Genome-wide Selective Sweeps in Natural Bacterial Populations Revealed by Time-series Metagenomics

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

[1]DOE Joint Genome Institute, [2]Uppsala University, [3]University of Georgia, [4]Vale University
[5]University of Wisconsin-Milwaukee, [6] University of Wisconsin-Madison

June 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

LBNL-xxxxxx

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.

Genome-wide selective sweeps in natural bacterial populations revealed by time-series metagenomics

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

[1]DOE Joint Genome Institute [2]Uppsala University [3]University of Georgia [4]Yale University [5]University of Wisconsin-Milwaukee [6]University of Wisconsin-Madison

ABSTRACT

Multiple evolutionary models have been proposed to explain the formation of genetically and ecologically distinct bacterial groups. Time-series metagenomics enables direct observation of evolutionary processes in natural populations, and if applied over a sufficiently long time frame, this approach could capture events such as gene-specific or genome-wide selective sweeps. Direct observations of either process could help resolve how distinct groups form in natural microbial assemblages. Here, from a three-year metagenomic study of a freshwater lake, we explore changes in single nucleotide polymorphism (SNP) frequencies and patterns of gene gain and loss in populations of *Chlorobiaceae* and *Methylophilaceae*. SNP analyses revealed substantial genetic heterogeneity within these populations, although the degree of heterogeneity varied considerably among closely related, co-occurring *Methylophilaceae* populations. SNP allele frequencies, as well as the relative abundance of certain genes, changed dramatically over time in each population. Interestingly, SNP diversity was purged at nearly every genome position in one of the *Chlorobiaceae* populations over the course of three years, while at the same time multiple genes either swept through or were swept from this population. These patterns were consistent with a genome-wide selective sweep, a process predicted by the 'ecotype model' of diversification, but not previously observed in natural populations.

APPROACH

- **Shotgun sequenced freshwater community at 45 times points from 2007-2009**
- **Assembled 2 genomes from *Chlorobiaceae* and 2 from *Methylophilaceae***
- **Mapped metagenomic reads to genomes at >95% nucleotide identity to identify:**
  1) **'sequence-discrete' populations**
  2) **Allele frequencies at SNP loci**
  3) **Relative gene abundance within populations**

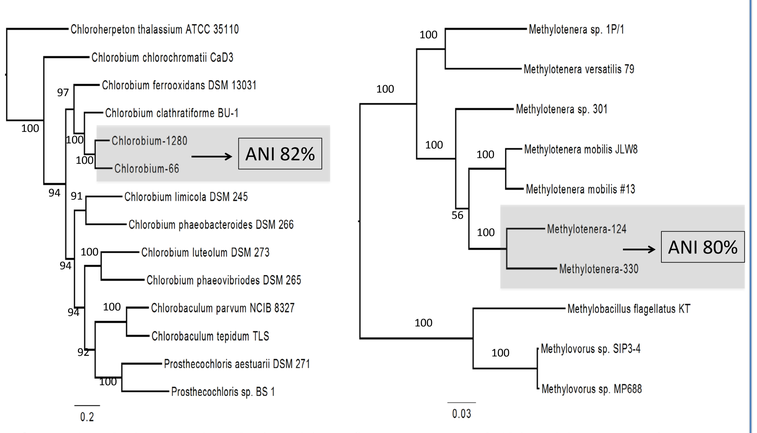

Figure 1: Phylogenetic relationships of reconstructed *Chlorobiaceae* and *Methylophilacea* genomes. Maximum likelihood tree of 27 conserved marker genes extracted using Phylosift. Bootstrap values generated from 100 replicates. Scale bar indicates substitutions per site. Grey boxes identify reconstructed genomes from this study.

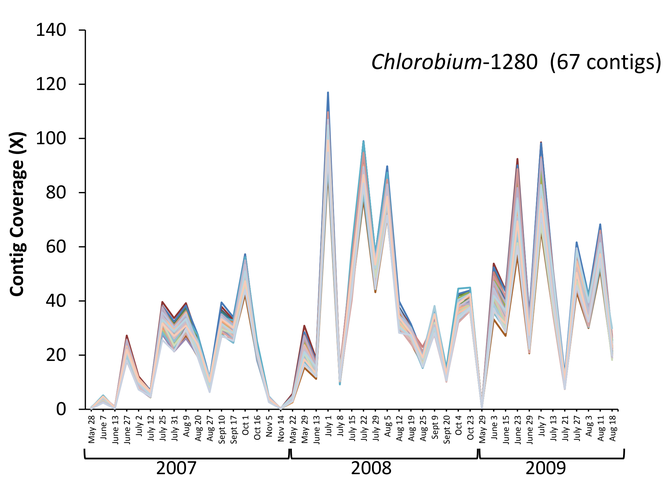

Figure 2: Temporal contig coverage patterns. Each contig is represented by a different colored line. The tight synchronization of contig coverage within each genome bin indicates these contigs were derived from the same organism.

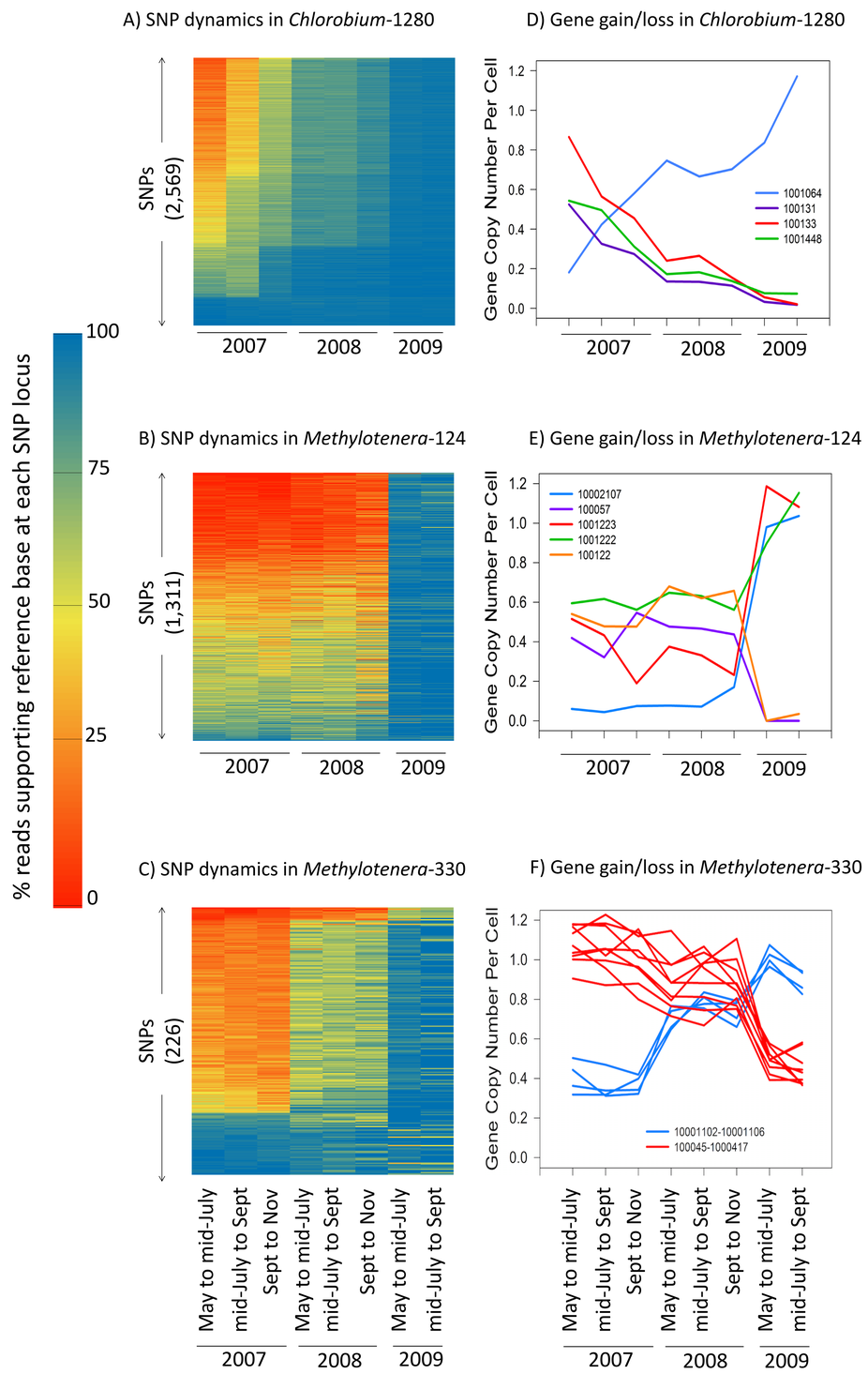

Figure 3: Temporal trends in SNP allele frequencies and gene content in natural *Chlorobium* and *Methylotenera* populations. Broad patterns were determined by combining metagenomic mapping results from all 45 time points into eight seasonal time periods.

- **Allele Frequencies (A-C):** SNPs are arrayed along the y-axis, one row equals one SNP locus, with the total number indicated in parentheses. SNP color indicates allele frequency, i.e. the percentage of metagenomic reads supporting the reference allele during each time period. SNP loci dominated by a single allele appear either as red (few reads matching reference base) or blue (most reads matching reference base). SNPs were distributed evenly throughout each genome (data not shown).
- **Gene Gain and Loss (D-F):** Relative abundance of genes gained or lost from *Chlorobium* and *Methylotenera* populations. Copy number per cell determined as coverage of a gene divided by median coverage of all other genes in genome. Gene locus id's are indicated in the legends. Two sets of contiguous genes were gained and lost from *Methylotenera*-330, and genes in each set were plotted with the same line color.

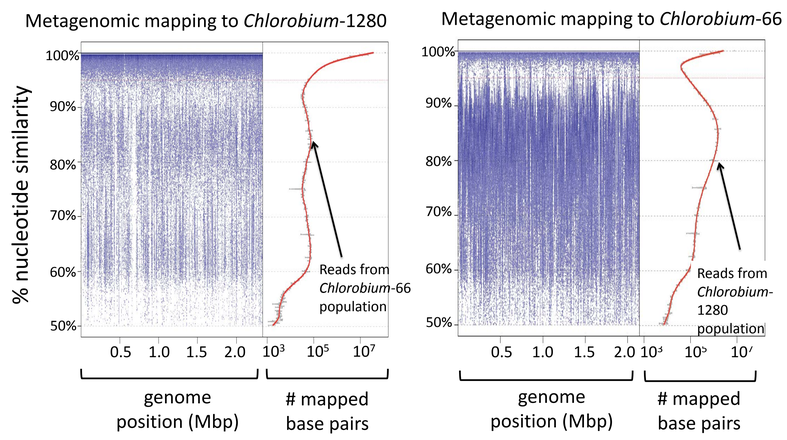

Figure 4: Metagenomic read recruitment to *Chlorobium* genomes.

- 'Sequence-discrete' populations revealed by reads that mapped with >95% nucleotide identity
- Closely related, co-occurring populations separated by coverage discontinuity at ~95% identity

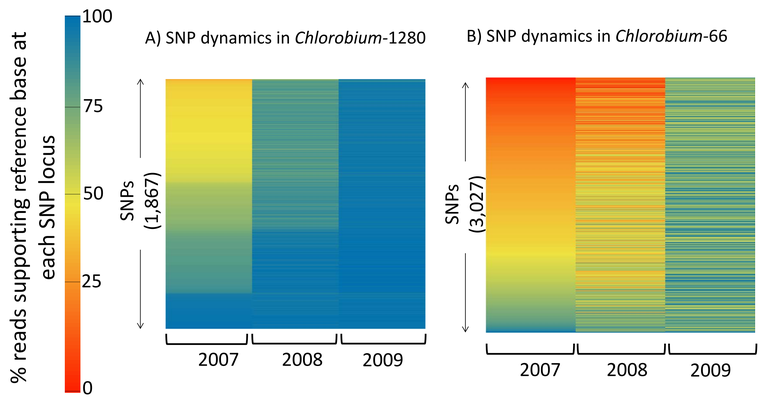

Figure 5: Temporal trends of SNP allele frequencies in *Chlorobium*-1280 (A) and *Chlorobium*-66 (B) populations. Allele frequencies were calculated per year due to coverage limitations in the less abundant *Chlorobium*-66 population. Coverage of the higher abundance *Chlorobium*-1280 was informatically reduced to comparable levels, thus the total number of detected SNPs was lower than reported when all data were used (see Figure 3). The genome-wide purge of SNP diversity in *Chlorobium*-1280 was still apparent with lower coverage and temporal resolution. In contrast, *Chlorobium*-66 did not experience a similar purge; only 414 of the SNP loci had an allele frequency >95% in 2009.

CONCLUSIONS

- **The dramatic loss of SNP diversity and the patterns of gene gain and loss in the *Chlorobium*-1280 population were consistent with a genome-wide selective sweep.**
- ***Methylotenera*-124 population may have experienced a 'soft sweep,' whereas most SNP diversity in *Methylotenera*-330 population was lost prior to the start of this study.**
- **'Sequence-discrete' populations behave like theoretically defined 'ecotypes'**
  - **Displacement of many co-existing strains by a single strain/lineage within the same population implies that all population members shared the same ecological niche.**
  - **Closely related, co-occurring sequence-discrete populations experience sweeps independently**

This work was supported by U.S. Dept. of Energy's Office of Science (DE-AC02-05CH11231), and the National Science Foundation's Microbial Observatory, Long Term Ecological Research, INSPIRE, and CAREER programs.