

ORIGINAL ARTICLE

Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations

Matthew L Bendall^{1,8}, Sarah LR Stevens^{2,8}, Leong-Keat Chan¹, Stephanie Malfatti¹, Patrick Schwientek¹, Julien Tremblay¹, Wendy Schackwitz¹, Joel Martin¹, Amrita Pati¹, Brian Bushnell¹, Jeff Froula¹, Dongwan Kang¹, Susannah G Tringe¹, Stefan Bertilsson³, Mary A Moran⁴, Ashley Shade⁵, Ryan J Newton⁶, Katherine D McMahon^{2,7} and Rex R Malmstrom¹

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

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.

The ISME Journal (2016) 10, 1589–1601; doi:10.1038/ismej.2015.241; published online 8 January 2016

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, 2014). In fact, recent comparative genomic analyses support an alternate

Correspondence: RR Malmstrom, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.
E-mail: rrmalmstrom@lbl.gov

⁸These authors contributed equally to this work.

Received 15 June 2015; revised 22 October 2015; accepted 18 November 2015; published online 8 January 2016

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; Deneff 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, 2014).

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.

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 $\sim 11\,000\text{ m}^2$, a maximum depth of 9 m 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 S4) and filtered on 0.2- μm pore-size polyethersulfone Supor filters (Pall Corp., Port Washington, NY, USA) prior to storage at $-80\text{ }^\circ\text{C}$. 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 $2 \times 150\text{ bp}$ 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 ≤ 0.25 and a minimum of 10 overlapping bases from paired sequences, resulting in merged read lengths of 150–290 bp. Metagenomic sequence reads are publicly available on the JGI Genome Portal (<http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=TroutBogmetagenomicdata>).

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 S5). 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 ≥ 2.5 kbp 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 $\geq 95\%$ 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 ≥ 10 -fold coverage in ≥ 3 years of the time-series study were then manually curated to ensure all contigs shared similar abundance patterns (Supplementary Figure S2). 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 S6). Accession numbers for publically

available genomes deposited in IMG are listed in Supplementary Table S7.

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 S6; 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). This coverage

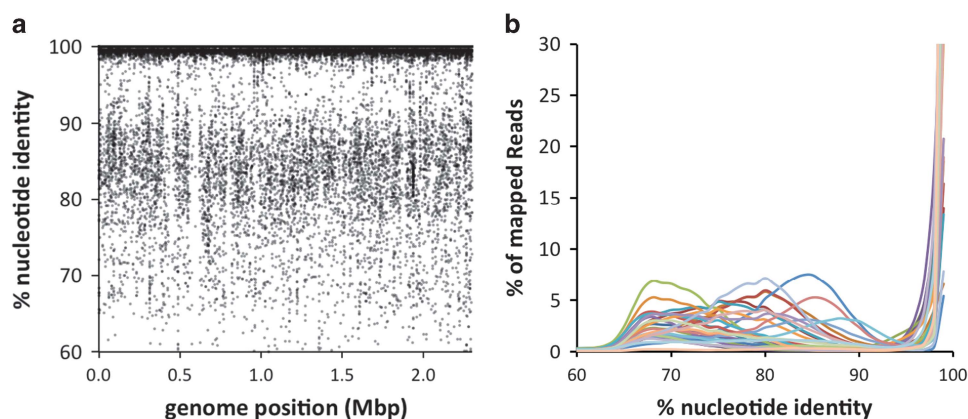


Figure 1 ‘Sequence-discrete’ populations revealed by metagenomic read mapping. (a) An example recruitment plot of 50 000 shotgun reads mapping across the *Chlorobium*-111 genome at various nucleotide identity levels. Each dot represents a read. (b) Summary of reads mapping at each percentage of nucleotide 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.

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 $\geq 99\%$ identity.

SNP identification and analysis

SNPs were discovered by first mapping reads with $\geq 95\%$ nucleotide sequence identity from each time point to the reference genomes using BWA. The vast majority of recruited reads actually mapped with $\geq 99\%$ 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 S8). 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 reference 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 $\geq 95\%$ sequence identity to each gene at each time point. Gene coverage was normalized by gene length, and spurious short gene annotations (<450 bp) 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 ≤ 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 999 000, 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 $P < 0.0001$.

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 S1), and the tight synchronization of contig coverage within bins (Supplementary Figure S2), 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). 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 on availability of related reference genomes (Supplementary Figure S3). Estimates of genome completeness ranged from ~50 to 100% (Table 1).

Genetic heterogeneity in natural populations

The recovered genomes were assembled from sequences collected at several time points and do

Table 1 Genomes reconstructed from metagenomic-combined assembly

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

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

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; Deneff 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 $\geq 99\%$ nucleotide identity to reference genomes and closely related populations whose reads mapped with $< 90\%$ identity (Figure 1). A large drop in coverage around 95% sequence identity was observed in all genomes (Figure 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 $\geq 99\%$ nucleotide identity, we found numerous SNPs in each population, ranging from 8501 SNPs in *Holophagales*-254 to only 3 SNPs in TM7-1225 (Table 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

Table 2 Summary of single-nucleotide polymorphisms (SNPs)

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

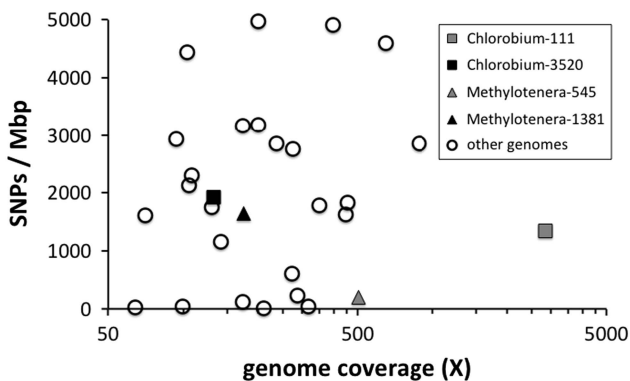


Figure 2 Differences in SNP-level heterogeneity among coexisting populations. The number of SNPs found in each sequence-discrete population, normalized to genome size (SNPs per Mbp), varied by three orders of magnitude among populations with similar coverage levels. Although the power to identify low-frequency 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.

(Figure 2; Supplementary Table S1). For example, *Methylothera*-1381 had eightfold more SNPs per Mbp than its close relative *Methylothera*-545 even though *Methylothera*-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 S3).

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 2).

Most SNPs within the sequence-discrete populations did not result in amino-acid substitutions (Table 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 2). The small proportion of nonsynonymous SNPs might indicate that purifying selection was driving 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 3a and c; Supplementary Figure S4). 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 3b and d; Supplementary Figure S4), 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 3b and d). Diversity levels also shifted substantially from year to year within *Methylobacter*-1380, *Methylotenera*-1381 and *Sulfurimonas*-1998 (Supplementary Figure S4).

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 4a; Supplementary Figure S4). These SNP sites were not localized to specific genomic regions (Supplementary Figure S5). This pattern did not result from differences in coverage (Supplementary Figure S1) or differences

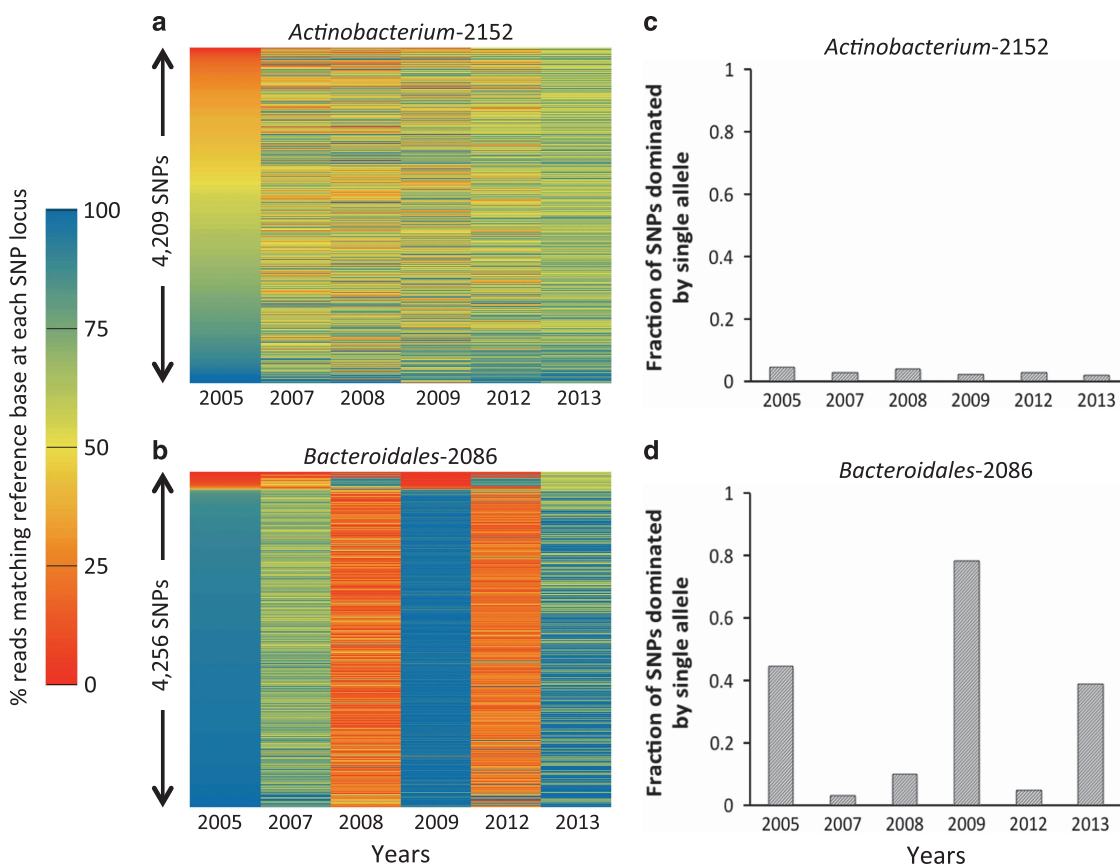


Figure 3 Temporal dynamics of SNP allele frequencies within different populations. (a, b) Two examples of populations with different SNP dynamics. SNPs are arrayed along the y axis, with each row representing one SNP locus. SNP color indicates allele frequency, that is, the percentage of metagenomic 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 ($\geq 95\%$ frequency) in each year. Broad patterns of allele frequencies were determined by combining sequence data for each year.

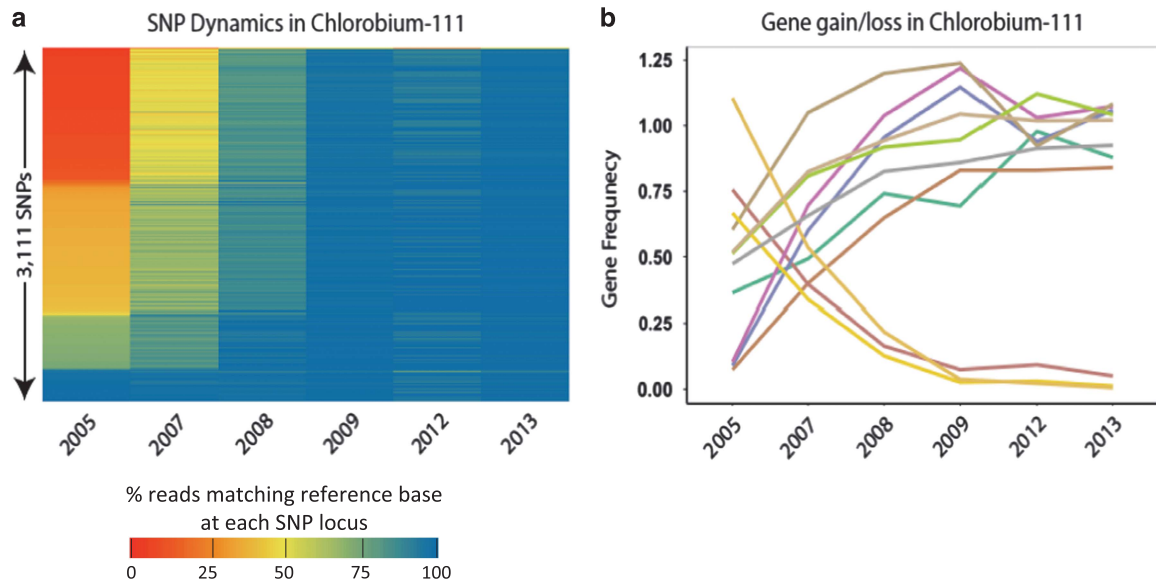


Figure 4 Temporal trends in SNP allele frequencies and gene content in a natural *Chlorobium* population. (a) SNPs are arrayed along the y axis, with each row representing one SNP locus. SNP color indicates allele frequency, that is, the percentage of metagenomic reads supporting the reference allele 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 S2. Broad patterns of allele frequencies and gene abundances were determined by combining sequence data for each year.

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 increased until they were encoded by nearly every cell in the population (Figure 4b; Supplementary Table S2). 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 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). In 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 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 S2).

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, 2014). 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 2), suggest that 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 *Synechococcus* (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 S4). 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 21 kbp region that lacked SNPs entirely ($P < 0.0001$; Supplementary Figure S5; Supplementary Table S3). Large SNP-free regions of 41 kbp, 9–25 kbp, 22–23 kbp, 11 kbp and 12 kbp were also found in *Methylobacter*-2062, *Holophagales*-254, *Opitutae*-1800, *Opitutae*-1301

and *Methylophilaceae*-913, respectively ($P < 0.0001$; Supplementary Table S3). 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 S3), 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 *et al.*, 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 3b and 4, Supplementary Figure S4). 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 S4). The *Chlorobium* populations were separated in sequence space by the coverage discontinuity around 95% nucleotide sequence identity—for example, metagenomic reads mapping with $\geq 99\%$ sequence identity to *Chlorobium*-111 also mapped with $\sim 70\text{--}90\%$ similarity to *Chlorobium*-3520, and vice versa—indicating that these populations could not be more similar and still remain sequence discrete (Figure 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; Dahllöf *et al.*, 2000; Walsh *et al.*, 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.

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, 2014). 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.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

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).

References

- Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.
- Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. (2007). Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci USA* **104**: 1883–1888.
- Barrick JE, Lenski RE. (2009). Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb Symp Quant Biol* **74**: 119–129.
- Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D *et al.* (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**: 1243–1247.
- Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML *et al.* (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol* **10**: e1001265.
- Caro-Quintero A, Deng J, Auchtung J, Brettar I, Hofle MG, Klappenbach J *et al.* (2011). Unprecedented levels of horizontal gene transfer among spatially co-occurring *Shewanella* bacteria from the Baltic Sea. *ISME J* **5**: 131–140.
- Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**: 347–355.
- Clausen J, Keck DD, Hiesey WM. (1940). *Effects of Varied Environments on Western North American plants. (Experimental Studies on the Nature of Species)*. Carnegie Institute of Washington.
- Cohan FM. (1994). The effects of rare but promiscuous genetic exchange on evolutionary divergence in prokaryotes. *Am Nat* **143**: 965–986.
- Cohan FM. (2001). Bacterial species and speciation. *Syst Biol* **50**: 513–524.
- Cohan FM, Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373–R386.
- Coleman ML, Chisholm SW. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* **15**: 398–407.
- Cordero OX, Polz MF. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**: 263–273.
- Dahllof I, Baillie H, Kjelleberg S. (2000). *rpoB*-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl Environ Microbiol* **66**: 3376–3380.
- Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.
- Denef VJ, Banfield JF. (2012). In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* **336**: 462–466.

- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Harbl C *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.
- Doolittle WF. (2012). Population genomics: how bacterial species form and why they don't exist. *Curr Biol* **22**: R451–R453.
- Eisen JA. (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. *J Mol Evol* **41**: 1105–1123.
- Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M *et al.* (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA* **98**: 15056–15061.
- Fraser C, Hanage WP, Spratt BG. (2007). Recombination and the nature of bacterial speciation. *Science* **315**: 476–480.
- Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**: 741–746.
- Guttman DS, Dykhuizen DE. (1994). Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* **138**: 993–1003.
- Herron MD, Doebeli M. (2013). Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol* **11**: e1001490.
- Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.
- Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**: 587–590.
- Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.
- Kang DD, Froula J, Egan R, Wang Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**: e11165.
- Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.
- Konstantinidis KT, DeLong EF. (2008). Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**: 1052–1065.
- Lee H, Popodi E, Tang HX, Foster PL. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* **109**: E2774–E2783.
- Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18.
- Magic T, Salzberg SL. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.
- Maharjan RP, Ferenci T, Reeves PR, Li Y, Liu B, Wang L. (2012). The multiplicity of divergence mechanisms in a single evolving population. *Gene Biol* **13**: R41.
- Majewski J, Cohan FM. (1999). Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* **152**: 1459–1474.
- Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER *et al.* (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J* **4**: 1252–1264.
- Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y *et al.* (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115–D122.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
- Messer PW, Petrov DA. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28**: 659–669.
- Moore LR, Chisholm SW. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. *Limnol Oceanogr* **44**: 628–638.
- Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo CW, Poretsky R *et al.* (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* **77**: 6000–6011.
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.
- Paulson JN, Pop M, Bravo HC. (2011). Metastats: an improved statistical method for analysis of metagenomic data. *Genome Biol* **12**: P17.
- Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.
- Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.
- Rodriguez-Brito B, Li LL, Wegley L, Furlan M, Angly F, Breitbart M *et al.* (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739–751.
- Rosen MJ, Davison M, Bhaya D, Fisher DS. (2015). Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* **348**: 1019–1023.
- Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G *et al.* (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**: 48–51.
- Shapiro BJ, Polz MF. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* **22**: 235–247.
- Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120.

- Simmons SL, Dibartolo G, Denev VJ, Goltsman DS, Thelen MP, Banfield JF. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6**: e177.
- Sommer DD, Delcher AL, Salzberg SL, Pop M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**: 64.
- Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
- Thingstad TF. (1998). A theoretical approach to structuring mechanisms in the pelagic food web. *Hydrobiologia* **363**: 59–72.
- Thingstad TF. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.
- Turesson G. (1922). Species and the variety as ecological units. *Hereditas* **3**: 100–113.
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al*. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.
- Walsh DA, Baptiste E, Kamekura M, Doolittle WF. (2004). Evolution of the RNA polymerase B' subunit gene

(rpoB') in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. *Mol Biol Evol* **21**: 2340–2351.

- Whitaker RJ, Grogan DW, Taylor JW. (2005). Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol* **22**: 2354–2361.

- Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC *et al*. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661–1665.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Supplementary Information accompanies this paper on The ISME Journal website (<http://www.nature.com/ismej>)