nature microbiology

LETTERS

https://doi.org/10.1038/s41564-018-0276-6

Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient

Jackson W. Sorensen[1], Taylor K. Dunivin[1,2], Tammy C. Tobin[3] and Ashley Shade[1,4*]

Small bacterial and archaeal genomes provide insights into the minimal requirements for life[1] and are phylogenetically widespread[2]. However, the precise environmental pressures that constrain genome size in free-living microorganisms are unknown. A study including isolates has shown that thermophiles and other bacteria with high optimum growth temperatures often have small genomes[3]. It is unclear whether this relationship extends generally to microorganisms in nature[4,5] and more specifically to microorganisms that inhabit complex and highly variable environments, such as soils[3,6,7]. To understand the genomic traits of thermally adapted microorganisms, here we investigated metagenomes from a 45 °C gradient of temperate-to-thermal soils that lie over the ongoing Centralia, Pennsylvania (USA) coal-seam fire. We found that hot soils harboured distinct communities with small genomes and small cell sizes relative to those in ambient soils. Hot soils notably lacked genes that encode known two-component regulatory systems, and antimicrobial production and resistance genes. Our results provide field evidence for the inverse relationship between microbial genome size and temperature in a diverse, free-living community over a wide range of temperatures that support microbial life.

Centralia, Pennsylvania is the site of a slow-burning, near-surface coal-seam fire that ignited in 1962. The heat from the fire vents through overlying soils, thus causing surface soil temperatures to reach up to >400 °C[8], but more recently in the range of 40–75 °C[9,10]. Centralia offers an interesting model press disturbance[11] that can be used to directly compare the traits of microorganisms that can withstand thermal temperatures to traits of microorganisms from proximal soils at ambient temperature.

We recently assessed the compositional changes in Centralia soil microbial communities along an ambient-to-thermal temperature gradient overlying the fire[10]. We collected surface soils that were hot due to the fire ('fire-affected'), previously hot but now recovered to ambient temperatures ('recovered') and never impacted by the fire ('reference'). Fire-affected soils had a starkly different community structure from ambient soils. These hot soils also had overlapping 16S rRNA gene compositions, but the abundances of taxa varied. However, after the fire advanced, soils recovered reasonably towards the reference community structure. This suggested a considerable capacity for resilience of soil microbiomes, even after exposure to a severe and unanticipated stressor, and prompted us to investigate which microbial attributes underlie the community changes in fire-affected soils.

From twelve metagenomes (six fire-affected, five recovered and one reference; Supplementary Table 1), we calculated the average genome sizes inclusive of chromosomes and plasmids. The average genome sizes were negatively and strongly correlated with temperature (Fig. 1a; Pearson's $R=-0.910$, $P<0.001$; $n=12$ metagenomes). This relationship was not due to changes in eukaryotes or plasmids along the gradient (Supplementary Table 2). We used three additional methods to assess the changes in genome size with soil temperature and found them all to be in agreement (Supplementary Fig. 1). Although other variables that were not measured might provide additional information, only temperature was explanatory out of the thirteen soil variables measured in this study (Supplementary Table 3). Here we report a decrease in average genome size across an in situ temperature gradient that spans the physiological requirements from mesophiles to thermophiles.

We next compared the average genome sizes estimated from Centralia metagenomes to those from 22 public soil metagenomes (Fig. 2a and Supplementary Table 4). Generally, hot Centralia soils had small genomes relative to other soils, whereas ambient Centralia soils were closer to the average size observed among this set. The average genome sizes from ambient Centralia soils were in agreement with sizes reported from other soils and calculated using comparable methods[7,12,13].

We compared average genome sizes in Centralia to the sizes of a collection of soil isolate genomes (RefSoil; Fig. 2b). Genome sizes from RefSoil did not differ across several soil types (Fig. 2c), which suggests that soil type has a minimal influence on genome size. Although the average genome size in hot Centralia soils is not as small as the soil oligotroph *Candidatus* Udaeobacter (2.81×10^6 base pairs (Mbp)[6]), it is significantly smaller than directly comparable ambient Centralia soils and small relative to other soils (Fig. 2a). Together, these results support comparably small genomes in Centralia soils and more generally provide a range of expected soil genome sizes. Moreover, the average genome sizes in Centralia ambient soils are not remarkably large. This suggests that the inverse relationship between genome size and soil temperature in Centralia soils is an ecologically meaningful outcome of abiotic filtering.

It was hypothesized by Sabath and colleagues that small cells may be selected for at high temperatures to minimize cellular maintenance costs and that small cells indirectly select for small genomes[3]. We re-analysed microscope images from soil cell counts in Centralia[10] to extract size information. Average cell sizes were also negatively correlated with temperature (Fig. 1b; Pearson's $R=-0.65$, $P=0.021$, $n=12$ soils; Supplementary Table 5). Accordingly, cell size correlated with genome size (Fig. 1c; Pearson's $R=0.64$, $P=0.025$, $n=12$ soils). These results agree with reported in situ relationships between cell size and temperature observed

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA. [2]Environmental and Integrative Toxicological Sciences, Michigan State University, East Lansing, MI, USA. [3]Department of Biology, Susquehanna University, Selinsgrove, PA, USA. [4]Department of Plant, Soil and Microbial Sciences, Program in Ecology, Evolutionary Biology and Behavior and the Plant Resilience Institute, Michigan State University, East Lansing, MI, USA. *e-mail: shade.ashley@gmail.com

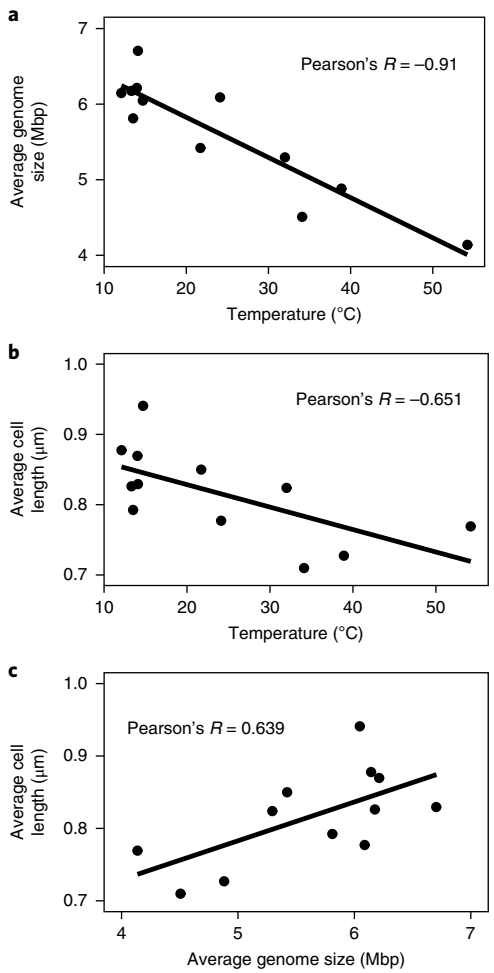

Fig. 1 | Changes in average genome and cell sizes across the soil temperature gradient in Centralia. a, The average genome size in each metagenome was calculated using MicrobeCensus and plotted against the site temperature (Pearson's correlation $P = 4.095 \times 10^{-5}$). **b**, The average cell length was measured from 44–910 cells from 3–9 replicate fields for each soil and plotted against the soil temperature (Pearson's correlation $P = 0.022$). **c**, The average genome size had a direct relationship with the average cell size (Pearson's correlation $P = 0.025$). All Pearson's correlations were two-tailed; $n = 12$.

in aquatic systems[4,5]. Our results extend the cell size–temperature trend to soils and also to a 45 °C temperature range.

Cell and genome sizes can be governed not only by environmental conditions but also by taxonomy (for examples, see refs [3,14]). As we previously reported[10] and as confirmed by this work using phylogenetic inference of genome size (Supplementary Fig. 1b), there were stark changes in community structure between fire-affected and ambient soils (Supplementary Fig. 2). This provides evidence that there was environmental filtering for taxa with small genomes in hot Centralia soils caused by compositional turnover. Using 104 high-quality, de novo metagenome-assembled genomes (MAGs; Supplementary Fig. 1c and Supplementary Table 6), which represent some of the most abundant taxa, we investigated whether small MAGs typical of hot Centralia soils were relatives of thermophiles or lineages that have characteristically small genomes (Fig. 3 and Supplementary Fig. 2b). Some of the MAGs assembled from hot soils were related to known thermophile lineages, such as Crenarcheota, Thaumarchaeota and Chloroflexi; however, other 'hot' MAGs cluster with lineages that are not described as thermotolerant (Fig. 3a). Taxonomy could not be assigned to 51 (out of 104) MAGs beyond the phylum level, and two bacteria could not be assigned beyond the domain level, suggesting previously undescribed taxa (Fig. 3b and Supplementary Table 6). For some phyla, Centralia MAGs trended small relative to the median genome sizes of isolate references (for example, Acidobacteria and Actinobacteria; Fig. 3b), although there were exceptions (for example, Chloroflexi). Other lineages did not have a sufficient number of reference genomes to make robust comparisons and point to phylogenetic gaps in soil reference genomes.

We used metagenome annotations from the KEGG module database to determine changes in functional genes with increasing temperature. KEGG modules are groups of KEGG orthologs that represent complexes, functional sets, metabolic pathways or signatures. Of the KEGG orthologs detected in Centralia metagenomes, 81% were found in all 12 soils and many patterns with temperature could be attributed to changes in normalized KEGG ortholog abundance rather than changes in KEGG ortholog detection. In total, 284 (out of 541 detected; 52.50%) were correlated with temperature (Fig. 4, Supplementary Table 8 and Supplementary Discussion).

Twenty-seven KEGG modules were positively correlated with temperature (Pearson's $R > 0.656$, false discovery rate < 0.05; Fig. 4a, Supplementary Table 8). Anaerobic processes, including dissimilatory sulphate reduction (M00596), dissimilatory nitrate reduction (M00530) and denitrification (M00529), were enriched in hot soils (Fig. 4a, cluster iii), aligning with known and expected environmental conditions in Centralia. Fire-affected soils from actively steaming vents had higher moisture than ambient soils (Pearson's $R = 0.714$, $P < 0.01$; $n = 12$ soils), which probably causes inundated and anaerobic microhabitats. Previous work in Centralia indicated an importance of sulphur, sulphate, nitrate and ammonium metabolisms because these compounds were commonly elevated at vents[8,9]. These results also agree with observations of thermophile metabolisms in other terrestrial and geothermal environments[15–18]. These anaerobic KEGG modules had similar dynamics to several archaeal proteins (Fig. 4a cluster iii; archaeal ribosome M00179, polymerase M00184 and exosome M00390). There was an increase in Crenarchaeota in fire-affected soils[10], an archaeal phylum that includes sulphate reducers[19] and has nine soil reference genomes that average 2.26 Mbp (Fig. 3b). Together, these data suggest that pathways enriched in small genomes from hot soils encode functions attuned to the Centralia habitat.

Temperature was negatively correlated with 257 KEGG modules (47.5%, Pearson's $R < -0.6$, false discovery rate < 0.05; Fig. 4b, Supplementary Table 8). In general, depleted KEGG modules were detected across ambient soils. Note that antimicrobial resistance and production and two-component regulatory systems comprised 32.7% of the KEGG modules negatively correlated with temperature (84 out of 257; Fig. 4b). This trend was striking, but some KEGG modules belonging to these categories had no relationship to temperature and these KEGG module categories were always detected in fire-affected soils.

Thirty-nine modules for antimicrobial production and resistance were negatively correlated with temperature, which is in agreement with our previous analysis of antibiotic resistance genes in Centralia[20]. Small genomes of host-associated symbionts often lack antimicrobial genes[21]. However, the free-living marine *Pelagibacter* clade, a model for genome streamlining attributed to oligotrophic conditions, has a multidrug transporter conserved across sequenced genomes[22]. The challenges in developing selectable antibiotic resistance markers for thermophiles[23,24] suggest that thermophiles might have fewer genes that encode resistance to described antimicrobials. Similar to most databases, KEGG is biased towards genomes and genes from fast-growing mesophiles and may miss annotation of poorly described thermophile antimicrobial genes. However,

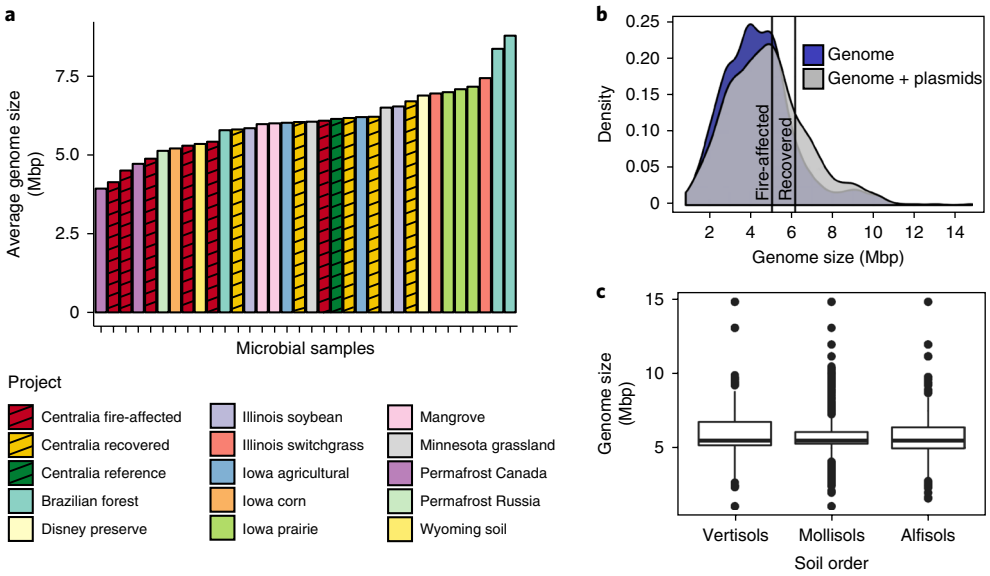

Fig. 2 | Comparison of Centralia genome sizes to other soils. a, Comparison to publicly available metagenomes of similar coverage and quality from the database MG-RAST. The average genome sizes in soil metagenomes were estimated using MicrobeCensus. The samples are ordered according to average genome size and colour coded according to the location from which the sample was obtained. **b**, Distribution of genome size from cultivable soil microorganisms (RefSoil) with and without plasmids. The mean genome size of Centralia fire-affected and recovered metagenomes are plotted over the distribution. **c**, The distribution of genome size (including plasmids) are not distinct across different soil orders. Previously published estimates of the abundance of RefSoil organisms in the soil Earth Microbiome Project[53] dataset were used to estimate the distribution of genome size of soil microbiomes in Alfisols, Vertisols and Mollisols. The midlines of each box plot correspond to the median values. The top and bottom of each boxplot represent the 75th and 25th percentiles, respectively. The upper and lower whiskers extend to the furthest values that are not outliers.

thermal conditions might present a strong environmental filter that reduces competition and the need for antimicrobial production and resistance. We previously reported decreased richness and phylogenetic diversity of fire-affected Centralia soils[10], which suggests that there is a smaller pool of potential competitors that inhabits the hot soils.

Forty-nine of the total detected modules were also negatively correlated with temperature (Pearson's $R < -0.6$). Two-component systems allow bacteria to respond to multiple stimuli with little genetic material[25,26]. Smaller genomes, including those that are reduced or streamlined, can have fewer regulatory components[5,7,27] and less regulation[22,28–31]. Our results suggest that thermophiles may have relatively low regulatory needs. It has been proposed that thermophiles with small genomes may be more likely to utilize global regulatory systems that mediate transcriptional responses to co-occurring environmental stimuli[29]. Environmental stability is also predicted to influence the relative benefit that an organism gains from investing in sensing its environment[32]. For example, obligate endosymbionts are thought to have drifted towards having small genomes in part because conditions are stable and sensing requirements are minimal[7]. In Centralia, seasonal temperature fluctuations in fire-affected and ambient soils are equivalent (Supplementary Fig. 3), providing evidence that the soils experience similar environmental stability in terms of temperature, albeit at different ranges. This suggests that wild small genomes are not necessarily conditional on stable environments[7] and begs the question of whether two-component regulatory systems are consistently less prevalent among thermophiles.

Our cultivation-independent field study supports cultivation-dependent studies that suggest that higher temperatures support the growth of bacteria and archaea with small genomes[3]. Surprisingly, it also suggests that microbial populations inhabiting complex environments, such as soils, may generally reflect similar overarching traits in genome size to those observed in laboratory studies.

These results add evidence that supports selection for both smaller genomes and cells at higher temperatures, but also offer a key point of distinction. Our study considers the ecological process of selection[33] through abiotic environmental filtering, not the evolutionary process of natural selection, towards streamlining. Although taxa that were enriched in hot soils characteristically had smaller genomes and cells, there is no evidence for contemporary genome streamlining in Centralia. Instead, we suspect that these thermotolerant cells were resuscitated from the vast dormant pool in the soil. This is supported by three lines of evidence. First, there was turnover in community membership across hot and ambient Centralia soils[10], thus providing evidence against contemporary streamlining within local lineages. Second, many of the lineages that we detected in high abundance at certain hot sites were also detected in low abundance at other sites, including ambient sites (Fig. 3a and ref. [10]), suggesting a role for the rare biosphere or dormant pool as a diversity reservoir for unanticipated thermal conditions. Finally, many other studies have described thermophile persistence and resuscitation from non-thermal environments, which suggests that thermophilic lineages are widespread but typically inactive (for example, see refs [16,34,35]). Therefore, we posit that the small genomes in Centralia are characteristic of previously dormant thermophiles in the soil and not the outcome of genome streamlining.

Centralia afforded a unique opportunity to directly compare the metagenomes of proximal soils along an extreme temperature range. It is unusual to observe such a wide temperature range in soils, especially one that is inclusive of thermal temperatures, historically and geologically comparable, and with shared exposure to the same regional pool of dispersed microorganisms. When more metagenomes are available, comparisons with other thermal soils will provide insights into the generality of the trends observed in Centralia.

There are many environmental factors that contribute to microbial genome size, including oligotrophic conditions[6,36], relative

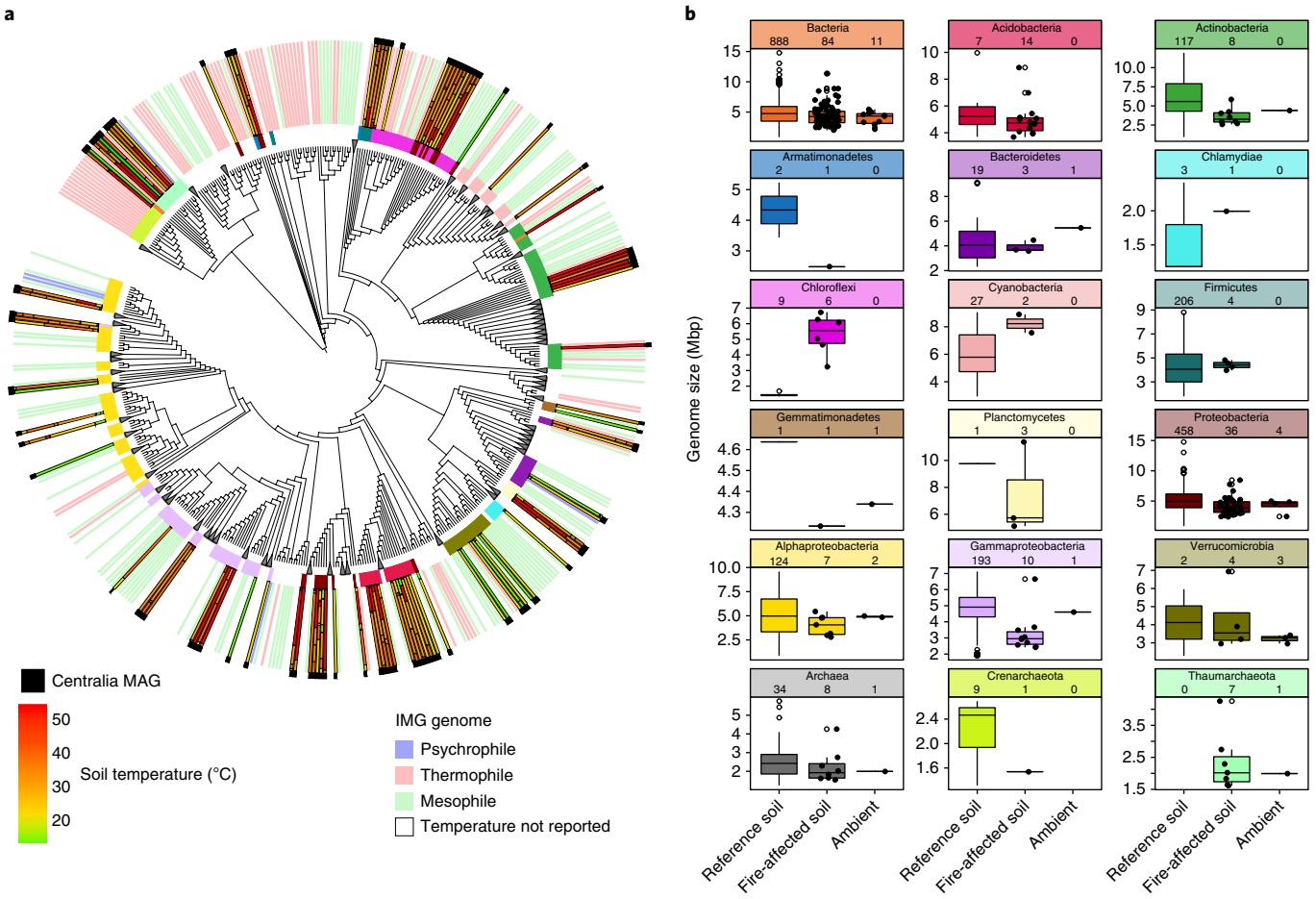

Fig. 3 | Temperature distribution and diversity of Centralia MAGs compared to reference soil genomes from IMG and RefSoil. a, Microbial reference phylogeny based on single-copy (or 'marker') genes[45] that was expanded to include Centralia MAGs. For clarity, large clades that did not contain MAGs are collapsed. The inner colour ring shows phylum-level taxonomy, matched to the phyla in **b**. The outer colour ring shows the temperature reported for Integrated Microbial Genomes (IMG) reference lineages (muted colours) and the distribution and measured soil temperatures for Centralia MAGs (brighter colours; black flags). **b**, Genome sizes of RefSoil isolates compared to 104 of the highest-quality Centralia MAGs from fire-affected and ambient soils (taxonomy was assigned by the Microbial Genome Atlas NCBI Prokaryote project[44]). The sample sizes indicated in the panel headers are the total number of RefSoil genomes or Centralia MAGs detected within each lineage. Note the differences in the *y* axis ranges. Because many of the highest-quality MAGs were assembled from hot soils, **b** does not provide robust MAG comparisons across the Centralia fire impact categories. The midlines of each box plot correspond to the median values. The top and bottom of each boxplots represent the 75th and 25th percentiles, respectively. The upper and lower whiskers extend to the furthest values that are not outliers.

environmental stability[7,32] and symbiotic lifestyle[28,31], and these factors are expected to interact with taxonomy[3,14]. Furthermore, there are evolutionary explanations as to why small genomes might trend with high temperatures, as discussed in detail by Sabath and colleagues[3]. Here, we provide evidence that many lineages of soil microorganisms that can thrive at thermal temperatures and have small genomes and cells, supporting the hypothesis that small cells constrain genome size[3]. Importantly, our results show that high temperature is one environmental factor that can drive overarching changes in the genomic and cellular traits of wild microbial communities.

Methods

DNA extraction and metagenome sequencing. DNA for metagenome sequencing was manually extracted using a phenol–chloroform extraction[37] and then purified using the MoBio DNeasy PowerSoil Kit according to the manufacturer's instructions. Briefly, as per a previously published protocol[5], after four freeze–thaw cycles, 10 ml phenol–chloroform-isoamyl alcohol (25:24:1) was added to each sample, mixed and centrifuged at 7,500 *g* for 10 min. After precipitation, DNA was pelleted through centrifugation at 7,500 *g* for 15 min. The pelleted DNA was resuspended in 100 µl TE buffer (10 mM Tris–HCl, 1 mM EDTA•Na_2). The extracted DNA was then purified using the MoBio DNeasy PowerSoil kit as per the manufacturer's instructions, omitting the 10 min vortexing step after the addition of solution C1. Total DNA sequencing was performed on all 12 samples by the Department of Energy's Joint Genome Institute (Community Science Project) using an Illumina HiSeq 2500. Libraries were prepared with a targeted insert size of 270 bp. Samples had between 19 Gbp and 50 Gbp of sequence data.

Quality control, assembly and annotation. Adapters were removed and quality trimmed at values smaller than 12 using BBDuk (https://sourceforge.net/projects/bbmap/). BBDuk was also used to remove reads that had more than one ambiguous base, a final length of less than 40 bp after trimming or an average quality score below eight. Reads that matched Illumina artifacts, spike-ins or phiX were also removed and the resulting reads mapped to human genome HG19 using BBMap, removing all reads that hit with >93% identity. These quality-controlled reads from each metagenome were assembled separately using megahit[6] with k-mer size ranging from 31–121 with a k-step of ten. The coverage of the resulting contigs was estimated using seal to map all reads onto the contigs.

To use all of the sequencing data, we worked with assembled and unassembled reads processed by IMG using their standard annotation pipeline[38]. After comparing several annotation methods (Supplementary Discussion), we chose to use the KEGG orthology database[39] to analyse the Centralia data due to its inherent structure and ability to integrate metabolic pathways. KEGG ortholog abundances were relativized to the median abundance at each site of a set of 36 single-copy genes that were published previously[40] (Supplementary Table 7). One single-copy

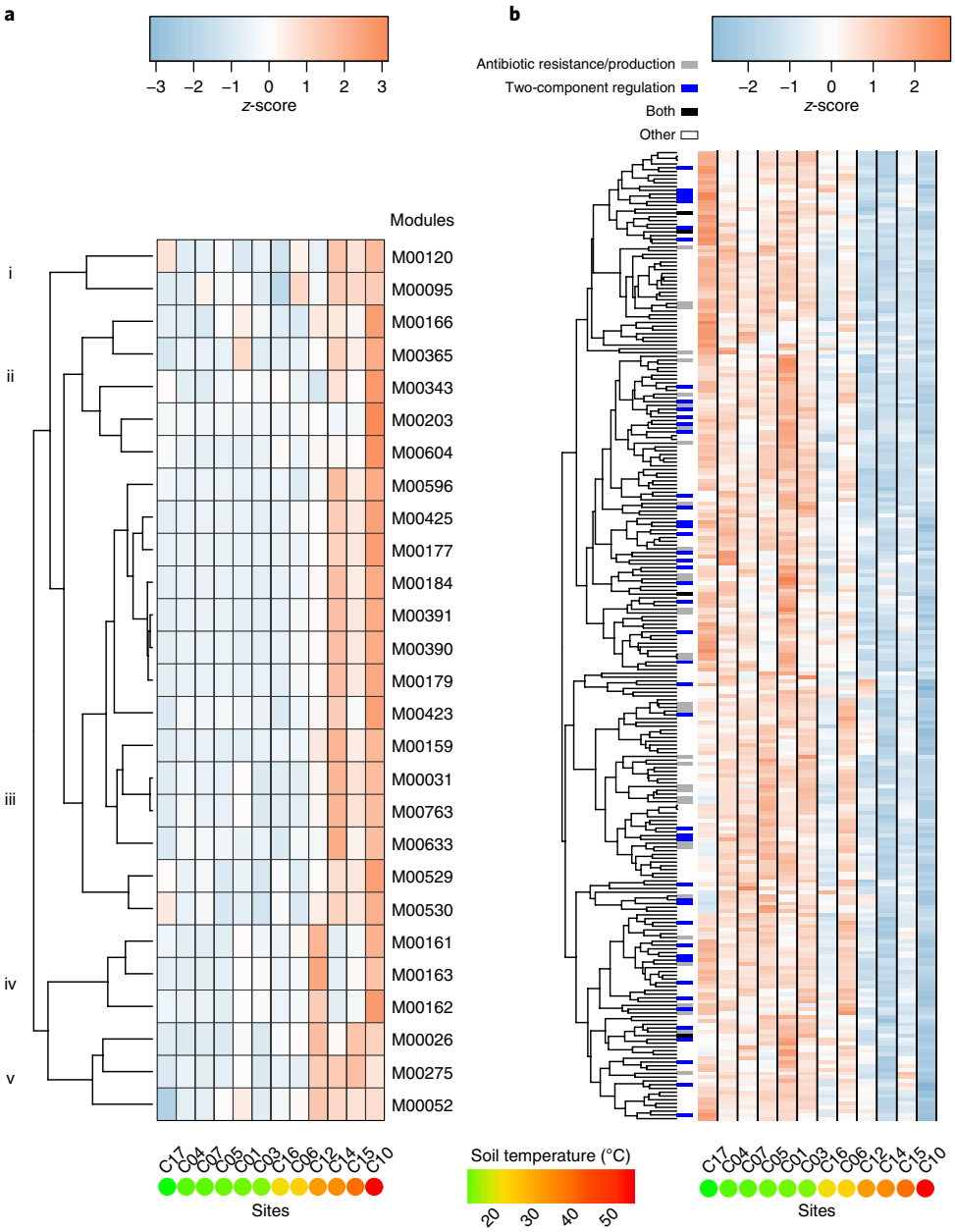

Fig. 4 | KEGG modules correlated with temperature. **a**, Twenty-seven KEGG modules correlated positively with temperature (Pearson's *R* range: 0.646 to 0.933). The roman numerals in **a** denote the clusters of KEGG modules with similar response patterns to the temperature gradient. **b**, Negative correlations between temperature and 257 KEGG modules were observed (Pearson's *R* range: −0.642 to −0.925). One-third of the KEGG modules that correlated negatively with temperature were two-component regulatory systems (blue dendrogram tips), the ability to resist or produce antimicrobials (grey tips) or both (black tips). Row dendrograms show the hierarchical clustering of KEGG modules according to response patterns to the temperature gradient. Modules (rows) are centred and standardized across Centralia metagenomes (columns), with warm colours showing relative enrichment and cool colours showing relative depletion. Note the differences in colour gradient ranges across **a** and **b**. The modules with significant relationships to temperature are shown. Sites are arranged according to temperature, increasing from left to right. Two-tailed Pearson's correlation false discovery rate < 0.05; $n = 12$ soils. The full information on correlation statistics for each KEGG module is listed in Supplementary Table 8.

gene (K01519) was an outlier in 7 out of 12 samples, as assessed by Grubb's test for outliers, and removed. We analysed patterns in KEGG Modules[39], a set of manually defined functional units made up of multiple KEGG orthologs. KEGG module abundances were calculated based on the median abundance of their constituent KEGG orthologs that were present in the metagenomes. KEGG modules were included in an analysis if 50% or more of their constituent KEGG orthologs were identified in the dataset. Approximately one-third of the open reading frames per sample could be annotated with KEGG (Supplementary Table 1). As a caveat to the study, unannotated open reading frames can result from erroneous reads and misassemblies but also could be previously undescribed and/or divergent genes that are critical for microbial processes. Thus, new annotations could impact the overarching patterns described here.

Average genome size. Average genome size was calculated from the quality-filtered DNA sequences using MicrobeCensus (run_microbe_census.y –n 2000000), which estimates the average genome size by calculating the percentage of sampled reads that match to a set of single-copy genes[41]. We also used three additional methods to calculate average genome size (Supplementary Fig. 1) and all were in agreement in detecting a significant, negative relationship between temperature and average genome size. Finally, eukaryotic sequence and plasmid contributions were

consistent and low across the metagenomes (Supplementary Table 2), showing that there was no systematic overestimation of genome size in ambient soils due to eukaryotic signal or characteristic changes in plasmids with temperature.

We calculated the odds ratios for each of the 36 single-copy gene KEGG orthologs, previously used by He et al.[40] to estimate average genome sizes. The odds ratios were determined at each site by comparing their abundance within a site to their average abundance across all 12 sites. One KEGG ortholog (K01519, triosephosphate isomerase) was an outlier in 7 out of 12 metagenomes, as determined by Grubb's test and was removed.

We used previously published 16S rRNA gene sequencing data[3] to estimate the average genome sizes. A mean phylum genome size was calculated for each phylum present in Centralia metagenomes using all complete or permanent draft genomes deposited in IMG. Outliers in genome sizes were identified using the Tukey method and omitted from the calculation of the mean phylum genome size[13]. Phyla that were present in Centralia metagenomes but lacked representative genomes in IMG were combined at the domain level and a mean domain genome size was calculated in the same manner. The weighted mean genome size of each site was calculated based on the relative abundance of the phyla at the respective site.

Quality-filtered metagenome reads were downloaded from the JGI GOLD database. Paired-end reads from all 12 soils were assembled using MEGAHIT (v1.0.2)[6] with a k-mer range of 27–107 and a k-step of 10. Reads were mapped to the resulting assembly using bbmap (v35.34) with a minimum identity of 76%. The resulting SAM files were converted to sorted BAM files using SAMTools (v1.3). Contigs that were larger than 2,500 bp were binned into MAGs with MetaBAT (v0.26.3) using the '—veryspecific' flag. Completeness and contamination were estimated for each MAG using CheckM (v1.0.5). MAGs with more than 90% completeness and less than 5% contamination were used to estimate genome size. The genome size of a MAG was estimated by multiplying the sum of the length of its constituent contigs by the inverse of its completeness[14]. The average MAG size at each site was calculated by taking the mean of the sizes of all MAGs detected at a site.

Average cell size. To calculate cell size, we re-analysed microscope images that were used in a previous report to count microbial cells for community size quantifications in the same soils[10]. We hand-curated a debris-free subset from the images and measured 44–910 cells from 3–9 replicate fields for each soil. The major and minor axes of cells were measured using a FIJI macro in ImageJ (version 2.0.0-rc-65/1.51s, build 961c5f1b7f). We found that the cell size range and deviations (Supplementary Table 5) were consistent with those reported in earlier work[42].

Construction of MAGs, taxonomic assignments and visualization. The assembled contigs from quality-filtered reads were binned into MAGs using MetaBAT[43] (v0.26.3) with the '—veryspecific' flag. A detailed description of assembly and binning procedures can be found in Supplementary Information. Completeness and contamination were estimated for each MAG using CheckM (v1.0.5). MAGs were assigned taxonomy using the Microbial Genome Atlas NCBI Prokaryote project[44]. The highest-quality MAGs with more than 90% completeness and less than 5% contamination were used to estimate genome size. The genome size of a MAG was estimated by multiplying the sum of the length of its constituent contigs by the inverse of its completeness[6]. The average MAG size at each site was calculated by taking the mean of the sizes of all MAGs detected at a site.

The CheckM[45] genome tree was extended to include Centalia high-quality MAGs. The Interactive Tree of Life[46] was used for visualization (https://itol.embl.de/tree/352041174435631527858534). The temperature range and taxonomy for each genome in the tree was collected from JGI IMG. MAGs were classified as fire-affected or ambient on the basis of which group of samples they had a higher coverage of and 95% of MAGs had at least 10× greater coverage in one soil category in comparison to the other.

Comparisons with other soil metagenomes and genomes. All metagenome datasets for comparison were obtained from MG-RAST (http://metagenomics.anl.gov/). The MG-RAST database was searched with the following criteria: material = soil, sequence type = shotgun and public = true. The resulting list of metagenome datasets were ordered according to the number of base pairs. Metagenomic datasets with the highest number of base pairs were included if they were sequenced using Illumina (to standardize sequencing errors), had an available FASTQ file (for internal quality control) and contained <30% low-quality reads as determined by MG-RAST. Within high-quality Illumina samples, priority for inclusion was given to projects with multiple samples. When a project had multiple samples, datasets with the greatest number of base pairs were selected. This search yielded 22 datasets from 12 locations and 5 countries (Supplementary Table 4). Sequences from MG-RAST datasets were quality checked using FastQC (v0.11.3[47]) and quality controlled using the FASTX toolkit (fastq_quality_filter, '−Q33 −q 30 −p 50'). The average genome size for each dataset was calculated from the quality-filtered DNA sequences using MicrobeCensus with default parameters.

RefSoil[48] was used to estimate the genome sizes of soil organisms. Genome and plasmid sizes from all 922 RefSoil organisms were extracted from GenBank files and read into R for analysis.

Statistical analyses. Statistics for the metagenome datasets were performed in the R environment for statistical computing[49]. The stats package was used to calculate two-sided Pearson's correlations[49]. The outliers package[50] was used to identify outlying KEGG orthologs. The ggplot2 package was used for visualization[51]. Heat maps were created with heatmap2 from the gplots package[52].

Code availability. All analysis workflows are available via GitHub (https://github.com/ShadeLab/PAPER_Sorensen_NatMicro_2018).

Data availability

Metagenome data are available on IMG under the GOLD Study ID GS0114513. MG-RAST data are available under Project IDs mgp3731, mgp252, mgp5588, mgp14596, mgp6377, mgp6368, mgp2592, mgp2076, mgp11628, mgp13948, mgp7176 and mgp15600.

Received: 2 March 2018; Accepted: 27 September 2018;
Published online: 5 November 2018

References

1. Hutchison, C. A. et al. Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
2. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
3. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).
4. Huete-Stauffer, T. M., Arandia-Gorostidi, N., Alonso-Sáez, L. & Morán, X. A. G. Experimental warming decreases the average size and nucleic acid content of marine bacterial communities. *Front. Microbiol.* **7**, 730 (2016).
5. Swan, B. K. et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl Acad. Sci. USA* **110**, 11463–11468 (2013).
6. Brewer, T. E., Handley, K. M., Carini, P., Gilbert, J. A. & Fierer, N. Genome reduction in an abundant and ubiquitous soil bacterium '*Candidatus* Udaeobacter copiosus'. *Nat. Microbiol.* **2**, 16198 (2016).
7. Giovannoni, S. J., Thrash, J. C. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
8. Janzen, C. & Tobin-Janzen, T. in *Microbiology of Extreme Soils* (eds Dion, P. & Nautiyal, C. S.) 299–316 (Springer, Berlin, 2008).
9. Tobin-Janzen, T. et al. Nitrogen changes and domain bacteria ribotype diversity in soils overlying the Centralia, Pennsylvania underground coal mine fire. *Soil Sci.* **170**, 191–201 (2005).
10. Lee, S.-H., Sorensen, J. W., Grady, K. L., Tobin, T. C. & Shade, A. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *ISME J.* **11**, 1447–1459 (2017).
11. Shade, A. Understanding microbiome stability in a changing world. *mSystems* **3**, e00157-17 (2018).
12. Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biol.* **8**, R10 (2007).
13. Tecon, R. & Or, D. Biophysical processes supporting the diversity of microbial life in soil. *FEMS Microbiol. Rev.* **41**, 599–623 (2017).
14. Schattenhofer, M. et al. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ. Microbiol.* **11**, 2078–2093 (2009).
15. Dodsworth, J. A., Hungate, B., de la Torre, J. R., Jiang, H. & Hedlund, B. P. Measuring nitrification, denitrification, and related biomarkers in terrestrial geothermal ecosystems. *Methods Enzymol.* **486**, 171–203 (2011).
16. Marchant, R. et al. Thermophilic bacteria in cool temperate soils: are they metabolically active or continually added by global atmospheric transport? *Appl. Microbiol. Biotechnol.* **78**, 841–852 (2008).
17. Reigstad, L. J. et al. Nitrification in terrestrial hot springs of Iceland and Kamchatka. *FEMS Microbiol. Ecol.* **64**, 167–174 (2008).
18. Santana, M., Gonzalez, J. & Clara, M. Inferring pathways leading to organic-sulfur mineralization in the Bacillales. *Crit. Rev. Microbiol.* **42**, 31–45 (2016).
19. Itoh, T., Suzuki, K., Sanchez, P. C. & Nakase, T. *Caldivirga maquilingensis* gen. nov., sp. nov., a new genus of rod-shaped crenarchaeote isolated from a hot spring in the Philippines. *Int. J. Syst. Bacteriol.* **49**, 1157–1163 (1999).
20. Dunivin, T. K. & Shade, A. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. *FEMS Microbiol. Ecol.* **94**, fiy016 (2018).
21. Gao, Z. M. et al. Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont '*Candidatus Synechococcus spongiarum*'. *mBio* **5**, e00079–14 (2014).
22. Grote, J. et al. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**, e00252–12 (2012).
23. Brouns, S. J. J. et al. Engineering a selectable marker for hyperthermophiles. *J. Biol. Chem.* **280**, 11422–11431 (2005).

24. Hoseki, J., Yano, T., Koyama, Y. & Kuramitsu, S. Directed evolution of thermostable kanamycin-resistance gene: a convenient selection marker for *Thermus thermophilus*. *J. Biochem.* **126**, 951–956 (1999).
25. Hoch, J. A. Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.* **3**, 165–170 (2000).
26. Whitworth, D. E. & Cock, P. J. A. Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids* **37**, 459–466 (2009).
27. Ranea, J. A. G., Grant, A., Thornton, J. M. & Orengo, C. A. Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet.* **21**, 21–25 (2005).
28. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
29. Wang, Q., Cen, Z. & Zhao, J. The survival mechanisms of thermophiles at high temperatures: an angle of omics. *Physiology* **30**, 97–106 (2015).
30. Yus, E. et al. Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326**, 1263–1268 (2009).
31. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).
32. Kussell, E. & Leibler, S. S. Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309**, 2075–2078 (2005).
33. Vellend, M. Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206 (2010).
34. Portillo, M. C., Santana, M. & Gonzalez, J. M. Presence and potential role of thermophilic bacteria in temperate terrestrial environments. *Naturwissenschaften* **99**, 43–53 (2012).
35. Müller, A. L. et al. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *ISME J.* **8**, 1153–1165 (2014).
36. Giovannoni, S. J. et al. Genetics: genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
37. Cho, J. -C., Lee, D. -H., Cho, Y. -C., Cho, J. -C. & Kim, S. -J. Direct extraction of DNA from soil for amplification of 16S rRNA gene sequences by polymerase chain reaction. *J. Microbiol.* **34**, 229–235 (2006).
38. Huntemann, M. et al. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand. Genomic Sci.* **11**, 17 (2016).
39. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
40. He, S. et al. Patterns in wetland microbial community composition and functional gene repertoire associated with methane emissions. *mBio* **6**, e00066-15 (2015).
41. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 51 (2015).
42. Balkwill, D. L. & Casida, L. E. Microflora of soil as viewed by freeze etching. *J. Bacteriol.* **114**, 1319–1327 (1973).
43. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
44. Rodriguez-R, L. M. et al. Microbial Genomes Atlas: standardizing genomic and metagenomic analyses for Archaea and Bacteria. *Nucleic Acids Res.* **46**, 282–288 (2018).
45. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
46. Letunic, I. & Bork, P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
47. Andrews, S. FastQC: a quality control tool for high throughput sequence data (2010); http://www.bioinformatics.babraham.ac.uk/projects/fastqc
48. Choi, J. et al. Strategies to improve reference databases for soil microbiomes. *ISME J.* **11**, 829–834 (2017).
49. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
50. Komsta, L. outliers: Tests for outliers. R package v.0.14 (2011); http://CRAN.R-project.org/package=outliers
51. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2009).
52. Warnes, G. R. et al. gplots: Various R programming tools for plotting data. R package v.3.0.1 (2016); https://rdrr.io/cran/gplots/
53. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

Acknowledgements

This research was supported by Michigan State University and the National Science Foundation grant no. DEB1749544. Computational resources were provided by the Michigan State Institute for Cyber-Enabled Research. Metagenome sequencing was supported by the Joint Genome Institute Community Science Project no. 1834. The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract no. DE-AC02-05CH11231. We thank K. L. Grady and S. -H. Lee for technical support and S. Yeh and J. Lee (REU-ACRES NSF grant no. 1560168) for their contributions to metagenome analyses.

Author contributions

A.S. and T.C.T. conceived the study and conducted the field work. J.W.S. and T.K.D. performed analyses with A.S. J.W.S., A.S. and T.K.D. wrote the manuscript. All authors discussed results, and commented on and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-018-0276-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Ashley Shade

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Illumina RTA software version 1.18.61 was used for sequence base calling.
Data analysis	Bbmap (v35.34) was used for adapter trimming, read filtering and read mapping. Megahit (v1.0.2) was used for assembly of metagenomes. SAMTools (v1.3) was used for sorting and converting SAM file to BAM files. MetaBAT (v0.26.3) was used for binning contigs. The outliers package (v0.14) was used for detecting outliers (grubbs.test). MicrobeCensus (v1.0.7) was used to estimate average genome size from metagenomes. ImageJ (Version: 2.0.0-rc-65/1.51s Build: 961c5f1b7f) was used to measure cell size. CheckM (v1.0.5) was used to estimate completeness and contamination of MAGs. FASTX toolkit (v0.0.14) was used for quality control of metagenomes downloaded from MG-RAST. R (v3.4.0) was used for statistical analyses. The vegan package (v2.4-6) was used for mantel tests(mantel), creating dissimilarity matrices(vegdist), and z-scoring(decostand). The stats package (v3.4.0) was used for Pearson's correlations (cor.test) and false discovery rate correction of p values (p.adjust). BBDuk was used trim and filter reads. bbduk.sh https://sourceforge.net/projects/bbmap/: ktrim=r, minlen=40, minlenfraction=0.6, mink=11, tbo, tpe, k=23, hdist=1, hdist2=1, ftm=5 FASTQC(v0.11.3) was used to quality check MG-RAST datasets MiGA(0.3.3.1) was used to identify and quality check Metagenome Assembled Genomes. Gplots (3.0.1) was used for plotting heatmaps.

ggplot2(2.2.1) was used for visualization.
iTOL (4.2.3) was used for visualizing phylogenetic trees.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All analysis workflows are available on GitHub (ShadeLab/PAPER_Sorensen_NatMicro_2018).
Metagenome data are available on IMG under the GOLD Study ID GS0114513.
MG-RAST data are available under Project IDs mgp3731, mgp252, mgp5588, mgp14596, mgp6377, mgp6368, mgp2592, mgp2076, mgp11628, mgp13948, mgp7176 and mgp15600.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This was an observational study. Soils overlying the Centralia coal mine fire were sampled to capture a gradient of measured surface soil temperature suitable for mesophiles and thermophiles. Samples were collected along a gradient of fire-impact along both fire fronts in Centralia, to span the range of environmental variability in temperature and historical knowledge of fire advancement at the time of sampling.
Data exclusions	No metagenomes were excluded from this analysis. When analyzing single copy genes abundance, a single KO was removed from analysis because it was an outlier in 7 out of 12 samples as assessed by Grubb's test for outliers. The KEGG Ortholog (KO) was removed (not a sample, but a housekeeping gene used as a point of reference with the suite of genes samples) were done so based on a statistical outlier test, which criteria are pre-established based on statistical expectations. This is described in detail in the methods. When calculating average genome size for individual phyla (Supporting Figure 1B), genomes that were outliers based on the Tukey method were omitted from the final calculation of that phylum's average genome size. For calculating average cell sizes using automated measurements with ImageJ, images with soil particles were excluded from analysis so that the software did not accidentally include soil particles in the calculation of cell size.
Replication	We have created computing workflow scripts to reproduce all of the analyses. Because this is an observational study of an environmental gradient, experimental replication is not applicable.
Randomization	Randomization was not relevant to our study. We were analyzing metagenome characteristics across a gradient of temperatures and not between groups. We controlled for covariates by checking for correlation of a suite of physico-chemical data with temperature across these samples; these covariates were previously reported in Sorensen et al. 2017 ISMEJ.
Blinding	Blinding was not relevant to this observational study. This is an observational study of an environmental gradient and there were no samples excluded from the dataset. There were no a priori treatment groups, etc.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	The only unique materials from this study are the soil samples. We have a limited amount of these soils that have been constantly stored at -80C (~200g per sample). Other researchers are welcome to visit the study site and collect their own soil samples using the GPS coordinates that we provide of the sampling locations.