A bacterial metapopulation adapts locally to phage predation despite global dispersal

Victor Kunin[1], Shaomei He[2], Falk Warnecke[1], S. Brook Peterson[3], Hector Garcia Martin[1], Matthew Haynes[4], Natalia Ivanova[3], Linda L. Blackall[5], Mya Breitbart[6], Forest Rohwer[4], Katherine D. McMahon[2] and Philip Hugenholtz[1] ¶

[1] Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[2] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI, USA.

[3] Department of Plant Pathology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI, USA.

[4] Department of Biology, San Diego State University, CA, USA.

[5] Advanced Wastewater Management Centre, University of Queensland, St Lucia, 4072, Queensland, AUSTRALIA.

[6] University of South Florida, St. Petersburg, Florida, United States of America

¶Corresponding author: fax 925-296-5720 • email: phugenholtz@lbl.gov

Abstract

Using a combination of bacterial and phage-targeted metagenomics we analyzed two geographically remote sludge bioreactors enriched in a single bacterial species Candidatus Accumulibacter phosphatis (CAP). We inferred unrestricted global movement of this species and identified aquatic ecosystems as the primary environmental reservoirs facilitating dispersal. Highly related and geographically remote CAP strains differed principally in genomic regions encoding phage defense mechanisms. We found that CAP populations were high-density, clonal and non-recombining, providing natural targets for 'kill-the-winner' phage predation. Community expression analysis demonstrated that phages were consistently active in the bioreactor community. Genomic signatures linking CAP to past phage exposures were observed mostly between local phage and host. We conclude that CAP strains disperse globally, but must adapt to phage predation pressure locally.

Introduction

Ecological theory is largely grounded on the study of macroscopic communities (Begon, Townsend and Harper 2006). Microbial communities are compelling alternative systems for testing ecological concepts because microorganisms have shorter generation times and can be studied under controlled conditions (Buckling and Rainey 2002; Jessup, Kassen, Forde, Kerr, Buckling, Rainey and B.J.M 2004). However, microbial ecology has been limited by technological hurdles, namely the inability to characterize most microbial species because of a cultivation bottleneck and the inability to distinguish

microorganisms at high resolution (species and strains) and track them *in situ* (Pace 1997).

Molecular methods developed over the past decade are addressing these limitations and allowing microbial ecology to mature as a discipline, and in the process are challenging long held assumptions about microbial populations. For example, multi-locus sequence typing (MLST) (Maiden, Bygraves, Feil, Morelli, Russell, Urwin, Zhang, Zhou, Zurth, Caugant et al. 1998) has challenged the notion of general asexuality of microbial populations by demonstrating high rates of homologous recombination in some bacterial species (Feil, Enright and Spratt 2000). Another widely held belief, the lack of geographic boundaries for microbial populations has also been challenged by MLST-based studies of extremophiles (Papke, Ramsing, Bateson and Ward 2003; Whitaker, Grogan and Taylor 2003).

Metagenomics, the application of shotgun sequencing to environmental samples, holds the promise of providing the least biased (culture-independent) and most comprehensive (genome-wide) resolution of sympatric populations (Whitaker and Banfield 2006). We analyzed metagenomic data from two Enhanced Biological Phosphorus Removal (EBPR) sludges dominated (up to 80% of the biomass) by an as-yet uncultured species, "*Candidatus* Accumulibacter phosphatis" (CAP) (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006). The sludge samples were obtained from two geographically remote lab-scale bioreactors, one from Madison, Wisconsin, USA (US sludge) and the other from Brisbane, Australia (OZ sludge). In addition, a phage-enriched sample of the US sludge was obtained for shotgun sequencing

seven months after sampling for the bacterial metagenome. A microarray was prepared from both US datasets to examine gene expression of the bacterial and phage communities. Here we report global dispersal of, and local predation pressure on, the CAP populations revealed through comparative genomic analysis and expression data.

Results and Discussion

We began by searching for evidence of geographic isolation of the CAP populations by analyzing the phylogenetic distribution of 48 single-copy genes (Table S1). Using single copy genes ensures that any given strain is not represented by more than one sequence and therefore minimizes the possibility of misinterpreting paralogs as orthologs (Venter, Remington, Heidelberg, Halpern, Rusch, Eisen, Wu, Paulsen, Nelson, Nelson et al. 2004). PCR clone libraries were prepared for an additional gene, polyphosphate kinase (ppk), which has previously been used to strain type CAP (McMahon, Dojka, Pace, Jenkins and Keasling 2002). We determined the presence of multiple CAP strains in both the US and OZ sludge samples whose genes typically only diverged by up to 4% at the nucleotide level (Fig. 1). Additional distinct *Accumulibacter* populations were also identified with an average nucleotide sequence divergence of 15% from CAP (Fig. 1).

Contrary to recent findings in hot springs using high-resolution molecular methods (Papke, Ramsing, Bateson and Ward 2003; Whitaker, Grogan and Taylor 2003), no phylogenetic separation based on geographic locale was observed. In all trees with adequate strain representation, the US and OZ CAP strains were intermingled. Furthermore, instances of identical US and OZ ppk genes were found (Fig. 1). This

indicates global dispersal of CAP strains since the two sampled EBPR sludges have not been in direct contact and both lab-scale reactors were inoculated from local full-scale EBPR sludges that had been operating in EBPR mode for over 5 years. To our knowledge, there was no intentional transfer of sludge between either the bioreactors or wastewater treatment plants from which they were derived.

To date, CAP has only been detected in activated sludges (Crocetti, Hugenholtz, Bond, Schuler, Keller, Jenkins and Blackall 2000; Hesselmann, Werlen, Hahn, van der Meer and Zehnder 1999; Wong, Mino, Seviour, Onuki and Liu 2005; Zilles, Peccia, Kim, Hung and Noguera 2002) which are sparse and tiny microbial reservoirs on a global scale, and have only been in operation for about a century (Tchobanoglous, Burton, Stensel and Metcalf & Eddy 2003). The relatively recent introduction of activated sludge systems suggests that CAP originated and therefore is able to survive in alternative environments. Indeed, we found that CAP has multiple genes encoding functions more likely to be used in oligotrophic habitats than in nutrient-rich activated sludge (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006). These include complete pathways for nitrogen and carbon fixation, high affinity phosphate transporters and flagellar and chemotaxis genes (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006). No CAP flagella have been observed in EBPR sludges in which this species forms large clusters of cells bound together by extracellular polymeric substances (EPS) (Crocetti, Hugenholtz, Bond, Schuler, Keller, Jenkins and Blackall 2000).

To identify CAP habitats, we surveyed a range of environmental samples using *Accumulibacter*-specific PCR targeting the 16S rRNA and ppk genes that were subsequently confirmed by sequencing. *Accumulibacter* species were detected in both fresh and estaurine waters and associated sediments but were rarely observed in soil samples (Table S2). We therefore suggest that CAP populations are distributed in the environment as sparse high-density point sources (EBPR sludges) linked by dispersal via widespread diffuse reservoirs (aquatic environments), conforming to the ecological definition of a metapopulation as a collection of contained populations connected by a small amount of gene flow (Hanski 1999).

The presence of multiple strains in each sludge sample allowed us to investigate CAP for evidence of homologous recombination between strains (Fig S1). Unlike recent studies in which microbial populations were found to be highly recombining (Tyson, Chapman, Hugenholtz, Allen, Ram, Richardson, Solovyev, Rubin, Rokhsar and Banfield 2004) (Nesbo, Dlutek and Doolittle 2006), CAP strains showed no compelling evidence for genomic mosaicism, or even modest levels of homologous recombination. This apparent asexuality would prevent homogenization of local and introduced strains and thereby highlight dispersal patterns (Fig. 1).

While most of the CAP strains were represented by unassembled reads or short contigs in the metagenomic data (indicating low abundance), one strain dominated each sludge producing large contigs with high read depths allowing assessment of within-strain heterogeneity. The dominant strain populations were found to be extremely homogeneous in the US and OZ sludges with an average of one confirmed single nucleotide

polymorphism (SNP) per 163.2 and 65.6 kb respectively (Table S3). This indicates that both dominant CAP strains are virtually clonal.

The near clonality of the dominant strains, and their inability to recombine, means that the bulk of the biomass in each lab-scale EBPR sludge is composed of genetically identical cells. Such populations are natural targets for phage predation, via the so-called "kill the winner" phenomenon (Pernthaler 2005; Thingstad and Lignell 1997). Comparison of the dominant CAP strain genomes in the US and OZ sludges provided clues supporting this scenario. These dominant strains were highly similar, sharing over 95% nucleotide identity across most of the genome (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006) implying that differences are the result of recent evolutionary dynamics. One striking difference was the variability of EPS gene cassettes (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006). EPS can provide a first line of defence against phage predation by masking attachment sites on the cell surface. In response, lytic bacteriophages are known to encode strain-specific polysaccharases to degrade host EPS and allow access to the cell surface (Sutherland 2001). The observed redundancy and variability of EPS gene cassettes in CAP genomes may impede strain-specific targeting of EPS by phage.

Another phage defence mechanism are CRISPR elements (Jansen, Embden, Gaastra and Schouls 2002) (Barrangou, Fremaux, Deveau, Richards, Boyaval, Moineau, Romero and Horvath 2007). CRISPR elements are rapidly evolving clusters of short repeats regularly interspersed by unique sequences 'spacers', derived from foreign DNA entering

the cell, including phages. It was recently demonstrated that spacers provide immunity to the phages from which they were derived (Barrangou, Fremaux, Deveau, Richards, Boyaval, Moineau, Romero and Horvath 2007). The bacterial metagenomes contained numerous CRISPR elements of which five could be unambiguously assigned to CAP strains (Table S4). Both substitutions and insertions of CRISPR elements were observed between CAP strains (Fig. 2A&B, Table S4). Only one type of repeat sequence and no spacers were common to the two datasets suggesting exposure to different local phage populations.

CRISPR elements and EPS gene clusters were among the most notable differences between closely related strains of *Streptococcus thermophilus,* which is used in co-culture with *Lactobacillus* species for industrial yogurt and cheese production (Bolotin, Quinquis, Renault, Sorokin, Ehrlich, Kulakauskas, Lapidus, Goltsman, Mazur, Pusch et al. 2004). Therefore rapid acquisition and substitution of EPS gene cassettes and CRISPR elements may be a widespread response in bacteria to the pressure of phage predation in low complexity engineered ecosystems.

To test the hypothesis that phage are playing a major role in structuring CAP populations in EBPR we sampled the phage virion metagenome of the US sludge 7 months after sampling the bacterial metagenome. Eleven US CRISPR spacers, 8 of which belonged to the dominant CAP strain, had matches to phage genome fragments, with some phages being targeted by multiple spacers (Fig 2C) and some spacers targeting multiple related phages. This provides a direct link between the uncultivated bacterial host and phage virions, and confirms that the CAP population had previously been

infected by these phages. Two CRISPR spacers found only in the dominant OZ CAP strain had matches to the US phage community supporting geographic dispersal of the host and/or phage.

To confirm that phages are active in the sludge ecosystem, we monitored the US sludge at 3 time points spanning 3 months using expression arrays targeting both phage and bacterial genes obtained from the metagenomic datasets. We found that large numbers of genes originating from the phage virion metagenome and some genes in the bacterial metagenome of putative prophage origin, were highly expressed (Tables 1 and S6). These included many hypothetical proteins, but also proteins associated with phage tail assembly, a phage-specific endonuclease and terminase (Table S6) suggesting that phages are continuously active in the sludge. Since the microarray was based on phage virion genes sampled almost 2 years prior to the expression analysis, some phage must persist for long periods in the sludge. These data imply that the bacterial community is under persistent local predation pressure by phages and live in a volatile but relatively stable co-existence.

In summary, we have shown that i) CAP is globally dispersed ii) highly related and geographically remote CAP strains differ principally in genomic regions encoding phage defence mechanisms iii) high-density, clonal, non-recombining CAP populations in EBPR bioreactors are natural targets of ‘kill-the-winner’ phage predation iv) phages are consistently active in EBPR bioreactor communities and v) signatures of past phage infections in CAP are observed mostly between local phage and host. We therefore conclude that CAP strains disperse globally, but must adapt to local persistent phage

predation pressure. The present study illustrates the value of combining high-throughput sequence and gene expression data from the bacterial and viral fractions of an ecosystem to elucidate population structure, biogeography and host-parasite interdependencies.

Methods

Metagenomic sequencing

Sludge samples for the US and Australian (OZ) bacterial metagenomes were obtained on July 3rd and August 18th 2004 respectively. Sequencing, assembly and gene prediction of the bacterial metagenomes are described elsewhere (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006). To obtain a phage virion metagenome, a sludge sample from the US bioreactor was taken on February 7th 2005, 7 months after sampling for the bacterial metagenome. Virion purification techniques, construction of shotgun libraries, sequencing, assembly and gene calling are described in the Supplemental Research Data.

Bioinformatic analyses

Single-copy gene analysis was performed to infer biogeographical patterns by 1) selecting 47 conserved single-copy gene families in isolate genomes in the IMG database (Markowitz, Korzeniewski, Palaniappan, Szeto, Werner, Padki, Zhao, Dubchak, Hugenholtz, Anderson et al. 2006) using PFAM (Bateman, Coin, Durbin, Finn, Hollich, Griffiths-Jones, Khanna, Marshall, Moxon, Sonnhammer et al. 2004) profile searches with rps-BLAST (Altschul, Madden, Schaffer, Zhang, Zhang, Miller and Lipman 1997),

2) identifying members of these families in the bacterial sludge metagenomes 3) aligning each family with ClustalX (Thompson, Higgins and Gibson 1994) 4) generating neighbor-joining trees using ClustalX. See Supplemental Research Data for details.

To refine the resolution of the single-copy gene analysis, we PCR-amplified the ppk gene from the sludge biomass and environmental samples. The amplification product was cloned into *E.coli* and 96 clones were picked from each library for sequencing. See Supplemental Research Data for further details.

Quantification of SNP frequency was done using CONSED program (Gordon, Abajian and Green 1998) on the largest 10 contigs and reported polymorphisms were manually re-checked. The screen for homologous recombination was done using SNP-VISTA (Shah, Teplitsky, Minovitsky, Pennacchio, Hugenholtz, Hamann and Dubchak 2005). CRISPR elements were identified using piler-cr (Edgar 2007) , and BLASTN (Altschul, Madden, Schaffer, Zhang, Zhang, Miller and Lipman 1997) was used to link between CRISPR spacers and genomic regions. See Supplemental Research Data for details.

Microarrays

Combimatrix CustomArray™ 12K microarrays were constructed from predicted genes from both bacterial and phage metagenomes. Samples from US sludge were extracted on Oct 30th, 2006, Jan 5th, 2007, and Jan 31st, 2007. RNA was purified, labeled and hybridized to the arrays. For each probe we calculated geometric average of all replicates, with the exclusion of dubious spots (**Table S6**). See Supplemental Research Data for further details.

Acknowledgments

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Figures

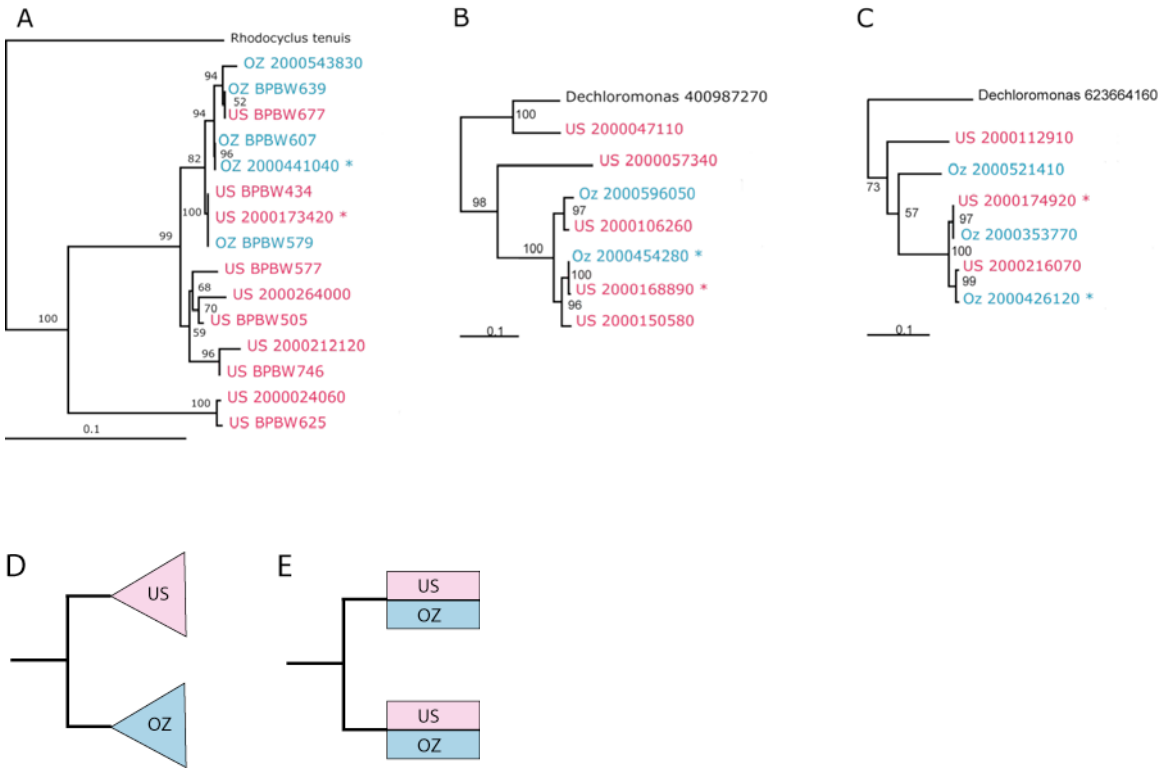

Fig. 1: Gene phylogenies reconstructed using nucleotide sequence show geographic intermingling of CAP strains. Sequences obtained from the US and OZ samples are shown in red and blue respectively. Asterisks mark dominant strains. IMG (Markowitz, Korzeniewski, Palaniappan, Szeto, Werner, Padki, Zhao, Dubchak, Hugenholtz, Anderson et al. 2006) gene object identifiers (beginning with 2000) are given for genes derived from metagenomic data. Support for interior nodes are indicated by bootstrap resampling percentages. Trees shown are: ***A.*** polyphosphate kinase ; PCR-amplified clones begin with BPBW. ***B.*** Ribosomal protein L9 and ***C.*** Holiday junction resolvase, DNA-binding subunit RuvA. Schematics are provided for reference to show the expected tree topologies for endemic (***D***) and freely migrating (***E***) populations. Note that in the

latter case, high recombination frequencies between local and introduced strains may mask dispersal patterns.

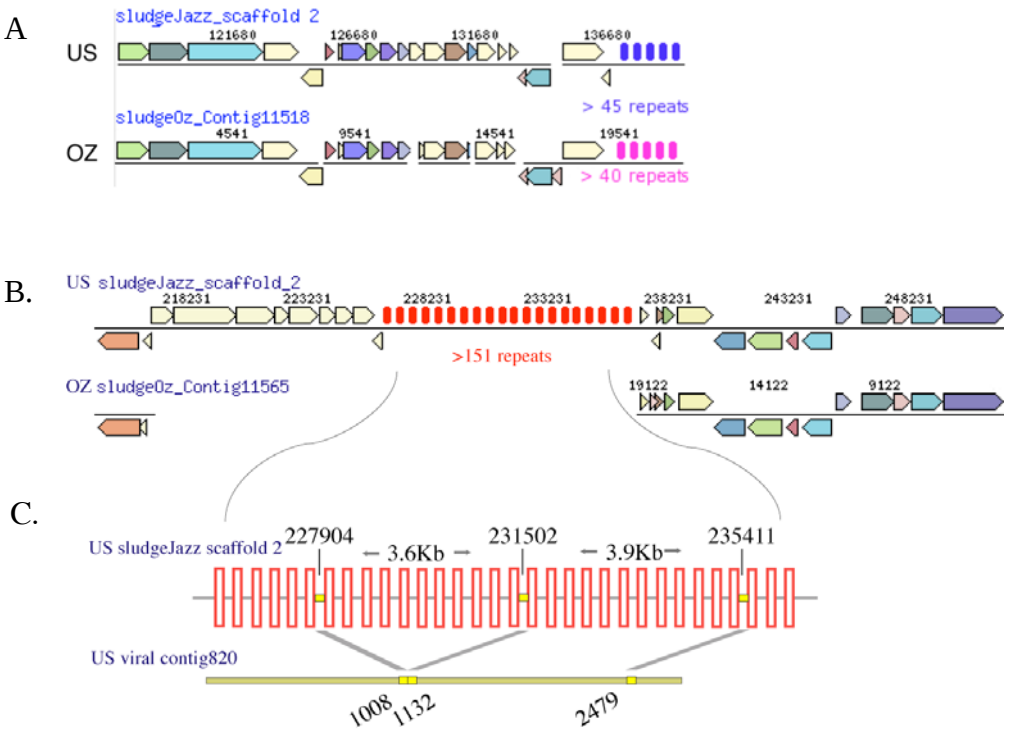

Fig 2. Sample alignments of homologous regions in the dominant US and OZ strains showing substitution (A) and insertion (B) of CRISPR elements. CRISPR repeats are indicated by sets of vertical bars with colors denoting different repeat sequences. Total number of repeats for each CRISPR element is unknown because of incomplete sequence information in the draft assembly, therefore a minimum estimate is given. (C) Schematic magnification of dominant CAP CRISPR element and a contig from the phage virion metagenome revealing a phage that has previously infected CAP. All spacers targeting the phage had the same orientation. The starting positions of each spacer and the

matching segments in the phage are indicated. The drawing is not to scale, and the actual number of repeats is significantly higher.

Table 1. A selection of several highly-expressed phage (top) and bacterial (bottom) genes. Geometric averages of 2 replica experiments, and 3 replicas for each probe are given. The results of expression array experiments are available in more detail as Table S6. In all cases, the negative control and levels are below 1000.

Probe source	30-Oct-06	5-Jan-07	31-Jan-07
Mu-like prophage protein gp29	60299	39397	56948
Bacteriophage tail assembly protein	21786	13617	23252
Phage-related protein, predicted endonuclease	18063	11717	12990
Phage terminase-like protein, large subunit	11817	11000	10039
Phage-related minor tail protein	9763	9279	7048
Acetyl-CoA acetyltransferase	11366	16536	16367
Ribosomal protein L16/L10E	4255	10686	10968
Polyphosphate kinase	4555	5010	3475

References:

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, and P. Horvath. 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315:** 1709-1712.

Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer et al. 2004. The Pfam protein families database. *Nucleic Acids Res* **32:** D138-141.

Begon, M., C.R. Townsend, and J.L. Harper. 2006. *Ecology : from individuals to ecosystems*. Blackwell Pub., Malden, MA ; Oxford.

Bolotin, A., B. Quinquis, P. Renault, A. Sorokin, S.D. Ehrlich, S. Kulakauskas, A. Lapidus, E. Goltsman, M. Mazur, G.D. Pusch et al. 2004. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nat Biotechnol* **22:** 1554-1558.

Buckling, A. and P.B. Rainey. 2002. The role of parasites in sympatric and allopatric host diversification. *Nature* **420:** 496-499.

Crocetti, G.R., P. Hugenholtz, P.L. Bond, A. Schuler, J. Keller, D. Jenkins, and L.L. Blackall. 2000. Identification of polyphosphate-accumulating organisms and design of 16S rRNA-directed probes for their detection and quantitation. *Appl Environ Microbiol* **66:** 1175-1182.

Edgar, R.C. 2007. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**.

Feil, E.J., M.C. Enright, and B.G. Spratt. 2000. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae. *Res Microbiol* **151:** 465-469.

Garcia Martin, H., N. Ivanova, V. Kunin, F. Warnecke, K.W. Barry, A.C. McHardy, C. Yeates, S. He, A.A. Salamov, E. Szeto et al. 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24:** 1263-1269.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Hanski, I. 1999. *Metapopulation ecology*. Oxford University Press, Oxford ; New York.

Hesselmann, R.P., C. Werlen, D. Hahn, J.R. van der Meer, and A.J. Zehnder. 1999. Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced biological phosphate removal in activated sludge. *Syst Appl Microbiol* **22:** 454-465.

Jansen, R., J.D. Embden, W. Gaastra, and L.M. Schouls. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43:** 1565-1575.

Jessup, C.M., R. Kassen, S.E. Forde, B. Kerr, A. Buckling, P.B. Rainey, and B. B.J.M. 2004. Big questions, small worlds: microbial model systems in ecology. *Trends in Ecology and Evolution* **19:** 189–197.

Maiden, M.C., J.A. Bygraves, E. Feil, G. Morelli, J.E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D.A. Caugant et al. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95:** 3140-3145.

Markowitz, V.M., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson et al. 2006. The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34:** D344-348.

McMahon, K.D., M.A. Dojka, N.R. Pace, D. Jenkins, and J.D. Keasling. 2002. Polyphosphate kinase from activated sludge performing enhanced biological phosphorus removal. *Appl Environ Microbiol* **68:** 4971-4978.

Nesbo, C.L., M. Dlutek, and W.F. Doolittle. 2006. Recombination in Thermotoga: implications for species concepts and biogeography. *Genetics* **172:** 759-769.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276:** 734-740.

Papke, R.T., N.B. Ramsing, M.M. Bateson, and D.M. Ward. 2003. Geographical isolation in hot spring cyanobacteria. *Environ Microbiol* **5:** 650-659.

Pernthaler, J. 2005. Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* **3:** 537-546.

Shah, N., M.V. Teplitsky, S. Minovitsky, L.A. Pennacchio, P. Hugenholtz, B. Hamann, and I. Dubchak. 2005. SNP-VISTA: An interactive SNP visualization tool. *BMC Bioinformatics* **6:** 292.

Sutherland, I. 2001. Biofilm exopolysaccharides: a strong and sticky framework. *Microbiology* **147:** 3-9.

Tchobanoglous, G., F.L. Burton, H.D. Stensel, and Metcalf & Eddy. 2003. *Wastewater engineering : treatment and reuse*. McGraw-Hill, Boston.

Thingstad, T. and R. Lignell. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13:** 19-27.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Whitaker, R.J. and J.F. Banfield. 2006. Population genomics in natural microbial communities. *Trends Ecol Evol* **21:** 508-516.

Whitaker, R.J., D.W. Grogan, and J.W. Taylor. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301:** 976-978.

Wong, M.T., T. Mino, R.J. Seviour, M. Onuki, and W.T. Liu. 2005. In situ identification and characterization of the microbial community structure of full-scale enhanced biological phosphorous removal plants in Japan. *Water Res* **39:** 2901-2914.

Zilles, J.L., J. Peccia, M.W. Kim, C.H. Hung, and D.R. Noguera. 2002. Involvement of Rhodocyclus-related organisms in phosphorus removal in full-scale wastewater treatment plants. *Appl Environ Microbiol* **68:** 2763-2769.

A bacterial metapopulation adapts locally to phage predation despite global dispersal

Online supporting material

Victor Kunin[1], Shaomei He[2], Falk Warnecke[1], S. Brook Peterson[3], Hector Garcia Martin[1], Matthew Haynes[4], Natalia Ivanova[3], Linda L. Blackall[5], Mya Breitbart[6], Forest Rohwer[4], Katherine D. McMahon[2] and Philip Hugenholtz[1] ¶

[1] Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[2] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI, USA.

[3] Department of Plant Pathology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI, USA.

[4] Department of Biology, San Diego State University, CA, USA.

[5] Advanced Wastewater Management Centre, University of Queensland, St Lucia, 4072, Queensland, AUSTRALIA.

[6] University of South Florida, St. Petersburg, Florida, United States of America

¶Corresponding author: fax 925-296-5720 • email: phugenholtz@lbl.gov

Sequencing, assembly and gene prediction of metagenomic datasets

Sludge samples for the US and Australian (OZ) bacterial metagenomes were obtained on July 3rd and August 18th 2004 respectively. Sequencing, assembly and gene prediction of the bacterial metagenomes are described elsewhere (Garcia Martin, Ivanova, Kunin, Warnecke, Barry, McHardy, Yeates, He, Salamov, Szeto et al. 2006).

The analysis of the bacterial metagenome (see main text) implicated phage as significant determinants of CAP population structure and dynamics. To validate this hypothesis, a sludge sample from the US bioreactor was obtained for the phage metagenome on February 7th 2005, 7 months after sampling for the bacterial metagenome. Phage virions were purified using a combination of filtration and density-dependent centrifugation (Breitbart, Miyake and Rohwer 2004; Breitbart, Salamon, Andresen, Mahaffy, Segall, Mead, Azam and Rohwer 2002). Approximately 200 ml of the sludge sample was filtered through a 0.2 μm Sterivex filter to remove bacteria and large particles. Phages in the filtrate were then concentrated by polyethylene glycol (PEG) precipitation. PEG 8000 was added to a final concentration of 10%, the samples were incubated for 12 hours at 4°C, and then centrifuged at 13,000*g* for 30 minutes. The phage pellet was resuspended into 0.02 μm-filtered PBS and loaded on to a caesium chloride step gradient consisting of 1 ml each of 1.7, 1.5, and 1.35 g/ml. The gradient was ultracentrifuged at 55,000*g* for 2 hours and DNA was isolated from the 1.35 – 1.5 g/ml fraction using a formamide and CTAB extraction (Sambrook, Fritsch and Maniatis 1989).

Linker-amplified shotgun libraries (LASLs) were constructed from the total phage community DNA by Lucigen Corporation (Middleton, WI) as described previously (Breitbart, Salamon, Andresen, Mahaffy, Segall, Mead, Azam and Rohwer 2002). Briefly, the phage DNA was

randomly sheared using a HydroShear, end-repaired, and double-stranded DNA linkers were ligated to the ends. The fragments were then amplified using the high-fidelity Vent DNA polymerase, ligated into the pSMART vector, and electroporated into MC12 cells. Random clones were sequenced from one end only using the AmpL1 primer. The resulting reads were assembled using the phrap program . Genes were predicted on phage fragments using fgenesv and fgenesb .

Single-copy gene analysis to infer biogeographical patterns

In metagenomic assemblies, genomes are incomplete and present in multiple pieces, making it difficult to ensure that studied regions are indeed orthologous (genes in different organisms diverged by speciation) and not paralogous (duplicated) regions in the same genome. Therefore for the population structure analysis it was essential to identify orthologous genes and remove potential paralogs. Consequently, we focused exclusively on single copy genes broadly distributed in completely sequenced genomes.

These single-copy genes were identified in the IMG database (Markowitz, Korzeniewski, Palaniappan, Szeto, Werner, Padki, Zhao, Dubchak, Hugenholtz, Anderson et al. 2006) as follows. All proteins in the IMG database were assigned to Pfam domains (Bateman, Coin, Durbin, Finn, Hollich, Griffiths-Jones, Khanna, Marshall, Moxon, Sonnhammer et al. 2004) using RPS-BLAST (reverse PSI-BLAST) (Altschul, Madden, Schaffer, Zhang, Zhang, Miller and Lipman 1997), e-value cut-off e-5. Only Pfam domains identifiable in at least half of the IMG isolate genomes (excluding draft genomes) in no more than a single copy per genome were selected. Forty-eight single-copy genes were identified in this manner and all orthologs of these genes were obtained from the bacterial metagenomic datasets for comparative analysis.

Alignments for each of the protein families were generated with ClustalW (Thompson, Higgins and Gibson 1994) using default parameters, and neighbor-joining trees were constructed in ClustalX excluding positions with gaps, correcting for multiple substitutions and 1000 bootstrap iterations were applied to establish robustness of interior nodes. These trees were manually examined and short metagenomic sequences were removed where they compromised the integrity of the phylogenetic inference, and trees were then regenerated. The metagenomic sequences were assigned manually to taxonomic groups based on phylogenetic neighborhood (**Table S1**).

We used assignments to Pfam families as opposed to calling orthologs or *de novo* protein families for the following reasons: i) applying algorithms to identify orthologs is meaningless for metagenomic datasets containing multiple species ii) Pfam assignments were performed using profile-to sequence comparisons (RPS-BLAST), which are more sensitive than pairwise alignment, iii) the calling of genes on genomes and metagenomes followed the same protocol, excluding the possibility of paralogous multiple-copy genes.

To measure the level of divergence between CAP strains, we identified contigs containing single-copy genes belonging to this species. In most cases, genes from the dominant strain in each sample could be unambiguously identified based on contig size and depth. The contigs were aligned using blastn (Altschul, Madden, Schaffer, Zhang, Zhang, Miller and Lipman 1997) and percentage divergence recorded. The average nucleotide sequence divergence between homologous regions in contigs derived from different CAP strains was 4%; the average nucleotide sequence divergence between homologous regions in contigs derived from different *Accumulibacter* populations was 15%.

Table S1. Single-copy gene assignments to Pfam families. The multiplicity of strains can be inferred from multiple copies of single-copy genes found in each species category. However, in most cases loci cannot be linked to each other and the total number of strains and species present cannot be inferred. Genes are denoted by IMG gene object identifiers (Markowitz, Korzeniewski, Palaniappan, Szeto, Werner, Padki, Zhao, Dubchak, Hugenholtz, Anderson et al. 2006).

		CAP		**Other *Accumulibacter* populations**	
Name	**Pfam id**	**US**	**OZ**	**US**	**OZ**
		2000096240			
Ribosomal_S3_C	00189	2000100000	2000407470		
		2000096230			
Ribosomal_L16	00252	2000169490	2000407460		
		2000096240			
Ribosomal_S3_N	00417	2000169500	2000407470		
		2000023690	2000643640		
Ribosomal_L20	00453	2000195230	2000470270		
			2000411200		
IGPD	00475	2000188120	2000347550	2000279830	
Ribosomal_L12	00542	2000163900	2000411830	2000078330	
SecE	00584	2000163950	2000411780		2000480260
		2000208510			
GlutR_dimer	00745	2000126780	2000476150	2000092990	
		2000096220			
Ribosomal_L29	00831	2000169480	2000407450		
Ribosomal_S16	00886	2000189560	2000468710	2000333350	
EF_TS	00889	2000181050	2000420960	2000314470	2000571780
		2000208310			
Ribosomal_L27	01016	2000112830	2000455860	2000119900	2000519180
		2000016100			
		2000194290			
		2000024010			
RNA_pol_Rpb6	01192	2000109920	2000385380	2000045330	
		2000169300			
		2000129610			
Ribosomal_L17	01196	2000078520	2000417370		
Ribosomal_L19	01245	2000189530	2000468740		
		2000017270			
		2000168920	2000596020		
Ribosomal_S6	01250	2000106290	2000454310		2000603770
		2000168890			
		2000150580			
		2000106260	2000454280		
Ribosomal_L9_N	01281	2000283080	2000596050	2000057340	
Ribosomal_L15	01305	2000169370	2000407340		
		2000174920	2000353770		
RuvA	01330	2000216070	2000426120	2000112910	2000521410
				2000133330	
HrcA	01628	2000207230	2000476580	2000217240	

		CAP		Other *Accumulibacter* populations	
Name	**Pfam id**	**US**	**OZ**	**US**	**OZ**
Ribosomal_S20p	01649	2000177910	2000406250		2000644480
		2000027400			
SmpB	01668	2000168530	2000447590		2000556360
		2000189540			
tRNA_m1G_MT	01746	2000256870	2000468730		
RRF	01765	2000181030	2000420940		
RimM	01782	2000189550			
RBFA	02033	2000175430	2000431860		
			2000426110		
			2000400840		
RuvC	02075	2000174930	2000353760	2000112920	
		2000158510			
UPF0054	02130	2000182540	2000407020		
		2000229140			
RecR	02132	2000327380			
NusG	02357	2000163940	2000411790		
		2000023310			
		2000193350			
DUF143	02410	2000250810	2000452350	2000002150	2000581610
		2000198340			
RecO	02565	2000141460	2000446350	2000113990	
			200043183		
DUF150	02576	2000175460	2000401420		
		2000189390			
Tyr_Deacylase	02580	2000021100	2000468850	2000110470	
Exonuc_VII_S	02609	2000189490	2000468780	2000140300	
		2000101260			
DUF173	02616	2000175490	2000401450	2000223230	
DUF177 poor		2000167520			
alingment	02620	2000237330	2000459870		
			2000475190		
Glu-tRNAGln	02686	2000199540	2000626120	2000223730	
Phe_tRNA-					
synt_N	02912	2000195240	2000470260		2000643650
		2000170320			
SRP_SPB	02978	2000123760	2000398270	2000317420	
B5 long gene	03484	2000195250	2000470250		
		2000197480			
		2000131630			
PNPase	03726	2000329340	2000434450		
SecG	03840		2000557670		
		2000168890			
		2000150580			
		2000106260	2000454280		
Ribosomal_L9_C	03948	2000283080	2000596050	2000057340	
		2000208510			
GlutR_N	05201	2000126780	2000476150	2000092990	2000346730

		2000122340	2000426140
RuvB_N	05496	2000174900	2000352270
		2000168820	
Trigger_C	05698	2000140640	2000454210

Polyphosphate kinase (ppk) PCR clone libraries

Environmental shotgun sequencing can easily miss low abundance species and strains (Tyson, Chapman, Hugenholtz, Allen, Ram, Richardson, Solovyev, Rubin, Rokhsar and Banfield 2004) (Tyson and Banfield 2005). Therefore, we used PCR amplification and cloning of the polyphosphate kinase (ppk) gene to provide a broader overview of *Accumulibacter* diversity in the two lab-scale EBPR samples (McMahon, Dojka, Pace, Jenkins and Keasling 2002). The ppk gene is an appropriate marker for strain diversity since it is single copy in most microbial genomes and has a faster evolutionary rate than 16S rRNA; the divergence between the dominant CAP strains and other *Accumulibacter* populations was 3.1% for 16S rRNA and 13.2% for ppk, suggesting that the latter evolves ~4 times faster than the former in *Accumulibacter* species.

Two sets of primers targeting overlapping regions of the *ppk* gene of *Accumulibacter* sp. and related organisms *Dechloromonas aromatica* and *Rhodocyclus tenuis* (ppk274f: ACCGACGGCAAGACSG and ppk1156r: CGGTAGACGGTCATCTTGAT; ppk734f: CTCGGCTGCTACCAGTTCCG and ppk1601r: GATSCCGGCGACGACGTT) were designed based on a multiple alignment of *ppk* sequences from the two sludge metagenomes, plus the *ppk* sequence of *D. aromatica.* We amplified *ppk* gene fragments using each of the primer sets from the DNA samples used to build each of the sludge metagenomic libraries. All reactions contained 1 μl of a 1:100 dilution of the template DNA, 0.25 μm of each primer, 50 μM of each dNTP and 1 unit *Taq* polymerase in 20 μl total volume. Three reactions for each primer set and sample combination were performed, using a 58°C annealing temperature and 17, 19, or 20

cycles of replication. All three reactions for each primer/sample combination were combined and the product of the expected size (~800 bp for both of the primer sets) was purified by agarose gel electrophoresis followed by gel extraction using the MinElute™ gel extraction protocol (Qiagen, Valencia, CA). We cloned purified PCR products using the TOPO II™ kit (Invitrogen, Carlsbad, CA) and transformed cloned products by electroporation into One Shot™ electrocompetent *E. coli* DH5α (Invitrogen). From each library (one for each primer set for each DNA sample), 96 clones were picked and sequenced with two-fold coverage using vector primers.

Survey for CAP in environmental samples

Sample collection and processing

We tested water, sediment and soil samples from several watersheds around Contra Costa County, CA (see **Table S2**) for the presence of *Accumulibacter* species using PCR. Sample sites ranged from 0.3 to 17.1 km distant from the Contra Costa wastewater treatment plant (CCWWTP), where wastewater is treated using EBPR. From each sample site (with a few exceptions as noted in **Table S2**) we aseptically collected 1000 ml of stream, lake or bay water, 50 ml sediment from the bottom of the water body and 50 ml soil from an adjacent bank (above obvious high water marks), and samples were stored on ice prior to processing. Water samples (200-1000 ml of each, depending on particulate concentration) were concentrated either by centrifugation (Walnut Creek near CCWWTP) or by filtration using Sterivex 0.2 μm filter cartridges (Millipore, Billerica, MA) connected to a peristaltic pump. Soil and sediment samples were each mixed thoroughly then aliquoted and frozen at -20°C. DNA was extracted from all samples using the Power Soil™ DNA extraction kit (Mo Bio, Carlsbad, CA). For water

samples, the lysis buffer was flushed through the concentrating cartridges to remove attached cells, or was used to resuspend cell pellets obtained through centrifugation.

Primer design and PCR

To test for the presence of CAP in environmental samples, we used *Accumulibacter*-specific primers targeting the 16S rRNA gene and the *ppk* gene described above. The *Accumulibacter*-specific 16S primers (cap438f: GGTTAATACCCTGWGTAGAT and cap846r: GTTAGCTACGGCACTAAAAGG) were designed based on previously tested *Accumulibacter*-specific probes (Crocetti, Hugenholtz, Bond, Schuler, Keller, Jenkins and Blackall 2000), and were used with an annealing temperature of 55°C. Primers targeting the *Accumulibacter ppk* gene (ppk254f: TCACCACCGACGGCAAGAC and ppk1357r: GACGATCATCAGCATCTTGGC) are described elsewhere (He *et al,* in preparation), and were used with an annealing temperature of 58°C . For each sample, we tested for the presence of *Accumulibacter* using both standard PCR with each set of *Accumulibacter*-specific primers, and with a 16S-targeted nested PCR approach to obtain enhanced sensitivity. All reactions contained the reagent concentrations described above in 20 μl total volume. For the un-nested PCR, reactions contained 0.5 to 20 ng template DNA, and 40 thermocycles were performed. In the first round of nested PCR, reactions contained 0.5 to 20 ng template DNA and 35 thermocycles were performed. In the second round of nested PCR, 1 μl of a 1:10 dilution of the first reaction was used as template and 35 thermocycles were again performed. In nested PCR reactions targeting the *Accumulibacter* 16S rRNA gene, we used primers specific for the 16S rRNA gene of organisms from the *Rhodocylales* order (RHC175F and RHC+1289R (Loy, Schulz, Lucker,

Schopfer-Wendels, Stoecker, Baranyi, Lehner and Wagner 2005)) in the first round, and *Accumulibacter*-specific primers cap438f and cap846r in the second round. Semi-nested PCR targeting the *ppk* gene used primers ppk274f and ppk1601r in the first round, followed by primers ppk274f and ppk1156r in round two. All samples were tested a minimum of three times with each approach and primer set, and 5 μl each reaction was analyzed by agarose gel electrophoresis. The specificity of the ppk primers was confirmed by cloning and sequencing selected PCR amplicons; these data will be described in more detail elsewhere (Peterson et al. in preparation).