Dynamics of Sequence-Discreet Bacterial Populations Inferred Using Metagenomics

Sarah Stevens[1]**,** Matthew Bendall[2], Dongwan Kang[2], Jeff Froula[2], Rob Egan[2], Leong-Keat Chan[2], Susannah Tringe[2], Katherine McMahon[2], Rex Malmstrom[2]

[1]University of Wisconsin – Madison, Department of Bacteriology
[2]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA, USA

March 2014

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Dynamics of Sequence-Discrete Bacterial Populations Inferred Using Metagenomes

Sarah Stevens[1], Matthew Bendall[2], Dongwan Kang[2], Jeff Froula[2], Rob Egan[2], Leong-Keat Chan[2], Susannah Tringe[2], Katherine McMahon[1], and Rex Malmstrom[2]

[1]University of Wisconsin - Madison, Dept. of Bacteriology; [2]Department of Energy Joint Genome Insitute, Walnut Creek, CA, USA

Abstract

From a multi-year metagenomic time series of two dissimilar Wisconsin lakes we have assembled dozens of genomes using a novel approach that bins contigs into distinct genomes based on sequence composition, e.g. kmer frequencies, and contig coverage patterns at various times points. Next, we investigated how these genomes, which represent sequence-discrete bacterial populations, evolved over time and used the time series to discover the population dynamics. For example, we explored changes in single nucleotide polymorphism (SNP) frequencies as well as patterns of gene gain and loss in multiple populations. Interestingly, SNP diversity was purged at nearly every genome position in some populations during the course of this study, suggesting these populations may have experienced genome-wide selective sweeps. This represents the first direct, time-resolved observations of periodic selection in natural populations, a key process predicted by the 'ecotype model' of bacterial diversification.

Introduction

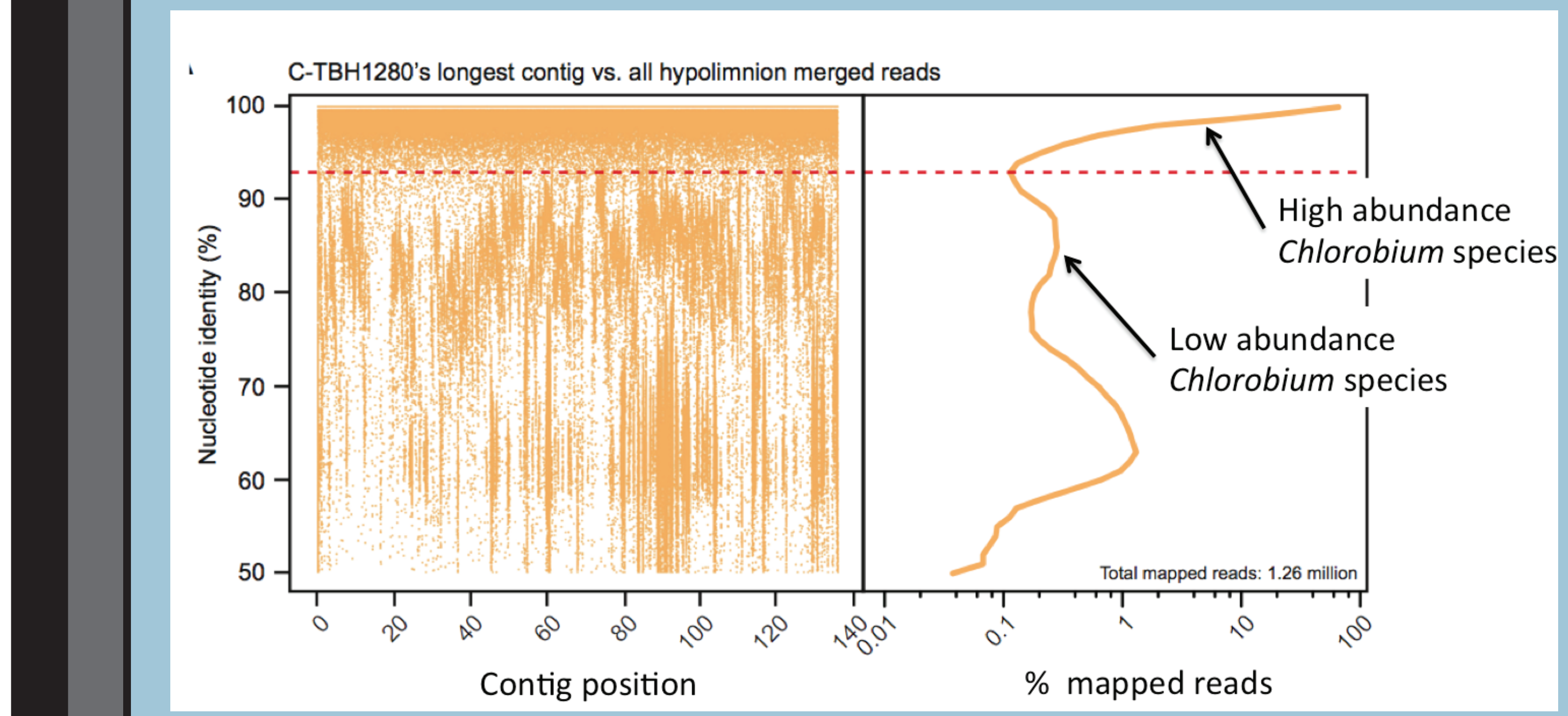

Fig. 1 - Sequence discrete populations

Left: Metagenomic read recruitment to the largest contig in the reconstructed Chlorobium-1280 genome using various sequence similarity levels.
Right: Summary of % reads mapped at each nucleotide identity in left panel.

For this study, we used a set of 96 metagenomes over the course of three years from two layers of Trout Bog, a dystrophic bog lake in Northern Wisconsin, and a set of 96 metagenomes over five years from Lake Mendota, a eutrophic lake in Madison, Wisconsin. Metagenomics reads from all time points were binned into a single pan-assembly for each environment. Assembled contigs were grouped into genome bins based on sequence composition, e.g. k-mer frequency, and similar coverage levels at various time points (see Methods). The metagenome reads were then mapped back to these contigs. As shown in figure 1, there is coverage discontinuity demarcating a sequence discrete population at ~95% nucleotide identity. We then examined these populations using single nucleotide polymorphism(SNP) analysis.

Binning Methods

Using the combined assembly of all metagenomes for the hypolimnion of Trout Bog, genomes were initially binned manually by first using sequence-composition-based classifiers PhylopythiaS and Classifier for Metagenomic Sequences (ClaMS) (Patil, Roune, & McHardy, 2012). Contigs grouped at the family level were further separated into genome bins based on differences in overall coverage. Contigs in the same bin showed a strikingly similar coverage pattern at all time points, thus validating the bins (Figure 2).

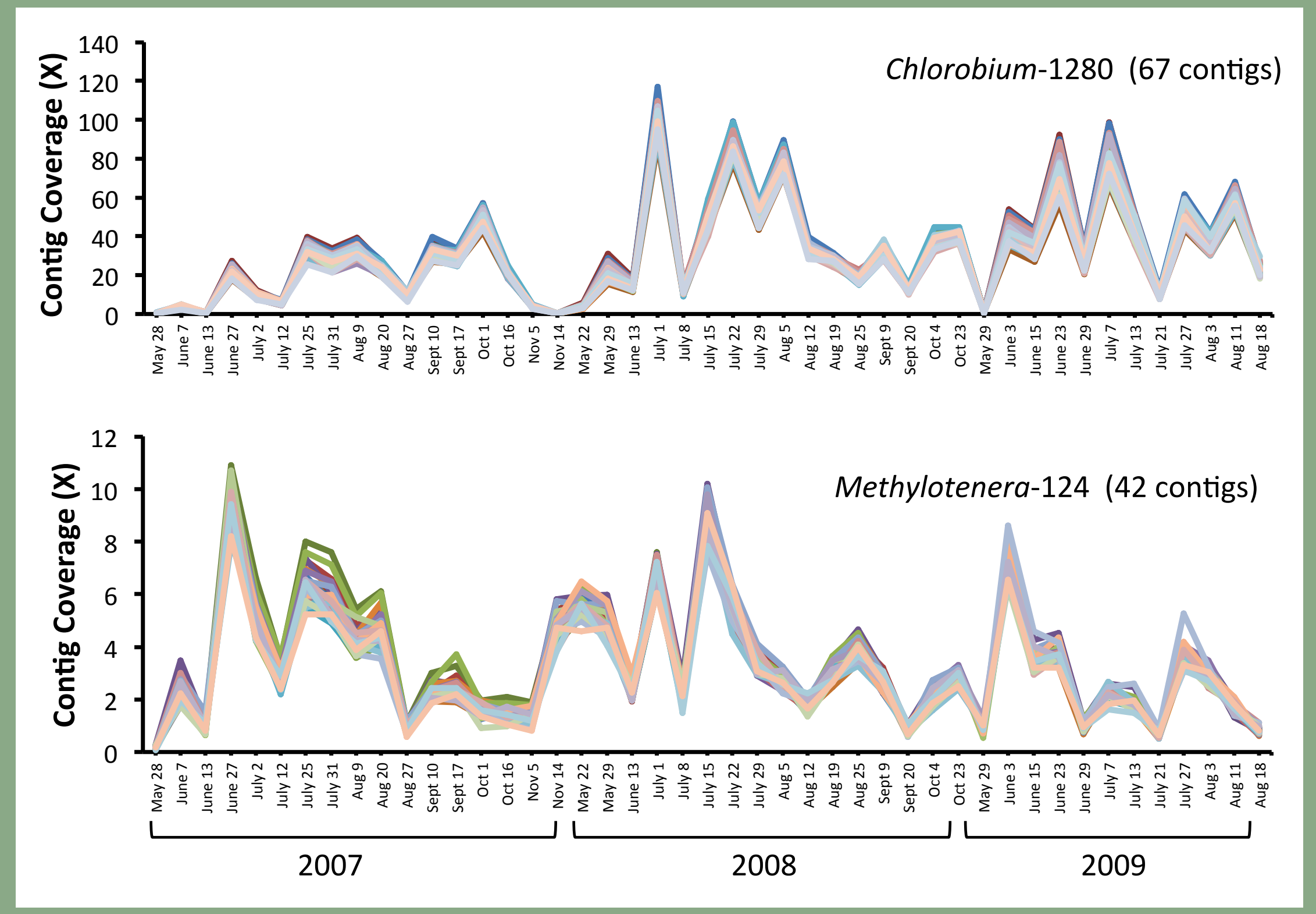

Fig. 2 – Tight synchronization of contig coverage levels within genome bins

For two genome bins, the coverage pattern across time for each contig in the genome bin

	Mendota	Trout Bog Epi	Trout Bog Hypo
# of genome bins	102	36	64
labeled at phylum level	98	36	60
phyla represented	8	6	8
labeled at class level	78	35	56
classes represented	14	11	15

Phylum	Mendota	Trout Bog Epi	Trout Bog Hypo
Acidobacteria	0	2	3
Actinobacteria	17	9	9
Bacteroidetes	33	3	9
Chlamydiae	1	0	0
Chlorobi	0	2	2
Chloroflexi	1	0	0
Elusimicrobia	0	0	1
Ignavibacteria	0	0	2
Planctomycetes	13	0	0
Proteobacteria	12	17	26
Alphaproteobacteria	2	3	4
Betaproteobacteria	7	9	11
Deltaproteobacteria	1	1	4
Epsilonproteobacteria	0	0	1
Gammaproteobacteria	2	4	4
Tenericutes	2	0	0
Verrucomicrobia	8	3	8

Fig. 3 - Genome Bin Summary Table

Top: Statistics for the genome bins and their classification; Bottom: Classification distribution for the genome bins
Epi short for Epilimnion, the upper warmer, oxic layer of the lake when stratified; Hypo short for Hypolimnion, the bottom, cooler, anoxic layer of the lake when stratified

Using Metabat, a software tool developed at JGI, a program that automatically and robustly bins contigs based on sequence composition and coverage patterns, we recovered 202 genomes from Trout Bog and from Lake Mendota. We estimate these genomes to be 50-100% complete based on a set of 139 single copy genes conserved among nearly all bacteria. These genomes were then classified based parsing the results of Phylosift(Darling et al., 2014). The results of the classification are summarized in figure 3.

Results and Future Directions

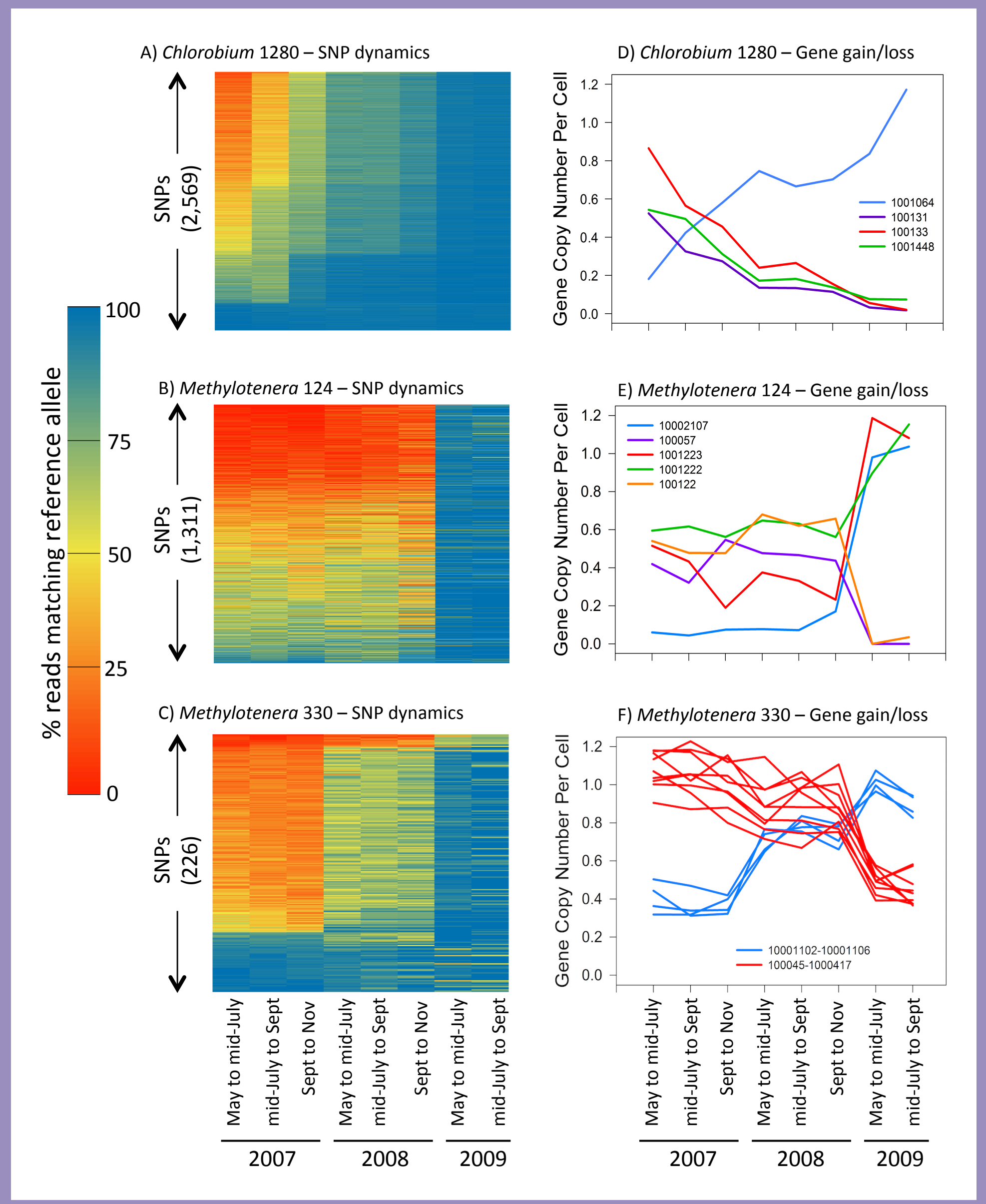

Fig. 4 – Selective sweeps in natural communities revealed by patterns in SNP frequencies and gene gain/loss

Left: Dynamics of allele frequencies at hundreds to thousands of SNP sites over a three-year period;
Right: Genes that were gained or lost from the population

Data for the three previously binned, sequence discrete populations shows all SNPs tending toward fixation in the three-year time series. This provides the first direct evidence supporting the 'ecotype model' of bacterial diversification which predicts selective pressure will periodically purge diversity, genome wide, for an ecotype(Cohan & Perry, 2007). Genes that were gained or lost follow the same pattern as the SNPs, suggesting these genes were present or absent, respectively, in the strain most dominate at the last time points. Next, we intend to investigate the population dynamics of the larger set of genomes binned with Metabat. We will look at SNP diversity across time as well as persistence vs. transience and seasonal dynamics.

References

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

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. a., Bik, H. M., & Eisen, J. a. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ, 2, e243.

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

Patil, K. R., Roune, L., & McHardy, A. C. (2012). The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PloS one, 7(6), e38581.

Acknowledgements

North Temperate Lakes
Microbial Observatory
Long-Term Ecological Research at the University of Wisconsin