Raymond and Alsop *Environ Syst Res (2015) 4:14*
DOI 10.1186/s40068-015-0038-x

Environmental Systems Research
a SpringerOpen Journal

RESEARCH ARTICLE

Open Access

Microbial evolution in extreme environments: microbial migration, genomic highways, and geochemical barriers in hydrothermal ecosystems

Jason Raymond* and Eric B Alsop

Abstract

Background: Recent advances in microbial ecology are providing unprecedented opportunities to test Baas Becking's oft-cited "everything is everywhere, environment selects" axiom. A number of recent studies have brought together genomic, ecological, and physico-chemical approaches that are, for the first time, beginning to test and quantify this axiom, providing fundamental shifts in our understanding of microbial ecology. Here we integrate environmental sequencing with biogeochemistry to interrogate patterns in abundance and community composition—as well as dispersal mechanisms and timing—that underlie microbial migration in natural ecosystems. Our analysis focuses on the presence of and similarities across high identity genomic DNA scaffolds and fragments, thousands of which are distributed across over two dozen communities sampled from hydrothermal ecosystems from Yellowstone National Park, Wyoming and Great Boiling Springs, Nevada.

Results: Despite their geographical isolation from one another and physico-chemical isolation from surrounding mesophilic environments, a large number (>43,000) of long, high identity DNA scaffolds were conserved across two or more hot springs communities. This widespread distribution of nearly identical DNA fragments suggests active mechanisms driving microbial migration and genomic information sharing. Genes encoded on these scaffolds encompass a broad spectrum of metabolic capabilities from diverse thermophilic taxa, but include revealing biases in the functions and taxonomic distribution of shared genes. Evolutionary rate analysis suggests that genomic migration and sharing is not only recent and ongoing, but that very different mechanisms are driving chemotrophic versus phototrophic community information exchange—mechanisms that include both biological and abiotic vectors and catastrophic events that have acted as evolutionary bottlenecks in particular on sunlight-driven photosynthetic communities.

Conclusions: The intersection of biology and environment is privy to an unprecedented level of interrogation as a result of advances in ecosystems biology, in particular through the integration of data from analysis across multiple scales and disciplines. Both the methodologies developed herein, and the findings our results support, help advance our understanding of microbial ecology and dispersal mechanisms in natural environments.

Keywords: Geobiology, Evolution, Microbial ecology, Biogeography, Extreme environments

Background

During the past decade, environmental sequencing has ushered in a golden era in microbial ecology, providing an unprecedented glimpse into the structure, function, and evolution of microbial communities. Earlier analyses were transformative but were necessarily piecemeal, focused largely on single genes and organisms, or occasionally microbial microcosms, as proxies for interrogating and understanding the complex, natural communities

*Correspondence: jason.raymond@asu.edu
School of Earth and Space Exploration, Arizona State University, Box 871404, Tempe, AZ 85254, USA

© 2015 Raymond and Alsop. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

that comprise the majority of Earth's biomass and evolutionary history. As the quality and availability of these environmental datasets continues to increase, so do opportunities for integration and meaningful comparisons between them to be made.

Hydrothermal systems are famous for their remarkable physico-chemical and concomitant biological variations: orders of magnitude changes in pH and elemental concentrations, steep temperature gradients, and dynamic temporal shifts support taxonomic and functional biodiversity that is unparalleled in most other environments on the planet (Barns et al. 1994; Shock et al. 2010). However, our and other studies reveal that these systems—many of which are separated by tens of kilometers and are geographically isolated—can display striking levels of genetic and genomic conservation (Ward et al. 1998; Reno et al. 2009; Miller-Coleman et al. 2012; Swingley et al. 2012; Inskeep 2013). While this hydrothermal diversity was anticipated even in early microbial biogeochemical studies (Brock 1967), what has been surprising is the remarkable propensity for microbes to distribute not just locally, but along global biogeographic patterns (Whitaker et al. 2003; Martiny et al. 2006; Whitaker and Banfield 2006). Once the domain of macroecologists, spatial scaling laws have recently been advanced for the microbial majority and are providing powerful, scale-invariant models of local-to-global distributions of key biodiversity metrics (Fenchel and Finlay 2004; Green and Bohannan 2006).

These advances owe much to recent progress in environmental sequencing of microbial communities. Seminal progress—and, arguably, the dawn of the current golden age of microbiology—came from PCR amplifying and sequencing the gene encoding the small ribosomal RNA subunit from complex, natural microbial communities, most members of which were recalcitrant to cultivation-dependent microbiological approaches (Olsen et al. 1994). More recently, microbial metagenome and metatranscriptome sequencing has unveiled an almost unmanageable abundance of information on the distribution and evolution of genes, species, and communities from all manner of hosts and environments (Handelsman 2004; Gilbert and Dupont 2011). These transformative studies have reshaped our understanding of how microbes and their environments interact and coevolve, yet very few studies have yet to investigate how these interactions change dynamically over Earth's history or across its biogeochemically diverse ecosystems. The handful of studies so far conducted suggest that integrating biogeography and evolutionary modeling into the wealth of metagenome analyses presently available is poised to provide enormous insights into how the Earth system evolves over time (Whitaker et al. 2003; Martiny et al. 2006; Whitaker and Banfield 2006; Hanson et al. 2012).

Here we make use of recently available metagenomic datasets from studies of over two-dozen terrestrial hydrothermal systems in Yellowstone National Park (YNP) and Great Boiling Springs, Nevada, USA (GBS) (Fig. 1) to understand how patterns of gene and genome flow have shaped community architecture. Our findings incorporate evolutionary rate data from thousands of genes in each community. Taken together, our results suggest that the prokaryotic communities inhabiting YNP+GBS hydrothermal systems show striking and unexpected levels of genomic similarity across their genomes.

Results

Gene, genome, and microbial migration: biological information flow in natural systems

Previous analyses have identified distinct thermophilic community compositions across YNP hot springs (Blank et al. 2002; Meyer-Dombard et al. 2005; Takacs-Vesbach et al. 2008; Inskeep et al. 2010; Swingley et al. 2012; Inskeep 2013), showing that the architecture of these communities—both their overall function and taxonomic make-up—is strongly dependent on physical and geochemical conditions (though with notable deviations, even in springs with nearly identical conditions, e.g. (Skirnisdottir et al. 2000; Reysenbach and Shock 2002; Reno et al. 2009; Inskeep et al. 2010; Meyer-Dombard et al. 2011; Cox et al. 2011)). These studies have begun to identify molecular geobiological interactions: clear cases where a microbial or community phenotype—the ensemble of genetically encoded traits as determined by DNA sequencing—is directly coupled to one or more environmental features. Because of this close coupling between community function and the physical chemistry of hydrothermal systems, they provide unparalleled opportunities for interrogating how environment shapes and constrains biology and evolution.

While recent studies are beginning to decipher microbe-environment interactions in hydrothermal systems, what remains to be resolved are the underlying evolutionary patterns that explain how these interactions came to be. Are hydrothermal communities genetically unique, suggesting isolated populations that each independently adapt to a particular environmental niche? Or is their 'cross-pollination'—transfer of biological information across hydrothermal ecosystems through gene and genome migration—so that genetic and genomic similarities in microbial communities correlate with physico-chemical similarities in their environments? Differentiating between these possibilities is of key importance in understanding how microbes evolve and adapt to diverse environmental conditions, as well

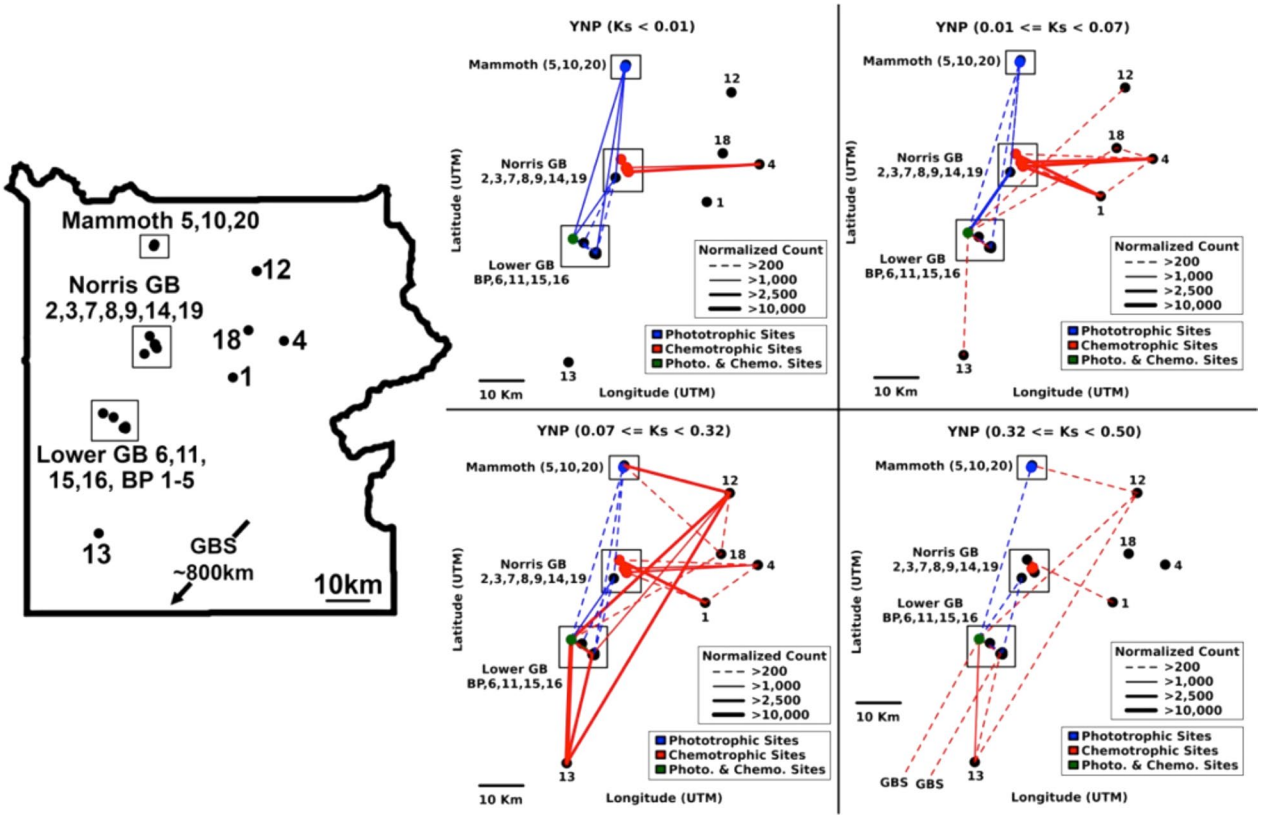

Fig. 1 *Left* map of sample sites (YNP1-2 and BP1-5) within YNP and approximate distance and direction of GBS sites and *right* number of long, high %id scaffolds shared between pairs of sites. Colors are based on types of sites that scaffolds are shared between (for instance, *blue lines* indicate transfers between two phototrophic communities) and *line thickness* is proportional to the number of shared scaffolds. The *four panels* show different ranges of synonymous substitution rates (Ks), a proxy for how conserved shared scaffolds are and how recently scaffold sharing most likely occurred (from most recent, Ks < 0.01 at *top left*, to most distant, Ks 0.32–0.50, at *bottom right*).

as how environmental constraints influence the patterns and rates of biological information flow in natural ecosystems.

To better understand how environment shapes and constrains community architecture and to what extent genes, species, and populations migrate across hydrothermal systems, we began by identifying highly conserved genomic DNA scaffolds shared across twenty five metagenomes from YNP and GBS (Additional file 1: Table S1) (Costa et al. 2009; Swingley et al. 2012; Inskeep 2013). Of 659,351 scaffolds that had some degree of homology between two or more metagenomes, 43,532 long (>1,000 bp) and high identity (>90% identity) DNA scaffolds were identified in at least two, and up to seven, distinct hot springs sampled from GBS and across ~100 km of YNP. These highly conserved scaffolds are a small subset of the total scaffolds with homologs in two or more YNP metagenomes (~6.6% of 659,351 scaffolds, with the remainder of scaffolds having only 30–50% DNA identity). That these scaffolds are retaining high identities despite substantial geographic—and in some cases geochemical—barriers between these ecosystems suggests a mechanism or mechanisms for preserving evolutionary relatedness. Here we integrate a number of approaches to identify plausible underlying mechanisms that could facilitate migration and sharing of genomic DNA across geographically isolated springs (Fig. 1), thereby sustaining highways of information exchange across YNP ecosystems.

Taxonomic, geochemical, and functional constraints on microbial migration

We used comparative metagenomic and rRNA analyses to identify biases in the taxonomy, function, and biogeochemical environments that correlated most strongly with the number of shared scaffolds between different metagenomes. A large number of shared, highly conserved scaffolds encode multiple genes whose taxonomic identity converges on a single genus or family (Fig. 2). This suggests that detected scaffolds are fragments of contiguous genomes that are dispersed by microbial migration: closely related species being sampled in two—or in some

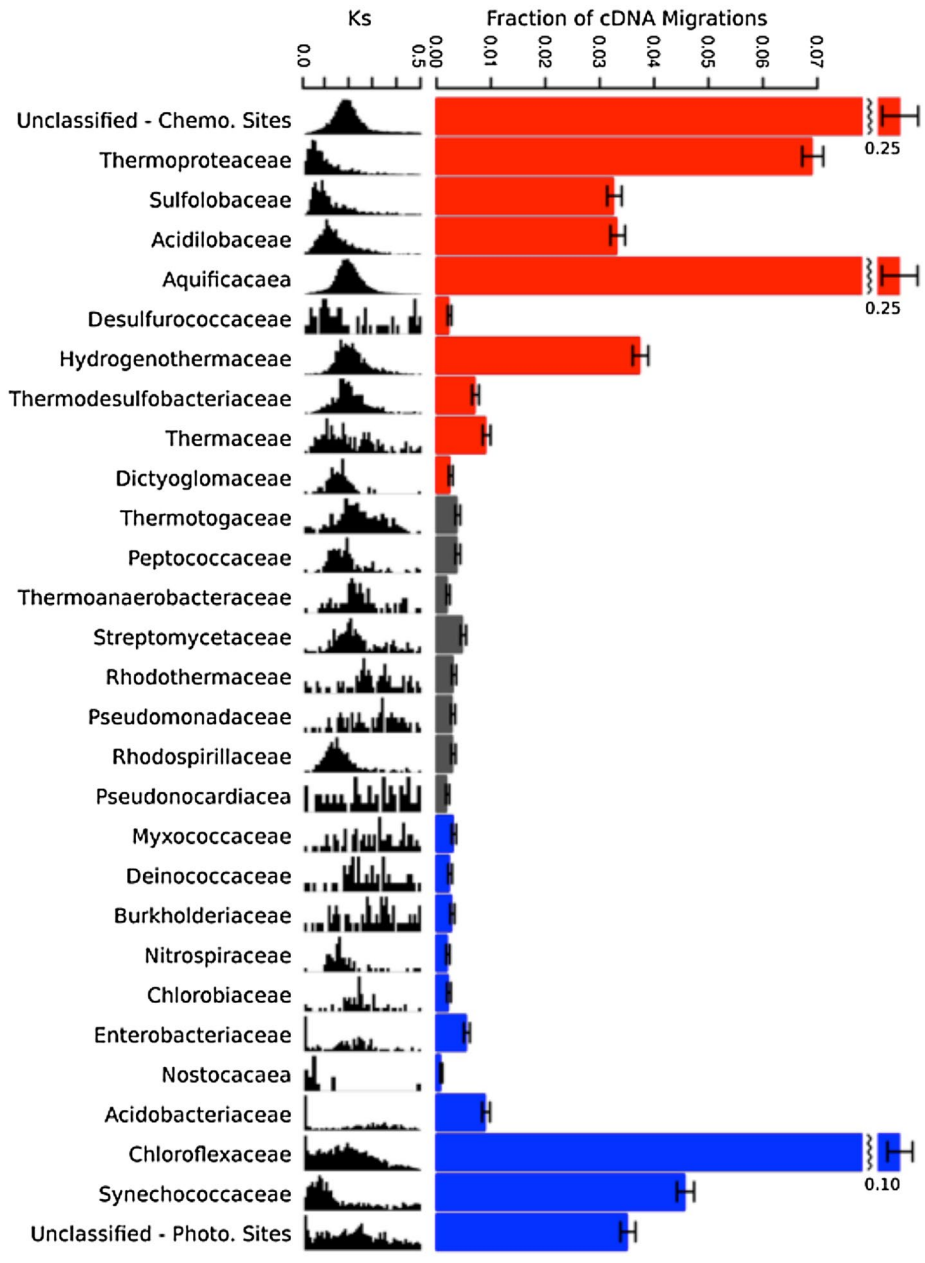

Fig. 2 Fraction of total cDNAs shared across two or more YNP/GBS communities, by most likely taxonomic family (or 'unclassified', in cases of ambiguous taxonomy). *Bar color* and *top-to-bottom* ordering is based on families being present predominantly (>80%) in chemotrophic (*red*) or phototrophic (*blue*) communities (or *grey* for families present in both). Histograms give the synonymous substitution rate (Ks) distributions of cDNAs for that family ("noisy" distributions correspond to families with fewer cDNA sequences available for calculations). These data support that trends in Ks distributions (Fig. 3)—in particular the chemotroph peak at Ks ~0.18 and phototroph peak near Ks ~0—extend across multiple families and are not the result of a single, abundant family dominating the distribution.

cases across many—hot springs ecosystems. A subset of scaffolds could represent de facto horizontal gene transfer between distinct species, though rigorous identification of such cases ideally requires longer sequence data (from high coverage metagenomes) showing DNA from one species present in the genome of another. However, and more importantly, both mechanisms—horizontal transfer or microbial migration—underscore that genetic/genomic mobility and habitat ranges of microbial species are substantially greater than previously known.

To identify specific environmental factors that influence the probability of finding long, high identity

metagenome fragments in two or more communities, we performed iterative multiple linear regression (iMLR) on 20 available physical and geochemical measurements from each YNP metagenome (Additional file 1: Table S2). iMLR extends classical multiple linear regression by permuting the predictor variables being fit to a response variable (or, more strictly speaking, assigning predictor coefficients/weights of 0), either exhaustively or through some manner of subsampling. The technique is useful as a method of dimensionality reduction as an alternative to e.g. principal components or factor analysis, especially when the relationship between predictor and response variables is initially unconstrained but may be 0 (for instance abiotic factors, such as the solution concentration of aluminum, that were measured but have no known mechanism of interaction with life). iMLR has the advantage of explicitly removing "poor" predictor variables from a model, whereas other dimensionality reduction methods leave them as part of the model, either minimized or combined with other (e.g. latent) variables—but nonetheless increasing the variance of the model fit.

iMLR identified eight environmental parameters that best accounted for variability in the number of DNA fragments shared between communities ($p <= 0.01$), including pH, temperature, and concentrations of NO_3, SO_4, $O_{2[aq]}$, and K^+—all of which are known to have roles in biological pathways. Of the eight, only pH and nitrate and potassium concentrations showed individually significant correlations at a $p < 0.05$, underscoring the interdependence of biological and environmental factors (Additional file 1: Fig. S5A) and the importance of multivariate analyses for integrating both (Alsop et al. 2014). Notably, proximity did not significantly correlate ($p \sim 0.38$) with number of observed shared scaffolds; high identity DNA scaffolds are found even across great distances within the YNP+GBS systems (Additonal file 1: Figs. S1–S4). While some co-local communities do show increases in the shared DNA fragments (discussed below), our analysis suggests that the benefit of proximity drops off sharply as a result of the steep geochemical gradients present in hot springs. Indeed, "environment selects" what genes and pathways are most frequently shared between hydrothermal communities.

Extending on this idea of environmental selection, functional assignments of genes encoded by scaffolds shared across two or more communities yielded 16,056 scaffolds with Enzyme Commission (EC) assignable functions. These include functions that were both overrepresented and underrepresented compared to their overall distribution in YNP/GBS metagenomes (Additional file 1: Table S3), further suggesting environmental selection on the types of functions being shared. Among these, carboxylases and oxidases were notably overrepresented, consistent with communities dependent on autotrophic pathways in environments known to be highly oxidizing (Rothschild et al. 2002). Biotin carboxylase in particular was encoded by 930 scaffolds involved in these transfers; this enzyme has recently been discovered to be involved in CO_2 fixation in thermophiles, and suggests that high temperature autotrophy—above the temperature limit of photosynthesis and the Calvin Cycle—may be an important enzyme for primary production in hydrothermal systems (Menendez et al. 1999; Auguet et al. 2008).

Mutation rates suggest rapid migration of biological information across YNP

Next we investigated whether rates of evolution within these 24 communities, and in the subset of DNA fragments most frequently shared between them, could provide insight into the relative timescales or mechanisms of genomic migration across YNP/GBS ecosystems. Importantly, proper calibration and temporal constraint of microbial evolutionary rates requires experimental measurement in specific strains and species, as rates have been shown to vary across even closely related organisms. Here, we instead focus on conservation of neutral mutation rates between major identifiable clades in these hydrothermal ecosystems. While differences in neutral mutation rates may arise due to a variety of mechanisms (e.g. variations in population sizes, reproductive success/adaptive fitness to a specific niche, or differences in DNA repair fidelity) (Ochman et al. 1999; Kuo and Ochman 2009; Barrick et al. 2009; Hanson et al. 2012), observed similarities in neutral mutation rates are most consistent with a single, null hypothesis: isolation of and divergence between two microbial communities (Hanson et al. 2012). Agreement of neutral mutation rates across many genes common to two communities (as opposed to e.g. just 16S rRNA) as well as in multiple species within each community further supports and points to the isolation and subsequent divergence of those communities as the most likely cause of observed genomic differences—even if the absolute timing of that divergence can not be measured due to absence of calibration points.

To identify and measure neutral mutation rates across these hydrothermal communities, we examined DNA substitution rates in the protein coding portions of the best-conserved scaffolds present in two or more communities. Figure 3 shows synonymous substitution rates (Ks)—which, for well-conserved genes, can be equated with background DNA mutation rates—for 30,668 cDNA sequences found on scaffolds in two or more communities. Figure 3 separates Ks distributions for the dominant types of metabolism in YNP + GBS communities: phototroph-dominated communities, occurring below a

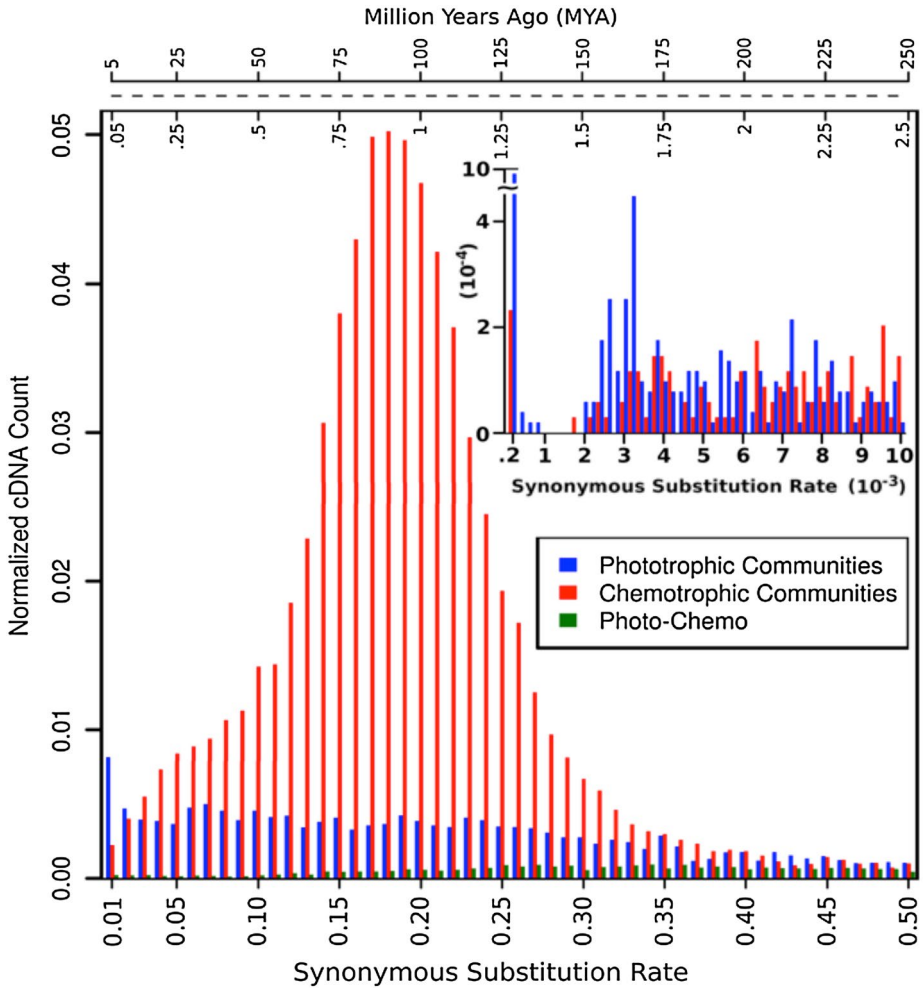

Fig. 3 Distribution of synonymous substitution rates (Ks) for all cDNAs transferred between chemotrophic (*red*), between phototrophic (*blue*), and from chemotrophic-to-phototrophic (*green*) communities. The inset expands the lowest Ks (<0.01) bin, illustrating the low Ks peak (Ks ~0 to 0.2) in chemotrophic and especially phototrophic communities. The top axis uses recently published prokaryotic mutation rates to bracket likely upper (*top*) and lower (*bottom*) limits on cDNA transfer times.

temperature of ~73°C, and higher temperature chemotrophic communities. While intraspecies background mutation rates are indeed similar between community members across this well-conserved subsampling of protein-coding genes, notably, significant differences in Ks distributions and statistics were observed between phototroph- versus chemotroph-dominated communities (Fig. 3). Chemotrophic communities show a distinct Poisson distribution centered at a Ks of 0.18, whereas phototroph communities exhibit non-Poisson distributions.

Importantly, taxonomic assignment of scaffolds shows that Ks distributions are relatively constant across prokaryotic phyla in each metabolic category (Fig. 2). This important result supports the null hypothesis that Ks distributions reflect isolation and subsequent diversification of these hydrothermal communities. Moreover, there is no clear evidence for a single, abundant taxon dominating Ks distributions after correcting for the overall abundance of each taxa in metagenomic data, as might be the case in communities dominated by one or a few species (i.e. the probability of observing shared scaffolds increases proportionately with taxa abundance).

Recent models shed light on these distributions of background mutation rates. For instance, the chemotroph Poisson distribution is consistent with models of community evolution where mutations occur at a relatively constant background frequency but fixation occurs at different rates across genomes (Aris-Brosou and Excoffier 1996; Patwa and Wahl 2008; Wielgoss et al. 2013). The chemotroph Poisson distribution (and Ks

~0.18 peak) represents the average evolutionary distance across sampled chemotroph communities and points to a strong bottleneck that, in effect, "reset the clock" on the evolutionary divergence between chemotrophs in Yellowstone hydrothermal communities. Using a spectrum of recent measurements of microbial background mutation rates, it becomes possible to at least loosely estimate the average elapsed time since isolation between YNP chemotrophic communities (discussed below) as a way to speculate on plausible bottleneck events. Also worth highlighting is the unexpected increase in the lowest Ks bin (Fig. 1b), suggesting a subset of genes that have been *very recently* shared between YNP chemotroph communities.

This distinct Poisson peak is absent in phototrophic communities, which show a polyphasic distribution with an abrupt peak/mode at the lowest Ks bin and a 'sawtooth' tail extending to higher Ks values (Fig. 3, blue). These fundamental differences suggest very different mechanisms and/or bottlenecks are driving dispersal and evolution in chemotrophic vs. phototrophic communities. The low Ks peak observed among chemotroph communities is, however, present in the phototrophic communities, pointing to a subset of nearly identical genes and scaffolds shared across these communities. This nearly identical subset appears to be moving against the tide of evolution, and sustaining it requires a mechanism whereby genes can be redistributed between geographically isolated communities (thereby 'resetting' the evolutionary distance between some subset of genes in those communities) (Martiny et al. 2006).

The observed Ks peaks correspond to the amount of time that, on average, will elapse between microbial migration events across YNP communities (keeping in mind that long, high %ID scaffolds represent ~6.6% of total scaffolds and, therefore, migration events are still minor contributors to overall community differences). Because studies of prokaryotic mutation rates are still limited in taxonomic scope, we scoured recent literature to apply a range of plausible rates (1×10^{-7} to 1×10^{-9} substitutions per synonymous site per year) (Drake 1991; Ochman et al. 1999; Kuo and Ochman 2009; Wielgoss et al. 2013). While frustrating in that these calibrations extend across two orders of magnitude of plausible microbial evolutionary rates, it is worth noting that their estimation of community divergence times—250,000 to 25 million years for phototrophic communities and 1.25 million to 125 million years for chemotrophic communities (Fig. 3, top)—spans a well-known range of YNP catastrophic events that may indeed constitute strong evolutionary bottlenecks that, as discussed below, would present quite differently to phototrophic versus chemotrophic communities.

Discussion

The Yellowstone caldera has been a hotbed of geological activity throughout its history. Three major eruptions have occurred with 600,000–800,000 year periodicity, culminating most recently in the Lone Creek eruption cf. 640,000 years ago (Lanphere et al. 2002). The region has experienced major episodic glaciations, most recently the Pinedale/Great Pleistocene glaciation, thought to have covered >90% of the YNP surface between 15,000 and 20,000 years ago. These evolutionary bottlenecks carry clear but distinct implications for microbial life at the YNP surface (phototrophic and chemotrophic) versus those in the subsurface (exclusively chemotrophic). Glaciations and supereruptions would have been cataclysmic for photosynthetic life. If not entirely sterilized, photosynthetic habitability and primary productivity would have been dramatically abated, with subsequent recolonization giving the appearance of genetically 'young' communities, as observed here in metagenome comparisons and supported by diffusion models (below). Conversely, subsurface hyperthermophiles would be the one group of microbes that might survive such a catastrophic incursion, supporting a genomic record harboring much older evolutionary divergence.

To test whether the observed patterns are compatible with ecosystem mutation/migration/extinction models, we developed jump diffusion simulations to integrate Fickian diffusion of mutating species across 2D and 3D lattices. These simulations track populations comprised of individual cells as they migrate and mutate across lattices while varying three key parameters: the probability, distance, and dimensionality (2D/3D) of allowed jumps between lattice points. Out to ~10^5 simulations (Fig. 4), the Poisson curve of chemotrophs appears to most clearly be correlated with models where jumps between lattice points occur with high probability but are restricted to adjacent jump: migration must occur between adjacent communities. Conversely, the polyphasic distributions only emerge at higher jump distances, where skipping multiple lattice points becomes the norm. The polyphasic distribution is, in effect, the summation of multiple Poisson curves spread out across lattice space, and sampling across this space would result in 'sawtooth' Ks distributions analogous to that observed for YNP/GBS phototrophs.

Notably, YNP/GBS photosynthetic mats are teeming with life. Their relatively low temperatures make them suitable for grazing by insects and arachnids and intrusion by mammals, all of which are potential vectors for transporting mat floccules between proximal hot springs. The idea that eukaryotes simply don't survive above ~56°C is a misnomer; transient interactions between macroscopic life and high temperature hot springs are easily observed in the field (perhaps most identifiably

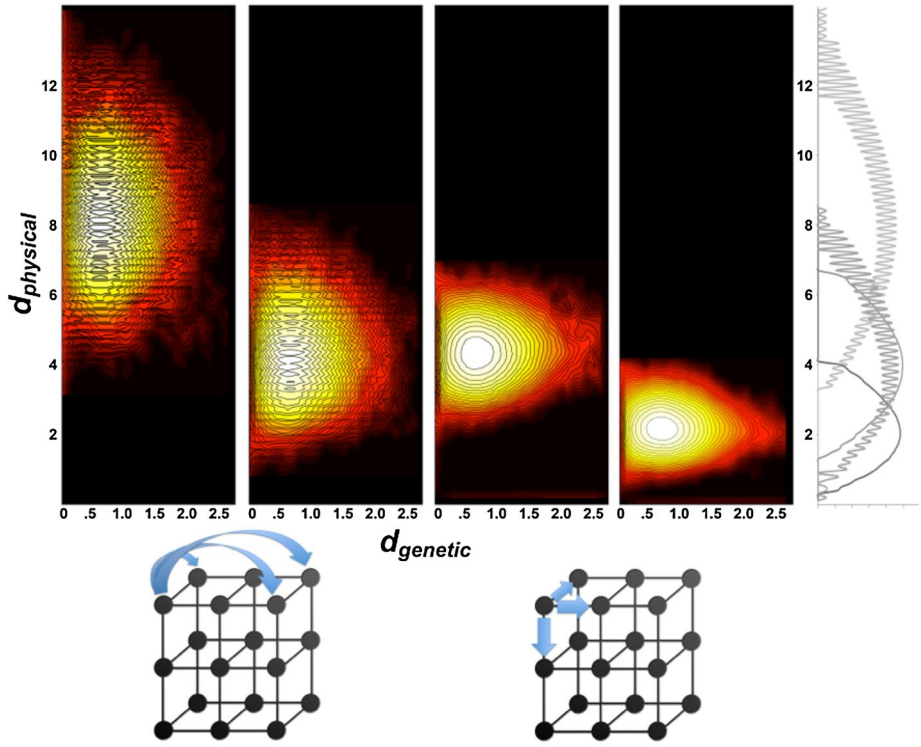

Fig. 4 Four exemplary jump diffusion models of microbial migration vs. mutation across a 3D lattice. The two *leftmost panels* show the polyphasic distributions resulting from allowing longer distance jumps ($d_{jump} > 1$, here $d_{jump} = 2$), while the two right panels show Poisson distributions that emerge from restricting jumps to adjacent lattice cells ($d_{jump} = 1$). $d_{physical}$ and $d_{genetic}$ correspond, respectively, to lattice and genetic Euclidean distances, plotted here as histogram heat maps for all pairs of cells in each population after 100 time steps (2^{100} cells). Panels 1 vs. 2 and 3 vs. 4 illustrate the (linear) effect of halving the jump probability, which effectively cuts the average physical distance between populations in half. A clocklike assumption of mutation rates and identical number of time steps for each simulation results in essentially identical $d_{genetic}$ (*x axis*) distributions. The combined graph at right shows the $d_{physical}$ sum for each of the *four panels*.

among insects and vertebrates including *Home sapiens*). Chemotrophic communities thrive at temperatures that do rapidly dissuade or sterilize these macroscopic vectors, and their propagation is likely more dependent on non-biological (e.g. Aeolian, climatic, groundwater) mechanisms for surface and subsurface transport and migration. These observations are consistent both with the decreased frequency of low Ks values among chemotroph communities (Fig. 3) as well as with the potential for vectors to carry mat floccules across long distances. The additional requirement that communities are transported between physico-chemically compatible environments is exactly consistent with the low probability/long distance requirement of jump diffusion models.

While phototroph communities are constrained by their metabolic lifestyle to surface migrations, chemotroph communities can thrive independent of sunlight and, theoretically, access much of the YNP subsurface. The YNP subsurface provides contiguous, anastomosing hydrothermal channels that may be enabling chemotroph migration: their contiguity amounts to higher frequency, shorter jump length migrations consistent with jump diffusion models responsible for observed Poisson distributions. While it remains to be directly sampled e.g. through drill cores, the notion of a substantial chemotrophic biosphere is supported by YNP hydrogeology, which suggests a prolific network of subsurface hydrothermal channels (White et al. 1975; Fournier 1989; Chapelle et al. 2002). Though water within hydrothermal chimneys is at temperatures well above the limit for life (Fournier 1989; Fournier et al. 2014), hydrothermal gradients radiating through pore spaces around these channels establish a range of temperatures and physico-chemical conditions compatible with thermophilic communities (Fig. 5), entirely independent of photosynthesis (Gold 1992; Fisk et al. 1998). Furthermore, these hydrothermal networks may extend over large extents of YNP, establishing interconnected subsurface 'highways' for microbial migration and genomic information exchange extending through shallow, permeable glacial sediments and possibly into deeper active faults extending 4–6 km below the YNP surface (White et al. 1975; Fournier 1989) (Fig. 5).

Conclusions

While it is well known that environment has shaped the evolutionary trajectory of microbial life, in few places is it as clear as in the greater Yellowstone hydrothermal systems where microorganisms are directly tied to environmental energy and nutrient sources. By integrating data on the distribution of metagenomes sampled from a broad range of hydrothermal settings with genetic analysis of the most highly conserved genes in those systems, we have begun to recustruct a history of the co-evolution of life and environment in the greater YNP system.

Taken together, our data show that thermophilic communities from these remarkably diverse YNP + GBS hydrothermal settings are far from being biogeographically isolated islands: the extant microbial communities are the result of millions of years of shuttling and recombination between hydrothermal ecosystems by both biological and environmental/climatic vectors. These vectors, combined with several catastrophic mass extinctions driven by YNP eruptions, have left distinct signatures in the rates of evolution of microbial communities, most notably in phototrophic communities bound to sun lit surface environments versus chemotrophs, whose utilization of geochemical energy sources apparently expanded their habitable zones well into the YNP subsurface. Simulations suggest the evolutionary rate signatures are a combination of surface transport mechanisms, which are occurring with irregular periodicity but over longer distances across YNP, and on a relatively slow but persistent hydrothermal "conveyer" that would drive dispersal of chemotrophic communities through a YNP subsurface biosphere. Our results have begun to map distinct highways for microbial migration and genomic exchange across hydrothermal systems, and beckon for much more extensive sampling of the YNP biosphere, particularly what may be a prolific subsurface reservoir of chemotrophic life.

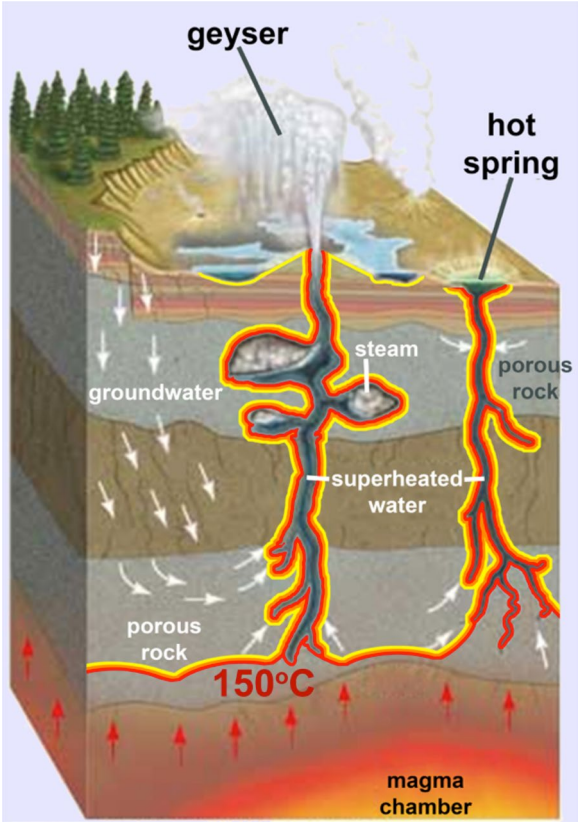

Fig. 5 Proposed contiguous isotherms compatible with chemotrophic life in the YNP subsurface. These anastomosing channels are established by thermal radiation from superheated water as well as deeper magmatic interactions [due to the latter, water temperature exceeds 150°C at depth in the YNP caldera subsurface (White et al. 1975; Fournier 1989)]. *Colors* illustrate regions of higher to lower temperature (*red*→*orange*→*yellow*). Whereas surface migration is restricted to 'jumps' between adjacent hot springs, subsurface migration can occur along highways established by these isotherms. Adapted from Encyclopedia Britannica (2006).

Methods

DNA scaffolds for the 24 metagenomes analyzed were downloaded from the Joint Genome Institute IMG/M web server (Grigoriev et al. 2011). All metagenomic scaffolds were combined into a single. FASTA file and an all-versus-all BLASTN (Altschul et al. 1990) was performed using this file as input. BLASTN output file was parsed to find all instances of scaffold overlap >1,000 bp in length with >90%ID between scaffolds from different metagenomes. A tally of overlaps (migrations) was determined between each sampled location and counts were normalized based on metagenome sizes. Normalization factors were determined by dividing the average sizes of the metagenomes involved by the average size of all metagenomes, or by dividing the number of scaffolds involved in migration by the total number of scaffolds obtained from a specific metagenome. Migration counts between sites were then normalized using this factor.

Iterative multiple linear regression was performed using the stepAIC() subroutine from R, in both the forward and reverse directions over 1,000 steps. Initial multiple linear regression was performed using R's lm() command to generate an equation such that the number of DNA migrations between any two communities was set to equal the difference between all geochemical and physical parameters (DNA migrations $= m_1\Delta pH + m_2\Delta Temperature + m_3Km + m_4\Delta[Al] + \ldots + b$) across communities. On ensuing iterations, model variances from observed data were progressively minimized through coefficient optimization

across all possible pairs of communities. R's stepAIC() command was then invoked on the optimized multiple linear regression equation to systematically add and/or remove variables from the overall MLR equation with the optimality criterion of maximizing goodness of fit of the model predictions vs. observed values of DNA migrations and physico-chemical measurements.

For Ka/Ks estimation across shared scaffolds, all protein encoding (cDNA) sequences encoded by scaffolds from the 24 metagenomes, as well as their transcribed amino acid sequences, were downloaded from JGI's IMG/m server. cDNA sequences were combined into a single. FASTA file and an all-verses-all NCBI BLASTN was performed using the combined file as input. As with the "all scaffolds" approach above, the BLASTN output was parsed to find all instances of overlap >1,000 bp in length with >90% identity between cDNA sequences from different metagenomes. From this subset, cDNA sequences and translated amino acid sequences were aligned and subsequently Perl scripts were used to downselect sequences that were predicted to be fully in frame and beginning with a start codon. Ks values between overlapping cDNA sequences were calculated using the Ka/Ks calculator (Zhang et al. 2006) software package, using both aligned cDNA sequences and their aligned, translated amino acid sequences as input. The Ka/Ks software package was run using the model averaging (MA) settings and correcting for multiple substitutions. MA extends the well known Goldman–Yang model, accounting for evolutionary features including transition-transversion ratios and nucleotide frequencies, and applies them to codon-based substitution frequencies. The Ka/Ks software package outputs a text file containing the Ks values between all pairs of high identity cDNA sequences.

Jump diffusion models tracked cellular mutation and migration across 100 timesteps, with each cell doubling at each timestep (1 initial cell, 2^{100} total cells at termination). Each cell was assigned a 20 position, 10 bit genetic code allowed to mutate with probabilities randomly sampled from a range of 0→1.0 mutation per site per generation. Each cell was allowed to migrate with a likelihood of migrating N steps randomly sampled from an equiprobable range of 0–0.5. The number of lattice migration steps N was set at time 0 in each simulation at an integral value of either 0, 1, 2, or 3. The dimensionality of migration was set at time 0 to either allow migrations in all three dimensions, or restrict migrations to 2D steps across the lattice surface. At the termination of each simulation, all cells were compared pairwise to calculate Euclidean distances between their 20-mer 'genomes', as well as between their respective [*xyz*] positions in the lattice, and the resultant histogram showing the distribution of these distances is plotted in Fig. 4, averaging distances across 10^5 independent simulations at four different values of model parameters.

Ks histograms were generated by binning Ks values into 0.01 Ks unit bins, separating sites based on community metabolic type that migrations occurred between (between chemotrophic, between phototrophic, between phototrophic and chemotrophic). The 0–0.01 Ks histogram (Fig. 3—inset) was generated by binning all Ks values <0.01 into 0.002 Ks unit bins, also separated by the type of community the migrations occurred between. Ks counts were normalized by dividing observed counts by the total number of >1,000 bp cDNA sequences with each corresponding metagenome. Chemotrophic and phototrophic communities were normalized separately so that community types could be compared based on the fraction of cDNA migrations at a given Ks value (bin).

cDNAs from the 24 metagenomes were compared to the KEGG enzyme database (Kanehisa and Goto 2000) using NCBI BLASTX, with best BLAST e-values (1^{-20} or better) used for EC assignments. EC counts within each metagenome and within the shared scaffold subset determined and compared using Pearson's Chi squared test, implemented in R's chisq.test(). Expected EC counts used in Chi squared were based on overall EC frequencies within each metagenome, and compared with observed EC assignments for shared ECs with a p value threshold of 0.05.

cDNAs from the 24 metagenomes were compared to the NCBI nt database using NCBI BLASTN, with taxonomic assignment based on best hit BLAST e-values better than 1^{-20}. The NCBI taxonomy spreadsheet was then used to determine the Linnean hierarchy (e.g. genus, family) of all taxonomically assigned cDNA sequences. Data presented are at the family level for robustness of assignment, as determined by dividing the count of cDNA sequences assigned to a family by the total count of cDNA sequences in each metagenome, as well as in shared scaffold subsets. Ks histograms for each family were determined by dividing all cDNA migrations assigned to a taxonomic family into 0.01 Ks unit bins from 0.01 to 0.50.

Additional file

Additional file 1. Figures S1–S4: Number of cDNA migrations (line thickness) detected across four ranges of Ks values (increasing in each panel from top left to lower right) among both phototroph and chemotroph communities (blue and red lines, respectively). **S1** shows cDNA migrations between Lower Geyser Basin and other hot springs, **S2** between Bison Pool and other hot springs, **S3** between Norris Geyser Basin and other hot springs, and **S4** between Mammoth and other hot springs. **Figure S5A:** PCA of measured geochemical parameters vs. detected cDNA migrations and **Figure S5B:** cDNA migrations mapped in the PCA space determined in S5A. **Table S1:** YNP/GBS physical and geochemical metadata. **Table S2:** Multiple regression analysis results. **Table S3:** Functions of over- and under-represented genes among detected cDNA migrations.

Authors' contributions
JR and EA conceived of, designed, and carried out the methodology and data analysis, and co-wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
We thank Matt Kellom, Everett Shock, Jack Farmer, and Ariel Anbar for extensive discussion and helpful feedback in carrying out this research. This work was funded by a NASA Astrobiology Institute (NAI) "Follow the Elements" Grant (JR) and Grant NNX08AP61G (JR) from the NASA Exobiology and Evolutionary Biology program.

Compliance with ethical guidelines

Competing interests
The authors declare that they have no competing interests.

Received: 30 June 2015 Accepted: 30 July 2015
Published online: 27 August 2015

References

Alsop EB, Boyd ES, Raymond J (2014) Merging metagenomics and geochemistry reveals environmental controls on biological diversity and evolution. BMC Ecol 14:16. doi:10.1186/1472-6785-14-16
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2
Aris-Brosou S, Excoffier L (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol Biol Evol 13:494–504
Auguet J-C, Borrego CM, Bañeras L, Casamayor EO (2008) Fingerprinting the genetic diversity of the biotin carboxylase gene (accC) in aquatic ecosystems as a potential marker for studies of carbon dioxide assimilation in the dark. Environ Microbiol 10:2527–2536. doi:10.1111/j.1462-2920.2008.01677.x
Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci 91:1609–1613. doi:10.1073/pnas.91.5.1609
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. doi:10.1038/nature08480
Blank CE, Cady SL, Pace NR (2002) Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. Appl Environ Microbiol 68:5123–5135. doi:10.1128/AEM.68.10.5123-5135.2002
Brock TD (1967) Life at high temperatures. Science 158:1012–1019. doi:10.1126/science.158.3804.1012
Chapelle FH, O'Neill K, Bradley PM, Methé BA, Ciufo SA, Knobel LL et al (2002) A hydrogen-based subsurface microbial community dominated by methanogens. Nature 415:312–315. doi:10.1038/415312a
Costa KC, Navarro JB, Shock EL, Zhang CL, Soukup D, Hedlund BP (2009) Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. Extrem Life Extreme Cond 13:447–459. doi:10.1007/s00792-009-0230-x
Cox A, Shock EL, Havig JR (2011) The transition to microbial photosynthesis in hot spring ecosystems. Chem Geol 280:344–351. doi:10.1016/j.chemgeo.2010.11.022
Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci 88:7160–7164. doi:10.1073/pnas.88.16.7160
Fenchel T, Finlay BJ (2004) The ubiquity of small species: patterns of local and global diversity. Bioscience 54:777–784. doi:10.1641/0006-3568(2004)054[0777:TUOSSP]2.0.CO;2
Fisk MR, Giovannoni SJ, Thorseth IH (1998) Alteration of oceanic volcanic glass: textural evidence of microbial activity. Science 281:978–980. doi:10.1126/science.281.5379.978
Fournier RO, Heasler HP, Hinckley B, Ingebritsen SE, Lowenstern JB, Susong DD (2014) Hydrogeology of the old faithful area, Yellowstone National Park, Wyoming, and its relevance to natural resources and infrastructure (No. 2014-1058). US Geological Survey
Fournier RO (1989) Geochemistry and dynamics of the yellowstone national park hydrothermal system. Annu Rev Earth Planet Sci 17:13–53. doi:10.1146/annurev.ea.17.050189.000305
Gilbert JA, Dupont CL (2011) Microbial metagenomics: beyond the genome. Annu Rev Mar Sci 3:347–371. doi:10.1146/annurev-marine-120709-142811
Gold T (1992) The deep, hot biosphere. Proc Natl Acad Sci 89:6045–6049
Green J, Bohannan BJM (2006) Spatial scaling of microbial biodiversity. Trends Ecol Evol 21:501–507. doi:10.1016/j.tree.2006.06.012
Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D et al (2011) The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. doi:10.1093/nar/gkr947
Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685. doi:10.1128/MMBR.68.4.669-685.2004
Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10:497–506. doi:10.1038/nrmicro2795
Inskeep WP (2013) The YNP metagenome project: environmental parameters responsible for microbial distribution in the Yellowstone geothermal ecosystem. Front Microb Physiol Metab 4:67. doi:10.3389/fmicb.2013.00067
Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH et al (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. PLoS One 5:e9773. doi:10.1371/journal.pone.0009773
Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. doi:10.1093/nar/28.1.27
Kuo C-H, Ochman H (2009) Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol Direct 4:35. doi:10.1186/1745-6150-4-35
Lanphere MA, Champion DE, Christiansen RL et al (2002) Revised ages for tuffs of the Yellowstone Plateau volcanic field: assignment of the Huckleberry Ridge Tuff to a new geomagnetic polarity event. Geol Soc Am Bull 114:559–568. doi:10.1130/0016-7606(2002)114<0559:RAFTOT>2.0.CO;2
Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL et al (2006) Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4:102–112. doi:10.1038/nrmicro1341
Menendez C, Bauer Z, Huber H, Gad'on N, Stetter KO, Fuchs G (1999) Presence of acetyl coenzyme A (CoA) carboxylase and propionyl-CoA carboxylase in autotrophic Crenarchaeota and indication for operation of a 3-hydroxypropionate cycle in autotrophic carbon fixation. J Bacteriol 181:1088–1098
Meyer-Dombard DR, Shock EL, Amend JP (2005) Archaeal and bacterial communities in geochemically diverse hot springs of Yellowstone National Park, USA. Geobiology 3:211–227. doi:10.1111/j.1472-4669.2005.00052.x
Meyer-Dombard DR, Swingley W, Raymond J et al (2011) Hydrothermal ecotones and streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. Environ Microbiol 13:2216–2231. doi:10.1111/j.1462-2920.2011.02476.x
Miller-Coleman RL, Dodsworth JA, Ross CA et al (2012) Korarchaeota diversity, biogeography, and abundance in Yellowstone and Great Basin hot springs and ecological Niche modeling based on machine learning. PLoS One 7:e35964. doi:10.1371/journal.pone.0035964
Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci 96:12638–12643. doi:10.1073/pnas.96.22.12638
Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6
Patwa Z, Wahl LM (2008) The fixation probability of beneficial mutations. J R Soc Interface 5:1279–1289. doi:10.1098/rsif.2008.0248
Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the Sulfolobus islandicus pan-genome. Proc Natl Acad Sci 106:8605–8610. doi:10.1073/pnas.0808945106
Reysenbach A-L, Shock E (2002) Merging genomes with geochemistry in hydrothermal ecosystems. Science 296:1077–1082. doi:10.1126/science.1072483
Rothschild LJ, Purcell D, Rogoff D, Wilson C, Brass JA (2002) The indirect effect of UV: some good news for microbes? In: Annual Meeting for the International Society for Evolutionary Protistology, vol 19, issue 24. Vancouver, British Columbia
Shock EL, Holland M, Meyer-Dombard D, Amend JP, Osburn GR, Fischer TP (2010) Quantifying inorganic sources of geochemical energy in

hydrothermal ecosystems, Yellowstone National Park, USA. Geochim Cosmochim Acta 74:4005–4043. doi:10.1016/j.gca.2009.08.036
Skirnisdottir S, Hreggvidsson GO, Hjörleifsdottir S, Marteinsson VT, Petursdottir SK, Holst O et al (2000) Influence of sulfide and temperature on species composition and community structure of hot spring microbial mats. Appl Environ Microbiol 66:2835–2841. doi:10.1128/AEM.66.7.2835-2841.2000
Swingley WD, Meyer-Dombard DR, Shock EL, Alsop EB, Falenski HD, Havig JR et al (2012) Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem. PLoS One 7:e38108. doi:10.1371/journal.pone.0038108
Takacs-Vesbach C, Mitchell K, Jackson-Weaver O, Reysenbach A-L (2008) Volcanic calderas delineate biogeographic provinces among Yellowstone thermophiles. Environ Microbiol 10:1681–1689. doi:10.1111/j.1462-2920.2008.01584.x
Ward DM, Ferris MJ, Nold SC, Bateson MM (1998) A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol Mol Biol Rev 62:1353–1370
Whitaker RJ, Banfield JF (2006) Population genomics in natural microbial communities. Trends Ecol Evol 21:508–516. doi:10.1016/j.tree.2006.07.001
Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301:976–978. doi:10.1126/science.1086909
White DE, Fournier RO, Muffler LJP, Truesdell AH (1975) Physical results of research drilling in thermal areas of Yellowstone National Park. Wyoming, United States Geological Survey
Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S et al (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci USA 110:222–227. doi:10.1073/pnas.1219574110
Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4:259–263. doi:10.1016/S1672-0229(07)60007-2

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com