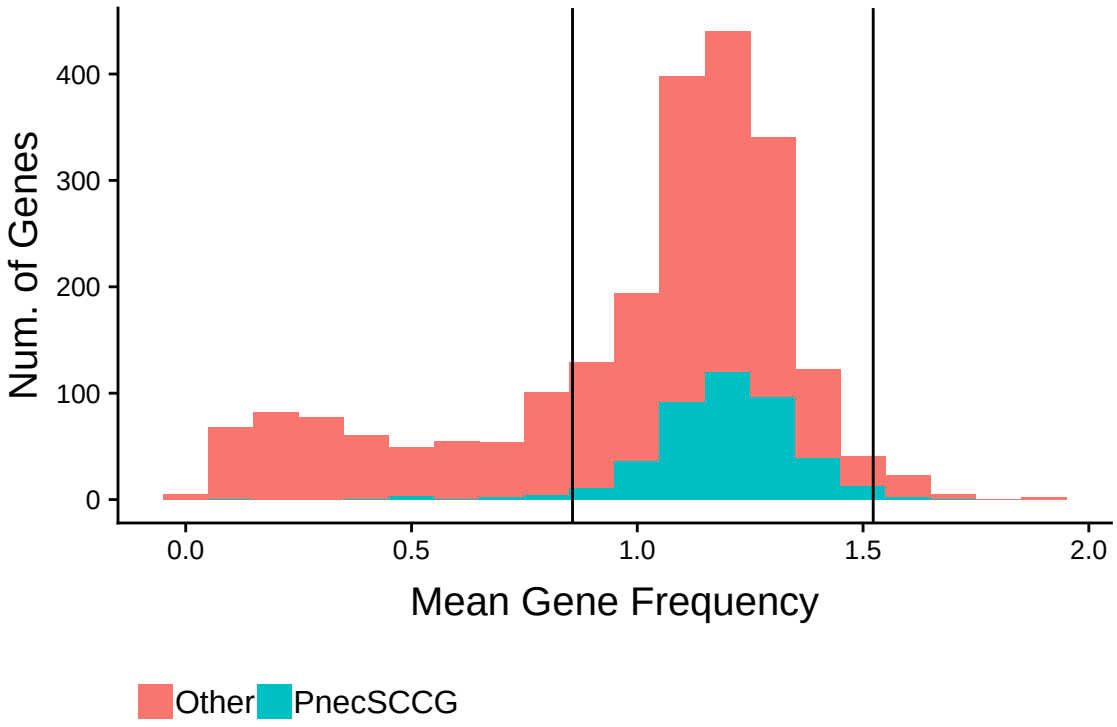

Figure 3.9: HMS19 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS19. PnecSCCGs are in blue. Black lines represent the 95% confidence interval for the Pnec_SCCGs

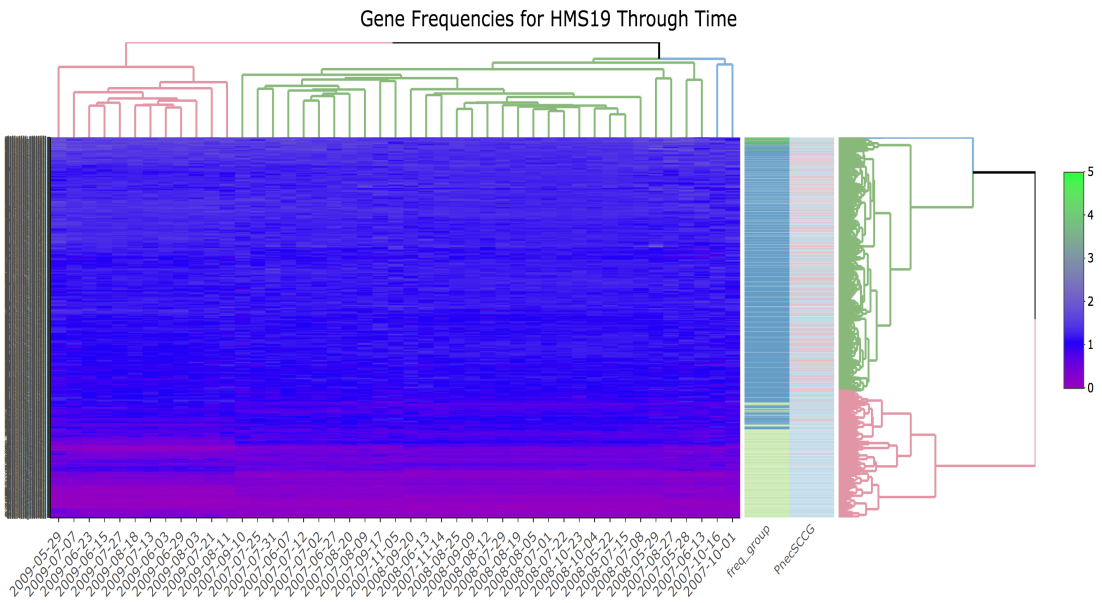

Figure 3.10: Gene Frequencies for HMS19 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multicopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot.

tween the SNV density and the gene frequency for the more abundant populations (HMS19 and HMS23). We choose these two populations since their higher abundance makes SNV and low frequency gene detection more sensitive. In both, we found a significant nonlinear relationship with a quadratic term, where the 'high' frequency genes have higher SNV densities than the 'low' or 'multicopy' genes. The relationship was much stronger for HMS19 ($R^2 = 0.357$, $p < 0.001$) than for HMS23 ($R^2 = 0.00893$, $p < 0.001$) (Figures 3.11 and 3.12).

We also hypothesized that high frequency genes would be under greater purifying/negative selection. We found that most genes in HMS19 had a percentage of non-synonymous SNVs, which suggests purifying selection (Figure 3.13). There was a slight linear trend indicating that low frequency genes had a slightly higher chance of being under directional selection.

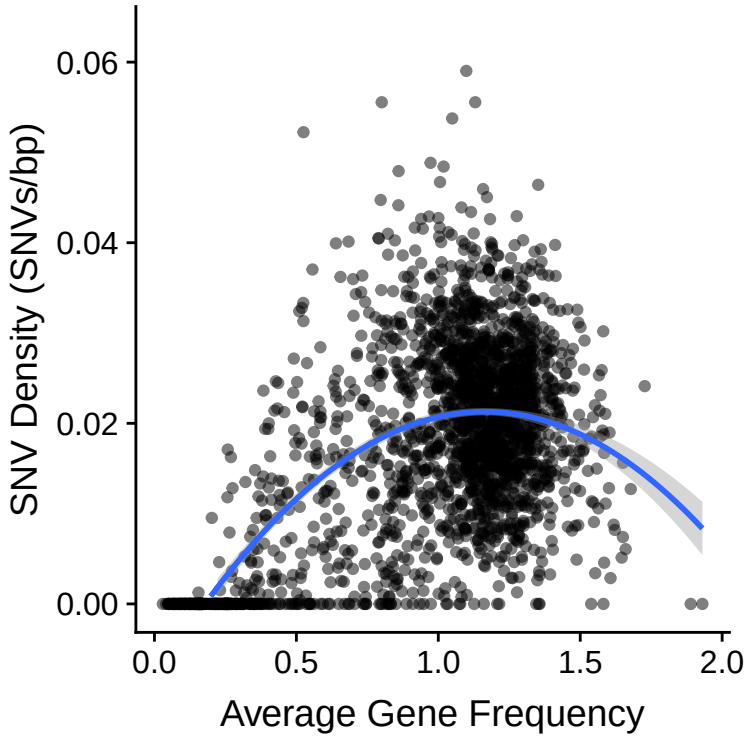

Figure 3.11: HMS19 SNVs per bp V Gene Frequency. There is a nonlinear relationship with a significant quadratic term between SNVs per basepair and average gene frequency, fit line shown in blue ($R^2 = 0.357$, $p < 0.001$). SNVs per basepair is low for genes with both high and low gene frequency. The blue line shows the linear relationship and the grey shaded area indicates the 95% confidence interval of the model

3.5 Discussion

From environmental samples, we can identify sequence-discrete populations where many cells share high identity due to their shared ancestry (Konstantinidis and DeLong, 2008; Caro-Quintero et al., 2011; Caro-Quintero and Konstantinidis, 2012; Bendall et al., 2016). Using metagenomics, we build upon the idea of the 'pangenome', which represents the full set of genes, both core and accessory, for a taxon, and extend it to examining the gene relationships within these populations (Delmont and Eren, 2018). In metagenomic analysis, individual assemblies and binning are often done on more than one metagenome. Then the resulting genome bins which are considered the 'same', usually based on sharing >95% gANI, are de-replicated and one genome is chosen to represent the population (Olm et al.,

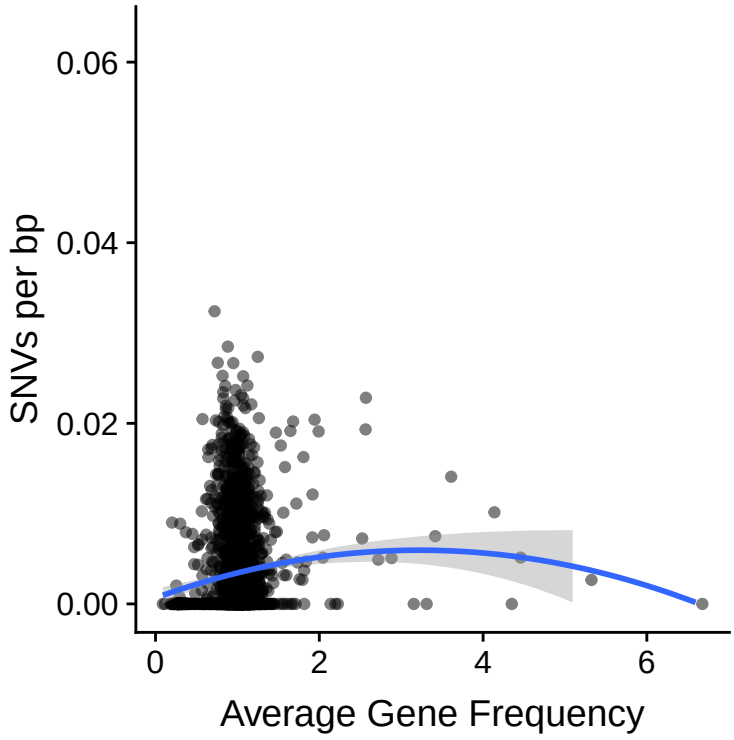

Figure 3.12: HMS23 SNVs per bp V Gene Frequency. There is a nonlinear relationship with a slight but significant quadratic term between SNVs per basepair and average gene frequency, fit line shown in blue ($R^2 = 0.00893$, $p < 0.001$). The number of SNVs per basepair is low for genes with both high and low gene frequency. The blue line shows the nonlinear relationship and the grey shaded area indicates the 95% confidence interval of the model.

2017; Sieber et al., 2018). To maximize our detection of accessory, or flexible, genes for a given population, we grouped high matching sets (HMSs) of metagenome-assembled genomes (MAGs) and unified their genes into a representative gene set. We included genes from all medium to high quality MAGs in an HMS and clustered all the genes into a non-redundant representative set of genes for each HMS. It is unsurprising that the number of MAGs recovered for each HMS is related to average abundance across the time series since more abundant populations will have a greater depth of coverage and are more likely to assemble (Figure 3.1). Once unified, we found a similar number of genes among most of the HMSs despite the greater numbers of MAGs recovered from the more abundant populations. However, the two HMSs with the smallest number of representative genes (HMS10, HMS18) had less than 5 MAGs representing them and are less complete

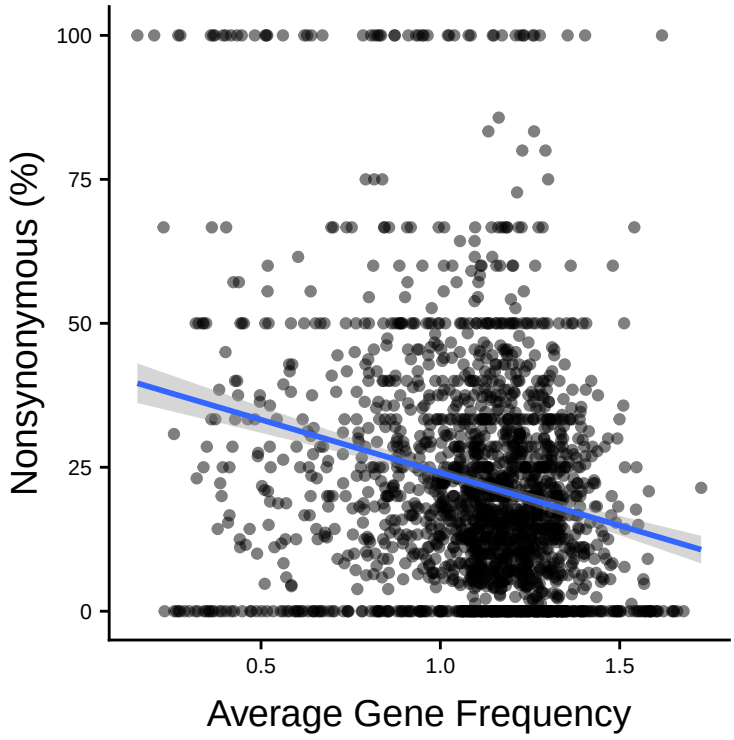

Figure 3.13: HMS19 Non-synonymous SNV fraction V Gene Frequency. There is a slight but significant linear relationship between the fraction of non-synonymous to and average gene frequency within HMS19 ($R^2 = 0.0500$, $p < 0.001$).

representations of their respective populations. With the exception of HMS10 and HMS18, the other HMSs recovered had a similar number of genes to known *Polynucleobacter* isolate genomes (Hoetzinger et al., 2017).

Polynucleobacter is one of the most studied freshwater microbes due to considerable genomic analysis of isolates from central Europe by Hahn and colleagues (Hahn, 2003; Wu and Hahn, 2006a,b; Jezberová et al., 2010; Hoetzinger and Hahn, 2017; Hoetzinger et al., 2017). Based on 16S rRNA gene analysis, we know that at least some species of *Polynucleobacter* are cosmopolitan (Zwart et al., 2002; Newton et al., 2011) and that *Polynucleobacter* are among the most highly abundant and persistent populations in bogs (Linz et al., 2017). With our time series we can track changes in the abundance and diversity of each population and observe the genomic changes of many community members. The observed *Polynucleobacter* populations seem to be relatively persistent members of the

community as each population is usually detected at some level throughout our metagenomic time-series (Figure 3.2), consistent with the findings of Linz et al. (2017). However, there was differential abundance between the populations, with HMS19 being the most abundant throughout the time series and had an average depth of coverage of ~65 across the 45 timepoints (Figure 3.2). The patterns of abundance through time have similar trends overall yet different magnitudes (Figure 3.2). This is curious due to the genetic separation (Table 3.4) of each of these populations. It may suggest that the overlapping dimensions of their niches are the limiting factors in their growth, but their non-overlapping niche space maintains the multiple populations. Interestingly, all of the populations are relatively more abundant in 2008 than in 2007 and 2009. The increase in abundance for all HMSs indicates an environmental or community shift which favored all the recovered *Polynucleobacter* populations in 2008.

The group of coexisting HMSs in this study are distributed across the known *Polynucleobacter* species included in our phylogenetic analysis (Figure 3.3). HMS19 groups closely with 4 SAGs, also collected from the hypolimnion of Trout Bog, which would be considered part of the same HMS if they were MAGs and would be considered part of the same species base on gANI(Konstantinidis and Tiedje, 2005; Kim et al., 2014; Varghese et al., 2015). Since this population is the most abundant of the ones recovered in this study, it is unsurprising that many of the SAGs sequenced from the same environment come from this group. After clustering the genes from these SAGs with those from HMS19, we found that only 17.2% of the total gene clusters were found in the SAGs but not in HMS19 (Figure A.1 and Supplementary Figure A.3). This suggests that the HMS19 is a better representative of the genes in this population than the SAGs but is still missing some gene content. Genes which were assembled in the SAGs but missing in the HMS may be caused by a variety of circumstances. We found that these missing genes had lower average coverage than the genes in the HMS19 (Figure 3.5). There were three genes with much higher average coverage (> 300) which were not assembled in the MAG (not shown in Figure 3.5). These genes are annotated as

transposases, which are mobile genetic elements and often distributed through a wide range of *Polynucleobacter* genomes which may account for their higher coverage. We would not expect these genes in MAGs since their mobility and presence in multipule populations would make their abundance patterns difficult to bin.

We also looked into if there was a threshold for a single sample which allowed the genes in HMS19 which were lower coverage to assemble but found no such threshold (data not shown). Other possible explanations, not interrogated here and which require further study, are that these genes missing from HMS19 could have higher strain heterogeneity or different tetranucleotide frequency due to relatively recent horizontal gene transfer into the population. Alternatively, these genes could also be explained by contamination, a common problem for SAGs (Alneberg et al., 2018).

In previous work, we saw that sequence-discrete populations had changes in their SNV frequencies through time. Unlike the co-existing *Chlorobium* population reported in Bendall et al. (2016), the *Polynucleobacter* populations in this study all had relatively low homogeneity (< .20) throughout the time-series (Figure 3.7). High SNV homogeneity values indicates that due to selection or genetic drift one or several strains dominated the population. When we broke down these values across time, the HMSs had different trends across years (Figure A.2). HMS19 has variable SNV homogeneity across time with a period of surprisingly low values in the latter half of 2007 indicating a possible rise in the strain diversity during that time. However, the homogeneity of HMS19 rose again in 2008 and remained highly variable. To further study the effect of abundance on SNV, we plotted the abundance of each HMS against SNV homogeneity across time (Figure 3.8). Interestingly the most abundant population, HMS19 has a significant linear relationship with SNV heterogeneity ($R^2 = 0.437$, $p < 0.001$). This seems unlikely to be a clonal expansion of a single strain since the heterogeneity value is still relatively low but does suggest that perhaps a few strains within the population bloom together. Future work should be done to compare the homogeneity of populations from the *Polynucleobacter* genus to other genera

in Trout Bog to see if this is a common trend among freshwater bacterial populations or corresponds to some aspect of their lifestyle. This would require repeating these analyses using many more MAGs and SAGs and initial comparisons to the other MAGs in Bendall et al. (2016) suggests that the low levels of SNV homogeneity in these *Polynucleobacter* populations are rather typical.

To identify the 'metapangenome' for each *Polynucleobacter* population, we first looked at the distribution of gene frequencies within each population. For all HMSs, we saw relatively normal gene frequency distribution with a peak at around 1 indicating those genes were present approximately once in every cell (Figure 3.9 and Supplementary Figures A.10, A.11, A.12, A.13, A.14). For the more abundant populations (HMS19 and HMS23) we also saw a secondary peak below 1 of lower frequency genes. The lack of this secondary peak in low abundance populations is likely due to limited recovery of these genes and limited detection in the metagenomic reads. We noted that the Pnec-SCCGs also had a relatively normal distribution with peak around 1 but no secondary peak and their average frequency followed the average gene frequency for the whole population throughout the time series (Supplementary Figures A.4, A.5, A.6, A.7, A.8, A.9). Based on this relationship, we defined frequency groups ('low', 'high', and 'multicopy') using the 95% confidence interval of the Pnec-SCCG frequencies. To verify that these gene frequency groups based on average gene frequency, held true throughout the time series, we clustered the genes based on their frequencies through time. For most and especially the more abundant populations, genes from the gene frequency groups clustered together (Figure 3.10 and Supplementary Figures A.10, A.11, A.12, A.13, A.14). Unexpectedly, the clustering of metagenomic samples based on their gene frequency distribution within each population showed that 2007 and 2008 typically grouped together and separately from 2009. This result is surprising given the difference in abundance observed in all populations in 2008 and indicates that the 2009 drop in abundance impacted the gene frequency relationships within the population. Taken together, the categorization based on gene frequency and clustering can be used to

group the metapangenome for a population into the core (multicopy, high) and accessory (low) for populations with sufficient coverage. This method provides a foundation for future functional analysis and can be applied to other populations in the time series as well.

As shown in a model proposed in Cordero and Polz (2014), we expected that these different gene frequencies may have different ecological interactions and thus evolutionary consequences. To probe these differences, we looked at the relationships between gene frequency and SNV density for HMS19 and HMS23. For HMS19 we found that the gene frequency had a nonlinear trend with a significant quadratic term where the density of SNVs tended to be higher for genes with frequencies around 1 (Figure 3.11). HMS23 showed a similar but less extreme pattern. One possibility is that the lower density of SNVs for the low frequency genes is due to difficulty detecting rare SNVs in low frequency genes. However, the lower density of SNVs for the multicopy gene group is rather surprising. We expected that multicopy genes might be in the process of gene duplication and thus have higher levels of SNVs as they diverge. However, it is also possible that these genes are in multicopy for increased gene expression and lower SNV density may instead indicate that these genes are under very strong purifying selection. We also found that most of genes in HMS19 had low non-synonymous percentages, indicating purifying selection (Figure 3.13). The low frequency genes were slightly more likely to be under directional selection. Perhaps this subset of genes is truly 'flexible' for the population, not under negative-frequency dependent selection, and thus new mutations are less likely to be deleterious.

3.6 Conclusions

The populations of *Polynucleobacter* tracked in this work are all relatively persistent though one population, HMS19, is much more abundant than the others. The genes which did not assemble in HMS19 but were recovered in the SAGs are mostly explained by lower coverage.

All of the *Polynucleobacter* populations have relatively low strain homogeneity, due to either a large number of co-existing strains or high levels of recombination within the population. However, the relationship observed between abundance and SNV homogeneity in the most abundant population may be due to a bloom of a subset of strains. Average gene frequency over time allows for identification of frequency groups (low, high, and multicopy) which are mostly maintained through the time series. The single copy core genes shared between all the populations generally follow a normal distribution among the high frequency gene group. Genes with high frequency have a higher SNV density than the low and multicopy genes. Most genes in the population seem to be under purifying selection, although low frequency genes have a slightly higher chance of being under directional selection.

4 CONCLUSIONS AND FUTURE DIRECTIONS

4.1 Conclusions

The work in this thesis begins the process of tracking the population dynamics of freshwater microbes through the lens of genomics. Using a metagenomic timeseries, we looked for changes in diversity through time in a number of different common freshwater microbes. Those timeseries in conjunction with metagenome-assembled genomes (MAGs) revealed a genome-wide sweep occurring in a natural population and showed evidence of a prior gene sweeps in other populations. Observing both types of sweeps suggests that these different populations are controlled by different evolutionary forces.

We also observed that at least one common freshwater population, LD12 shows an unusual population structure using single-cell amplified genomes (SAGs) as references. The two populations which where identified had a less stark coverage discontinuity separating them and had highly correlated relative abundance patterns through the timeseries. It seems like these two populations may be in the early stages of differentiation from one another. We also found that the populations of acI seem to have different seasonal abundance patterns and are likely ecologically distinct. While both acI and LD12 are common, abundant, and streamlined bacteria, they seem to have different underlying population structure and differentiation.

In our final chapter, we looked deeply at six different populations from the same genus. We found that one of the six was by far the most abundant through out the time series. All the populations investigated had somewhat low SNV homogeneity values through the time series. However the abundance of the dominant population had a significant linear relationship with SNV homogeneity, which suggests a subset of the strains within the population are responsible for a given increase in abundance. By comparing MAGs and SAGs from the same population, we found that genes from the population that are missing from MAGs are often lower coverage, rarer genes. Finally, we characterized the

core and accessory genome of the populations and found that high frequency genes had higher SNV density than low frequency or multicopy genes.

Overall this work provides a beginning and framework for looking at population dynamics using metagenomic timeseries.

4.2 Future Directions

Although enough time always remains elusive, I have a number of ideas about how I would follow up on the work presented in this thesis. To follow up on the results from chapter one, I would design qPCR primers for two genes or sets of genes: one which had no change in abundance through time and one which rose in abundance with as the 'winning' strain swept. This would allow us to easily quantify the portion of the *Chlorobium*-111 population which remains the the 'winning' strain from the sweep through the rest of the time series and in current samples. There likely remains some level of diversity below our detection limit, and in fact, in 2012 we did see a slight increase in diversity from the level in 2009. Since *Chlorobium* has been successfully cultured in other labs, we could isolate and cultivate *Chlorobium*-111 allowing us to perform biochemical tests on the resulting strain. If the resulting strain could be genetically manipulated, we could recapitulate some of the lost mutations and compare how these strains compete in the lab. With an isolated set of *Chlorobium* strains, we could isolate phage and begin to understand how viral predation impacts this homogenous population.

In chapter two, the LD12 populations tracked in this chapter showed different population structures than the other populations tracked. This may be indicative of the early stages of differentiation between the populations. LD12 is already an unusual taxonomic group as there are no closely related populations of freshwater bacteria at roughly the family level Newton et al. (2011). This is unusual and might be due to LD12's transition into freshwater causing it to loose access to the wide accessory gene pool shared by its

marine sister clade SAR11 (Eiler et al., 2016). To investigate this further a biogeographical study across a salt water gradient may be needed. When LD12 was isolated, it was shown that some strains of LD12 could grow in brackish water (Henson et al., 2018a,b). Since LD12 has been isolated since the publication of this chapter, we could also use these techniques to isolate the populations found in this study. This would allow us to perform more comparative genomics analysis on their genomes, compete populations against one another, and perform biochemical tests which may help elucidate the differences between these two diverging groups.

In chapter three, we investigated the changes in *Polynucleobacter* populations recovered from Trout Bog. To follow up, I would do an analysis of the environmental and phytoplankton community and how that might affect the *Polynucleobacter* populations. I would also follow up by further investigating the differences between the HMS19 genes and the genes found in the 4 most closely related SAGs. So far, it has been rare for us to find more than one SAG which belongs to the same sequence discrete population. These new *Polynucleobacter* SAGs provide an example of many from the same population. Others have compared SAGs and MAGs from the same environments (Alneberg et al., 2018). However we have the added advantage of being able to compare across the metagenomic time series.

While I did find that coverage explains much of why the SAG genes are missing from the MAG, I would further investigate why the SAG genes with higher coverage didn't assemble. There are several factors I think might yet be possible, higher strain heterogeneity in those genes/regions, different tetranucleotide frequencies (TNF) due to recent horizontal gene transfer, or contamination. I would continue my analysis of this by calculating the nucleotide identities for reads mapped to this set of genes and test if they are lower than the rest of the genes from the HMS, as I did with coverage in chapter 3. In parallel, I could compare TNF between the genes which did assemble in the MAGS and those which only assembled in the SAGs. Contamination is a bit more difficult. A phylogenetic analysis can be done for each gene to find its taxonomy but it is likely that many of these genes are

unique and not have a related best hit. While some leniency in the level of classification can be helpful, it is not always possible to be confident in labeling these genes as contamination.

Another project building upon this work would be an in-depth analysis of the functional predictions of the gene frequency groups characterized in this study for each of the *Polynucleobacter* populations. From that analysis we might begin to understand the metabolic functions or ecological interactions which allow these groups to coexist. The analysis set forth in chapter three, could also be applied to within other genera or taxonomic groups of microbes from the two lakes studied in this thesis. Using the framework provided in this study, we can begin to characterize the metapangenome for a range of phylogenetically diverse populations.

Our understanding of all the bacterial populations studied in this thesis could benefit from recent computational advances in deconvolution of strains in metagenomes Quince et al. (2017). The prediction of strains within each of the sequence-discrete populations and tracking of the genotypes which make up the population may provide additional insight into the intra-population dynamics. The four SAGs and the HMS from the most abundant *Polynucleobacter* population reported in chapter three provides a good opportunity for validation of the genetic linkages predicted by strain deconvolution.

A.1 Chapter 1 Supplementary Figures

- Supplementary Figure 1
- Supplementary Figure 2
- Supplementary Figure 3
- Supplementary Figure 4
- Supplementary Figure 5
- Supplementary Figure 6

A.2 Chapter 1 Supplementary Tables

- Supplementary Table 1
- Supplementary Table 2
- Supplementary Table 3
- Supplementary Table 4
- Supplementary Table 5
- Supplementary Table 6
- Supplementary Table 7
- Supplementary Table 8

A.3 Chapter 2 Supplementary Figures

- Supplementary Figure 1
- Supplementary Figure 2
- Supplementary Figure 3
- Supplementary Table 4

- Supplementary Figure 5
- Supplementary Figure 6
- Supplementary Figure 7
- Supplementary Figure 8
- Supplementary Figure Text

A.4 Chapter 2 Supplementary Tables

- Supplementary Table 1
- Supplementary Table 2
- Supplementary Table 3
- Supplementary Table 4
- Supplementary Table 5
- Supplementary Table 6

A.5 Chapter 3 Supplementary Figures and Tables

Bin Name	HMS	# of contigs	Size (bp)	GC content	N50	L50
3300020704.bin.21	HMS3	245	1414743	0.447	66	6402
3300020715.bin.9	HMS3	267	1253311	0.445	85	4934
3300020717.bin.5	HMS3	229	1359862	0.448	65	6618
3300020719.bin.2	HMS3	100	1781098	0.447	19	32868
3300020720.bin.1	HMS3	156	1679634	0.447	32	15176
3300020723.bin.5	HMS3	255	1569064	0.447	69	7361
3300020734.bin.15	HMS3	208	1326225	0.449	54	7368
3300021135.bin.8	HMS3	102	1980604	0.446	20	33566
3300020679.bin.1	HMS19	152	1623737	0.466	34	15430

Bin Name	HMS	# of contigs	Size (bp)	GC content	N50	L50
3300020680.bin.4	HMS19	189	1609823	0.466	48	11043
3300020681.bin.5	HMS19	243	1370058	0.465	67	6106
3300020682.bin.6	HMS19	169	1475039	0.467	43	11251
3300020684.bin.8	HMS19	218	1598049	0.466	57	9381
3300020687.bin.3	HMS19	210	1627550	0.465	52	9149
3300020688.bin.7	HMS19	217	1636841	0.465	53	9451
3300020690.bin.3	HMS19	162	1815124	0.465	35	15852
3300020691.bin.11	HMS19	200	1682871	0.465	40	11885
3300020692.bin.4	HMS19	134	1866841	0.464	31	19768
3300020697.bin.9	HMS19	167	1583648	0.466	33	14059
3300020699.bin.11	HMS19	158	1823723	0.465	34	16171
3300020700.bin.2	HMS19	134	1837328	0.465	28	19170
3300020701.bin.14	HMS19	214	1684618	0.465	51	10554
3300020703.bin.10	HMS19	123	1879086	0.465	26	22815
3300020704.bin.14	HMS19	255	1498954	0.464	71	6686
3300020706.bin.5	HMS19	227	1435263	0.466	64	7091
3300020707.bin.9	HMS19	208	1732714	0.465	50	11304
3300020708.bin.4	HMS19	141	1857435	0.464	31	20957
3300020709.bin.11	HMS19	178	1638127	0.464	38	12943
3300020711.bin.15	HMS19	160	1470393	0.468	41	12599
3300020713.bin.8	HMS19	208	1706462	0.465	51	10513
3300020715.bin.12	HMS19	278	1539919	0.464	80	6462
3300020721.bin.8	HMS19	155	1646454	0.466	30	14876
3300020722.bin.7	HMS19	197	1610479	0.466	47	10616
3300020723.bin.27	HMS19	256	1457840	0.464	66	6441

Bin Name	HMS	# of contigs	Size (bp)	GC content	N50	L50
3300020724.bin.5	HMS19	204	1643626	0.465	47	10211
3300020726.bin.21	HMS19	188	1927319	0.464	46	13214
3300020729.bin.21	HMS19	253	1183692	0.463	84	4920
3300020730.bin.6	HMS19	239	1370761	0.466	68	6446
3300020734.bin.1	HMS19	131	1743177	0.466	30	19116
3300020735.bin.10	HMS19	226	1820378	0.463	57	10933
3300021113.bin.5	HMS19	145	1892224	0.464	33	17715
3300021116.bin.1	HMS19	146	1706037	0.465	32	17556
3300020679.bin.4	HMS23	73	1786518	0.457	11	45452
3300020682.bin.3	HMS23	110	1725738	0.457	14	26529
3300020683.bin.2	HMS23	108	1732723	0.457	16	30877
3300020687.bin.5	HMS23	119	1818311	0.456	11	43955
3300020688.bin.11	HMS23	113	1756877	0.456	12	33749
3300020691.bin.5	HMS23	101	1788213	0.457	10	45070
3300020697.bin.12	HMS23	87	1748815	0.458	10	39925
3300020698.bin.6	HMS23	119	1714411	0.457	16	26128
3300020701.bin.11	HMS23	96	1852899	0.455	9	68057
3300020707.bin.6	HMS23	107	1851432	0.457	10	39574
3300020709.bin.10	HMS23	99	1895790	0.453	12	44067
3300020721.bin.2	HMS23	104	1784579	0.456	11	32619
3300020722.bin.6	HMS23	104	1885812	0.455	10	52264
3300020724.bin.7	HMS23	104	1915987	0.454	11	52458
3300020734.bin.20	HMS23	77	1693699	0.457	9	52135
3300021116.bin.5	HMS23	84	1865246	0.457	14	44093
3300020682.bin.1	HMS18	196	1285252	0.47	51	7578

Bin Name	HMS	# of contigs	Size (bp)	GC content	N50	L50
3300020683.bin.6	HMS18	200	1310003	0.471	53	7305
3300020698.bin.10	HMS18	191	1296402	0.472	50	8518
3300020711.bin.5	HMS18	178	1301155	0.47	44	9267
3300020683.bin.10	HMS28	125	2300397	0.444	22	31216
3300020687.bin.9	HMS28	271	2288631	0.442	57	11557
3300020691.bin.10	HMS28	245	2232834	0.442	49	12495
3300020697.bin.11	HMS28	166	2570966	0.442	20	28686
3300020698.bin.11	HMS28	85	2409168	0.443	14	63413
3300020701.bin.4	HMS28	210	2424786	0.443	36	19558
3300020707.bin.4	HMS28	312	2117314	0.441	80	7833
3300020709.bin.12	HMS28	198	2342758	0.443	33	18908
3300020711.bin.13	HMS28	273	2334059	0.442	64	10929
3300020721.bin.13	HMS28	139	2393274	0.443	16	36589
3300020722.bin.5	HMS28	273	2292746	0.441	64	10468
3300020724.bin.8	HMS28	258	2371896	0.443	53	12838
3300020734.bin.14	HMS28	74	2476275	0.441	6	102299
3300020707.bin.3	HMS10	280	1304046	0.451	89	4918

Table A.1: Medium quality MAG assembly stats

Bin Name	HMS	CheckM Taxon	Completeness (checkM)	Redundancy (checkM)
3300020704.bin.21	HMS3	o__Burkholderiales	71.52	2.44
3300020715.bin.9	HMS3	o__Burkholderiales	56.72	0.72
3300020717.bin.5	HMS3	o__Burkholderiales	71.05	0.96
3300020719.bin.2	HMS3	o__Burkholderiales	89.01	0.23

Bin Name	HMS	CheckM Taxon	Completeness (checkM)	Redundancy (checkM)
3300020720.bin.1	HMS3	o__Burkholderiales	85.59	0.55
3300020723.bin.5	HMS3	o__Burkholderiales	75.59	2.06
3300020734.bin.15	HMS3	o__Burkholderiales	67.5	1.98
3300021135.bin.8	HMS3	o__Burkholderiales	93.21	1.03
3300020679.bin.1	HMS19	o__Burkholderiales	84.51	0.55
3300020680.bin.4	HMS19	o__Burkholderiales	82.87	0.29
3300020681.bin.5	HMS19	o__Burkholderiales	60.52	1.4
3300020682.bin.6	HMS19	k__Bacteria	71.9	0.86
3300020684.bin.8	HMS19	o__Burkholderiales	79.47	0.7
3300020687.bin.3	HMS19	k__Bacteria	75.86	1.72
3300020688.bin.7	HMS19	o__Burkholderiales	78.1	1.92
3300020690.bin.3	HMS19	o__Burkholderiales	88.12	1.03
3300020691.bin.11	HMS19	o__Burkholderiales	82.6	1.21
3300020692.bin.4	HMS19	o__Burkholderiales	91.85	1.5
3300020697.bin.9	HMS19	k__Bacteria	70.69	0
3300020699.bin.11	HMS19	o__Burkholderiales	90.51	0.96
3300020700.bin.2	HMS19	o__Burkholderiales	89.43	0.4
3300020701.bin.14	HMS19	o__Burkholderiales	84.73	0.45
3300020703.bin.10	HMS19	o__Burkholderiales	89.7	0.78
3300020704.bin.14	HMS19	k__Bacteria	68.28	1.72
3300020706.bin.5	HMS19	k__Bacteria	65.52	0
3300020707.bin.9	HMS19	k__Bacteria	77.59	0
3300020708.bin.4	HMS19	o__Burkholderiales	90.58	0.98
3300020709.bin.11	HMS19	o__Burkholderiales	78.03	1.85

Bin Name	HMS	CheckM Taxon	Completeness (checkM)	Redundancy (checkM)
3300020711.bin.15	HMS19	k__Bacteria	77.07	2.59
3300020713.bin.8	HMS19	o__Burkholderiales	83.56	0.73
3300020715.bin.12	HMS19	o__Burkholderiales	71.16	3.11
3300020721.bin.8	HMS19	o__Burkholderiales	82.03	1.58
3300020722.bin.7	HMS19	o__Burkholderiales	82.75	2.3
3300020723.bin.27	HMS19	o__Burkholderiales	71.83	2.54
3300020724.bin.5	HMS19	k__Bacteria	78.62	3.45
3300020726.bin.21	HMS19	o__Burkholderiales	90.65	1.48
3300020729.bin.21	HMS19	o__Burkholderiales	55.64	1.15
3300020730.bin.6	HMS19	o__Burkholderiales	67.43	0.75
3300020734.bin.1	HMS19	k__Bacteria	82.76	1.72
3300020735.bin.10	HMS19	o__Burkholderiales	87.29	0.97
3300021113.bin.5	HMS19	o__Burkholderiales	91.02	1.13
3300021116.bin.1	HMS19	k__Bacteria	75.34	2.59
3300020679.bin.4	HMS23	o__Burkholderiales	92.86	1.65
3300020682.bin.3	HMS23	o__Burkholderiales	92.39	0.39
3300020683.bin.2	HMS23	o__Burkholderiales	91.43	0.75
3300020687.bin.5	HMS23	o__Burkholderiales	91.05	1.79
3300020688.bin.11	HMS23	o__Burkholderiales	88.17	1.64
3300020691.bin.5	HMS23	k__Bacteria	74.14	2.59
3300020697.bin.12	HMS23	o__Burkholderiales	87.52	2.43
3300020698.bin.6	HMS23	o__Burkholderiales	90.87	1.04
3300020701.bin.11	HMS23	o__Burkholderiales	90.52	1.4
3300020707.bin.6	HMS23	o__Burkholderiales	93.86	1.06

Bin Name	HMS	CheckM Taxon	Completeness (checkM)	Redundancy (checkM)
3300020709.bin.10	HMS23	o__Burkholderiales	92.15	1.92
3300020721.bin.2	HMS23	o__Burkholderiales	88.8	2.21
3300020722.bin.6	HMS23	o__Burkholderiales	93	1.91
3300020724.bin.7	HMS23	o__Burkholderiales	95.66	2.93
3300020734.bin.20	HMS23	o__Burkholderiales	88.85	0
3300021116.bin.5	HMS23	o__Burkholderiales	96.69	2.08
3300020682.bin.1	HMS18	o__Burkholderiales	76.25	5.36
3300020683.bin.6	HMS18	o__Burkholderiales	78.13	3.55
3300020698.bin.10	HMS18	o__Burkholderiales	80.74	7.77
3300020711.bin.5	HMS18	o__Burkholderiales	78.53	3.7
3300020683.bin.10	HMS28	o__Burkholderiales	93.41	1.68
3300020687.bin.9	HMS28	o__Burkholderiales	89.54	2.92
3300020691.bin.10	HMS28	o__Burkholderiales	89.85	2.7
3300020697.bin.11	HMS28	o__Burkholderiales	93.78	5.69
3300020698.bin.11	HMS28	o__Burkholderiales	94.82	1.23
3300020701.bin.4	HMS28	o__Burkholderiales	91.56	6.12
3300020707.bin.4	HMS28	o__Burkholderiales	80.52	2.58
3300020709.bin.12	HMS28	o__Burkholderiales	90.39	3.13
3300020711.bin.13	HMS28	o__Burkholderiales	87.75	1.81
3300020721.bin.13	HMS28	o__Burkholderiales	96.78	3.52
3300020722.bin.5	HMS28	o__Burkholderiales	89.55	2.54
3300020724.bin.8	HMS28	o__Burkholderiales	89.02	3.84
3300020734.bin.14	HMS28	o__Burkholderiales	98.81	3.38
3300020707.bin.3	HMS10	o__Burkholderiales	62.55	5.88

Bin Name	HMS	CheckM Taxon	Completeness (checkM)	Redundancy (checkM)

Table A.2: Medium quality MAG completeness stats

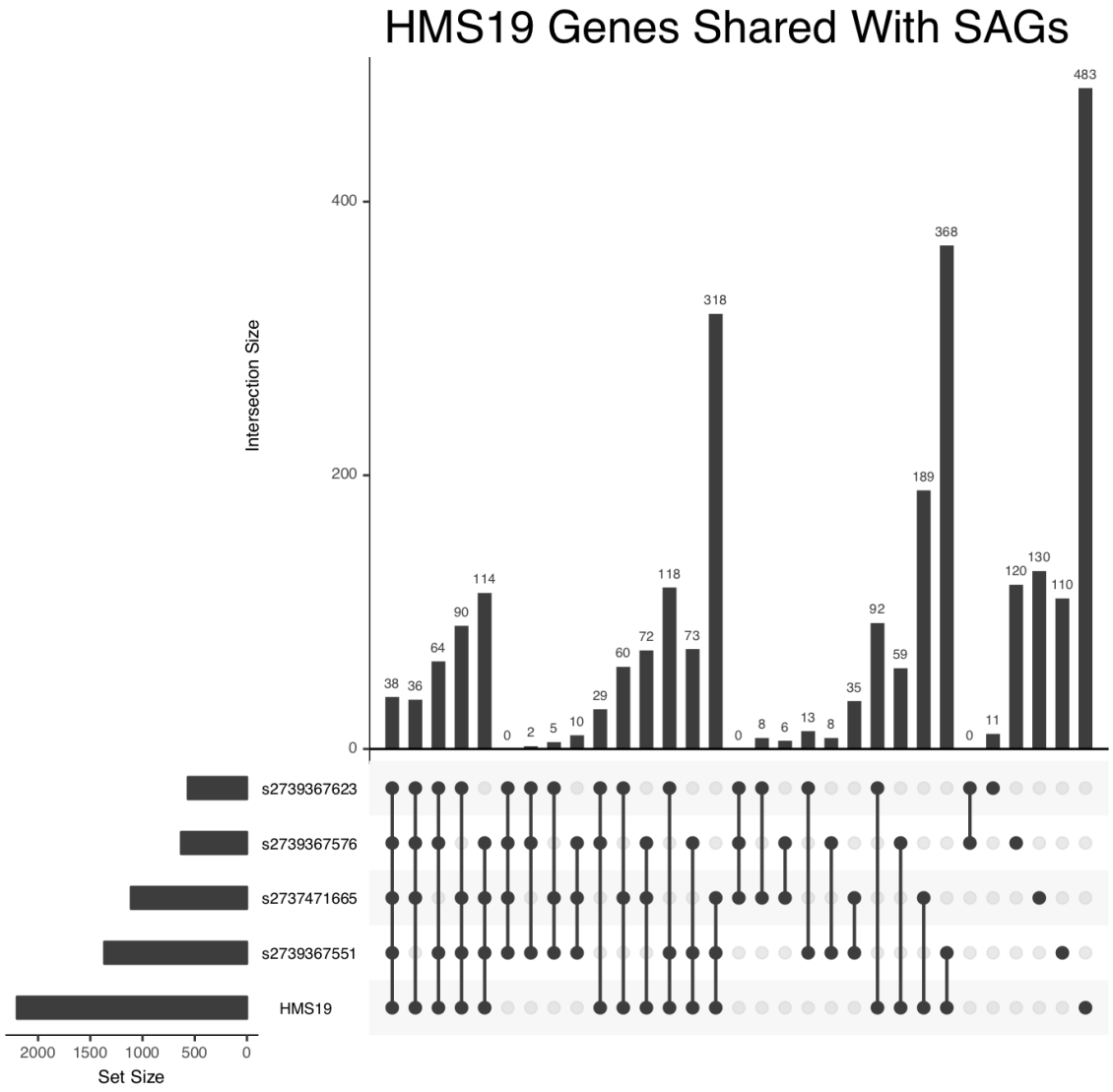

Figure A.1: Grouped Histogram of Genes Shared Between HMS19 and Pnec SAGs. Genes were clustered by using blastp (Camacho et al., 2009) on all pairwise comparisons of the amino acid sequences for each predicted gene then the results were clustered with MCL (van Dongen, 2000; Enright et al., 2002). The set size histogram on the bottom left shows the number of genes in each genome. The upper histogram shows the number of genes in the group highlighted by the black dots on the x axis. Groups are ordered by the number of genomes in the group.

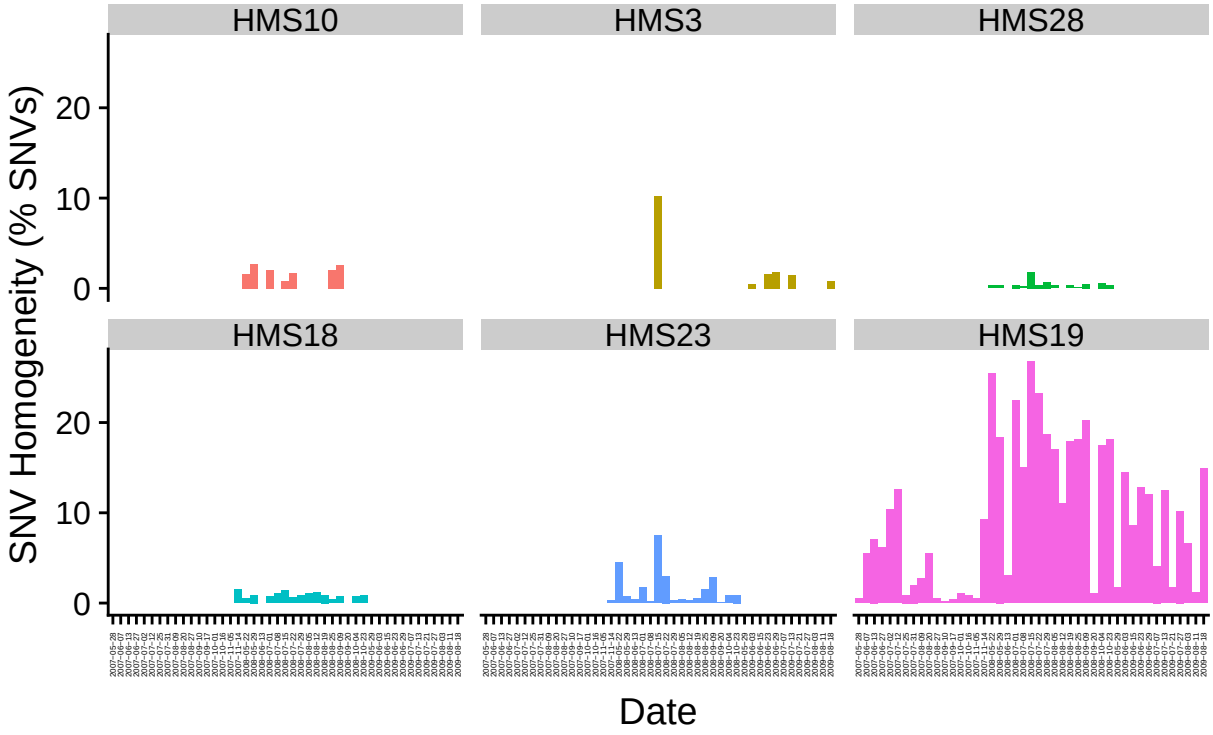

Figure A.2: HMS SNV Homogeneity Through Time. For each HMS the SNV homogeneity across the samples. SNV homogeneity is the percentage of SNVs which are $>$ 90% one variant. Homogeneity was not calculated for samples where average gene coverage for the HMS was below 10.

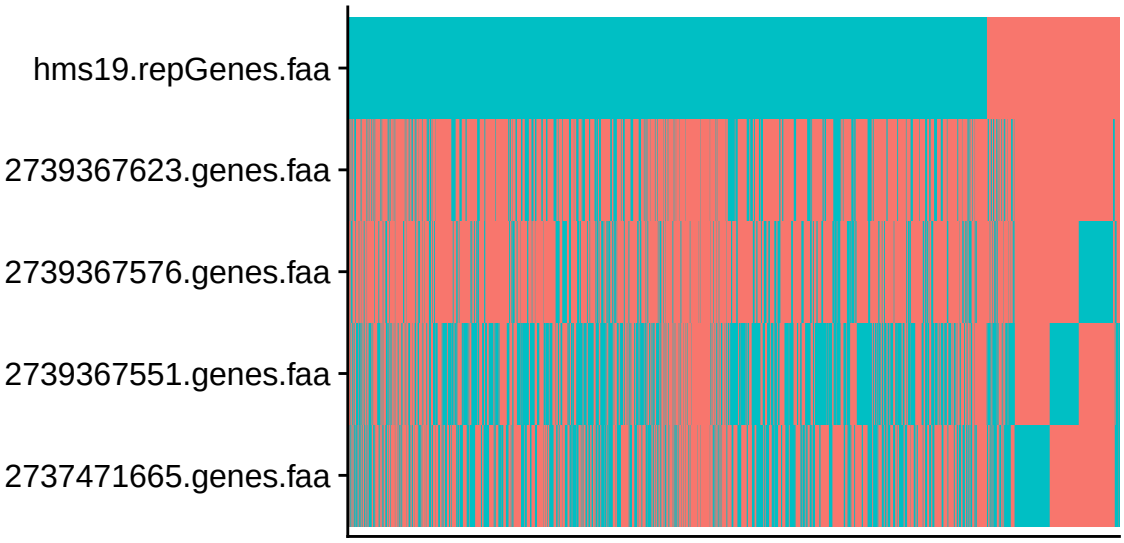

Figure A.3: Presence/Absence Map of Genes Shared Between HMS19 and Pnec SAGs. Genes were clustered by using blastp (Camacho et al., 2009) on all pairwise comparisons of the amino acid sequences for each predicted gene then the results were clustered with MCL (van Dongen, 2000; Enright et al., 2002). Each column represents a gene cluster which is colored blue if present or red if absent in the genome for that row. The clusters are ordered by HMS19 to highlight the group of genes found in the SAGs but not in HMS19.

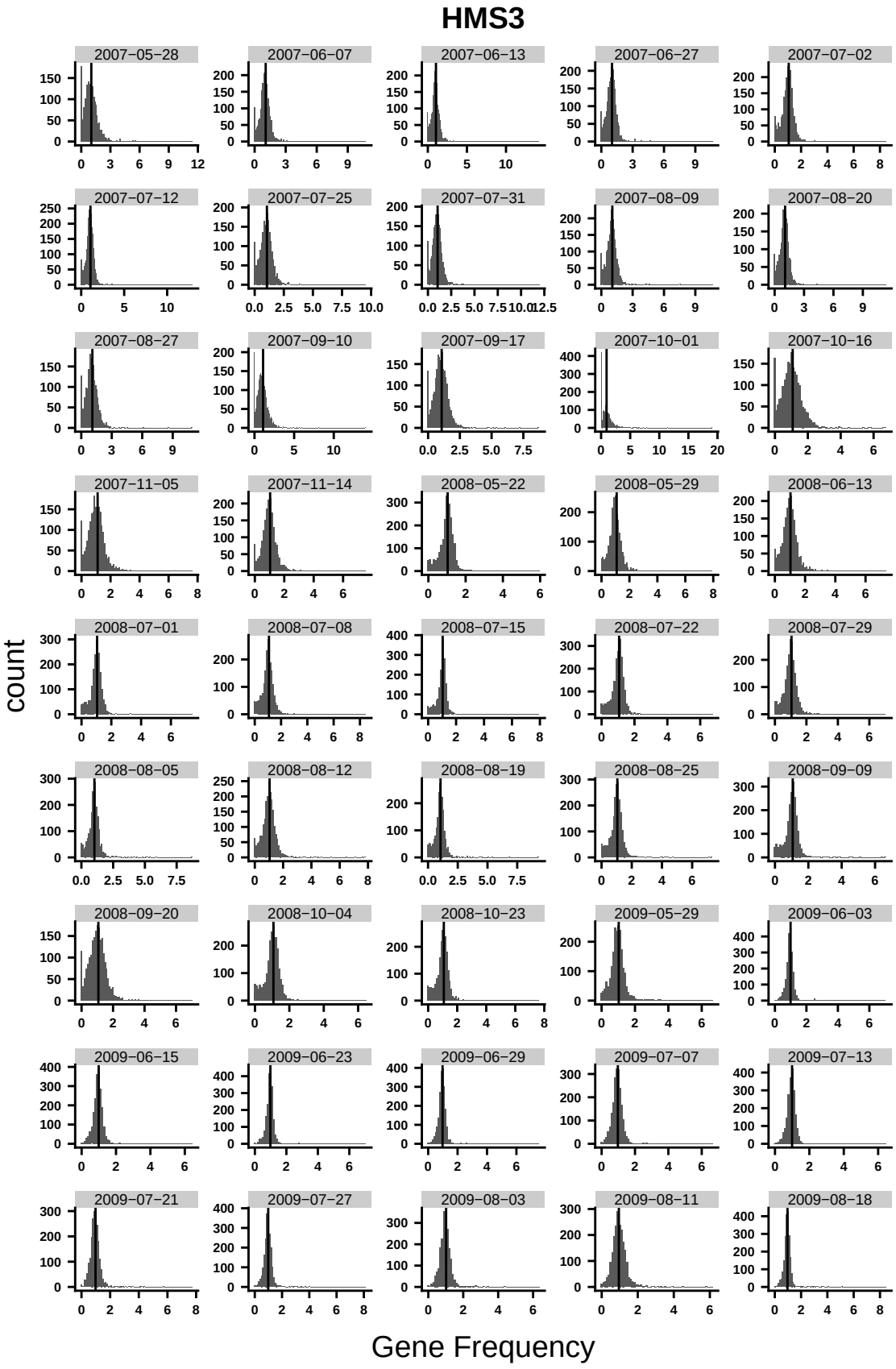

Figure A.4: HMS3 Gene Frequency Histogram By Sample

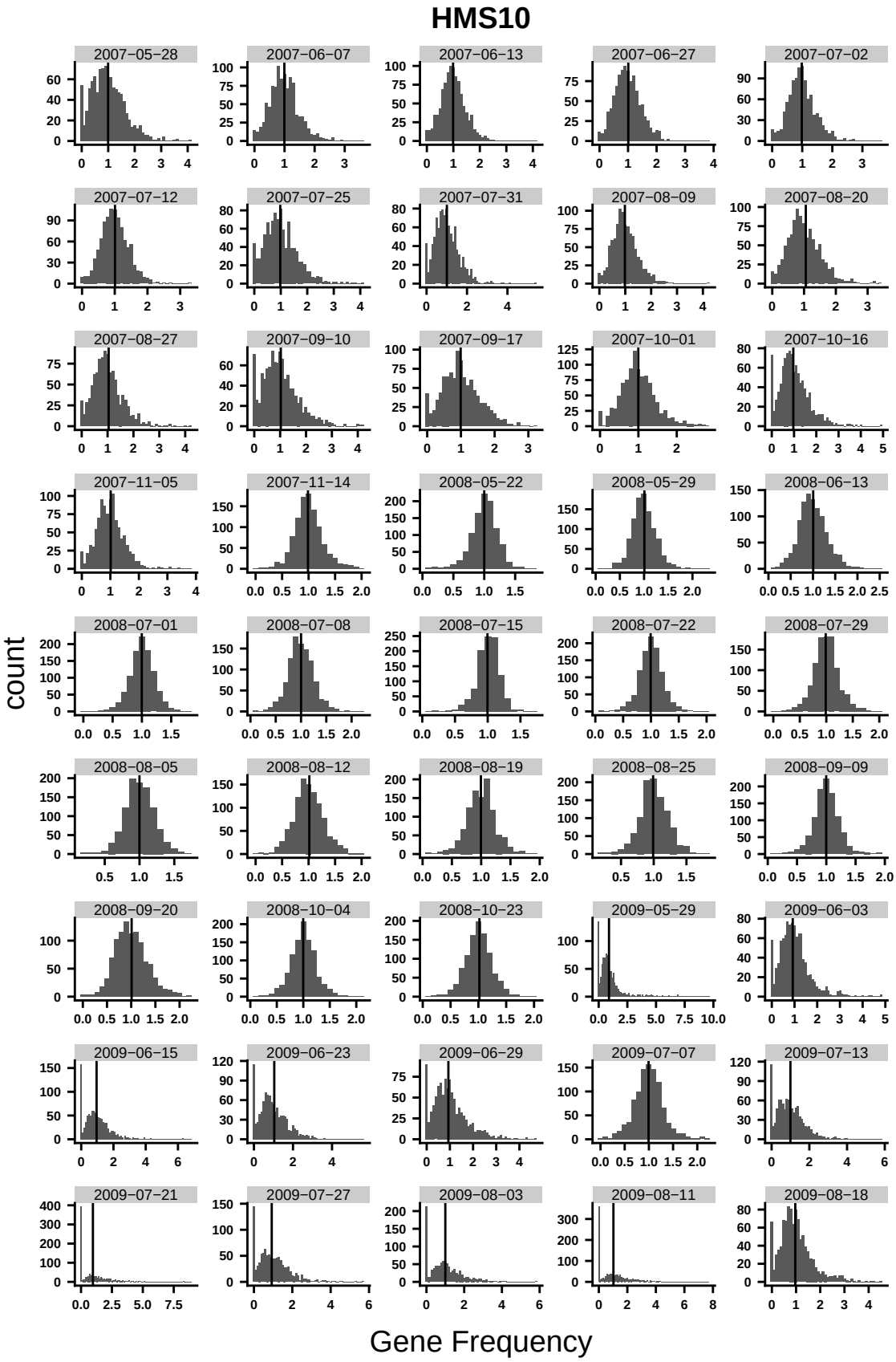

Figure A.5: HMS10 Gene Frequency Histogram By Sample

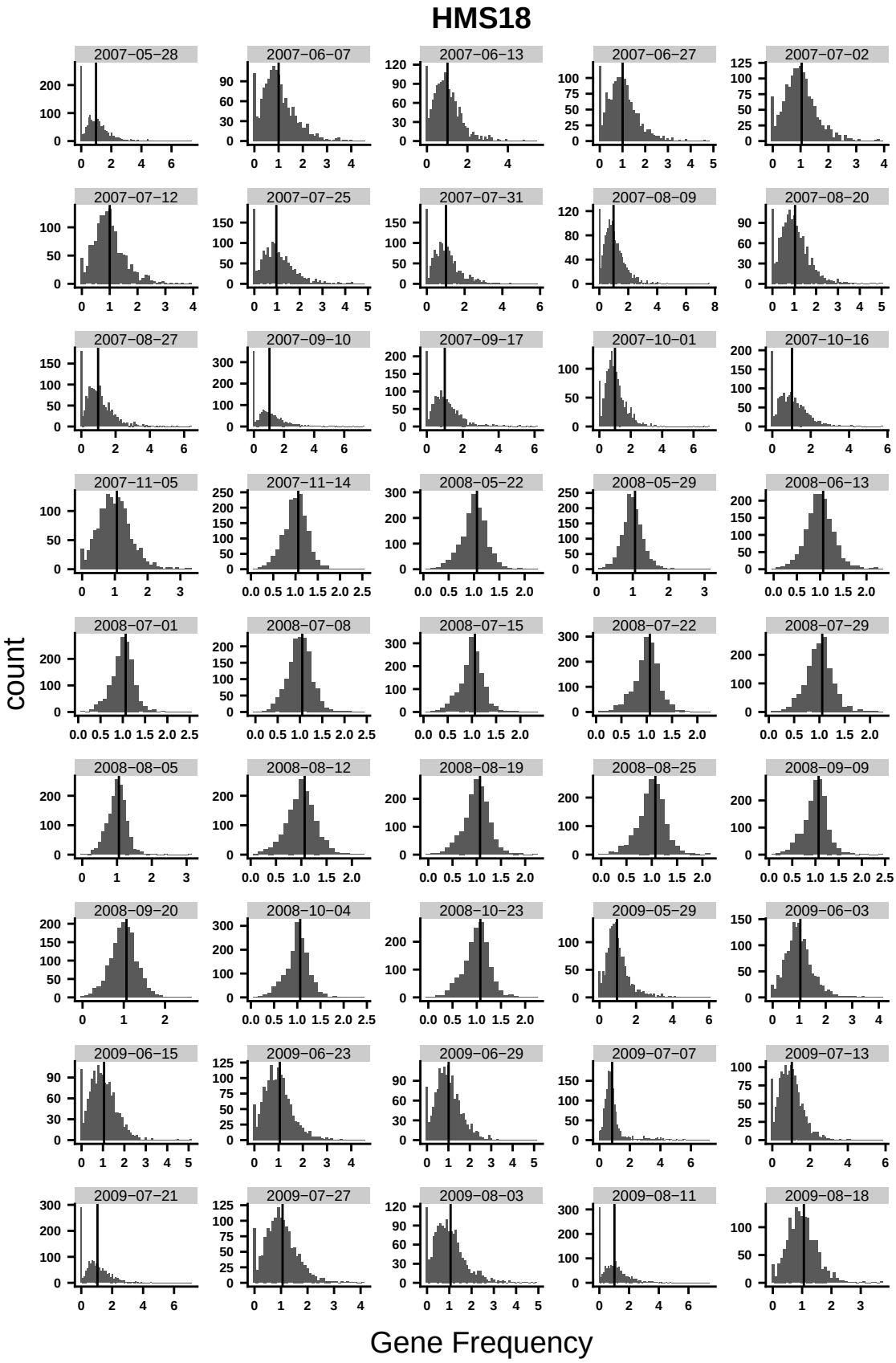

Figure A.6: HMS18 Gene Frequency Histogram By Sample

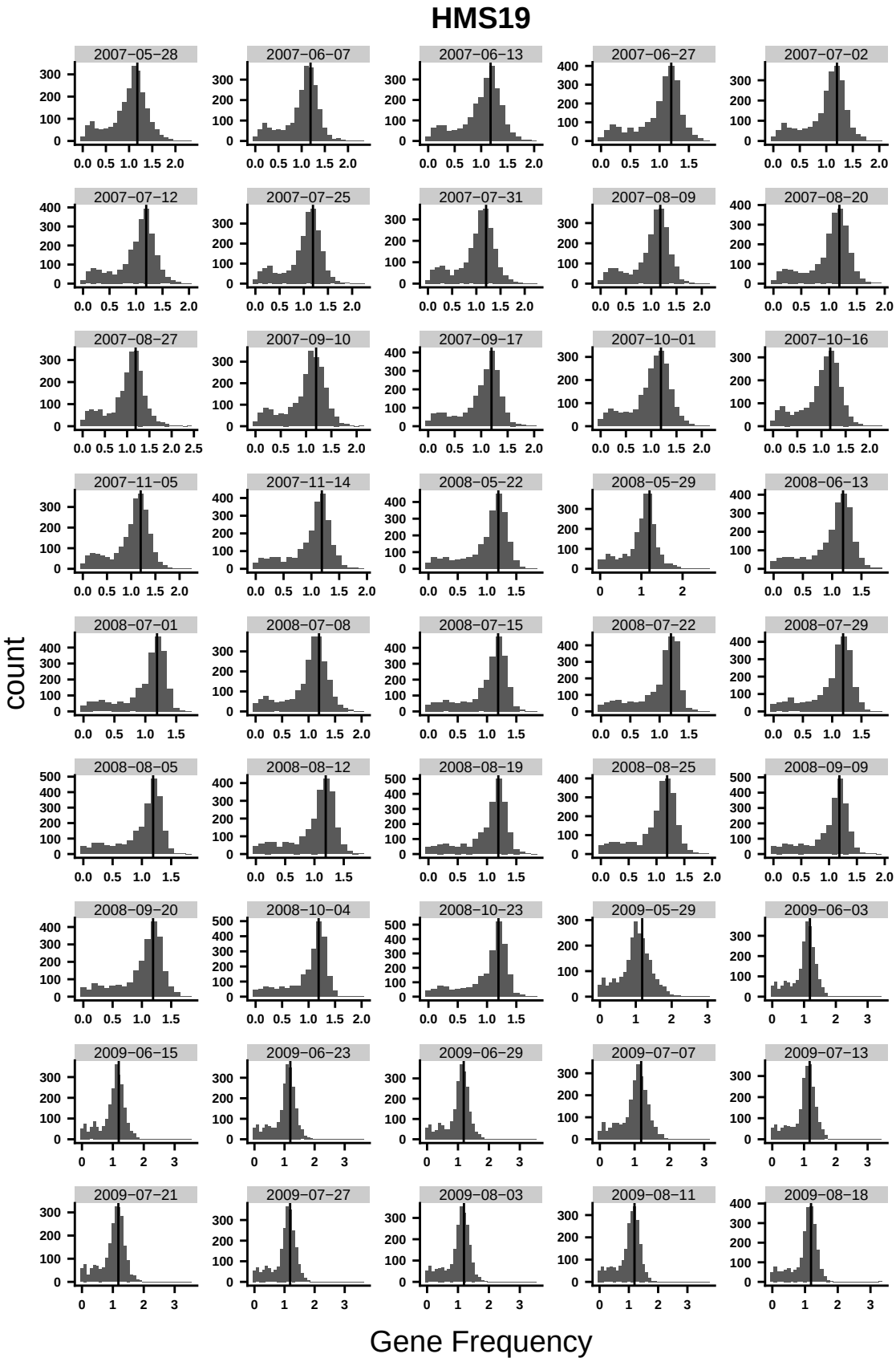

Figure A.7: HMS19 Gene Frequency Histogram By Sample

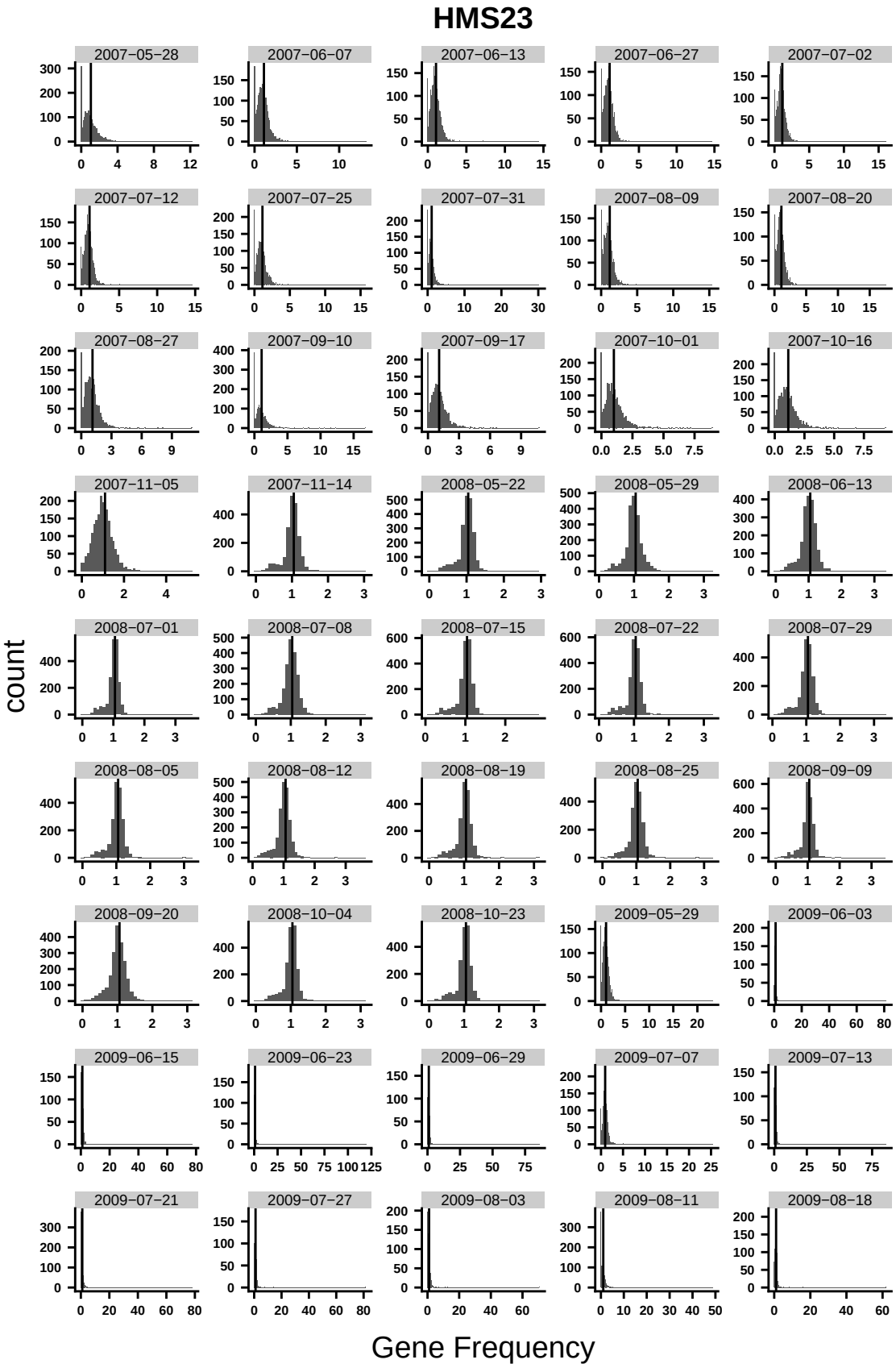

Figure A.8: HMS23 Gene Frequency Histogram By Sample

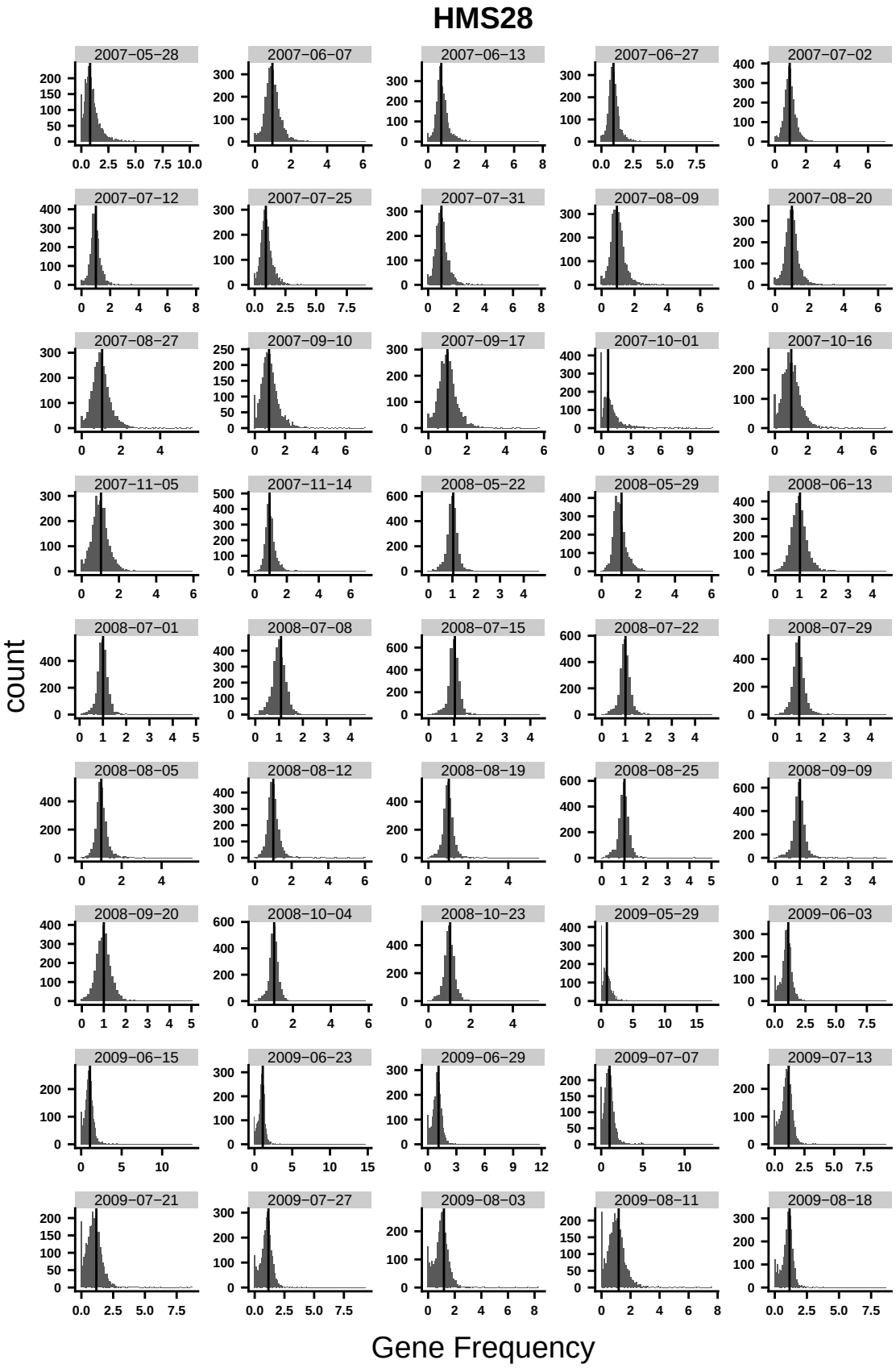

Figure A.9: HMS28 Gene Frequency Histogram By Sample

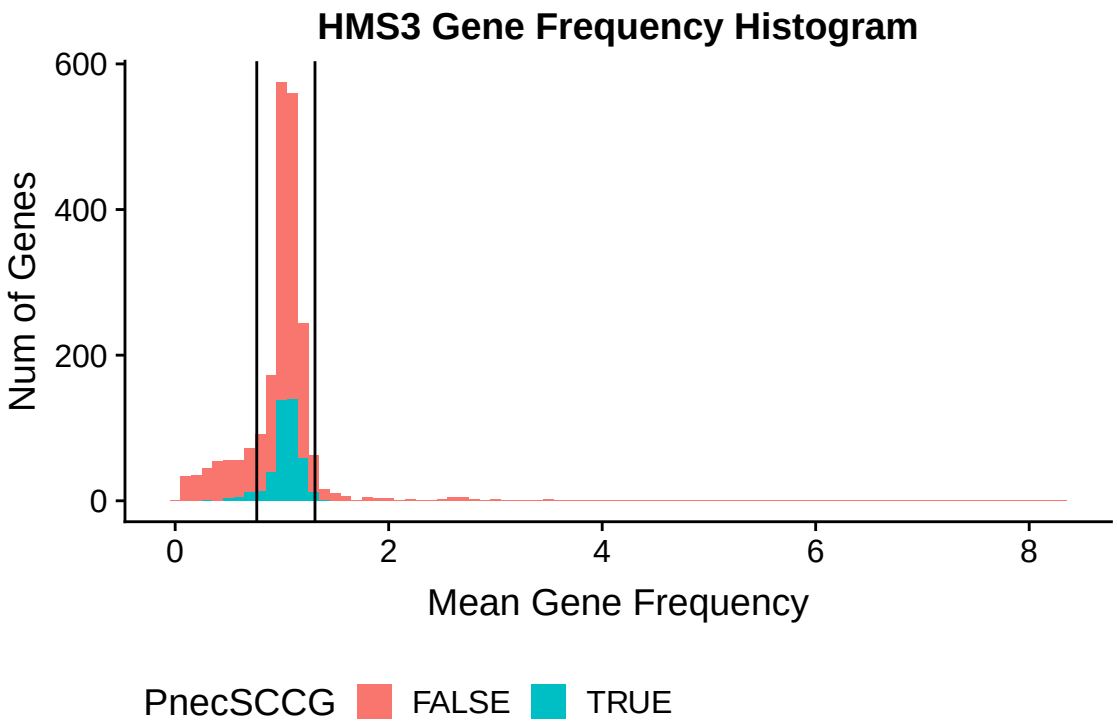

Figure A.10: HMS3 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS3. PnecSCCGs are in blue. Black horizontal lines represent the 95% confidence interval for the Pnec_SCCGs.

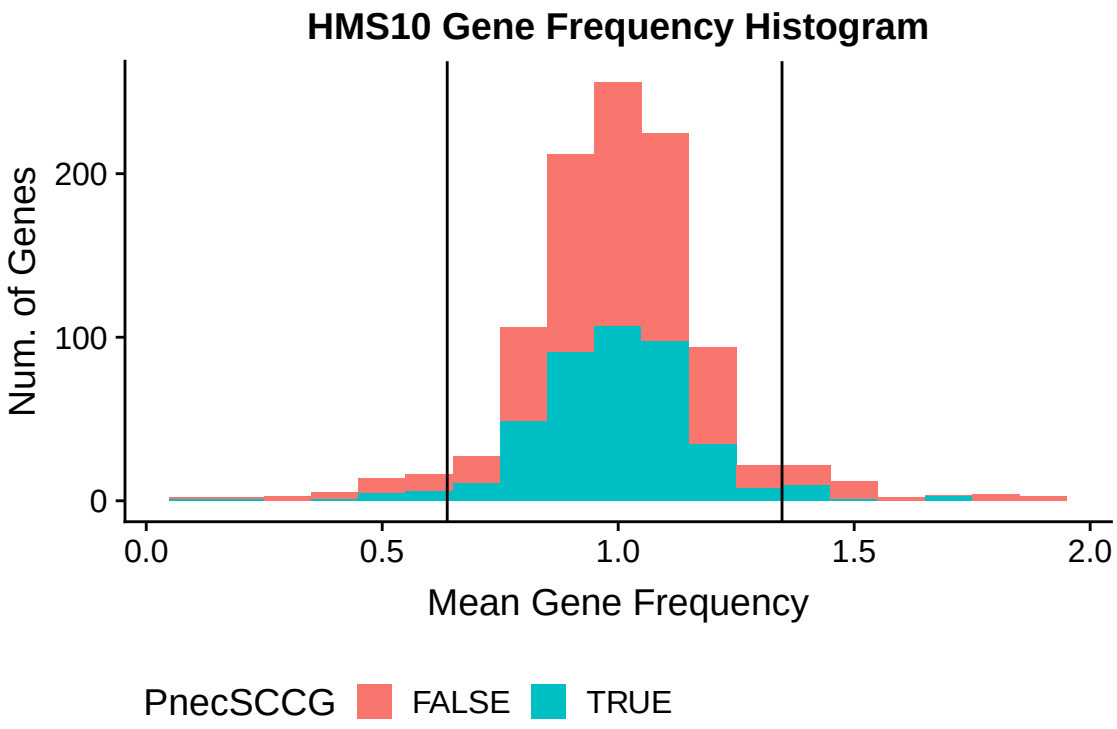

Figure A.11: HMS10 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS10. PnecSCCGs are in blue. Black horizontal lines represent the 95% confidence interval for the Pnec_SCCGs.

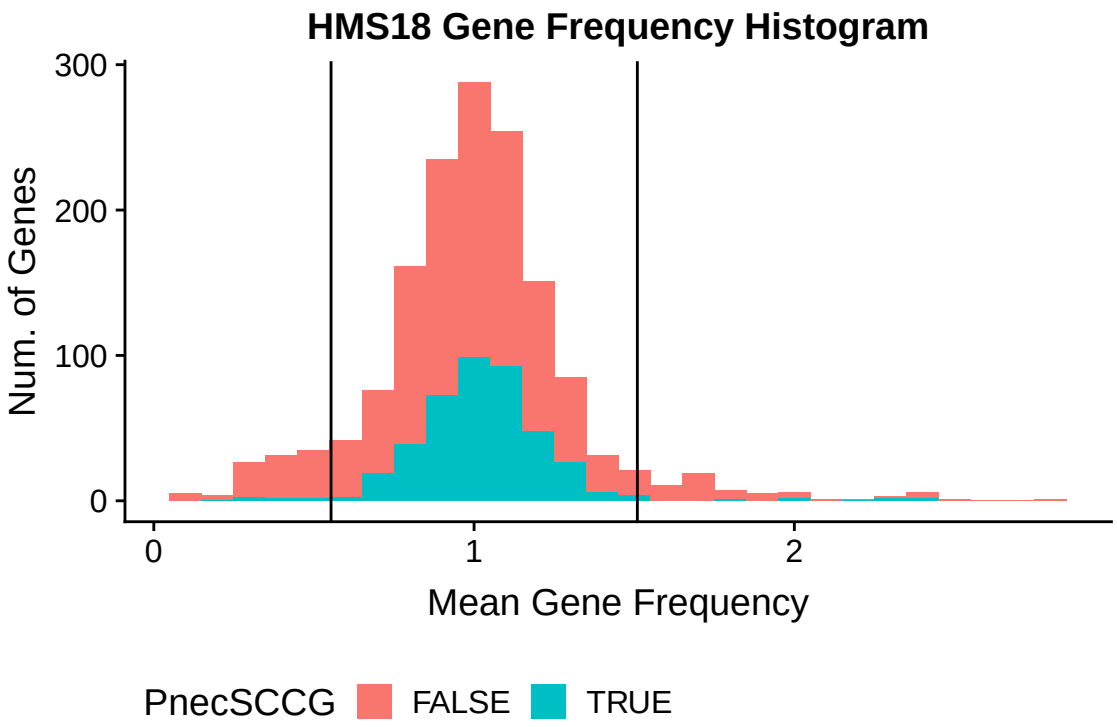

Figure A.12: HMS18 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS18. PnecSCCGs are in blue. Black horizontal lines represent the 95% confidence interval for the Pnec_SCCGs.

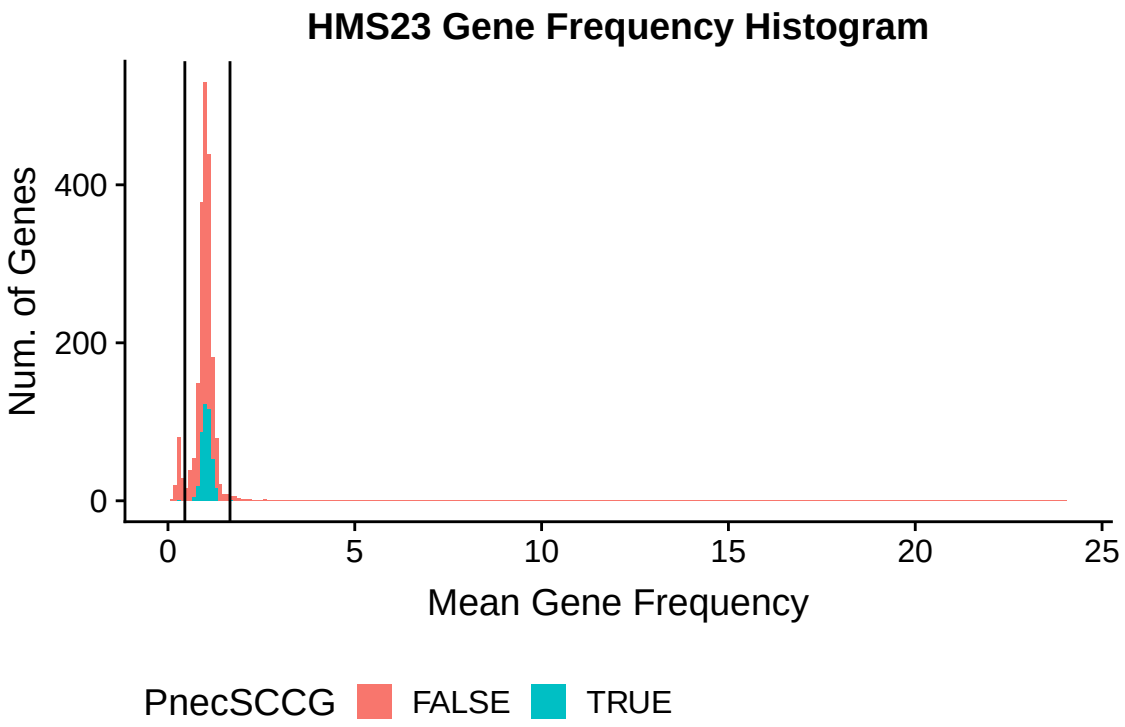

Figure A.13: HMS23 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS23. PnecSCCGs are in blue. Black horizontal lines represent the 95% confidence interval for the Pnec_SCCGs.

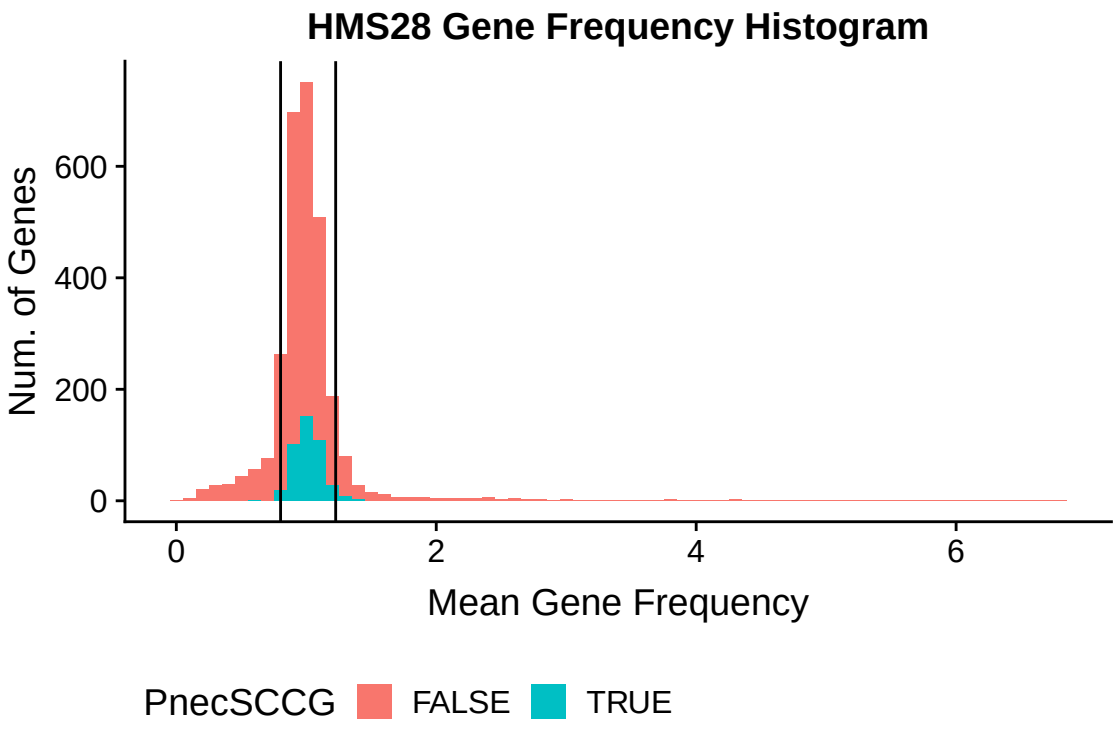

Figure A.14: HMS28 Average Gene Frequency Histogram. Average gene frequency for each gene in HMS28. PnecSCCGs are in blue. Black horizontal lines represent the 95% confidence interval for the Pnec_SCCGs.

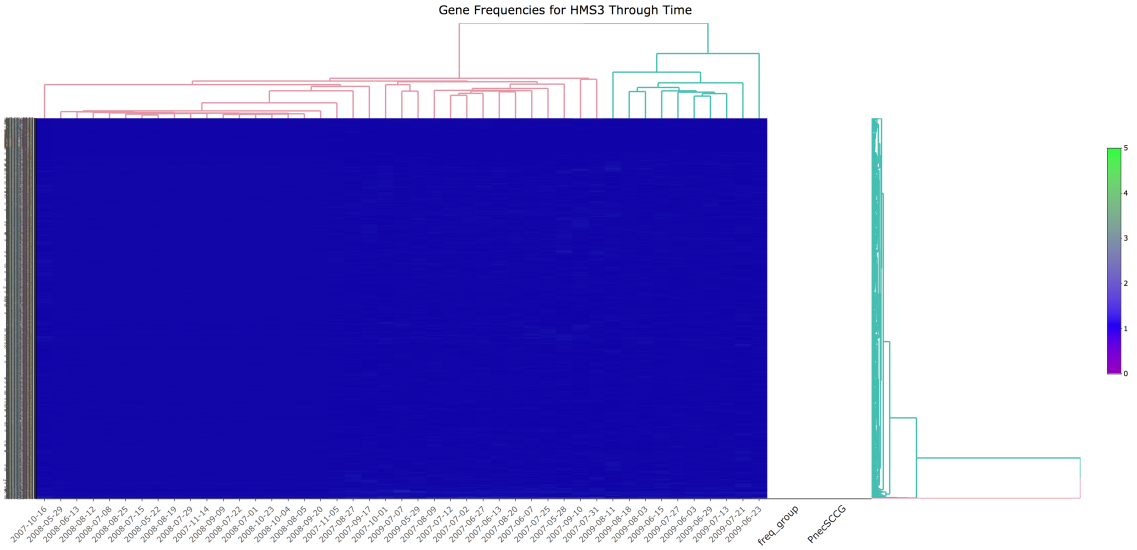

Figure A.15: Gene Frequencies for HMS3 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multipcopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot.

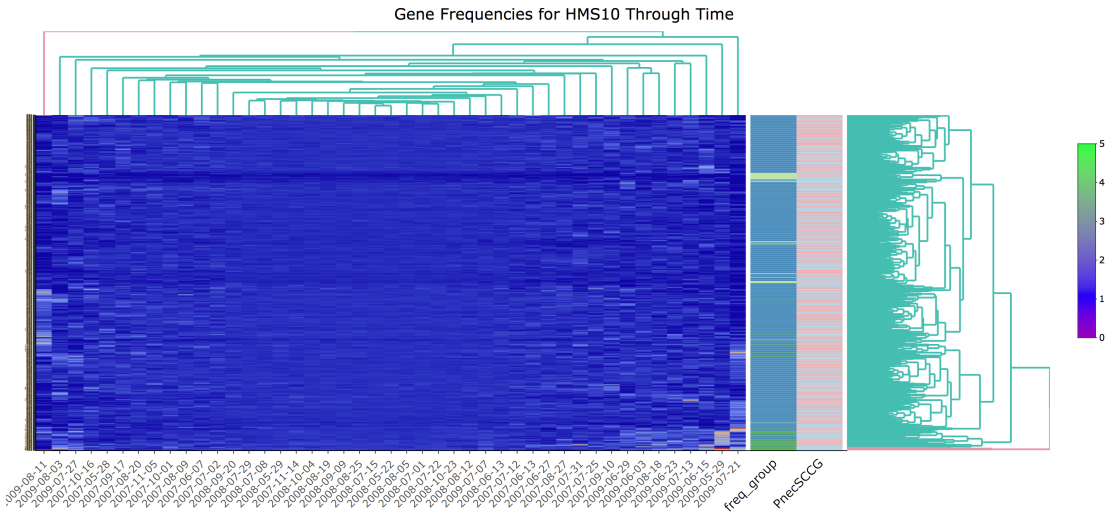

Figure A.16: Gene Frequencies for HMS10 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multipcopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot. Heatmap color scale set to max of 5. Above 5 gene frequencies are red.

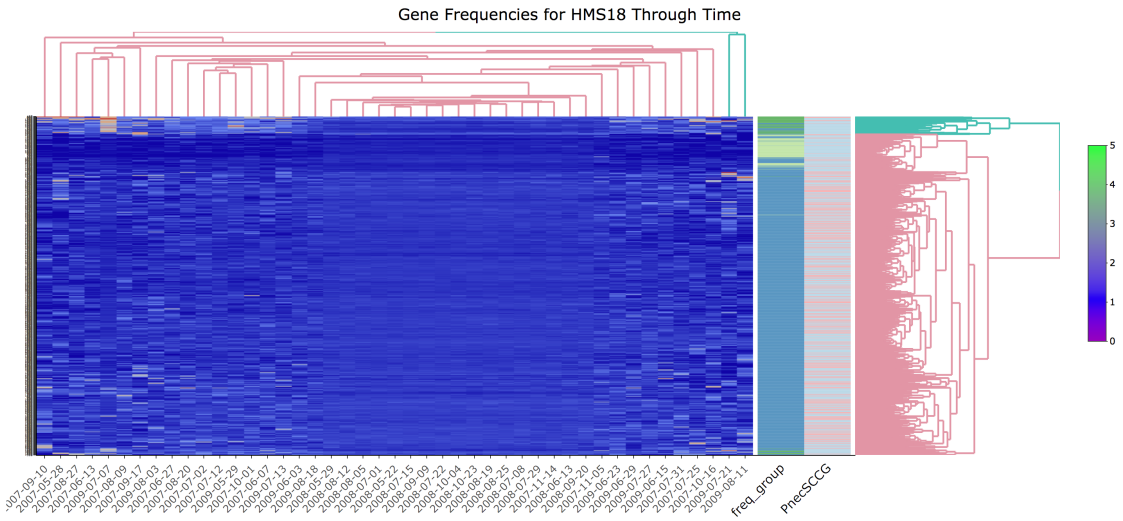

Figure A.17: Gene Frequencies for HMS18 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multipcopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot.

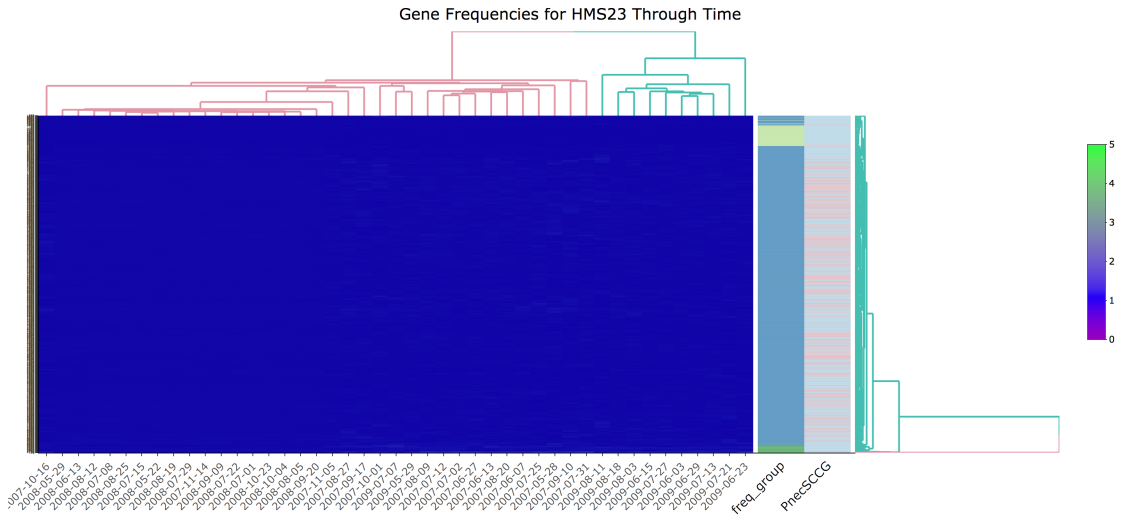

Figure A.18: Gene Frequencies for HMS23 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multipcopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot.

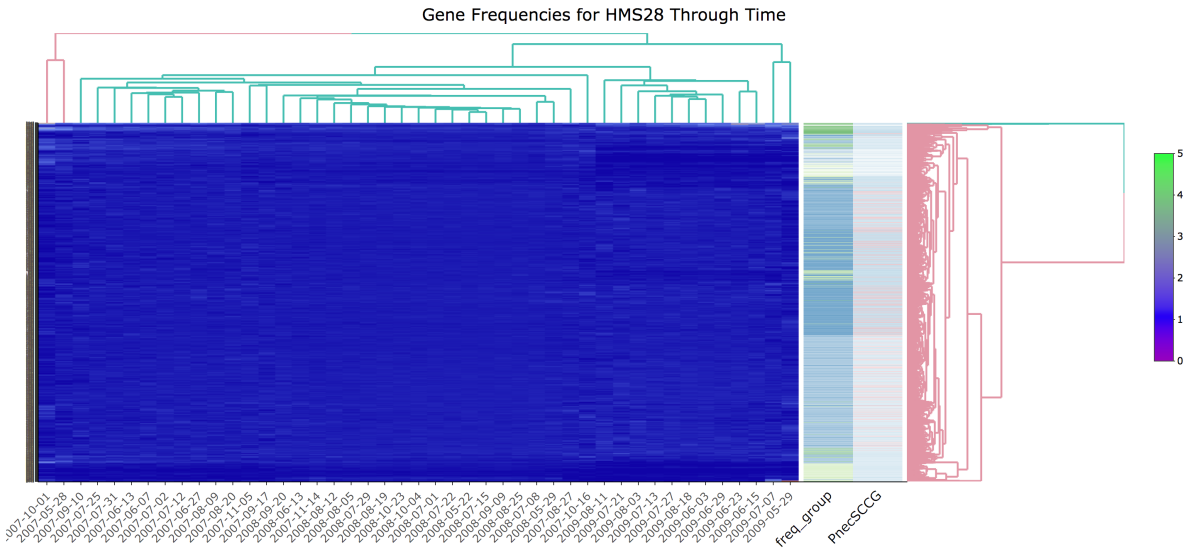

Figure A.19: Gene Frequencies for HMS28 Through Time. Color of each square for the sample dates represents the gene frequency for that sample. Dendrograms were constructed by clustering Euclidean distance between genes and sample patterns. Freq_group column shows the frequency group determined by average coverage through the time series, dark green for multipcopy, blue for high, and light green for low frequency. Pnec-SCCGs are denoted in red in the PnecSCCG column. Click here for interactive version of the plot. Heatmap color scale set to max of 5. Above 5 gene frequencies are red.

Bin Name	Classification
3300020704.bin.21	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;
3300020715.bin.9	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020717.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020719.bin.2	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020720.bin.1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020723.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020734.bin.15	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300021135.bin.8	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020679.bin.1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020680.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020681.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;
3300020682.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020684.bin.8	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020687.bin.3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020688.bin.7	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020690.bin.3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020691.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020692.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;

Bin Name	Classification
3300020697.bin.9	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020699.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020700.bin.2	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020701.bin.14	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020703.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020704.bin.14	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020706.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020707.bin.9	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020708.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020709.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020711.bin.15	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020713.bin.8	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020715.bin.12	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020721.bin.8	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020722.bin.7	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020723.bin.27	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020724.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020726.bin.21	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;

Bin Name	Classification
3300020729.bin.21	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020730.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020734.bin.1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020735.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300021113.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300021116.bin.1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020679.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020682.bin.3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020683.bin.2	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020687.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020688.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020691.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020697.bin.12	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020698.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020701.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020707.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020709.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020721.bin.2	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;

Bin Name	Classification
3300020722.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020724.bin.7	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020734.bin.20	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300021116.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020682.bin.1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020683.bin.6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020698.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020711.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020683.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;unclassified Betaproteobacteria;
3300020687.bin.9	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;
3300020691.bin.10	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;
3300020697.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020698.bin.11	Bacteria;Proteobacteria;Betaproteobacteria;
3300020701.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;
3300020707.bin.4	Bacteria;Proteobacteria;Betaproteobacteria;
3300020709.bin.12	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;
3300020711.bin.13	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;
3300020721.bin.13	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;

Bin Name	Classification
3300020722.bin.5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;
3300020724.bin.8	Bacteria;Proteobacteria;Betaproteobacteria;unclassified Betaproteobacteria;
3300020734.bin.14	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;
3300020707.bin.3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Polynucleobacter;

Table A.3: Medium Quality MAG classifications.

B BUILDING A LOCAL COMMUNITY OF PRACTICE IN SCIENTIFIC PROGRAMMING FOR LIFE SCIENTISTS

Sarah L. R. Stevens[1], Mateusz Kuzak[2], Carlos Martinez[3], Aurelia Moser[4], Petra Bleeker[5] and Marc Galland[5]

[1]*Department of Bacteriology, University of Wisconsin–Madison, Microbial Sciences Building, 1550 Linden Drive, Madison, WI 53706, USA*; [2]*Dutch Techcentre for Life Sciences (foundation office), Catharijnesingel 54, 3511 GC Utrecht, Netherlands*; [3]*Netherlands eScience Center, Science Park 140, Amsterdam, 1098 XG, Netherlands*; [4]*Mozilla Foundation, 331 East Evelyn Avenue, Mountain View, CA 94041, USA*; [5]*Department of Plant Physiology, Swammerdam Institute for Life Sciences, University of Amsterdam, 1098 XH Amsterdam, Netherlands*

All authors helped write and revise the manuscript.

Preprint: Stevens, S. L. R., Kuzak, M., Martinez, C., Moser, A., Bleeker, P. M., & Galland, M. 2018. Building a local community of practice in scientific programming for Life Scientists. *BioRxiv*, 265421. https://doi.org/10.1101/265421

Accepted for publication in PLOS Biology. Stevens S.L.R., Kuzak, M., Martinez, C., Moser, A., Bleeker, P., Galland, M. 2018. Building a local community of practice in scientific programming for life scientists. *PLoS Biol* 16(11): e2005561. https://doi.org/10.1371/journal.pbio.2005561

B.1 Abstract

In this paper, we describe why and how to build a local community of practice in scientific programming for life scientists that use computers and programming in their research. A community of practice is a small group of scientists that meet regularly to help each other and promote good practices in scientific programming. While most life scientists are well-trained in the laboratory to conduct experiments, good practices with (big) datasets and their analysis are often missing. We propose a model on how to build such a community of practice at a local academic institution, present two real-life examples and introduce challenges and implemented solutions. We believe that the current data deluge that life scientists face can benefit from the implementation of these small communities. Good practices spread among experimental scientists will foster open, transparent and sound scientific results beneficial to society.

B.2 Introduction

Life Sciences is becoming a data-driven field

In the last ten years, since the advent of the first next-generation sequencing (NGS) technologies, DNA and RNA sequencing costs have plunged to levels that make genome sequencing an affordable reality for every life scientist (Hayden, 2014; Hiraoka et al., 2016). Yet the vast majority of wet lab biologists need tailor-made, practical training to learn scientific programming and data analysis (Batut et al., 2018; Watson-Haigh et al., 2013; Friesner et al., 2017; Welch et al., 2014; Corpas et al., 2015; Schneider et al., 2012). Current efforts in bioinformatics and data science training for life scientists have been initiated worldwide to cope with these training demands (Morgan et al., 2017; Wilson, 2016; Pawlik et al., 2017; Corpas et al., 2015; Schneider et al., 2012).

Good practices in scientific programming are needed to increase research reproducibility

Modern biology is facing reproducibility issues (Baker, 2016). While evidence suggests this might not be as bad as it sounds (Fanelli, 2018), there is clearly a need for increased reproducibility. For instance, out of 400 algorithms presented at two conferences, only 6% had published their corresponding code (Hutson, 2018). Thus, most research code remains a "black box" (Morin et al., 2012) although programming is a central tool in research(Hettrick et al., 2014). Use of laboratory notebooks is widely taught in biology but not emphasized for coding. Both code documentation and better practices in data management are needed so anyone can redo or understand the analyses later on. Part of the solution lies in dedicated training to researchers to promote good programming practices (Wilson et al., 2017). One of the recent relevant initiatives is the FAIR (Findable, Accessible, Interoperable and Reusable) principles initiative which provides guidelines to boost reproducibility and reuse of datasets (Wilkinson et al., 2016). Therefore, the long term goal of any programming scientist should be to steward good practices in code-intensive research by promoting open science, reproducible research and sustainable software development.

Part of the solution: building a local community of practice

Training workshops in scientific programming are often offered as one-time courses but researchers would benefit from a more permanent support. Fueled by Etienne Wenger's idea that learning is usually a social activity (Wenger, 1998; Lave and Wenger, 1991), we propose to build a local community of practice in scientific programming for life scientists. This community fulfills the three requirements of Wenger's definition: it has a specific domain i.e. bioinformatics and data science, its members engage in common activities e.g. training events, and they are practitioners i.e. researchers currently engaged

in research that involves scientific programming. Community building and organization is a field in itself that has been considerably reviewed (Webber, 2018; Brown, 2007; Wenger, 2011; Budd et al., 2015; Li et al., 2009b; Wenger et al., 2002). Requirements include a few motivated leaders and a safe environment where participants can experiment with their new knowledge (Webber, 2018). As stated by Wenger and Snyder (Wenger and Snyder, 2000), communities of practice "help to solve problems quickly", "transfer best practices" and "develop professional skills". While short-term immediate issues ("help me now to debug my code") can be solved, the community also has the capacity to steward solutions for long-term data-related problems ("how do I comply with the FAIR guidelines?") and can therefore help to solve reproducibility issues. Communities of practice can also foster the adoption of good practices (Bauer et al., 2015a) since by co-working with their peers, scientists are probably more likely to compare their methods and embrace best practices. This paper will explicitly describe why and how to build a local community of practice in scientific programming. We propose a model of how to build such a community that we exemplify in two case studies. Finally, we discuss the challenges and possible solutions that we encountered when building these communities. Overall, we believe that building these local communities of practice in scientific programming will support and speed-up scientific research, spread good practices and, ultimately, help to tackle the data deluge in the life sciences.

B.3 Why do we need to build up local community of practice in scientific programming?

Isolation

Wet lab biologists are increasingly being asked by their supervisors to analyze a set of pre-existing data in labs where their peers have little to no coding experience. Without access

to experienced bioinformaticians, they can lead to a sentiment of isolation deleterious to their work.

Self-learning and adoption of bad practices

In such a scenario, most researchers tend to invent their own solution sometimes re-inventing the wheel. While wasting time, it also leads to the adoption of bad practices (lack of version control) and irreproducible results. While some compiled easy-to-use software such as samtools (Li et al., 2009b) can help to get started, typically researchers need to build their own collection of tools and scripts. For instance, version control is essential: we believe that using git[1] and github[2] for instance should be considered a mandatory, good practice just like accurate pipetting in the molecular lab.

Apprehension

Researchers may also fear the breadth of knowledge they need before achieving anything which may lead to `impostor syndrome`: the researcher feels like he will be exposed as a fraud and someone more competent will unveil his lack of knowledge of coding and bioinformatics. This also inhibits continued learning since the researcher is then afraid to ask for help.

The issue of how to get started

Learning to code in a research team is akin to an apprenticeship. The 'apprentice' will benefit from the experience and knowledge of more experienced team members. For instance, a researcher working on RNA-Seq for several years will be able to demonstrate the use of basic QC tools, short-read aligners, differential gene expression calls, etc. Yet, many research teams do not have an experienced bioinformatician on staff. Even in the

[1] `https://git-scm.com/`
[2] `https://github.com/`

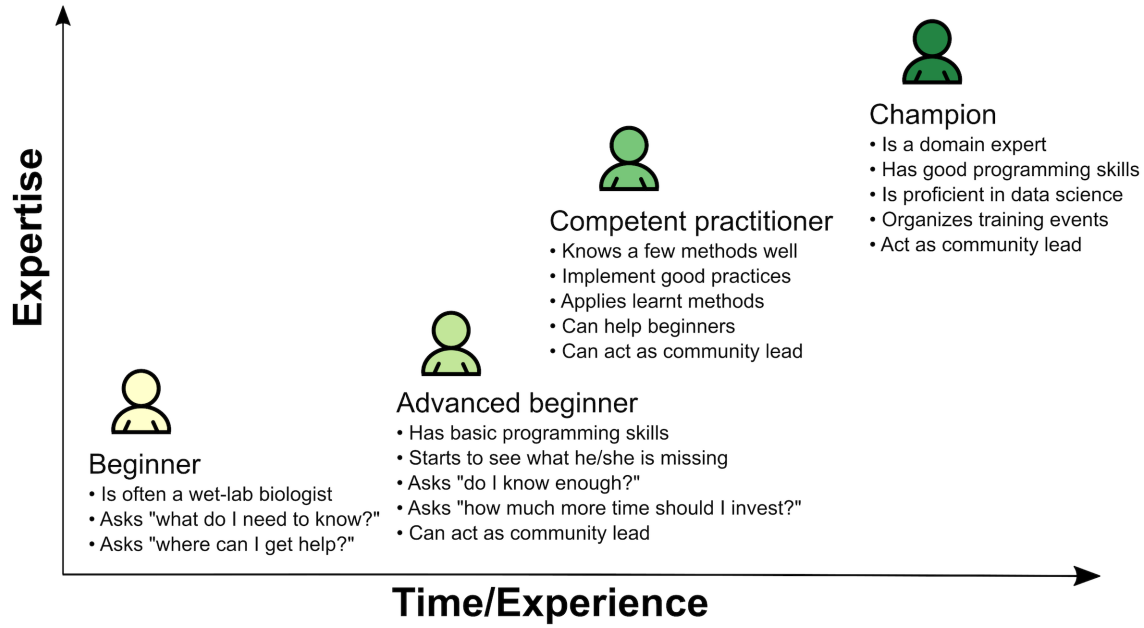

Figure B.1: Different learning stages in scientific programming. This figure displays the different stages of learning encountered by experimental biologists.

best case where an expert bioinformatician is available, it may be problematic for beginners to get all their knowledge in one field from one person. Instead, we propose that building a community to spread good practices and help to connect novices and experts. Ideally, a novice should make progress toward increased skill levels, as illustrated in Fig~B.1 (Dreyfus and Dreyfus, 1980).

B.4 How do we build local communities in scientific programming? A model inspired by experience

Here, we propose a three-stage working model (Fig. B.2) to create a local community of practice in scientific programming composed of life scientists at any given institution without any prior community structure.

In stage 1, we form the "primer" of a local community of practice by first running basic programming workshops organized by local community leads ("champions") and

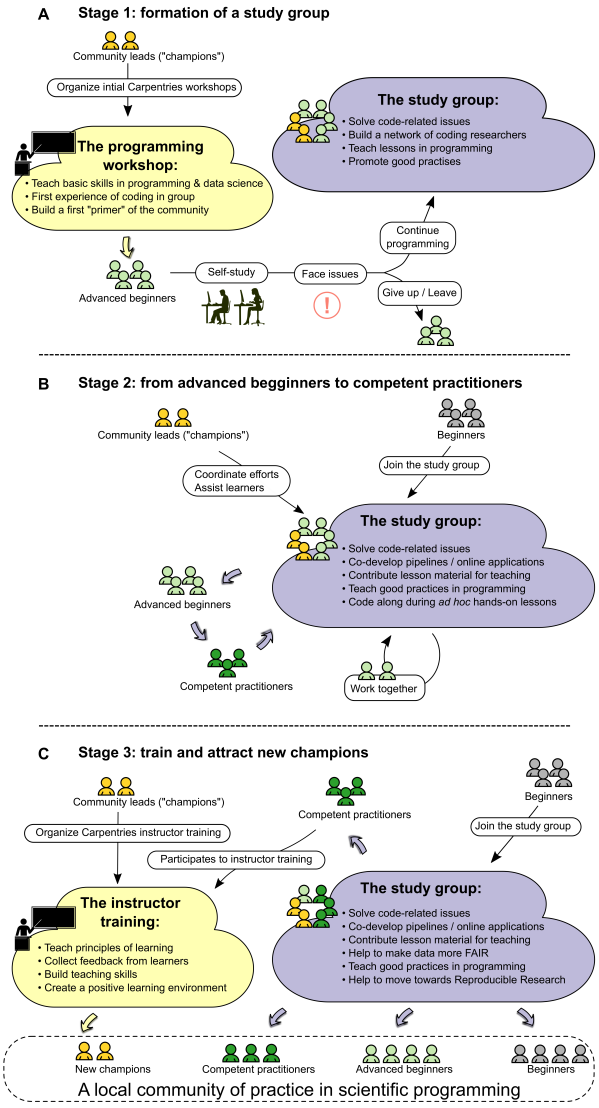

Figure B.2: A three-step model to build a local community of practice in scientific programming for life scientists. (A) First, a few scientists acting as community leads set up one or more Carpentries workshops to impart basic programming and data science skills to wet lab life scientists. After completion of the workshop, the novices will often face programming issues that need to be solved frequently. Furthermore, they need to continue to learn new programming skills. Therefore, a local study group such as a Mozilla Study Group can be formed by community leads ("champions") and "advanced beginners" to foster a regular meeting place for solving programming issues together and discovering new tools. (B) By attending a regularly scheduled study group, advanced beginners start to work together and make progress. Together with additional guidance and *ad hoc* assistance by community leads, some advanced beginners become "competent practitioners". (C) Finally, as some "competent practitioners" attend the Carpentries' instructor training sessions, new community leads ("champions") are trained. In addition, the local study group keeps attracting new beginners. Study group sessions together with optional Carpentries events help to educate community members and help them to become "advanced beginners" and "competent practitioners". As "competent practitioners" become community "champions", this closes the loop and help the local community of practice become fully mature with all categories of learners present.

then coupling them to formation of a study group. Champions do not necessarily have to be experts themselves. In our experience, Carpentries workshops work well since they provide training aimed at researchers and possess a long history of teaching programming to scientists (Wilson et al., 2017; Wilson, 2016). These programming workshops serve as a starting point for both learning and gathering researchers together in one room where people are actively paired and invited to learn about each other. Often beginners and bioinformaticians who might have never met despite working at the same institution will connect and engage afterwards.

When absolute beginners join these workshops, they become "advanced beginners" once they gain some programming notions. During their daily work, "advanced beginners" often lack the support needed to face programming issues that they may encounter frequently. Community "champions" and "advanced beginners" can "seed" a local community of practice (Fig~B.2) which meet regularly to continue practicing the skills they learned at these programming workshops. Therefore, a local co-working group that follow a well documented handbook such as that of the Mozilla Study Group[3] should be set-up with a regular meeting schedule. Other forms of co-working groups can be used but we believe that Mozilla Study Groups offer the best existing model.

In stage 2, the study group becomes a regular practice for advanced beginners where they progressively become competent practitioners (Fig~B.2). This study group also welcomes new novice members as they join the research institution or as they hear about the existence of the group. The community leads will provide guidance, specific lessons, and assistance during hands-on practicals which will nurture the community and raise the community global scientific programming level. Again, leading sessions is not restricted to champions and any motivated individual can lead. Also, champions do not necessarily have to be experts themselves but can instead invite experts and facilitate discussions. At the

[3] `http://mozillascience.github.io/studyGroupHandbook/`

end of this stage, most advanced beginners will likely have become competent practitioners.

In stage 3, a subset of the competent practitioners from the local community will become community leads ("champions", Fig~B.2) by increasing their teaching and facilitating skills and recognizing the skill level of their audience (Fig~B.1). These competencies can be attained by becoming a Carpentries instructor which requires attending an instructor training event: these sessions can be organized by initial community champions since they usually have both the network and know-how to set-up these specific workshops. Once again, it is not mandatory to rely on the Carpentries Foundation organization as long as competent practitioners get a deeper knowledge of teaching techniques where they improve their own skills. However, we now have a good perspective on the long-term experience and success of the Carpentries Foundation with over 500 workshops organized and 16,000 attendees present (Wilson, 2016; Pawlik et al., 2017).

B.5 Case studies

The Amsterdam Science Park example

In October 2016, Mateusz Kuzak, Carlos Martinez and Marc Galland organized a two-day Software Carpentry workshop in Amsterdam to teach basic programming skills (Shell, version control and Python) to a group of 26 wet lab biologists. This started a dialog about the skills life scientists need in their daily work. After a few months, a subset of the workshop attendees made progress but most of them did not continue to program either because (i) they did not need it at the time, (ii) they felt isolated and could not get support from their peers or (iii) they did not make time for practice alongside regular lab work. Thus, a regular meetup group was needed so that researchers with different programming levels

could help and support each other. Hence, in March 2017, we started up the Amsterdam Science Park Study Group following the Mozilla Study Group guidelines. We quickly decided to stick to the guidelines suggested by the Mozilla Science Lab[4]. Originally, we started with one scientist from the University of Amsterdam (Marc Galland) and two engineers in software engineering (Mateusz Kuzak and Carlos Martinez). But after five months, we decided to gather more scientists to build up a community with expertise in R and Python programming as well as from different scientific fields (genomics, statistics, ecology). Most study group members came from two different institutes which helped the group to be more multidisciplinary. At the same time, a proper website[5] was set-up to streamline communication and advertise events.

The University of Wisconsin-Madison example

At the University of Wisconsin-Madison, Sarah Stevens started a community of practice in the fall of 2014 centered around Computational Biology, Ecology and Evolution ("ComBEE"). It was started as a place to help other graduate students to learn scientific coding, such as Python and discuss scientific issues in computational biology, such as metagenomics. The main ComBEE group meets once a month to discuss computational biology in ecology and evolution. Under the ComBEE umbrella, there are also two spin-off study groups, which alternate each week so that attendees can focus on their favorite programming language. Later in ComBEE's development, Sarah transitioned to being a part of the Mozilla Study Group community, taking advantage of the existing resources to, for instance, build their web page[6].

Early in the development of ComBEE, the facilitating of the language-specific study groups was delegated on a semester by semester basis: this helped to keep more members involved in the growth and maturation of the local community. One of the early members

[4] https://mozillascience.github.io/study-group-orientation/

[5] https://scienceparkstudygroup.github.io/studyGroup/

[6] https://combee-uw-madison.github.io

of ComBEE was a life sciences graduate student who had recently attended a Software Carpentry Workshop and had no other experience doing bioinformatics. He wanted to continue his development and was working on a very computationally intensive project. He has since run the Python Study Group for several semesters and is now an exceedingly competent computational biologist. He continued to contribute back to the group through the end of his PhD, lending his expertise and experience to the latest study group discussions. The ComBEE study group is now more than three years old and acts as a stable resource center for new graduate students and employees.

B.6 Room for improvement: challenges and solutions learned from experience

Below we describe essential components of a successful community of practice based on both literature (Webber, 2018; Brown, 2007; Wenger, 2011; Budd et al., 2015) and experience.

Gather a core group of motivated individuals

One of the first tasks for setting up a community of practice is to gather a team of motivated individuals that will act as leaders of the community (Brown, 2007; Webber, 2018). To recruit these leaders, one can:

- Rely on existing communities e.g. "R lunch group" since these informal groups are often lead by motivated individuals.
- Recruit scientists that share similar values such as:
  - Advocating Open Science
  - Having a collaborative attitude
  - Show tolerance towards cultural and scientific differences

  - Being supportive of beginners and lifelong learners in general

- Search within institutions with a reasonably big size e.g. Universities.

Keeping participants coming and engaging into the community

For someone who is part of the "core team" of a study group, the challenge is to attract experts or new members and ensure that they regularly participate in activities (lessons, co-working sessions, organizational meetings) (Brown, 2007; Webber, 2018; Budd et al., 2015). Among possible incentives to keep new members and leaders engaging, we suggest to tell them that they can:

- Reach out to a wider audience by participating to lessons, workshops, etc.
- Improve their teaching skills and eventually become a Carpentries instructor
- Solve basic issues for several beginners simultaneously through workshops
- Lead the community for a semester and thereby develop their leadership
- Tailor topics to their interests
- Increase their group management, communication and networking capacities

How to deal with the ever-ongoing turnover at academic institutions

The constant turnover of students and temporary staff remains a continual challenge. Keeping the local community ongoing requires a critical mass both for the core team and for the audience. Yet, the high turnover of students and staff also has its positive sides: a dynamic environment brings in new people eager to learn and with relevant knowledge to share in the group. We recommend using the turnover of people to your advantage by making an effort to recruit both new members and champions. Some practical solutions include:

- Advertising the community through its leaders: people bring people through word to mouth
- Invite permanent staff to sustain the community development
- Use the turnover to your advantage: quickly invite newcomers to join the community

Dealing with the impostor syndrome

Creating a safe learning environment is one of the requirement for a thriving community of practice (Webber, 2018). To encourage beginners and newcomers to participate and feel welcome, we recommend to:

- Enforce a Code of Conduct following an existing example[7] to set-up expectations and promote a welcoming atmosphere
- Promote all questions and forbid surprise reactions to very basic questions ("What is the Shell?", "Oh you don't know?")
- Ban in-depth technical discussions that alienate novices

Community leadership and institutional support

An effort should be made to assign clear and specific roles to administration members of the local community based on their expertise and interest. Another challenge is to secure funding and people support from the local institution (Brown, 2007; Webber, 2018). To do so, we advise to:

- Delegate as much as possible to promote leadership: appoint someone to lead the community for a semester for instance
- Get support from the local institution as soon as possible in terms of money, time and/or staff

[7] `https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html`

Community composition

Another important aspect to consider is the composition of the community. We have identified the following types of community members as common components of the community:

- Absolute and advanced beginners: these are people with the most basic level of knowledge. For them, the motivation to be part of a community is obvious: they want to learn programming and often need rapid assistance to complete their research.
- Competent practitioners: these are people who already competent in a particular bioinformatics/data science domain. For them, contributing to the community is a good way to reinforce their set of talents. Often, competent practitioners make excellent teachers, as they are able to easily relate to the beginner state of mind. In turn, this increases their learning and teaching skills.
- Experts: these are people with the highest experience level on a particular skill. Experts usually reinforce their knowledge by 'going back to basics': it is useful for them to understand what are the usual *gotchas for novices*. Building a local community of practice provides experts with an opportunity to help novices in a more structural way instead of helping each one individually.

Practical considerations

In our experience, we have found the following practical tips to be useful:

- Gather a critical mass of at least 10 recurrent community members that regularly attend meetings and community sessions
- Send meeting notifications in advance and frequently enough: schedule the meetings well-in-advance and keep a consistent day, time and place to help people remember them.

- Have weekly or fortnightly meetings so that it is a compromise between researchers' schedules and community development.
- Organize meetings in a relatively quiet environment with a good Internet connection. Places such as a campus café outside of busy hours or a small conference room can be good places to start and help to keep an informal and welcoming atmosphere,

B.7 Conclusions

We hope that our model and the lessons learned from our experience described in this paper will save time and effort for future community leads when they start their own local community of practice in scientific programming. Building such a community is far from trivial and we, as scientists, are perhaps not the most proficient on community building and organization (Webber, 2018; Brown, 2007; Wenger, 2011; Li et al., 2009b; Wenger et al., 2002). Since "progress will not happen by itself" (Wilson et al., 2017), a community of practice in scientific programming will bring many benefits to its members and to their institution: it fosters the development of new skills for its members, breaks down "mental borders" between departments, networks domain experts at a local site and helps to retain knowledge that would otherwise be lost with the departure of temporary staff and students. The convergence of the "big data" avalanche in biology and new FAIR requirements for data management (Wilkinson et al., 2016) makes it more and more important for wet lab researchers to conduct good scientific programming, efficient data analysis, and proper research data management. Eventually, these local communities of practice in scientific programming should speed up code-intensive analyses, promote open science, research reproducibility and spread good practices at a given institution.

B.8 Acknowledgments

We are thankful to the Carpentries Foundation for assistance in workshop organization. We kindly acknowledge the Mozilla Foundation for assistance in starting and maintaining the Amsterdam Science Park and the Computational Biology, Ecology, & Evolution (ComBEE) Study Groups. The local community of the Amsterdam Science Park Study Group not in the author list is fully acknowledged and consist of Dr Emiel van Loon (UvA-IBED), Pietro Marchesi (UvA-SILS), Joeri Jongbloets (UvA-SILS), Dr Like Fokkens (UvA-SILS), Zsofia Koma (UvA-IBED) and Dr Susanne Wilkens (UvA-IBED). We are grateful to Dr Anita Schürch (UMC Utrecht) for training researchers through several Software and Data Carpentry workshops. We would also like to thank the members and leaders of the Computational Biology, Ecology and Evolution (ComBEE) Study Group and the Carpentry community at the University of Wisconsin-Madison.

REFERENCES

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* 31(6):533–538.

Allen, E. E., G. W. Tyson, R. J. Whitaker, J. C. Detter, P. M. Richardson, and J. F. Banfield. 2007. Genome dynamics in a natural archaeal population. *Proceedings of the National Academy of Sciences* 104(6):1883–1888.

Allgaier, M., and H.-P. Grossart. 2006. Diversity and Seasonal Dynamics of Actinobacteria Populations in Four Lakes in Northeastern Germany. *Applied and Environmental Microbiology* 72(5):3489–3497.

Alneberg, Johannes, Christofer M G Karlsson, Anna-Maria Divne, Claudia Bergin, Felix Homa, Markus V Lindh, Luisa W Hugerth, Thijs J G Ettema, Stefan Bertilsson, Anders F Andersson, and Jarone Pinhassi. 2018. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* 6(1):173.

Amann, RI, W Ludwig, and KH Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143–69.

Baker, Monya. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452–4.

Barrick, J. E., and R. E. Lenski. 2009. Genome-wide Mutational Diversity in an Evolving Population of Escherichia coli. *Cold Spring Harbor Symposia on Quantitative Biology* 74(0): 119–129.

Barrick, Jeffrey E., Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, and Jihyun F. Kim. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461(7268):1243–1247.

Batut, Bérénice, Saskia Hiltemann, Andrea Bagnacani, Dannon Baker, Vivek Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Guéguen, Martin Čech, John Chilton, Dave Clements, Olivia Doppelt-Azeroual, Anika Erxleben, Mallory Freeberg, Simon Gladman, Youri Hoogstrate, Hans-Rudolf Hotz, Torsten Houwaart, Pratik Jagtap, Delphine Lariviere, Gildas Le Corguillé, Thomas Manke, Fabien Mareuil, Fidel Ramírez, Devon Ryan, Florian Sigloch, Nicola Soranzo, Joachim Wolff, Pavankumar Videm, Markus Wolfien, Aisanjiang Wubuli, Dilmurat Yusuf, Rolf Backofen, James Taylor, Anton Nekrutenko, and Björn Grüning. 2018. Community-driven data analysis training for biology. *bioRxiv* 225680.

Bauer, Mark S., Laura Damschroder, Hildi Hagedorn, Jeffrey Smith, and Amy M. Kilbourne. 2015a. An introduction to implementation science for the non-specialist. *BMC Psychology* 3(1):32.

———. 2015b. An introduction to implementation science for the non-specialist. *BMC Psychology* 3(1):32.

Bendall, Matthew L, Sarah LR Stevens, Leong-Keat Chan, Stephanie Malfatti, Patrick Schwientek, Julien Tremblay, Wendy Schackwitz, Joel Martin, Amrita Pati, Brian Bushnell, Jeff Froula, Dongwan Kang, Susannah G Tringe, Stefan Bertilsson, Mary A Moran, Ashley Shade, Ryan J Newton, Katherine D McMahon, and Rex R Malmstrom. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal* 10(7):1589–1601.

Blainey, Paul C. 2013. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews* 37(3):407–427.

Blount, Zachary D, Richard E Lenski, and Jonathan B Losos. 2018. Contingency and determinism in evolution: Replaying life's tape. *Science (New York, N.Y.)* 362(6415):eaam5979.

Bowers, Robert M, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A

Eloe-Fadrosh, Susannah G Tringe, Natalia N Ivanova, Alex Copeland, Alicia Clum, Eric D Becraft, Rex R Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M Weinstock, George M Garrity, Jeremy A Dodsworth, Shibu Yooseph, Granger Sutton, Frank O Glöckner, Jack A Gilbert, William C Nelson, Steven J Hallam, Sean P Jungbluth, Thijs J G Ettema, Scott Tighe, Konstantinos T Konstantinidis, Wen-Tso Liu, Brett J Baker, Thomas Rattei, Jonathan A Eisen, Brian Hedlund, Katherine D McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W Tyson, Christian Rinke, Nikos C Kyrpides, Lynn Schriml, George M Garrity, Philip Hugenholtz, Granger Sutton, Pelin Yilmaz, Folker Meyer, Frank O Glöckner, Jack A Gilbert, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H Parks, A M Eren, Lynn Schriml, Jillian F Banfield, Philip Hugenholtz, and Tanja Woyke. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35(8):725–731.

Brown, MJ. 2007. *Building Powerful Community Organizations: A Personal Guide to Creating Groups that Can Solve Problems and Change the World*. Long Haul Press.

Budd, Aidan, Manuel Corpas, Michelle D. Brazas, Jonathan C. Fuller, Jeremy Goecks, Nicola J. Mulder, Magali Michaut, B. F. Francis Ouellette, Aleksandra Pawlik, and Niklas Blomberg. 2015. A Quick Guide for Building a Successful Bioinformatics Community. *PLOS Computational Biology* 11(2):e1003972.

Bushnell, Brian. 2018. BBTools User Guide - DOE Joint Genome Institute.

Cadillo-Quiroz, Hinsby, Xavier Didelot, Nicole L Held, Alfa Herrera, Aaron Darling, Michael L Reno, David J Krause, and Rachel J Whitaker. 2012. Patterns of gene flow define species of thermophilic Archaea. *PLoS biology* 10(2):e1001265.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421.

Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *BIOINFORMATICS APPLICATIONS NOTE* 25(15):1972–1973.

Caro-Quintero, Alejandro, Jie Deng, Jennifer Auchtung, Ingrid Brettar, Manfred G Höfle, Joel Klappenbach, and Konstantinos T Konstantinidis. 2011. Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea. *The ISME Journal* 5(1):131–140.

Caro-Quintero, Alejandro, and Konstantinos T Konstantinidis. 2012. Bacterial species may exist, metagenomics reveal. *Environmental microbiology* 14(2):347–55.

Carpenter, SR, RC Lathrop, P Nowak, DE Armstrong, EM Bennett, T Reed-Andersen, and Et Al. 2006. The ongoing experiment: restoration of lake mendota and its watershed. In *Long term dynamics of lakes in the landscape*. Oxford: Oxford Press.

Clausen, J. Keck, DD., and WM. Hiesey. 1940. Effects of Varied Environments on Western North American plants. *Experimental Studies on the Nature of Species*.

Cohan, Frederick M. 1994. The Effects of Rare but Promiscuous Genetic Exchange on Evolutionary Divergence in Prokaryotes. *The American Naturalist* 143(6):965–986.

———. 2001. Bacterial Species and Speciation. *Systematic Biology* 50(4):513–524.

Cohan, Frederick M, and Elizabeth B Perry. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Current biology : CB* 17(10):R373–86.

Coleman, Maureen L., and Sallie W. Chisholm. 2007. Code and context: Prochlorococcus as a model for cross-scale biology. *Trends in Microbiology* 15(9):398–407.

Cordero, Otto X., and Martin F. Polz. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology* 12(4):263–273.

Corpas, M., R. C. Jimenez, E. Bongcam-Rudloff, A. Budd, M. D. Brazas, P. L. Fernandes, B. Gaeta, C. van Gelder, E. Korpelainen, F. Lewitter, A. McGrath, D. MacLean, P. M. Palagi, K. Rother, J. Taylor, A. Via, M. Watson, M. V. Schneider, and T. K. Attwood. 2015. The GOBLET training portal: a global repository of bioinformatics training materials, courses and trainers. *Bioinformatics* 31(1):140–142.

Dahllof, I., H. Baillie, and S. Kjelleberg. 2000. rpoB-Based Microbial Community Analysis Avoids Limitations Inherent in 16S rRNA Gene Intraspecies Heterogeneity. *Applied and Environmental Microbiology* 66(8):3376–3380.

Darling, Aaron E., Guillaume Jospin, Eric Lowe, Frederick A. Matsen, Holly M. Bik, and Jonathan A. Eisen. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243.

Delmont, Tom O., and A. Murat Eren. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* 6:e4320.

Denef, V. J., and J. F. Banfield. 2012. In Situ Evolutionary Rate Measurements Show Ecological Success of Recently Emerged Bacterial Hybrids. *Science* 336(6080):462–466.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5):491–498.

van Dongen, Stijn. 2000. Graph Clustering by Flow Simulation. Phd thesis, University of Utrecht.

Doolittle, W. Ford. 2012. Population Genomics: How Bacterial Species Form and Why They Don't Exist. *Current Biology* 22(11):R451–R453.

Dreyfus, S.E., and H.L. Dreyfus. 1980. *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition.* Washington, DC: Storming Media.

Eiler, Alexander, Friederike Heinrich, and Stefan Bertilsson. 2012. Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal* 6(2):330–342.

Eiler, Alexander, Rhiannon Mondav, Lucas Sinclair, Leyden Fernandez-Vidal, Douglas G Scofield, Patrick Schwientek, Manuel Martinez-Garcia, David Torrents, Katherine D McMahon, Siv GE Andersson, Ramunas Stepanauskas, Tanja Woyke, and Stefan Bertilsson. 2016. Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria. *The ISME Journal* 10(8):1902–1914.

Eisen, JonathanA. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species. *Journal of Molecular Evolution* 41(6).

Enright, A J, S Van Dongen, and C A Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* 30(7):1575–84.

Falkowski, Paul G. 2016. *Life's Engines: How Microbes Made Earth Habitable*. Princeton University Press.

Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox, M. Achtman, and S. Suerbaum. 2001. Recombination and mutation during long-term gastric colonization by Helicobacter pylori: Estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences* 98(26):15056–15061.

Fanelli, Daniele. 2018. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America* 115(11):2628–2631.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt, and W. P. Hanage. 2009. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. *Science* 323(5915):741–746.

Fraser, C., W. P. Hanage, and B. G. Spratt. 2007. Recombination and the Nature of Bacterial Speciation. *Science* 315(5811):476–480.

Free Software Foundation. 2018. Bash - Unix shell program.

Friesner, Joanna, Sarah M Assmann, Ruth Bastow, Julia Bailey-Serres, Jim Beynon, Volker Brendel, C Robin Buell, Alexander Bucksch, Wolfgang Busch, Taku Demura, Jose R Dinneny, Colleen J Doherty, Andrea L Eveland, Pascal Falter-Braun, Malia A Gehan, Michael Gonzales, Erich Grotewold, Rodrigo Gutierrez, Ute Kramer, Gabriel Krouk, Shisong Ma, R J Cody Markelz, Molly Megraw, Blake C Meyers, James A H Murray, Nicholas J Provart, Sue Rhee, Roger Smith, Edgar P Spalding, Crispin Taylor, Tracy K Teal, Keiko U Torii, Chris Town, Matthew Vaughn, Richard Vierstra, Doreen Ware, Olivia Wilkins, Cranos Williams, and Siobhan M Brady. 2017. The Next Generation of Training for Arabidopsis Researchers: Bioinformatics and Quantitative Biology. *Plant physiology* 175(4):1499–1509.

Fuhrman, Jed A., Jacob A. Cram, and David M. Needham. 2015. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology* 13(3): 133–146.

Galperin, Michael Y, Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research* 43(Database issue):D261–9.

Garcia, Sarahi L, Katherine D McMahon, Manuel Martinez-Garcia, Abhishek Srivastava, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, and Falk Warnecke. 2013. Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *The ISME Journal* 7(1):137–147.

Garcia, Sarahi L., Sarah L. R. Stevens, Benjamin Crary, Manuel Martinez-Garcia, Ramunas Stepanauskas, Tanja Woyke, Susannah G. Tringe, Siv G. E. Andersson, Stefan Bertilsson,

Rex R. Malmstrom, and Katherine D. McMahon. 2018. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *The ISME Journal* 12(3): 742–755.

Ghai, Rohit, Carolina Megumi Mizuno, Antonio Picazo, Antonio Camacho, and Francisco Rodriguez-Valera. 2014. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* 23(24):6073–6090.

Ghylin, Trevor W, Sarahi L Garcia, Francisco Moya, Ben O Oyserman, Patrick Schwientek, Katrina T Forest, James Mutschler, Jeffrey Dwulit-Smith, Leong-Keat Chan, Manuel Martinez-Garcia, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, Falk Warnecke, Rex Malmstrom, Stefan Bertilsson, and Katherine D McMahon. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *The ISME Journal* 8(12):2503–2516.

Glockner, F. O., E. Zaichikov, N. Belkova, L. Denissova, J. Pernthaler, A. Pernthaler, and R. Amann. 2000. Comparative 16S rRNA Analysis of Lake Bacterioplankton Reveals Globally Distributed Phylogenetic Clusters Including an Abundant Group of Actinobacteria. *Applied and Environmental Microbiology* 66(11):5053–5065.

Gowda, G A Nagana, and Danijel Djukovic. 2014. Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods in molecular biology (Clifton, N.J.)* 1198:3–12.

Guttman, D S, and D E Dykhuizen. 1994. Detecting selective sweeps in naturally occurring Escherichia coli. *Genetics* 138(4):993–1003.

Hahn, Martin W. 2003. Isolation of strains belonging to the cosmopolitan Polynucleobacter necessarius cluster from freshwater habitats located in three climatic zones. *Applied and environmental microbiology* 69(9):5248–54.

Hahn, Martin W, Gerlinde Karbon, Ulrike Koll, Johanna Schmidt, and Elke Lang. 2017. Polynucleobacter sphagniphilus sp. Nov. a planktonic freshwater bacterium isolated from an acidic and humic freshwater habitat. *International Journal of Systematic and Evolutionary Microbiology* 67(9):3261–3267.

Hahn, Martin W, Thomas Scheuerl, Jitka Jezberová, Ulrike Koll, Jan Jezbera, Karel Šimek, Claudia Vannini, Giulio Petroni, and Qinglong L Wu. 2012. The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-living polynucleobacter population. *PloS one* 7(3):e32772.

Hahn, Martin W, Johanna Schmidt, Alexandra Pitt, Sami J Taipale, and Elke Lang. 2016. Reclassification of four Polynucleobacter necessarius strains as representatives of Polynucleobacter asymbioticus comb. nov., Polynucleobacter duraquae sp. nov., Polynucleobacter yangtzensis sp. nov. and Polynucleobacter sinensis sp. nov., and emended description of Polynucleobacter necessarius. *International journal of systematic and evolutionary microbiology* 66(8):2883–92.

Hall, Ed K., Emily S. Bernhardt, Raven L. Bier, Mark A. Bradford, Claudia M. Boot, James B. Cotner, Paul A. del Giorgio, Sarah E. Evans, Emily B. Graham, Stuart E. Jones, Jay T. Lennon, Kenneth J. Locey, Diana Nemergut, Brooke B. Osborne, Jennifer D. Rocca, Joshua P. Schimel, Mark P. Waldrop, and Matthew D. Wallenstein. 2018. Understanding how microbiomes influence the systems they inhabit. *Nature Microbiology* 3(9):977–982.

Hanage, William P, Christophe Fraser, and Brian G Spratt. 2005. Fuzzy species among recombinogenic bacteria. *BMC Biology* 3(1):6.

Handelsman, J., J. Tiedje, L. Alvarez-Cohen, M. Ashburner, I. Cann, E. Delong, and Et Al. 2007. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC, USA: The National Academies Press.

Hayden, Erika Check. 2014. Technology: The $1,000 genome. *Nature* 507(7492):294–295.

Heinrich, F, A Eiler, and S Bertilsson. 2013. Seasonality and environmental control of freshwater SAR11 (LD12) in a temperate lake (Lake Erken, Sweden). *Aquatic Microbial Ecology* 70(1):33–44.

Henson, Michael W., V. Celeste Lanclos, Brant C. Faircloth, and J. Cameron Thrash. 2018a. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *The ISME Journal* 12(7):1846–1860.

Henson, Michael Winslow, V. Celeste Lanclos, Brant C. Faircloth, and J. Cameron Thrash. 2018b. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *bioRxiv* 093567.

Herron, Matthew D., and Michael Doebeli. 2013. Parallel Evolutionary Dynamics of Adaptive Diversification in Escherichia coli. *PLoS Biology* 11(2):e1001490.

Hettrick, Antonioletti, Carr, Chue Hong, Crouch, De Roure, Emsley, Goble, Hay, Inupakutika, Jackson, Nenadic, Parkinson, Parsons, Pawlik, Peru, Proeme, Robinson, and Sufi. 2014. UK Research Software Survey 2014.

Hiraoka, Satoshi, Ching-Chia Yang, and Wataru Iwasaki. 2016. Metagenomics and Bioinformatics in Microbial Ecology: Current Status and Beyond. *Microbes and environments* 31(3):204–12.

Hoetzinger, Matthias, and Martin W. Hahn. 2017. Genomic divergence and cohesion in a species of pelagic freshwater bacteria. *BMC Genomics* 18(1):794.

Hoetzinger, Matthias, Johanna Schmidt, Jitka Jezberová, Ulrike Koll, and Martin W Hahn. 2017. Microdiversification of a Pelagic Polynucleobacter Species Is Mainly Driven by Acquisition of Genomic Islands from a Partially Interspecific Gene Pool. *Applied and environmental microbiology* 83(3):e02266–16.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro

Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. 2016. A new view of the tree of life. *Nature Microbiology* 1(5):16048.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science* 320(5879):1081–1085.

Huntemann, Marcel, Natalia N Ivanova, Konstantinos Mavromatis, H James Tripp, David Paez-Espino, Krishnaveni Palaniappan, Ernest Szeto, Manoj Pillay, I-Min A Chen, Amrita Pati, Torben Nielsen, Victor M Markowitz, and Nikos C Kyrpides. 2015. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Standards in genomic sciences* 10:86.

Hutson, Matthew. 2018. Artificial intelligence faces reproducibility crisis. *Science (New York, N.Y.)* 359(6377):725–726.

Iverson, V., R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust. 2012. Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* 335(6068):587–590.

Jezberová, Jitka, Jan Jezbera, Ulrike Brandt, Eva S. Lindström, Silke Langenheder, and Martin W. Hahn. 2010. Ubiquity of Polynucleobacter necessarius ssp. asymbioticus in lentic freshwater habitats of a heterogenous 2000 km2 area. *Environmental Microbiology* 12(3):658–669.

Johnson, Z. I. 2006. Niche Partitioning Among Prochlorococcus Ecotypes Along Ocean-Scale Environmental Gradients. *Science* 311(5768):1737–1740.

Kaeberlein, T. 2002. Isolating "Uncultivable" Microorganisms in Pure Culture in a Simulated Natural Environment. *Science* 296(5570):1127–1129.

Kang, D, J Froula, R Egan, and Z Wang. 2018. berkeleylab / MetaBAT — Bitbucket.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165.

Kang, Ilnam, Suhyun Kim, Md. Rashedul Islam, and Jang-Cheon Cho. 2017. The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. *Scientific Reports* 7(1):42252.

Kara, Emily L, Paul C Hanson, Yu Hen Hu, Luke Winslow, and Katherine D McMahon. 2013. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *The ISME Journal* 7(3): 680–684.

Kashtan, N., S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, M. J. Follows, R. Stepanauskas, and S. W. Chisholm. 2014. Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* 344(6182):416–420.

Kashtan, Nadav, Sara E. Roggensack, Jessie W. Berta-Thompson, Maor Grinberg, Ramunas Stepanauskas, and Sallie W. Chisholm. 2017. Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. *ISME Journal* 11(9):1997–2011.

Katoh, K., Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14):3059–3066.

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30(4):772–780.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–1649.

Kim, M., H.-S. Oh, S.-C. Park, and J. Chun. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* 64(Pt 2):346–351.

Kim, Suhyun, Ilnam Kang, Ji-Hui Seo, and Jang-Cheon Cho. 2018. Culturing the ubiquitous freshwater actinobacterial acI lineage by supplying a biochemical 'helper' catalase. *bioRxiv* 343640.

Koboldt, Daniel C, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22(3):568–76.

Konstantinidis, K. T., and J. M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* 102(7):2567–2572.

Konstantinidis, Konstantinos T, and Edward F DeLong. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *The ISME journal* 2(10):1052–65.

Krause, David J, Xavier Didelot, Hinsby Cadillo-Quiroz, and Rachel J Whitaker. 2014. Recombination shapes genome architecture in an organism from the archaeal domain. *Genome biology and evolution* 6(1):170–8.

Krause, David J, and Rachel J Whitaker. 2015. Inferring speciation processes from patterns of natural variation in microbial genomes. *Systematic biology*.

Lave, J, and E Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Lee, H., E. Popodi, H. Tang, and P. L. Foster. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences* 109(41):E2774–E2783.

Letunic, Ivica, and Peer Bork. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* 44(W1): W242–5.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li, Heng. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* 27(21):2987–93.

———. 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics* 31(17):2885–2887.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing 1000 Genome Project Data Processing Subgroup. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25(16):2078–9.

Li, Linda C, Jeremy M Grimshaw, Camilla Nielsen, Maria Judd, Peter C Coyte, and Ian D Graham. 2009b. Evolution of Wenger's concept of community of practice. *Implementation Science* 4(1):11.

Linz, Alexandra M., Benjamin C. Crary, Ashley Shade, Sarah Owens, Jack A. Gilbert, Rob Knight, and Katherine D. McMahon. 2017. Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog Lakes. *mSphere* 2(3):e00169–17.

Little, Ainslie E.F., Courtney J. Robinson, S. Brook Peterson, Kenneth F. Raffa, and Jo Handelsman. 2008. Rules of Engagement: Interspecies Interactions that Regulate Microbial Communities. *Annual Review of Microbiology* 62(1):375–401.

Logares, R., J. Brate, F. Heinrich, K. Shalchian-Tabrizi, and S. Bertilsson. 2010. Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Molecular Biology and Evolution* 27(2):347–357.

Luo, C., S. T. Walk, D. M. Gordon, M. Feldgarden, J. M. Tiedje, and K. T. Konstantinidis. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences* 108(17):7200–7205.

Luo, Chengwei, Luis M Rodriguez-R, and Konstantinos T Konstantinidis. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic acids research* 42(8):e73.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1(1):18.

Macintyre, Geoff, Magali Michaut, and Thomas Abeel. 2013a. The Regional Student Group Program of the ISCB Student Council: Stories from the Road. *PLoS Computational Biology* 9(9):e1003241.

———. 2013b. The regional student group program of the ISCB student council: stories from the road. *PLoS computational biology* 9(9):e1003241.

Magoc, T., and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957–2963.

Maharjan, Ram P, Thomas Ferenci, Peter R Reeves, Yang Li, Bin Liu, and Lei Wang. 2012. The multiplicity of divergence mechanisms in a single evolving population. *Genome Biology* 13(6):R41.

Majewski, J, and F M Cohan. 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152(4):1459–74.

Malmstrom, Rex R, Allison Coe, Gregory C Kettler, Adam C Martiny, Jorge Frias-Lopez, Erik R Zinser, and Sallie W Chisholm. 2010. Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *The ISME Journal* 4(10):1252–1264.

Markowitz, V. M., I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides. 2012. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* 40(D1):D115–D122.

Martinez-Garcia, Manuel, Brandon K Swan, Nicole J Poulton, Monica Lluesma Gomez, Dashiell Masland, Michael E Sieracki, and Ramunas Stepanauskas. 2012. High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal* 6(1):113–123.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit:

A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9):1297–1303.

Messer, Philipp W., and Dmitri A. Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution* 28(11):659–669.

Miller, Mark A, Wayne Pfeiffer, and Terri Schwartz. 2010. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. Tech. Rep., San Diego Supercomputer Center.

Moore, Lisa R., and Sallie W. Chisholm. 1999. Photophysiology of the marine cyanobacterium Prochlorococcus: Ecotypic differences among cultured isolates. *Limnology and Oceanography* 44(3):628–638.

Morgan, Sarah L, Patricia M Palagi, Pedro L Fernandes, Eija Koperlainen, Jure Dimec, Diana Marek, Lee Larcombe, Gabriella Rustici, Teresa K Attwood, and Allegra Via. 2017. The ELIXIR-EXCELERATE Train-the-Trainer pilot programme: empower researchers to deliver high-quality training. *F1000Research* 6:1557.

Morin, A, J Urban, P D Adams, I Foster, A Sali, D Baker, and P Sliz. 2012. Shining Light into Black Boxes. *Science* 336(6078):159–160.

Newton, R. J., S. E. Jones, A. Eiler, K. D. McMahon, and S. Bertilsson. 2011. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews* 75(1):14–49.

Newton, R. J., S. E. Jones, M. R. Helmus, and K. D. McMahon. 2007. Phylogenetic Ecology of the Freshwater Actinobacteria acI Lineage. *Applied and Environmental Microbiology* 73(22):7169–7176.

NTL-LTER. a. Lake Mendota | North Temperate Lakes.

———. b. Trout Bog | North Temperate Lakes.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome research* 27(5):824–834.

Nylander, Johan A. A. 2018. catfasta2phyml.

Oh, Seungdae, Alejandro Caro-Quintero, Despina Tsementzi, Natasha DeLeon-Rodriguez, Chengwei Luo, Rachel Poretsky, and Konstantinos T. Konstantinidis. 2011. Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Applied and Environmental Microbiology* 77(17):6000–6011.

Olm, Matthew R, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* 11(12):2864–2868.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25(7):1043–1055.

Paulson, Joseph N, Mihai Pop, and Hector Bravo. 2011. Metastats: an improved statistical method for analysis of metagenomic data. *Genome Biology* 12(Suppl 1):P17.

Paver, Sara F., Nicholas D. Youngblut, Rachel J. Whitaker, and Angela D. Kent. 2015. Phytoplankton succession affects the composition of P olynucleobacter subtypes in humic lakes. *Environmental Microbiology* 17(3):816–828.

Pawlik, Aleksandra, Celia W.G. van Gelder, Aleksandra Nenadic, Patricia M. Palagi, Eija Korpelainen, Philip Lijnzaad, Diana Marek, Susanna-Assunta Sansone, John Hancock, and Carole Goble. 2017. Developing a strategy for computational lab skills training through Software and Data Carpentry: Experiences from the ELIXIR Pilot action. *F1000Research* 6: 1040.

Pitt, Alexandra, Johanna Schmidt, Elke Lang, William B Whitman, Tanja Woyke, and Martin W Hahn. 2018. Polynucleobacter meluiroseus sp. nov., a bacterium isolated from a lake located in the mountains of the Mediterranean island of Corsica. *International journal of systematic and evolutionary microbiology* 68(6):1975–1985.

Poldrack, Russell A, and Krzysztof J Gorgolewski. 2014. Making big data open: data sharing in neuroimaging. *Nature neuroscience* 17(11):1510–7.

Python Core Team. 2018. Python: A dynamic, open source programming language.

Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biology* 18(1):181.

Quinlan, Aaron R., and Ira M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

———. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K. Swan, Esther A. Gies, Jeremy A. Dodsworth, Brian P. Hedlund, George Tsiamis, Stefan M. Sievert, Wen-Tso Liu, Jonathan A. Eisen, Steven J. Hallam, Nikos C. Kyrpides, Ramunas Stepanauskas, Edward M. Rubin, Philip Hugenholtz, and Tanja Woyke. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.

Rocap, Gabrielle, Frank W. Larimer, Jane Lamerdin, Stephanie Malfatti, Patrick Chain, Nathan A. Ahlgren, Andrae Arellano, Maureen Coleman, Loren Hauser, Wolfgang R.

Hess, Zackary I. Johnson, Miriam Land, Debbie Lindell, Anton F. Post, Warren Regala, Manesh Shah, Stephanie L. Shaw, Claudia Steglich, Matthew B. Sullivan, Claire S. Ting, Andrew Tolonen, Eric A. Webb, Erik R. Zinser, and Sallie W. Chisholm. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424(6952):1042–1047.

Rodriguez-Brito, Beltran, LinLin Li, Linda Wegley, Mike Furlan, Florent Angly, Mya Breitbart, John Buchanan, Christelle Desnues, Elizabeth Dinsdale, Robert Edwards, Ben Felts, Matthew Haynes, Hong Liu, David Lipson, Joseph Mahaffy, Anna Belen Martin-Cuadrado, Alex Mira, Jim Nulton, Lejla Pašić, Steve Rayhawk, Jennifer Rodriguez-Mueller, Francisco Rodriguez-Valera, Peter Salamon, Shailaja Srinagesh, Tron Frede Thingstad, Tuong Tran, Rebecca Vega Thurber, Dana Willner, Merry Youle, and Forest Rohwer. 2010. Viral and microbial community dynamics in four aquatic environments. *The ISME Journal* 4(6):739–751.

Rösel, Stefan, Martin Allgaier, and Hans-Peter Grossart. 2012. Long-Term Characterization of Free-Living and Particle-Associated Bacterial Communities in Lake Tiefwaren Reveals Distinct Seasonal Patterns. *Microbial Ecology* 64(3):571–583.

Rosen, M. J., M. Davison, D. Bhaya, and D. S. Fisher. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* 348(6238):1019–1023.

RStudio Team. 2016. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Salcher, Michaela M., Jakob Pernthaler, and Thomas Posch. 2010. Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limnology and Oceanography* 55(2):846–856.

Salcher, Michaela M, Jakob Pernthaler, and Thomas Posch. 2011. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *The ISME Journal* 5(8):1242–1252.

Salcher, Michaela M, Thomas Posch, and Jakob Pernthaler. 2013. In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake. *The ISME Journal* 7(5):896–907.

Schneider, M. V., P. Walter, M.-C. Blatter, J. Watson, M. D. Brazas, K. Rother, A. Budd, A. Via, C. W. G. van Gelder, J. Jacob, P. Fernandes, T. H. Nyronen, J. De Las Rivas, T. Blicher, R. C. Jimenez, J. Loveland, J. McDowall, P. Jones, B. W. Vaughan, R. Lopez, T. K. Attwood, and C. Brooksbank. 2012. Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Briefings in Bioinformatics* 13(3):383–389.

Segata, Nicola, Daniela Börnigen, Xochitl C. Morgan, and Curtis Huttenhower. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications* 4(1):2304.

Shade, Ashley, Angela D. Kent, Stuart E. Jones, Ryan J. Newton, Eric W. Triplett, and Katherine D. McMahon. 2007. Interannual dynamics and phenology of bacterial communities in a eutrophic lake. *Limnology and Oceanography* 52(2):487–494.

Shapiro, B Jesse. 2016. How clonal are bacteria over time? *Current Opinion in Microbiology* 31:116–123.

Shapiro, B. Jesse. 2018. What Microbial Population Genomics Has Taught Us About Speciation. In *Population genomics*, 1–17. Springer, Cham.

Shapiro, B. Jesse, Jonathan Friedman, Otto X. Cordero, Sarah P. Preheim, Sonia C. Timberlake, G. Szabo, Martin F. Polz, Eric J. Alm, Gitta Szabó, Martin F. Polz, and Eric J. Alm. 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* 336(6077):48–51.

Shapiro, B. Jesse, and Martin F. Polz. 2014a. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* 22(5):235–247.

———. 2014b. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* 22(5):235–247. NIHMS150003.

Sharon, I., M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research* 23(1):111–120.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 3(7):836–843.

Simmons, Sheri L, Genevieve DiBartolo, Vincent J Denef, Daniela S. Aliaga Goltsman, Michael P Thelen, and Jillian F Banfield. 2008. Population Genomic Analysis of Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLoS Biology* 6(7):e177.

Singh, Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh, and Soumik Sarkar. 2016. Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science* 21(2):110–124.

Sommer, Daniel D, Arthur L Delcher, Steven L Salzberg, and Mihai Pop. 2007. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8(1):64.

Stamatakis, Alexandros. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stepanauskas, Ramunas. 2012. Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology* 15(5):613–620.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102(39):13950–13955. `Masignani,Vega,2005,Genome`.

Thingstad, T. Frede. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* 45(6):1320–1328.

Thingstad, TF. 1998. A theoretical approach to structuring mechanisms in the pelagic food web. *Hydrobiologia* 363.

Turesson, Göte. 1922. The species and the variety as ecological units. *Hereditas* 3(1): 100–113.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37–43.

Varghese, Neha J., Supratim Mukherjee, Natalia Ivanova, Konstantinos T. Konstantinidis, Kostas Mavrommatis, Nikos C. Kyrpides, and Amrita Pati. 2015. Microbial species delineation using whole genome sequences. *Nucleic Acids Research* 43(14):6761–6771.

Walsh, D. A. 2004. Evolution of the RNA Polymerase B′ Subunit Gene (rpoB′) in Halobacteriales: a Complementary Molecular Marker to the SSU rRNA Gene. *Molecular Biology and Evolution* 21(12):2340–2351.

Warnecke, F., R. Sommaruga, R. Sekar, J. S. Hofer, and J. Pernthaler. 2005. Abundances, Identity, and Growth State of Actinobacteria in Mountain Lakes of Different UV Transparency. *Applied and Environmental Microbiology* 71(9):5551–5559.

Watson-Haigh, N. S., C. A. Shang, M. Haimel, M. Kostadima, R. Loos, N. Deshpande, K. Duesing, X. Li, A. McGrath, S. McWilliam, S. Michnowicz, P. Moolhuijzen, S. Quenette, J. N. D. L. Revote, S. Tyagi, and M. V. Schneider. 2013. Next-generation sequencing: a challenge to meet the increasing demand for training workshops in Australia. *Briefings in Bioinformatics* 14(5):563–574.

Webber, Emily. 2018. *Building Successful Communities of Practice*. TACIT.

Welch, Lonnie, Fran Lewitter, Russell Schwartz, Cath Brooksbank, Predrag Radivojac, Bruno Gaeta, and Maria Victoria Schneider. 2014. Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Computational Biology* 10(3):e1003496.

Wenger, E. 1998. *Communities of Practice: Learning, Meaning, And Identity*. Cambridge University Press.

———. 2011. Communities of practice: A brief introduction. *National Science Foundation (US)*.

Wenger, E, RA McDermott, and Snyder W. 2002. *Cultivating Communities of Practice*. Harvard Business School Press.

Wenger, Etienne, and William M. Snyder. 2000. Communities of Practice: The Organizational Frontier. *Harvard Business Review* 78(1):139–145. NIHMS150003.

Whitaker, Rachel J., Dennis W. Grogan, and John W. Taylor. 2005. Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon Sulfolobus islandicus. *Molecular Biology and Evolution* 22(12):2354–2361.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Wilson, Greg. 2016. Software Carpentry: lessons learned. *F1000Research* 3.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. Good enough practices in scientific computing. *PLOS Computational Biology* 13(6):e1005510.

Wrighton, K. C., B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams, P. E. Long, and J. F. Banfield. 2012. Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* 337(6102):1661–1665.

Wu, Qinglong L., and Martin W. Hahn. 2006a. Differences in structure and dynamics of Polynucleobacter communities in a temperate and a subtropical lake, revealed at three phylogenetic levels. *FEMS Microbiology Ecology* 57(1):67–79.

———. 2006b. High predictability of the seasonal dynamics of a species-like Polynucleobacter population in a freshwater lake. *Environmental Microbiology* 8(9):1660–1666.

Zaremba-Niedzwiedzka, Katarzyna, Johan Viklund, Weizhou Zhao, Jennifer Ast, Alexander Sczyrba, Tanja Woyke, Katherina McMahon, Stefan Bertilsson, Ramunas Stepanauskas, and Siv G E Andersson. 2013. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biology* 14(11):R130.

Zojer, Markus, Lisa N Schuster, Frederik Schulz, Alexander Pfundner, Matthias Horn, and Thomas Rattei. 2017. Variant profiling of evolving prokaryotic populations. *PeerJ* 5:e2997.

Zwart, G, BC Crump, MP Kamst-van Agterveld, F Hagen, and SK Han. 2002. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology* 28:141–155.