DISTURBANCE ECOLOGY OF SOIL MICROBIAL COMMUNITIES IN RESPONSE TO THE CENTRALIA, PA COAL FIRE

By

Jackson Winther Sorensen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics--Doctor of Philosophy

2019

ABSTRACT

DISTURBANCE ECOLOGY OF SOIL MICROBIAL COMMUNITIES IN RESPONSE TO THE CENTRALIA, PA COAL FIRE

By

Jackson Winther Sorensen

Microbial communities are ubiquitous in our world and play important roles in biogeochemical and ecosystems processes on Earth. The ability of these microbial communities to provide these different processes is frequently tied to their community structure, which can be thought of both in terms of membership (i.e. who is there) and the relative abundance of these members. Changes in environmental conditions often lead to changes in microbial community structure as well. Microbial communities are formed through the process of assembly, which in turn is driven by the four processes of 1) Selection 2) Dispersal 3) Drift and 4) Diversification. Understanding the relative importance of each of these processes in different systems is important for predicting how microbial communities will change in response to disturbances.

This dissertation presents work that uses the coal fire in Centralia, PA as a model press disturbance for understanding soil microbial community responses to and recovery from disturbance. The experiments herein aim to shed light the relative roles of Selection, Dispersal, and Drift in governing these responses in soil microbial communities experience a temperature disturbance. An observation study of a chronosequence of fire disturbance in Centralia, PA is used to generate hypotheses as to the relative roles of Selection, Dispersal, and Drift in the assembly of soil microbial communities experiencing a temperature disturbance. Further, an in depth look at some of these communities using shotgun metagenomics is used to observe specific microbial traits and characteristics selected for by the temperature disturbance. Finally, a

laboratory soil mesocosm warming experiment investigates the relative influence of Dispersal and dormancy in governing responses to and recovery from disturbance.

This dissertation is dedicated to my partner in all things, Jenny, who pushes me everyday to be a better scientist, communicator, and person.

ACKNOWLEDGEMENTS

I would first like to acknowledge my mentor Dr. Ashley Shade for her relentless support during all of this work. Her willingness to give me a chance in her lab right as she arrived at Michigan State and to spend countless hours in impromptu end of the day discussions were critical to building my confidence and enthusiasm as a scientist. As I became a more seasoned graduate student and scientist, her trust in me to present at conferences in her place and lead bioinformatics tutorials when I had no trust in myself was invaluable to me. It was through these experiences that I began to cultivate a deep appreciation for teaching and effective communication of research within our own scientific community and to the community at large.

I'd also like to thank all of my academic advisors and mentors past and present. I have gained something from each and every one of you that makes me the scientist and person I am today. And thank you not only to my advisors and mentors, but also to my peers and mentees, as each of you has challenged me to be a better scientist, communicator, and educator.

Finally, thank you to my friends and family (though some of you are now getting thanked twice). You've provided a great deal of financial, mental, emotional, and physical support through these years and I hope that I have done as much for you all as you have done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

DNA - deoxyribonucleic acid

RNA - ribonucleic acid

rRNA - ribosomal ribonucleic acid

KEGG - Kyoto Encyclopedia of Genes and Genomes

KO - KEGG Ortholog

KM – KEGG Module

MAG – metagenome assembled genome

IMG – Integrated Microbial Genomes

JGI – Joint Genome Institute

DOE – Department of Energy

PA – Pennsylvania

TCRS – Two component regulatory system

PCoA – Principle Coordinates Analysis

OTU – Operational Taxonomic Unit

PCR – polymerase chain reaction

qPCR – quantitative polymerase chain reaction

rpm – rotations per minute

MiGA – Microbial Genomes Atlas

MG-RAST – metagenomic rapid annotations using subsystems technology

Ca – Calcium

S – Sulfur

NH_4^+ - Ammonium

NO_3^- - Nitrate

NO_2^- - Nitrite

ppm – parts per million

Fe – Iron

As – Arsenic

P – Phosphorus

K – Potassium

Mg – Magnesium

RS – Resistance

RL – Resilience

dn – *de novo*

CHAPTER 1: Introduction

Microbial communities and structure-function relationships

Microbial communities are ubiquitous in our world and are important players in important geochemical processes on Earth (1, 2). Microbes play important roles in the carbon cycle (3), and carry out key steps of the nitrogen and sulfur cycles (4, 5). Their functional potential is not limited to these environmental processes either, as they can play key roles in pathogen defense and growth promotion in plants (6, 7). There are an estimated 4-6 $X10^{30}$ microbial cells on Earth, equaling anywhere from 60-100% of the total carbon of plants and estimated at nearly 10^{12} species (1, 8). Overall, these microbial communities provide crucial functions for Earth

The functional output of microbial communities is often dictated by their community structures. The relationship between the composition of a microbial community (taxonomic membership and relative abundances of those members) and the functions that it can perform is referred to as structure-function relationships. Early investigations into this relationship reported correlations between community structure and specific ecosystem functions (9–13) and others have shown strong causal relationship between community structure and ecosystem function (14, 15). However, community structure does not always appear to be intrinsically linked to all functions. Some ecosystem processes may be more dependent on environmental context than on community structure, though even for these processes there appears to be some relationship between structure and function (14, 16).

Community assembly: Vellend's synthesis of community ecology and Nemergut's extension to microbes

Given the relationship between community structure and function, understanding what shapes community structure is an important field of study. The processes by which communities are formed is often referred to as assembly. Studies that have aimed to determine what governs assembly have often looked at the four sub-processes of Selection, Drift, Dispersal and, Speciation. Vellend proposed a synthesis of these four processes as a model for community assembly (17) and Nemergut and authors extended this synthesis to microbial communities, substituting Diversification for Speciation since microbiology lacks a strong species definition (18). Selection refers to natural selection from abiotic and biotic factors on fitness differences between species. Drift refers to slight stochastic changes in the abundance of different members. Dispersal is the process by which species travel between locations. Diversification is the process by which new genetic variation arises. The relative importance of each of these processes is dependent upon the environment.

Selection is the most common assembly processes that has been studied to date. Numerous studies have looked at how certain environmental parameters shape and influence community structure. Rainfall, temperature, and pH have all been identified as important factors shaping community structure through selection (19–21). The influence of dispersal on microbial communities tends to vary by habitat type, with soils showing little evidence for the importance of dispersal in shaping community structure and dispersal having a greater influence in water and air environments (22, 23). The role of diversification in community assembly has typically been a difficult process to study. However, advances in high throughput sequencing techniques have made it possible to recover population level genomes of microbes from environmental

samples(24). A recent study used genome reconstruction from lake metagenomes across nine years to observe genome wide selective sweeps (25). Ecological drift appears to have its strongest influence when community size is small across, both for microbial and plant communities(26, 27).

Community assembly has historically borrowed terms from plant ecology to describe the different situations in which a community assembles. The first of these terms, primary succession, refers to assembly of a community on a blank slate environment, where there are no species to begin the assembly process. Some have suggested that this term is not as widely useful to microbial communities due to their larger phylogenetic and metabolic diversity in comparison to plants, and advocate defining different succession and assembly patterns based on the resources available at the environment(28). The second of the terms borrowed from plant ecology is secondary succession, or the assembly of communities after the occurrence of a disturbance. This type of succession occur when some type of disturbance shifts microbial community structure and allows new taxa to proliferate in the community. Some authors have advocated for calling microbial secondary succession "post-disturbance" succession, and splitting it into "*post-press* (after a long-term disturbance that impacts multiple generations) and *post-pulse* (after short-term disturbance)"(29).

Disturbance and disturbance response

Disturbances are events that cause some change in an ecosystem/environment. Historically, disturbances in abiotic factors which lead to different fit taxa being selected in the environment have been studied. These studies have shown that factors such warming, nutrient overload, rainfall/drought, and pH changes all have an influence on the resulting community

structure(14, 20, 30, 31). However, disturbances that influence any of the four processes of community assembly could have ramifications for the resulting community structure and function.

When studying a microbial community's response to a disturbance it helps to classify the potential outcomes of the disturbance. Allison and Martiny categorized these different types of responses(32). Given a disturbance, a community that does not change in either structure or function would be described as resistant. Given the same disturbance, a community that changes in structure but not in function could be described as functionally redundant. In this case despite an altered community composition, some metric about the community remains the same. This metric can be any function such as nitrogen fixation or decomposition, and so long as the metric remains the same while the community structure changes the community would be called functionally redundant. Finally, given the same disturbance, a community that changes in either structure or function, but returns to the original state would be called a resilient community. It is important to note that a single community could be functionally redundant for one metric while it may be sensitive to the disturbance for some other metric (i.e. an altered community structure may perform the same in regards to nitrogen fixation given a disturbance, but may perform differently in regards to primary production or respiration) .

It is also possible to calculate indices of resistance and resilience of a microbial community for a particular parameter. Resistance can be thought of as the extent to which a disturbed community's parameter of interest does not change in response to a disturbance after a given lag period(33, 34). Likewise, resilience indices can be calculated that represent the extent of recovery of a microbial community's parameter post disturbance, and is frequently calculated as a rate.

While any environmental factor can be disturbed, it can be helpful to classify types of disturbances. One such way of doing so is to classify the disturbance based on its duration relative to the generation time of the disturbed community. A pulse disturbance is an environmental stress that acts on less than one generation for the community being described(33) and results in ecological change. A press disturbance on the other hand represents a stress that persists for multiple generations of a community and may result in evolutionary changes.

Dormancy and its implications for community assembly

Microbial species can have a particular trait that can greatly influence both their response to disturbance through the different processes of community assembly, dormancy. Dormancy is a state of reduced metabolic activity. Microbes enter dormancy to persist in the face of harsh environmental conditions. Dormancy strategies are widespread throughout the microbial world, though there are particular strategies that are phylogenetically conserved. For instance, Gram positive bacteria of the phylum Firmicutes developed the ability to make endospores, a highly resistance cell that can persist and remain viable in environments for thousands of years(35). This particular form of dormancy is often initiated in response to a suite of environmental factors sensed by histidine kinases. Conversely, some bacteria spontaneously make persister cells, which are cells that have a reduced metabolic state. These persister cells were first observed as cells that were able to survive an antibiotic treatment but after regrowth, remained susceptible to the antibiotic(36).

Dormancy has the potential to influence the four processes of community assembly and consequently microbial community disturbance response. Dormancy can ease the process of selection on microbes by reducing their susceptibility to the abiotic and biotic conditions. Indeed,

viable spores of thermophilic microbes have been found in habitats that are non-permissive to their growth, and as mentioned before, persister cells wait out ephemeral antibiotic treatments(37–39). Likewise, this increased resistance and relaxed selection causes cells in a dormant state to be better passive dispersers as well. Thermophilic endospores have been used as markers of global currents because of their longevity(40). Likewise, the global distribution of *Polaromonas* species across 6 continents at high elevations and in polar environments is thought to be due to the presence of a gene allowing them to enter into a dormant state different from that of thermophilic endospores(41). Efforts have also been made to incorporate dormancy into the island biogeography theory. The island biogeography theory posits that the number of species on an island is governed by the rate of immigration and rate of extinction(42). Accounting for dormancy within this theory would causes higher rates of immigration and lower rates of extinction, causing higher richness of communities(43).

Dormancy's direct influence on diversification is slightly harder to untangle. While dormancy does help cells evade selection, and therefore may be thought to slow evolution by natural selection, it also helps maintain genetic diversity locally. This maintenance of genetic diversity has important ramifications in microbial communities due to the possibility of horizontal gene transfer. Finally, it is unclear what, if any, direct effects dormancy may have on the drift of an assembling community. One potential avenue for influence could be related to community size. Drift is hypothesized and shown to have its largest influence when population or community sizes are small(44, 45). Since dormancy can frequently help increase persistence of microbial cells, it's possible that dormancy actually lowers the impact of drift on microbial communities by maintaining large community sizes.

Some studies have made use of different molecular techniques to look at changes in dormant taxa through time or in response to different disturbances. One such technique is the use of heavy water stable isotope probing (46, 47). In this method, microbial communities are incubated in the presence of isotopically labeled water. Active microbes take up this labeled water and incorporate the "heavy" oxygen atom into their DNA allowing their DNA to be separated from the rest of the communities DNA through density gradient centrifugation. One such study used this method and saw that rare biosphere members were resuscitated from the soil during "rewetting" events (46). These rewetting events act as a disturbance of sorts to the "dried" microbial communities, and thus these findings support a role for dormancy and dormancy transitions in microbial community disturbance response.

Another method for investigating dormant and active communities of microbes is the 16S rRNA:16S rRNA gene ratio methods. This method requires isolating both RNA and DNA from a sample and sequencing both sets of nucleic acids separately. The relative recovery of sequences associated with a taxon in the total RNA of a community vs the total DNA of a community is used to infer the taxon's activity. While a relationship between cellular rRNA content and activity has been observed for pure culture isolates, it is important to note that there are exceptions to this relationship(48), and as such 16S rRNA is indicative more of activity potential, then pure activity(49). Despite these drawbacks, there have been studies showing an agreement between 16S rRNA gene ratio methods and other methods for assessing activity such as differential staining(50). Likewise, a long-term study on salt marshes used 16S rRNA sequencing to investigate the active and dormant communities in response to elevated nutrients, another form of disturbance. This study found that despite total community richness and structure remaining the same in the presence of the nutrient stress, the active microbial community changed

significantly(31). The authors suggested that in this case, nutrient stress induced dormancy in many of the microbial community members.

The Centralia, PA coal fire

This dissertation presents work centered on an atypical disturbance in the town of Centralia, PA. Centralia was originally a coal mining town, but the mines shut down and the locals used abandoned coal strip mines as landfills. These strip mines eventually became filled, and caught fire in 1962(51). The burning trash eventually spread to an exposed coal seam in the landfill. Despite several efforts to extinguish the coal seam fire the state was unsuccessful in controlling the fire and eventually purchased all the land in the area and relocated most of the residents.

The coal seam fire in Centralia burns to this day, and is expected to continue burning for another 100 years(52, 53). The fire slowly moves along the coal seam, warming the overlying soils and depositing them with coal combustion products. As the fire burns all the fuel in a given location, the overlying soils are allowed to cool back down to ambient temperatures and begin the process of recovery. Consequently, the coal fire has left behind a chronosequence of temperature disturbance, where there are currently areas that have never been affected by the fire, soils currently affected by the fire, and areas that at one point in time were affected but have since recovered to ambient temperatures. Previous studies of boreholes and fire affected soils in Centralia showed evidence for reductions in microbial diversity as temperatures increased, and also pointed to elevated levels of ammonium and nitrate in some of these boreholes(54).

Throughout this dissertation, the coal mine fire in Centralia PA is used as a model system for a press disturbance on microbial communities. The coal fire in Centralia, PA is useful model

for multiple reasons. First, it represents an intense disturbance on the soil microbial communities, with coal combustion products being deposited on the surface soils and their temperatures having been measured at >400°C (55). The disturbance is also an appealing system due to the length of disturbance each field site experiences. A given surface soil site may be affected by the fire for years before the temperature at a site begins to recover. Similarly, there are also sites that were at one point in time affected by the fire, but are currently recovered in temperature. This allows for the long term study of recovery dynamics. Finally, the fire is expected to continue burning for over a hundred years(51–53), and while the timeframe is beyond the scope of this dissertation, it provides an opportunity to study disturbance ecology for years to come.

Given the existing chronosequence of disturbance in Centralia, PA, in Chapter 2 we use 16S rRNA gene sequencing of microbial communities along this chronosequence to assess both how microbial communities assemble during disturbance and how well they recover post disturbance. In Chapter 3 shotgun metagenomics of the chronosequence is used to investigate particular traits selected for by the elevated temperature and disturbance in Centralia. Finally, Chapter 4 assesses the influence of dormancy and dispersal on disturbance response and recovery using a mesocosm warming experiment in conjunction with 16S rRNA and 16S rRNA gene sequencing designed to mimic the warming of soils in Centralia, PA. Together these works expand our understanding of how community assembly processes act and interact with one another to govern community disturbance response in soil environments and set the path for future work predicting community outcomes to disturbance.

REFERENCES

REFERENCES

1. Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: The Unseen Majority. Proc Natl Acad Sci USA 95:6578–6583.

2. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039.

3. Högberg P, Nordgren A, Buchmann N, Taylor AFS, Ekblad A, Högberg MN, Nyberg G, Ottosson-Löfvenius M, Read DJ. 2001. Large-scale forest girdling shows that current photosynthesis drives soil respiration. Nature 411:789–792.

4. Kuypers MMM, Marchant HK, Kartal B. 2018. The microbial nitrogen-cycling network. Nat Rev Microbiol 16:263–276.

5. Wasmund K, Mußmann M, Loy A. 2017. The life sulfuric: microbial ecology of sulfur cycling in marine sediments. Environ Microbiol Rep 9:323–344.

6. Pieterse CMJ, Zamioudis C, Berendsen RL, Weller DM, Van Wees SCM, Bakker PAHM. 2014. Induced Systemic Resistance by Beneficial Microbes. Annu Rev Phytopathol 52:347–375.

7. Lugtenberg B, Kamilova F, Lugtenberg B, Kamilova F. 2015. Plant-Growth-Promoting Rhizobacteria Plant-Growth-Promoting Rhizobacteria.

8. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci Early Edit:1–6.

9. Griffiths BS, Ritz K, Bardgett RD, Cook R, Christensen S, Ekelund F, Sørensen SJ, Bååth E, Bloem J, De Ruiter PC, Dolfing J, Nicolardot B. 2000. Ecosystem response of pasture soil communities to fumigation-induced microbial diversity reductions: An examination of the biodiversity-ecosystem function relationship. Oikos 90:279–294.

10. Webster G, Embley TM, Freitag TE, Smith Z, Prosser JI. 2005. Links between ammonia oxidizer species composition, functional diversity and ntrification kinetics in grassland soils. Environ Microbiol 7:676–684.

11. Waldrop MP, Firestone MK. 2004. Altered utilization patterns of young and old soil C by microoorganisms caused by temperature shifts and N additions. Biogeochemistry 67:235–248.

12. He S, Malfatti SA, McFarland JW, Anderson FE, Pati A, Huntemann M, Tremblay J, Glavina del Rio T, Waldrop MP, Windham-Myers L, Tringe SG. 2015. Patterns in wetland microbial community composition and functional gene repertoire associated with

methane emissions. MBio 6:e00066-15.

13. Waldrop MP, Firestone MK. 2006. Seasonal dynamics of microbial community composition and function in oak canopy and open grassland soils. Microb Ecol 52:470–479.

14. Strickland MS, Lauber C, Fierer N, Bradford MA. 2009. Testing the functional significance of microbial community composition. Ecology 90:441–451.

15. Reed HE, Martiny JBH. 2007. Testing the functional significance of microbial composition in natural communities. FEMS Microbiol Ecol 62:161–170.

16. Balser TC, Firestone MK. 2005. Linking microbial community composition and soil processes in a California annual grassland and mixed-conifer forest. Biogeochemistry 73:395–415.

17. Vellend M. 2010. Conceptual synthesis in community ecology. Q Rev Biol 85:183–206.

18. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and Processes of Microbial Community Assembly. Microbiol Mol Biol Rev 77:342–356.

19. DeAngelis KM, Pold G, Topçuoglu BD, van Diepen LTA, Varney RM, Blanchard JL, Melillo J, Frey SD. 2015. Long-term forest soil warming alters microbial communities in temperate forest soils. Front Microbiol 6:1–13.

20. Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A 103:626–631.

21. Evans SE, Wallenstein MD, Burke IC. 2014. Is bacterial moisture niche a good predictor of shifts in community composition under long-term drought. Ecology 95:110–122.

22. Lindström ES, Östman Ö. 2011. The importance of dispersal for bacterial community composition and functioning. PLoS One 6.

23. Bowers RM, Sullivan AP, Costello EK, Collett JL, Knight R, Fierer N. 2011. Sources of bacteria in outdoor air across cities in the midwestern United States. Appl Environ Microbiol 77:6350–6356.

24. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

25. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10:1–13.

26. Lankau EW, Hong PY, MacKie RI. 2012. Ecological drift and local exposures drive enteric bacterial community differences within species of Galápagos iguanas. Mol Ecol 21:1779–1788.

27. Gilbert B, Levine JM. 2017. Ecological drift and the distribution of species diversity. Proc R Soc B Biol Sci 284.

28. Fierer N, Nemergut D, Knight R, Craine JM. 2010. Changes through time: integrating microorganisms into the study of succession. Res Microbiol 161:635–642.

29. Kearns PJ, Shade A. 2018. Trait-based patterns of microbial dynamics in dormancy potential and heterotrophic strategy: case studies of resource-based and post-press succession. ISME J.

30. Bradford MA, Davies CA, Frey SD, Maddox TR, Melillo JM, Mohan JE, Reynolds JF, Treseder KK, Wallenstein MD. 2008. Thermal adaptation of soil microbial respiration to elevated temperature. Ecol Lett 11:1316–1327.

31. Kearns PJ, Angell JH, Howard EM, Deegan LA, Stanley RHR, Bowen JL. 2016. Nutrient enrichment induces dormancy and decreases diversity of active bacteria in salt marsh sediments. Nat Commun 7:1–9.

32. Allison SD, Martiny JBH. 2009. Resistance, resilience, and redundancy in microbial communities. PNAS 105:11512–11519.

33. Shade A, Peter H, Allison SD, Baho DL, Berga M, Bürgmann H, Huber DH, Langenheder S, Lennon JT, Martiny JBH, Matulich KL, Schmidt TM, Handelsman J. 2012. Fundamentals of microbial community resistance and resilience. Front Microbiol 3:1–19.

34. Orwin KH, Wardle DA. 2004. New indices for quantifying the resistance and resilience of soil biota to exogenous disturbances. Soil Biol Biochem 36:1907–1912.

35. Nicholson WL, Munakata N, Horneck G, Melosh HJ, Setlow P. 2000. Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. Microbiol Mol Biol Rev MMBR 64:548–572.

36. Lewis K. 2007. Persister cells, dormancy and infectious disease. Nat Rev Microbiol 5:48–56.

37. Portillo MC, Santana M, Gonzalez JM. 2012. Presence and potential role of thermophilic bacteria in temperate terrestrial environments. Naturwissenschaften 99:43–53.

38. Hubert C, Loy A, Nickel M, Arnosti C, Baranyi C, Brüchert V, Ferdelman T, Finster K, Christensen FM, de Rezende JR, Vandieken V, Jørgensen BB. 2009. A Constant Flux of Diverse Thermophilic Bacteria into the Cold Arctic Seabed. Science (80-) 325:1541–1544.

39. Marchant R, Franzetti A, Pavlostathis SG, Tas DO, Erdbrugger I, Unyayar A, Mazmanci M a., Banat IM. 2008. Thermophilic bacteria in cool temperate soils: Are they metabolically active or continually added by global atmospheric transport? Appl Microbiol Biotechnol 78:841–852.

40. Müller AL, de Rezende JR, Hubert CRJ, Kjeldsen KU, Lagkouvardos I, Berry D, Jørgensen BB, Loy A. 2014. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. ISME J 8:1153–65.

41. Darcy JL, Lynch RC, King AJ, Robeson MS, Schmidt SK. 2011. Global distribution of Polaromonas phylotypes - evidence for a highly successful dispersal capacity. PLoS One 6.

42. MacArthur R, Wilson E. 1967. The Theory of Island Biogeography. Princeton University Press, Princeton, NJ.

43. Locey KJ. 2010. Synthesizing traditional biogeography with microbial ecology: the importance of dormancy. J Biogeogr no-no.

44. Orrock JL, Watling JI. 2010. Local Community size mediates ecological drift and competition in metacommunities. Proc R Soc B Biol Sci 277:2185–2191.

45. Chase JM, Myers JA. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. Philos Trans R Soc B Biol Sci 366:2351–2363.

46. Aanderud ZT, Jones S, Fierer N, Lennon JT. 2015. Resuscitation of the rare biosphere contributes to pulses of ecosystem activity. Front Microbiol 6:1–11.

47. Aanderud ZT, Lennon JT. 2011. Validation of heavy-water stable isotope probing for the characterization of rapidly responding soil bacteria. Appl Environ Microbiol 77:4589–4596.

48. Kramer JG, Singleton FL. 1992. Variations in rRNA content of marine Vibrio spp. during starvation- survival and recovery. Appl Environ Microbiol 58:201–207.

49. Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J 7:2061–2068.

50. Bowsher AW, Kearns PJ, Shade A. 2019. 16S rRNA/rRNA Gene Ratios and Cell Activity Staining Reveal Consistent Patterns of Microbial Activity in Plant-Associated Soil. mSystems 4:1–14.

51. Nolter M a, Vice DH. 2004. Looking back at the Centralia coal fire: a synopsis of its present status. Int J Coal Geol 59:99–106.

52. Elick JM. 2011. Mapping the coal fire at Centralia, Pa using thermal infrared imagery. Int J Coal Geol 87:197–203.

53. Elick JM. 2013. The effect of abundant precipitation on coal fire subsidence and its implications in Centralia, PA. Int J Coal Geol 105:110–119.

54. Tobin-Janzen T, Shade A, Marshall L, Torres K, Beblo C, Janzen C, Lenig J, Martinez A, Ressler D. 2005. Nitrogen Changes and Domain Bacteria Ribotype Diversity in Soils Overlying the Centralia, Pennsylvania Underground Coal Mine Fire. Soil Sci 170:191–201.

55. Janzen C, Tobin-Janzen T. 2008. Microbial Communities in Fire-Affected Soils, p. 299–316. *In* Dion, P, Nautiyal, CS (eds.), Soil Biology Vol 13: Microbiology of Extreme Soils. Springer-Verlag Berlin Heidelberg.

CHAPTER 2: Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire

Work presented in the chapter has been published as

Lee SH*, Sorensen JW*, Grady KL, Tobin TC, and Shade A. Divergent extremes but convergent recovery of bacterial and archaeal soil microbial communities. *The ISME Journal* **11**, 1447-1459 (2017)

*Contributed Equally

Abstract

Press disturbances are stressors that are extended or ongoing relative to the generation times of community members, and, due to their longevity, have the potential to alter communities beyond the possibility of recovery. They also provide key opportunities to investigate ecological resilience and to probe biological limits in the face of prolonged stressors. The underground coal mine fire in Centralia, Pennsylvania has been burning since 1962 and severely alters the overlying surface soils by elevating temperatures and depositing coal combustion pollutants. As the fire burns along the coal seams to disturb new soils, previously disturbed soils return to ambient temperatures, resulting in a chronosequence of fire impact. We used 16S rRNA gene sequencing to examine bacterial and archaeal soil community responses along two active fire fronts in Centralia, and investigated the influences of assembly processes (selection, dispersal and drift) on community outcomes. The hottest soils harbored the most variable and divergent communities, despite their reduced diversity. Recovered soils converged toward similar community structures, demonstrating resilience within 10-20 years and exhibiting near-complete return to reference communities. Measured soil properties (selection), local dispersal, and neutral community assembly models could not explain the divergences of communities observed at temperature extremes, yet beta-null modeling suggested that communities at temperature extremes follow niche-based processes rather than null. We hypothesize that priority effects from responsive seed bank transitions may be key in explaining the multiple equilibria observed among communities at extreme temperatures. These results suggest that soils generally have an intrinsic capacity for robustness to varied disturbances, even to press disturbances considered to be “extreme”, compounded, or incongruent with natural conditions.

Introduction

Human interactions with and alterations of environmental systems are important components of global change (1). Anthropogenic disturbances are outcomes of human activity, and include land-use and land-cover changes, pollution, dispersal of invasive species, and over-harvesting of native animal or plant populations (2). Anthropogenic disturbances are typically classified as press disturbances, as they often impact multiple generations of organisms within their ecosystems (3). Because of their longevity, press disturbances have the capacity to alter ecosystems beyond the possibility of recovery (4).

Within every ecosystem, microbial communities underpin biogeochemical processes, sustain the bases of food webs, and recycle carbon and nutrients. In some situations of anthropogenic disturbance, such as pollution, native microbial communities also can provide bioremediative functions to support ecosystem recovery (5-8). Because of their foundational roles in driving important ecosystem processes, understanding how microbial communities respond to press disturbance can provide insights into the potential for ecosystems to recover. It may also help to uncover mechanisms by which environmental microbial communities may be managed to improve ecosystem outcomes. A better understanding of microbial responses to press disturbances, including examples of communities that have recovered or shifted to an alternative stable state, is necessary to move toward the goal of microbial community management (9) .

Recent work has highlighted the importance of understanding the relative contributions of community assembly processes to community changes (10-16), and these processes can also be informative for understanding community changes after a disturbance (e.g., secondary succession; (12)). According to Vellend, 2010, community assembly can be summarized by four

major processes: dispersal, diversification, drift, and selection. *Dispersal* is the movement of individuals between localities, *diversification* is the generation of new genetic variation (which can lead to speciation), *drift* encompasses the stochastic processes resulting in fluctuations in member abundances (e.g. births and deaths), and *selection* refers to deterministic fitness differences among taxa driven by abiotic conditions or biotic interactions (as summarized by (11)). Together, these processes complement and interact to drive community patterns, and together provide a foundation on which to build a predictive theoretical framework for microbial community ecology.

Because diversification processes are relatively more important at evolutionary scales, Vellend et al. 2014 focused on the remaining processes of ecological selection, drift, and dispersal. They asserted that selection processes are deterministic, that drift processes are stochastic, and that dispersal processes can be either or both, depending on the situation (14). Tucker and colleagues provided clarity to the distinction between deterministic/stochastic and niche/neutral processes, which are often used interchangeably. Niche/neutral refers to the ecological differentiation and equivalence of species, while deterministic/stochastic refers to non-probabilistic or probabilistic outcomes (15). Thus, neutrality concerns ecological equivalence of species, while stochasticity concerns demographic variability in birth, death, and dispersal.

We aimed to understand the responses of soil microbial communities to an anthropogenic press disturbance, and to apply the Vellend, 2010, Nemergut *et al.*, 2013, and Tucker *et al.*, 2016 conceptual frameworks of community assembly for interpretation of patterns. The town of Centralia, Pennsylvania is the site of an underground coal mine fire that has been burning since 1962. It is one of thousands of coal mine fires burning in the world today (17), which are

inconspicuously common anthropogenic disturbances. However, the Centralia fire is especially long-lived, and, after efforts to extinguish it failed, it was left to burn until it self-extinguished (18). The fire is expected to burn slowly until the coal reserves have been consumed. The fire currently underlies more than 150 acres and continues to spread slowly (3-7 m/yr (19)) through underground coal seams. Depending on the depth of the coal bed, it burns at an estimated 46-69 m below the surface (18,19). Heat, steam and combustion products vent upward from the fire through the overlying soils. The surface soil temperatures can exceed 80°C, scarring the landscape with dead vegetation that reveals the fire's subsurface trajectory. As steam and gasses pass through the overlying rock and soil, soil temperatures increase while soil chemical composition is altered by both spontaneous and microbial-mediated chemical reactions (20). As the fire expands into new areas, it also retreats from some affected sites, which then recover to ambient temperatures (18,19). Thus, the "end" of the disturbance can be delineated by temperature recovery. In this way, a chronosequence of fire-affected Centralia soils provides a space-for-time proxy of disturbance response and recovery.

Our research objectives were to understand the diversity and spatio-temporal dynamics of the surface soil bacterial and archaeal communities that have been impacted historically or are currently influenced by the ongoing subterranean coal mine fire in Centralia. We used a definition of disturbance response to include changes in member relative abundances as well as in composition. Previous work using terminal restriction fragment length polymorphism analysis showed that microbial diversity decreased at hotter sites, and that compositional changes were correlated with soil ammonium and nitrate concentrations (21). We move forward from this work to use high throughput sequencing of soil community 16S rRNA genes to quantify the

community dynamics along a chronosequence of fire response and recovery. We specifically investigated the community assembly processes of selection, dispersal, and drift.

Materials and Methods

Study site, soil sampling, soil biogeochemistry and microbial community DNA extraction

We undertook fieldwork in Centralia (GPS: 46°46”24’N, 122°50”36W) on 5-6 October 2014. We collected surface soils to capture the expected maximum changes along a chronosequence of fire recovery (**Figure C.1**). We sampled two fire fronts along gradients of historical fire activity. Fronts are trajectories of fire spread from the 1962 ignition site outward along near-surface coal seams (19). These fronts include surface soils that were previously hot and have cooled, as well as soils that are currently warmed by the ongoing fire. We collected soil from two unaffected, proximate sites as references, seven recovered sites along the gradient, and nine fire-affected sites (18 total soils), and these collections were distributed across both fire fronts. Soil samples were collected from the top 20 cm of surface soil (core diameter 5.1 cm), and were sieved through 4 mm stainless steel mesh. We collected cores only at bare surface soil locations (no vegetation) to minimize the influence of local vegetation and to maximize comparability between soils, as the thermal surface soils generally lacked vegetation. Collected soils were stored on ice up to 72 hr during transport to the laboratory, then stored at -80°C pending further processing. The physico-chemical characteristics of each soil sample (percent moisture, organic matter (500°C), NO_3^-, NH_4^+, pH, SO_4, K, Ca, Mg, P, As, and Fe) were assayed by the Michigan State Soil and Plant Nutrient Laboratory according to their standard protocols (East Lansing, MI, USA, http://www.spnl.msu.edu/). Gravimetric soil moisture was measured after drying the soil at 80°C for 2 days. Fire history was estimated as years since the surface soil

was first hot from the fire, at each sampling location. Fire history observations were measured using either winter snow cover, aerial vegetation photography, or thermal infrared imagery, as collated and reported by Elick, 2011(Figure 3 therein). Soil community DNA was extracted from 0.25 g of soil in three technical replicates using the MoBio Power Soil DNA Isolation Kit according to the manufacturer's protocol (MoBio, Solana Beach, CA, USA). The concentration of the extracted DNA was measured using the Qubit® dsDNA BR Assay Kit (Life Technologies, NY, USA), and DNA amount was standardized for sequencing to 1,000 ng/sample.

Soil cell counts

Direct bacterial and archaeal cell counts were conducted on frozen soil samples based on a protocol to separate cells from soil reported in (22). To dissociate the microbial cells from soil particles, 10 g of soil was mixed with 100 mL of phosphate buffered saline containing 0.5% Tween-20 (PBST). Soil samples were homogenized in a Waring blender three times for 1 min each, followed by a 5 min incubation on ice. Slurries were centrifuged at 1000 x g for 15 min to concentrate soil particulates. Supernatants were set aside and stored at 4°C, and the remaining soil pellets were re-suspended in 100 mL of fresh PBST and blended for an additional 1 min. The soil slurry was then transferred to sterile 250 mL centrifuge bottles and the blender was washed with an additional 25 mL of sterile PBST and added to the slurry before centrifugation at 1000 x g for 15 min. All resulting supernatants for each site were combined, then centrifuged at 10,000 x g for 30 min to pellet cells. Supernatants were discarded, and cell pellets were re-suspended in 10 mL of sterile Milli-q water and 400 mL of 37% formaldehyde to fix cells. 1 mL of cell suspension was then carefully layered over 500 µL of sterile Nycodenz solution (0.8 g/mL in 0.85% NaCl), then centrifuged at 10,000 x g for 40 min. The upper layer was then collected and

cells were pelleted by centrifugation at 20,000 x g for 15 min, then resuspended in 1 mL of sterile 0.85% NaCl. To dissociate remaining soil clumps, cell suspensions were sonicated for 10 s in a sonicating water bath.

Cell suspensions were stained with DTAF ((5-(4,6-Dichlorotriazinyl) Aminofluorescein)) according to (23). DTAF-stained smears were visualized on a Nikon Eclipse e800 microscope (Tokyo, Japan) equipped with a Photometrics Coolsnap Myo camera (Tuscon, AZ, USA), and images were collected using Micro-Manager software (24). Fiji image analysis software was used to adjust background, thresholding, and to conduct particle counts from images (25). Briefly, background correction was completed using an automated rolling ball subtraction with a 35-pixel radius, followed by automatic local thresholding using the Bernsen method with a 12-pixel radius to convert greyscale images to binary. Watershed segmentation was conducted to separate touching nuclei, then particles were counted using the ImageJ *"*Analyze Particles*"* function, excluding anything smaller than 0.1 micron (26).

Quantitative PCR

We performed quantitative PCR (qPCR) using bacterial and archaeal 16S rRNA gene universal primer sets (**Table B.1;** (27)**)**. The reaction mixtures consisted of 10 µL SYBR qPCR Master mix (Quanta Bioscience, Gaithersburg, MD, USA), 0.4 µL each of the forward and the reverse primers (0.4 pM), 2 µL of template DNA, and sterilized deionized water to adjust the final volume of 20 µL. The thermal profile was as follows: initial denaturation at 95°C for 10 s, followed by 40 cycles of denaturation at 95°C for 10 s, annealing at 50°C for 15 s, and extension at 72°C for 40 s. A final dissociation protocol (58°C to 94.5°C, increment 0.5°C for 10 s) was performed to ensure the absence of nonspecific amplicons. The reactions were conducted using

the Bio-Rad iQ5 real time detection system (Bio-Rad, Hercules, CA, USA). Please see the supporting materials for more details as to the qPCR methods.

16S rRNA amplicon sequencing

For each of the 54 DNA samples (18 soils, each with three replicate DNA extractions) and mock community DNA, paired-end sequencing (150 base pair) was performed on the bacterial and archaeal 16S rRNA gene V4 hypervariable region using the Illumina MiSeq platform (Illumina, CA, USA; **Table B.1**; (27). All of the sequencing procedures, including the construction of Illumina sequencing library using the Illumina TruSeq Nano DNA Library Preparation Kit, emulsion PCR, and MiSeq sequencing were performed by the Michigan State University Genomics Core sequencing facility (East Lansing, MI, USA) following their standard protocols. The Genomics Core provided standard Illumina quality control, including base calling by Illumina Real Time Analysis v1.18.61, demultiplexing, adaptor and barcode removal, and RTA conversion to FastQ format by Illumina Bcl2Fastq v1.8.4. Raw sequences were submitted to the GenBank SRA Accession SRP082686.

To estimate sequencing error, mock community DNA was prepared from six different type strains (*D. radiodurans* ATCC13939, *B. thailandensis* E264, *B. cereus* UW85, *P. syringae* DC3000, *F. johnsoniae* UW101, *E. coli* MG1655). The genomic DNA from these type strains were extracted separately using the EZNA Bacterial DNA Kit (Omega Bio-tek, GA, USA) according to the manufacturer's protocol, and then quantified using the Qubit® dsDNA BR Assay Kit (Life Technologies, NY, USA). Each isolates' 16S rRNA sequence was amplified using universal 27F and 1492R primers. Amplification was performed with the GoTaq Green Master Mix (Promega) with the following reaction conditions: 0.4uM each primer, 20-200 ng

template, 12.5ul 2X GoTaq Green Mastermix and nuclease free water to 25 uL final volume. The products were visualized on 1% agarose gels before being cleaned using the Promega Wizard SV Gel and PCR Cleanup System per manufacturer's instructions. Cleaned amplification products were sequenced using the 27F and 1492R primers using the ABI Prism BigDye Terminator Version 3.1 Cycle kit at Michigan State's Genomics Research Technology Support Facility (https://rtsf.natsci.msu.edu/genomics/). Forward and reverse reads were merged using the merger tool in the EMBOSS (V. 6.5.7) package (28). Based on the DNA concentration, size of genomic DNA, and 16S rRNA gene copy number, the final mixture contained 100,000 copies of 16S rRNA gene from each strain. The mock community was sequenced alongside the 54 soils' metagenomic DNA. All sequences are available in NCBI's Short Read Archive (https://www.ncbi.nlm.nih.gov/sra/SRP082686).

Sequence processing

Paired-end sequence merging, quality filtering, denoising, singleton-sequence removal, chimera checking, and open-reference Operational Taxonomic Unit (OTU) picking were conducted using a UPARSE workflow v8.1 (29,30). Open-reference OTU picking was modified for compatibility with the UPARSE pipeline but proceeded as described for open-reference workflows (31). We selected open-reference OTU picking because it allowed us to retain all high-quality sequences, even if they did not match to the reference database. In addition, we expected novel diversity in Centralia, and it was likely that many Centralia sequences would not hit to reference databases. Furthermore, we wanted to create consistent OTU definitions that could be tractable across this study and future work. In the open-reference OTU picking workflow, reference-based OTU clustering first was conducted using the usearch_global

command to cluster sequences with 97% identity to the greengenes database (v 13.8, http://greengenes.secondgenome.com/downloads). Second, de novo OTU picking was performed for any sequences that did not hit the greengenes reference; the usearch command cluster_otus was used to cluster sequences at 97% identity (this step includes chimera checking). The reference-based and de novo OTUs were combined together to create the final dataset. Finally, to reduce the potential effects of candidate contaminant sequences, any sequences in the final dataset that matched 100% to a database of extraneous sequences (found in the mock community) were removed.

Additional analyses were performed with QIIME v. 1.9.1 (32), including alignment with PyNAST (33), taxonomic assignment with the RDP Classifier (34), tree building with FastTree (35), subsampling/rarefaction to an equal sequencing depth, and within and comparative diversity calculations (e.g., UniFrac , (36)). Sequences identified as Chlorophyta, Streptophyta (i.e., Chloroplasts) and Mitochondria were removed before subsampling to an even sequencing depth. Our sequence analysis workflow and computing notes are available on GitHub (https://github.com/ShadeLab/PAPER_LeeSorensen_inprep/blob/master/Sequence_analysis/MockCommunityWorkflow.md). We used the UPARSE workflow (with the recommended 10% divergence filter) for error rate calculation using the mock community (http://drive5.com/usearch/manual/upp_tut_misop_qual.html).

Ecological statistics

We first assessed the reproducibility of evenly-sequenced technical replicates (DNA extraction and sequencing replicates), and found that replicates were similar to one another in measures of within-sample (alpha) and comparative diversity (beta diversity). The average and

standard deviation of weighted nonnormalized UniFrac distances between replicates was 0.319 ± 0.126 with a range from 0.105 to 1.29 (maximum distance between different samples was 4.49; **Figure C.2;** and alpha diversity among technical replicates provided in **Table B.2**). Given the low technical variability, unrarefied technical replicates were collapsed into one combined set of sequences for each soil core to provide more exhaustive sequencing of each soil; these collapsed samples were subsampled to an even sequencing depth (321,000 sequences per soil), and singleton OTUs (observed only once in the dataset) were removed before proceeding with analysis. Within sample-diversity of species richness, Faith's phylogenetic diversity (whole tree method), and comparative diversity of weighted and unweighted UniFrac distance (nonnormalized and normalized, (37,38) were calculated within QIIME. Non-normalized UniFrac distances can fall outside of 0 and 1, while normalized UniFrac distances are bound to 0 to 1; Lozupone *et al.*, 2007 reported no differences in overarching patterns in beta diversity between the nonnormalized and normalized UniFrac (37), and we have found that this holds for our dataset (**Table B.3**). The data were then moved into the R environment for statistical analyses. Briefly, we used vegan functions for multivariate hypothesis testing, fitting environmental vectors to ordinations (envfit), constrained ordination (capscale), and Mantel tests (mantel) and to calculate Pielou's evenness (39); the cmdscale function (stats) for principal coordinates analysis; custom code of neutral models of community assembly (40) as written and implemented by Burns *et al.*, 2015 ("sncm.fit_function.R"); custom R scripts for beta-null model fitting written by Tucker *et al.*, 2016, Appendix 2 therein) modified by our group to include weighted UniFrac beta-null modeling; and ggplot and ggplots2 for plotting (42). Our R script is available on GitHub ("R_analysis" repository in https://github.com/ShadeLab/PAPER_LeeSorensen_ISMEJ_2017)

Results and discussion

Soil physical-chemical characteristics and microbial population size

We measured a suite of contextual data for each sampling site, and asked whether any of those data were correlated with surface soil temperature (**Figure C.3)**. Centralia soils generally represented a wide range of soil chemistry. We did not find strong correlations between measured contextual data and temperature, with the exception of correlations with ammonium and nitrate (Pearson's R = 0.50 and 0.54, respectively; $p < 0.05$). This finding supports previous work in Centralia showing that ammonium and nitrate were elevated at active vents (21). In addition, the pH of recovered sites was consistently lower than reference sites (mean pH = 4.4 and 5.9, respectively), and the hottest soils were more likely to have extreme or disparate values. In two previous reports, soil ammonium, nitrate, and sulfur concentrations were not necessarily correlated with absolute soil temperature values at Centralia, nor to proximity to an active vent; though extreme or disparate chemistry values were sometimes observed at hot sites, values comparable to unaffected sites were also routinely observed (20,21). The authors suggested that duration of fire impact, whether the fire was advancing or receding from the site, and other complex environmental factors were likely contributing.

All soils were within one order of magnitude of 16S rRNA copies per dry mass of soil with fire-affected soils having the highest copy numbers and recovered soils having the lowest, but there were no statistical differences among groups (**Figure C.4A**, Student's t-test all pairwise $p \geq 0.09$). Total number of cells per dry mass of all soil ranged from 10^5 to 10^7 cells per gram of dry soil, but cell counts across fire classifications also were not statistically distinct (**Figure**

C.4B, Student's t-test all pairwise $p \geq 0.09$**)**. Together, these data indicate overall community size is relatively stable across the fire gradient and that any changes in community structure along the fire gradient are due to changes in member abundances rather than to differences in the total number of individuals (community size) among soils.

Sequencing efforts were near-exhaustive for these soils, as assessed by a clear asymptote achieved with rarefaction (**Figure C.5**). A summary of sequencing efforts, as well as a discussion of reference-based and *de novo* OTU taxonomic assignments for fire-affected and recovered soils, are provided in supporting materials.

Selection

To understand the influence of selection (deterministic) processes on community responses, we used surface soil temperatures measured in 2014 to designate categorical groups of communities according to their fire classification. Soils classified as reference and recovered had temperatures between 12 and 15°C (ambient air temperature was 13.3°C at the time of soil collection), while soils classified as fire-affected had temperatures ranging from 21 to 58°C. We hypothesized that within-sample diversity would be lower in fire-affected soils because of the extreme environmental filter of high temperatures, which we expected to result in lower richness and less phylogenetic breadth. Faith's phylogenetic diversity and OTU richness both were lowest and most variable for fire-affected soils, and highest for reference sites (**Figure 2.1;** Student's t-test all pairwise $p < 0.001$). Pielou's evenness had a similar trend, with fire-affected soils having lower evenness than other soils, suggesting that there are a small number of highly dominant OTUs in the fire-affected soils (all pairwise $p > 0.05$, not significant). These results generally agree with studies investigating soil microbial diversity after coal mine reclamation in China and

Brazil, respectively, where the most recovered/reconstructed soils (20 years post-mining in (43) and 19 years of reconstruction in (44)) had highest within-sample diversity and were most comparable to reference sites. Centralia soils are expected to share similar contamination from coal extraction with these mine reclamation soils, but also are distinct because of their thermal conditions and ongoing surface contamination by coal combustion products, such as inorganic gases containing arsenic, selenium, ammonium, sulfur, and hydrogen sulfide, and organic toxins like polycyclic aromatic hydrocarbons (20). Elements within inorganic gases mineralize and deposit around active vents (20). Some coal combustion products, like volatile sulfur and nitrogen compounds, may enrich for microorganisms capable of using them, while other combustion products, like organic toxins, may decrease microbial community size or diversity (20).

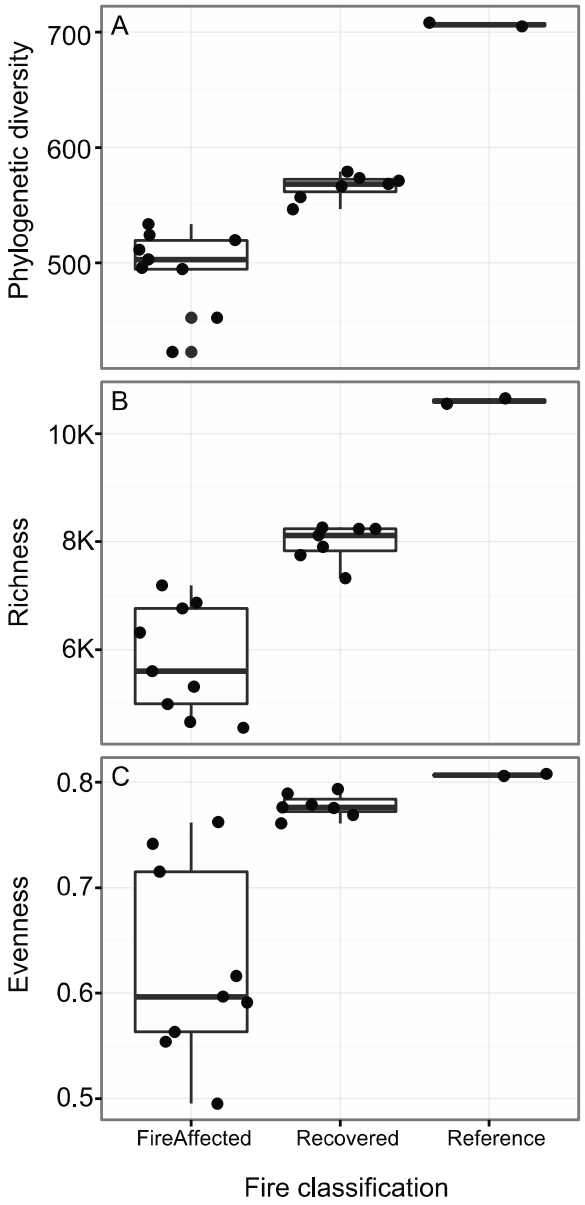

Figure 2.1. Alpha diversity of Centralia soils.

Within-sample (alpha) diversity of fire-affected, recovered, and reference soils in Centralia for bacterial and archaeal community (**A**) Faith's phylogenetic diversity (all $p < 0.001$); (**B**) richness (total no. observed OTUs clustered at 97% sequence identity, all $p < 0.001$); and (**C**) Pielou's evenness (all p not significant).

We used weighted UniFrac distance to assess comparative community diversity across the fire categories. Weighted UniFrac distance was chosen after considering multiple taxonomic and phylogenetic, and weighted and unweighted metrics. All resemblances revealed the same overarching patterns (all pairwise Mantel and PROTEST $p < 0.001$, **Table B.3**), demonstrating that these patterns were very robust. However, weighted UniFrac distance provided the highest explanatory value (**Table B.3**), suggesting that changes in both phylogenetic breadth and the relative abundances of taxa are important for interpreting community responses. As compared to recovered and reference sites, fire-affected soils were distinct (PERMANOVA pseudo $F = 16.10$, $R^2 = 0.50$ and $p = 0.001$ on 1000 permutations) and more variable in their community structure (difference in median dispersions = 0.53, $p = 0.008$; **Figure 2.2**). Differences in surface soil temperature had most explanatory value on Axis 1 (77.1% variance explained by Axis 1, temperature Axis 1 correlation = 0.97, $p = 0.001$, **Table B.4**), with nitrate and iron contributing; calcium and pH (and, to a lesser extent, soil moisture) explained variation on Axis 2 (12.7% variance explained by Axis 2, **Table B.4**). Notably, soil fire history (estimated years since the local soil surface was first measured hot as reported by (19)) was not correlated to community dynamics (**Table B.4**).

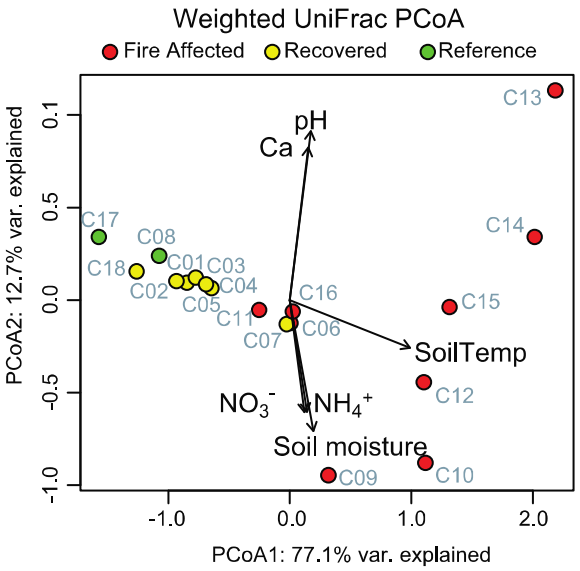

Figure 2.2. PCoA of Centralia Microbial communities based on weighted UniFrac.

Principal coordinate analysis (PCoA) based on weighted UniFrac distances of phylogenetic bacterial and archaeal community structure. Colors show the fire classification of the soil as fire-affected (red), recovered (yellow), or reference (green). The strength of statistically significant ($p < 0.10$) explanatory variables are shown with solid arrows.

Fire-affected soils were more variable in their community structure across soils, especially in soils at the most extreme temperatures observed (sites C13, C10 which were >50°C at the time of sampling and were at the opposite ends of PCoA2). In contrast, recovered soils were less variable, even though they spanned decades of difference in their years of peak fire activity (the earliest impacted soils that we sampled were last recorded to be hot in 1980; (19). Also, recovered soils were very similar in community structure to reference soils. These patterns show that Centralia soils achieve divergent community structures over the transition from ambient to extreme conditions, but then generally converge towards a consistent community structure after the fire subsides. These results also show resilience of soil communities impacted by an extreme press disturbance, with recovery occurring within 10-20 years after the stressor subsided.

We observed a temperature "threshold" effect among fire-affected soils, and soils with temperatures between 21 and 24.5°C (sites C06, C11, and C16) separated cleanly from soils with temperatures greater than 30°C (**Figure 2.2**). To better understand the divergence in community structure among fire-affected soils, we performed a PCoA with these communities (**Figure C.6A, Table B.5**), and also a constrained analysis to ask what variability remained after removing the influence of temperature (**Figure C.6B, Table B.6**). Even after removing the influence of temperature, three discrete subsets of fire-affected communities separated from each other along both axes, with C13 remaining as an outlying point. C13 had very different calcium and pH than the other soils, and both of these factors had high value in discriminating C13 from the other fire-affected soils ($p = 0.092$ and 0.014 respectively). There were no other measured abiotic factors that explained the divergence among the fire-affected soils. In addition, the constrained axes had high explanatory value (**Figure C.6B,** combined axes 1 and 2 = 90.0% var.

explained), suggesting that, given the measured conditions, there are additional processes beyond abiotic selection that explain the differences in these subsets.

We observed broad phylum-level changes in response to the fire (**Figure 2.3**, **Table B.8**). Not all OTUs affiliated with particular phyla had identical responses; however, our analysis of phylum-level responses points to some general trends. In particular, fire-affected soils were enriched for members of Chloroflexi, Crenarcheaota and many lineages of unidentified Bacteria. As compared to the fire-affected soils, recovered soils also were enriched for Parvarchaeota, Bacteroidetes, Elusimicrobia, Gemmatimonadetes, Planctomycetes, Spirochaetes, TM6, and Verrucomicrobia suggesting that members affiliated with this these phyla are able to persist after the fire subsides. Acidobacteria also had an increase in recovered soils (but less significant, $p = 0.10$), presumably because of the decrease in soil pH observed post-fire (**Figure C.3, pH panel: row 1, column 3**). Reference soils had higher representation of Proteobacteria and Verrucomicrobia, which suggests that members of these phyla may be sensitive to the fire.

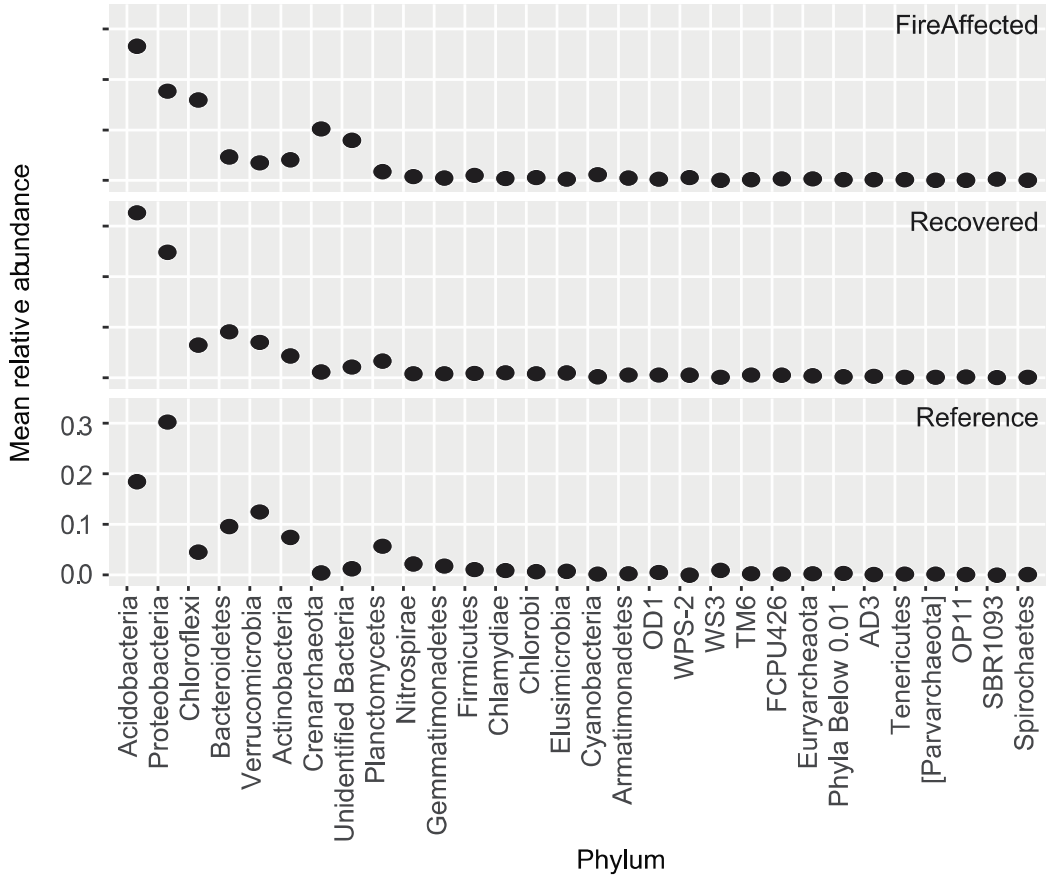

Figure 2.3. Phylum-level responses to the Centralia coal mine fire.

Phylum-level responses to the Centralia coal mine fire. Mean relative abundance of phyla summarized within soil fire classifications (fire-affected, recovered, and reference). Unidentified Bacteria are a combination of OTUs unable to be assigned taxonomy at the phylum level, and are not a monophyletic group. “Phyla Below 0.01” are all OTUs assigned to phyla that collectively comprise less than 0.01 relative abundance in, and also are not a monophyletic group.

Dispersal and drift

To investigate the relative importance of local dispersal, we assessed the value of spatial distance for explaining differences in community structure. If local dispersal were important, we would expect that soils in close proximity would have more similar community structures than soils that are distant from one another. We found no relationship in the measured spatial distances between soil collection sites and their corresponding differences in community structure for all sites (Mantel p = 0.66 on 999 permutations), nor for recovered sites only (after removing the fire-affected sites from analysis; Mantel p = 0.135 on 999 permutations). The lack of evidence for spatial autocorrelation suggests that local dispersal is not a key factor shaping community structure in Centralia soils.

To explore the relative importance of drift in fire-affected and recovered soils, we used two complementary approaches. First, we fitted a neutral model of community assembly. The model predicts taxon frequencies as a function of their metacommunity log abundances, which is one method to consider the influence of drift with the influence of dispersal (calculated as an immigration term, *m*, to the model). The neutral model fit better to the recovered sites than to fire-affected sites (R-squared = 0.53, 0.12 respectively; **Figure C.7, Table B.7**). Furthermore, we found a lower influence of dispersal (lower value of *m)* in the fire-affected sites (**Table B.7**). These differences in fit and generally minimal influence of dispersal suggest that neutral processes play a more minor role in the microbial community assembly of fire-affected sites than they do in the recovered sites.

Next, we asked how observed differences in beta diversity deviate from null expectations. We used abundance-based beta-null approaches to distinguish niche and null processes according to (15), and we extended their approach to also consider community differences in

phylogenetic breadth by applying it to weighted UniFrac distances. In this comparative approach, deviations to and from a permuted null expectation (neutral) are used to interpret the relative influences of neutral and niche processes, respectively. All Centralia communities deviated from neutral, with reference and recovered soils falling closer to neutral expectations than fire-affected soils (**Figure 2.4A**). Fire-affected soils had statistically higher beta-null deviations than recovered soils (both $p < 0.05$ for Bray-Curtis and weighted UniFrac). In the fire-affected soils, there was a consistent increase in niche processes with increasing soil temperature, and the hottest sites deviated furthest from the neutral expectation (**Figure 2.4B**). Accounting for phylogenetic breadth (using weighted UniFrac distance, **Figure 2.4B** suggested relatively less deviation from neutral than accounting for abundance alone (using Bray-Curtis dissimilarity, **Figure 2.4B**), but both resemblances had similar trends (Pearson's $R= 0.71$, $p = 0.001$) and produced identical statistical outcomes. These abundance null deviation results agree with the Sloan neutral model because they suggest that unmeasured niche processes structure soil communities at temperature extremes.

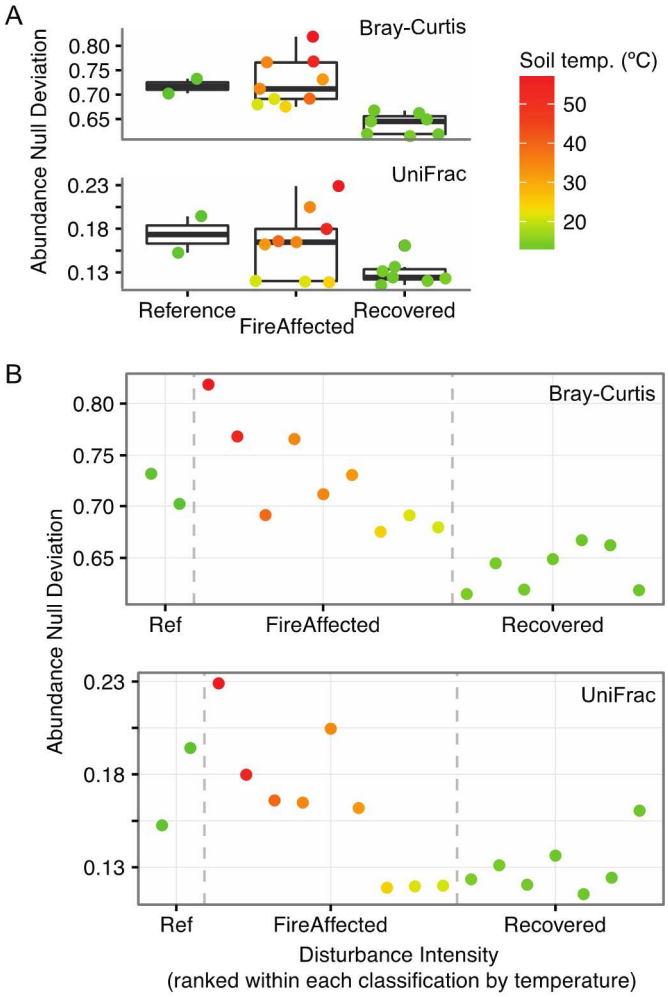

Figure 2.4. Beta-null model deviations in Centralia soil microbial communities.

The relative changes in niche and neutral processes assessed using deviations from abundance-weighted beta-null models. Color gradient shows the soil temperature, as a proxy for disturbance intensity. (**A**) Abundance null deviations by fire classification. For both Bray-Curtis and weighted Unifrac resemblances, recovered and fire-affected communities had distinct null deviations (both $p < 0.05$); (**B**) Trajectory of beta-null deviations ranked by disturbance intensity from reference to fire-affected to recovered soils. Weighted UniFrac and Bray-Curtis trajectories are correlated ($p = 0.71$, $p = 0.001$).

Understanding community divergences at temperature extremes

To dig deeper into the differences in the three subsets of fire-affected soil (**Figure C.6**) that were not well explained by measured abiotic selection, local dispersal, or drift as assessed by the Sloan neutral model of community assembly and beta-null modeling, we asked if there were notable differences in their dominant memberships. Fire-affected soils generally had more variability and greater phylogenetic breadth in their dominant membership than recovered soils, and each fire-affected subset harbored an exclusive membership among their most prevalent taxa. We examined the top 10 prevalent taxa from each of the nine fire-affected soils. Collectively, there were 68 unique top 10 OTUs in fire-affected soils (out of a possible 90, if each of the nine fire-affected soil harbored mutually exclusive membership across their top 10). These prevalent fire-affected OTUs spanned fourteen phyla or Proteobacteria classes, included 30 *de novo* OTUs, and included seven taxa of unidentified Bacteria and two taxa of unidentified Proteobacteria. Acidobacteria OTUs were detected among the top 10 for all fire-affected soils, and eight of nine fire-affected soils included Chloroflexi among the top 10 OTUs. In comparison, recovered soils included ten phyla or Proteobacteria classes among their collective top 10, had no unidentified Bacteria or Proteobacteria, and included four *de novo* OTUs. Acidobacteria and Alphaproteobacteria OTUs were among the top 10 for all recovered soils, and six of the seven recovered soils also included Deltaproteobacteria. Together, these results show that fire-affected soils were more divergent and diverse in their prevalent membership than recovered soils.

An analysis of occurrence patterns of prevalent OTUs also showed greater divergence among fire-affected soils than recovered (**Figure 2.5**), and further supported the distinction among the subsets of fire-affected soils revealed by the constrained ordination **Figure C.6B**). Fire-affected soils had more OTUs within their collective most prevalent taxa, and were more

heterogeneous as shown by the wider range represented by the color scale and the more divergent sample and OTU clustering. In fact, taxa that were among the top 10 in one fire-affected soil were likely to be among the rare biosphere in another fire-affected soil, exhibiting stark contrast in their abundances within these soils. However, most of the top 10 prevalent OTUs were detected within every fire-affected soil (**Table 1, Figure 2.5**), suggesting that changes in taxa relative abundances, rather than turnover in membership, were driving these patterns.

Table 2.1. **Ten most abundant OTUs in fire-affected Centralia soils.**

OTUs (defined at 97% sequence identity) were assigned to the most resolved taxonomic level possible; there were no taxonomic assignments that could be made to these prevalent OTUs below the family level (RDP Classifier confidence > 0.80).

OTU ID	Cumulative % abundance (out of total No. sequences in fire-affected samples)	% occurrence (out of 9 warm or venting fire-affected soils)	Taxonomic assignment
111933	5.5%	100%	Archaea; Crenarchaeota; MBGA
OTU_dn_1	2.5	100%	Bacteria; Chloroflexi; Ktedonobacteria;Thermogemmatisporales; Thermogemmatisporaceae;
OTU_dn_2	2.2	100%	Bacteria; Chloroflexi; Ktedonobacteria;Thermogemmatisporales Thermogemmatisporaceae
242467	2.0	100%	Bacteria; Acidobacteria; DA052;Ellin6513
174835	2.0	100%	Archaea; Crenarchaeota; Thermoprotei;YNPFFA; SK322
61819	1.7	100%	Bacteria; Acidobacteria; TM1
OTU_dn_17	1.5	78%	Bacteria; Proteobacteria; Deltaproteobacteria
215700	1.4	100%	Bacteria; Acidobacteria; Acidobacteriia;Acidobacteriales; Koribacteraceae
OTU_dn_8	1.3	100%	Bacteria
OTU_dn_3	1.2	100%	Bacteria

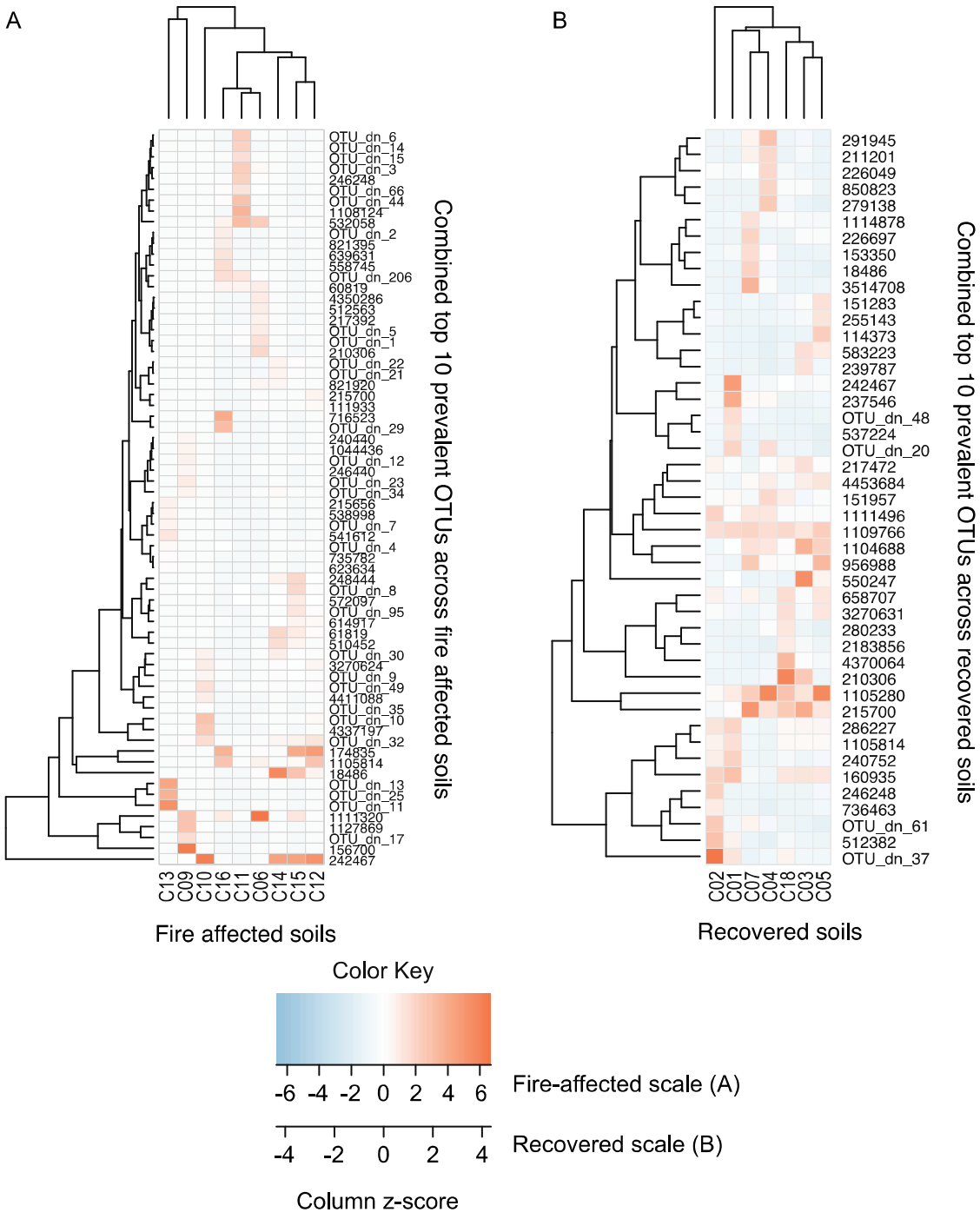

Figure 2.5. Heatmap of "top 10" prevalent taxa in Centralia soils.

Relative abundances of the collection of the most prevalent combined "top 10" taxa (rows) observed in (**A**) fire-affected or (**B**) recovered soils (columns) in Centralia. Color

Figure 2.5. (cont'd)

gradients indicate taxa relative abundances, with warm colors indicating prevalent taxa and cool colors indicating rare taxa within that soil. Note differences in color scale gradient between (**A**) and (**B**). Column labels are sample IDs, and OTU IDs are provided as row labels. OTU IDs that begin "OTU_dn" indicate that the taxon was clustered *de novo* in the open-reference OTU picking workflow; IDs that are numeric indicate that the taxon was assigned with high identity to a reference in the greengenes database. For reference-based OTUs, the numeric identifier corresponds to its representative sequence in the greengenes database. Top dendrograms cluster soils that have similar community structure, and side dendrograms cluster OTUs that have similar occurrence patterns.

This dominance analysis helps to explain the lower fit of the neutral model, and the relatively higher influence of niche processes with beta-null modeling, to fire-affected communities. Outliers to the neutral model that were below detection (taxa that were present in fewer sites than predicted given their relative abundance in the metacommunity) included these many lineages that were prevalent in few fire-affected soils. Taxa that fall below their neutral model prediction have been proposed to be "selected against" or particularly dispersal limited (41). However, in the Centralia extreme environment, we suggest these are taxa that were most successful locally given the thermal disturbance.

Community assembly processes given a press disturbance

Centralia soil communities were sensitive to the coal mine fire, and changed substantially from reference conditions. Selection processes, specifically abiotic soil conditions, offered high explanatory value for Centralia soil community dynamics. These communities first were constrained by environmental filters imposed by the press disturbance, such as thermal temperatures in fire-affected soils and low pH in recovered soils. The fire acts as a strong environmental filter, resulting in decreased diversity and a very different phylogenetic representation among the surviving lineages in fire-affected soils. These environmental filters, such as changes in pH, likely alter the functions of the community as well as its composition. However, even after removing the influence of temperature on fire-affected communities, the communities fell into three distinct subsets that could not be explained by the physico-chemical characteristics measured. Furthermore, neutral modeling, beta-null modeling and lack of spatial autocorrelation suggests that these particular assessments for drift and dispersal processes offer minimal explanation for fire-affected sites. Given the low explanatory value of unweighted

resemblances in describing patterns of comparative diversity (**Table B.3**), and the observation that many of the prevalent taxa detected in some fire-affected soils were rare in other fire-affected soils (**Figure 2.5A**), we can also attribute these patterns to changes in the relative abundances of taxa within a locality, rather than to changes in taxa turnover (differing memberships). Thus, given that neither assessed selection, dispersal, nor drift processes, nor their combination can provide a complete explanation for the divergence of fire-affected communities, the questions remain: why are fire-affected soils so divergent from each other, and how do they eventually manage to recover to the same post-disturbance community structure?

One hypothesis is that the remaining variability in community structure of fire-affected sites may be attributed to priority effects initiated from different local transitions between the dormant seed bank and the active community. The proportion of dormant cells in soils is estimated to be as high as 80% (45), and the importance of dormancy for microbial community assembly processes has been discussed at length (11). Specific to the Centralia coal mine fire disturbance, thermophiles are prime examples of microbial seed bank members that often have been found in environments that are improbable to permit their growth (46-48).

There are two aspects of seed banks that could help to explain Centralia community divergences at temperature extremes: membership and dynamics. If each soil harbored a different seed bank membership, different thermophilic taxa could become active and prevalent in each fire-affected soil, and would manifest as drift influences. This scenario is not well-supported by our data because we detect the dominant members of each fire-affected soil in the other fire-affected soils, albeit in lower abundances. Alternatively, awakenings from the microbial seed bank (49) could result in priority effects at temperature extremes, in which the first fit microorganisms to wake after the fire's local onset have important influence over the

community's ultimate trajectory (50). In our chronosequence study, the outcome of priority effects would appear as divergent community structures at high temperatures that are explained by niche processes. In addition, unknown nuances in local abiotic conditions at fire onset could also set communities onto parallel trajectories and result in multiple equilibria during the press, which would also be explained by niche processes. Our data indirectly support either of these last two scenarios, as the three separate clusters of fire-affected communities suggest multiple equilibria (**Figure C.6B**). It could be that the most similar fire-affected communities began either from the same (or functionally equivalent) waking pioneer taxon, or from the same abiotic conditions (that are similar beyond reaching thermal temperatures), or from some combination of both, which initiated distinct trajectories towards each equilibrium.

Diversification is a fourth community assembly process discussed by Vellend, 2010 and Nemergut *et al.*, 2013. At ecological time scales, diversification was suggested by Vellend *et al.*, 2014 to have relatively lower influence than the other community assembly processes. We do not directly address diversification in this study, focusing instead on ecological processes. Aside from a consistent observation of Acidobacteria and Chloroflexi among the dominant taxa in fire-affected soils, there is no evidence that different but closely related lineages are most prevalent across all fire-affected soils, which may have hinted at distinct but parallel trajectories of diversification within a locality. However, we cannot reject the hypothesis that diversification processes also contribute to divergences in community structure at temperature extremes.

Conceptual model

Extending the conceptual models of (16) and (12), we present a hypothesis of the assembly processes shaping communities before, during, and after an extreme press disturbance. Our model is based on our chronosequence trajectory for beta-null data presented in **Figure 2.4B**, and includes a phase encompassing the press disturbance, which extends beyond the representation of a pulse disturbance as a single time point as typical in previous conceptual models. Our model also incorporates a hypothesis of multiple transient equilibria within the press disturbance phase. We apply the advice of (15) to not use the direction of the change from neutral (positive or negative) to infer specific ecological processes.

We hypothesize that weak variable selection drives stability in heterogeneous Centralia soil communities before the fire (reference sites in **Figure 2.4**; phase 1 in **Figure 2.6**). This is additionally supported by the literature demonstrating generally high heterogeneity and diversity in mature soil microbial communities (51). Next, strong environmental filtering from thermal temperatures (homogeneous selection, phase 2) decreases community diversity at the onset of the press disturbance. The lower diversity and prolonged disturbance conditions permit priority effects initiated by taxa fit in the thermal environment (e.g., thermophiles waking from the seedbank), which set communities onto distinct deterministic trajectories with multiple equilibria during the fire (phase 2). Alternatively, the distinct trajectories and multiple equilibria could have been initiated by unmeasured nuances in abiotic conditions at thermal onset. Finally, weak environmental filtering from increased soil acidity relaxes communities back towards neutral in post-fire conditions (homogeneous selection, phase 3).

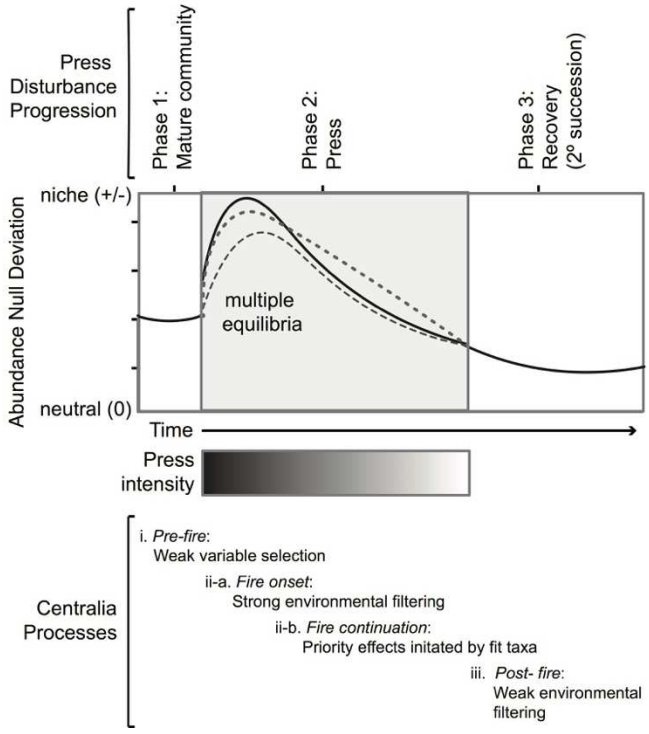

Figure 2.6. Conceptual model of Centralia community assembly.

Hypothesized conceptual model of Centralia community assembly following press disturbance. Phase 1 represents the stable soil community pre-fire, and is characterized by weak variable selection from typical soil heterogeneity and high community diversity. Because the disturbance is a press, phase 2 occurs concurrent with the fire, when strong environmental filters (homogenizing selection) imposed by the extreme conditions drive a sharp increase in niche processes away from neutral conditions at the onset of the fire. Within phase 2, multiple equilibria result from priority effects of pioneer taxa that are fit to survive in the extreme press environment. Phase 3 is post-fire, characterized by relatively weak environmental filtering (e.g., increased in soil acidity) that relaxes communities towards neutral. Complete neutrality was not observed in pre-fire or post-fire soils.

Regardless of the interim dynamics that resulted in community divergence to the stressor, Centralia soils eventually recovered to a community structure very similar to reference soils, and these community structures were explained by the ultimate post-fire soil environment. Our results show that Centralia soil communities, though sensitive to this extreme, complex, and arguably unnatural stressor, had near-complete return to pre-disturbance conditions, and were resilient within ten to twenty years after the stressor subsides. We have no reason to suspect that temperate soils in Centralia are exceptional as compared to other soils. Thus, these results suggest that soils may have an intrinsic capacity for robustness to varied disturbances, even to those disturbances considered to be "extreme", compounded, or incongruent with natural conditions. Understanding the precise functional underpinnings of soil microbial community resilience, including the roles of seed banks in determining that resilience, is a next important step in predicting and, potentially, managing, microbial community responses to disturbances.

APPENDICES

APPENDIX A

Supplemental methods and results

Supplemental Methods

We performed quantitative PCR (qPCR) using bacterial and archaeal 16S rRNA gene universal primer sets (**Table B.1; (1)**). The qPCR was conducted in 20 µL reactions, consisting of 10 µL SYBR qPCR Master mix (Quanta Bioscience, Gaithersburg, MD, USA), 0.4 pM each of the forward and the reverse primers, and 2 µL of template DNA. Triplicate qPCR reactions for each DNA sample was performed. The thermal profile was as follows: initial denaturation at 95°C for 10 s, followed by 40 cycles of denaturation at 95°C for 10 s, annealing at 50°C for 15 s, and extension at 72°C for 40 s. A final dissociation protocol (58°C to 94.5°C, increment 0.5°C for 10 s) was performed to ensure the absence of nonspecific amplicons. The reactions were conducted using the Bio-Rad iQ5 real time detection system (Bio-Rad, Hercules, CA, USA).

To create the standard curve for the primer set, extracted *E. coli* K-12 MG1655 genomic DNA was used to amplify 16S rRNA genes with the 515F and 806R universal primer set (1). The reaction mixtures consisted of 1X final concentration GoTaq® Green Master Mix (Promega), 1 pM each of the forward and the reverse primers, and 1 µL of *E. coli* template DNA, in a 50 µL final volume. The thermal profile was as follows: initial denaturation at 95°C for 10 s, followed by 30 cycles of denaturation at 95°C for 10 s, annealing at 50°C for 15 s, and extension at 72°C for 40 s. Amplified *E.coli* PCR products were purified using Promega Wizard SV Gel and PCR Cleanup System per manufacturer's instructions. Purified PCR amplicons were cloned into the TOPO cloning vectors with a TOPO TA cloning kit (Invitrogen, Carlsbad, Calif.) according to the manufacturer's protocol. Cloned plasmid DNA was extracted using QIAPrep Spin Plasmid Miniprep kit (Qiagen) following manufacturer's protocol, and the concentration was measured using Qubit® dsDNA BR Assay Kit (Life Technologies, NY, USA). A standard curve was then constructed using a 10-fold dilution series of cloned plasmid DNA. Based on the

DNA size for plasmid DNA clone and Avogadro's number (6.02 x 10^{23} molecules per mole), we calculated the copy number of cloned plasmid DNA (where 4.52 x 10^{-3} fg is equal to one plasmid copy). qPCR amplifications were performed in triplicate with a range of concentrations from 18.8 to 1.88 x 10^{8} copies of plasmid DNA using Bio-Rad iQ5 real time detection system, and the observed C_T values were plotted with regression curve using Sigma plot software (**Figure C.8**). Copy number of 16S rRNA genes in each DNA sample was determined based on the observed C_T values calculated by function of regression curve [Y = -3.13x + 41.81, where x is observed C_T value and Y is converted copy number of 16S rRNA gene. The qPCR efficiency, *E*, was calculated based on the slope in the qPCR standard curves as described by (2):

$$E = 10^{\left[-1/slope\right]}$$

According to this calculation, the qPCR amplification efficiency of 16S rRNA gene using EMP primers was 2.08.

To calculate 16S rRNA copies per gram of dry soil, the average copies of the three qPCR technical replicates per DNA extraction was multiplied by the dilution factor (the elution volume of the DNA extraction divided by the microliters added to the qPCR reaction), and then that value was divided by the dry mass of the soil used for the DNA extraction to get copies per gram of dry soil.

Supplemental Results

After quality filtering, our 16S rRNA amplicon dataset produced 5,778,000 high-quality reads (5,776,626 sequences after omitting singletons OTUs) with a UPARSE-calculated error rate of 0.469%. In total, we observed 28,220 OTUs (26,846 when omitting singleton OTUs) defined at 97% sequence identity; approximately one-third of OTUs were defined based on high-

identity matches to the greengenes v13.8 reference database (8,967 OTUs; 8,794 when omitting singleton OTUs), while two-thirds were defined *de novo* after unsuccessful attempts to match the database (19,253 OTUs; 18,052 when omitting singleton OTUs). We observed 65 phyla in Centralia soils.

Though it was not unexpected in a soil ecosystem impacted by an unusual disturbance, the observation of a large proportion on *de novo* OTUs (with the open-reference OTU picking workflow) suggests that Centralia soils may harbor substantial undescribed microbial diversity and functions. Coal mine fire ecosystems have been sources of novel microbial functions, including reported aerobic nitrogen fixation (3) and novel antibiotics (4,5). Furthermore, thermophiles are of interest for bioprospecting for natural products such as thermally-stable enzymes (e.g., for biomass deconstruction from lignocellulosic crops (6) and novel antibiotics (7). Among the *de novo* lineages of interest were several archaeal taxa tentatively identified as Crenarcheaota and Parvarcheaota, and several minor bacterial lineages tentatively assigned as TM6, TM7, OD1, OP11, LD1, WPS-2, and WS-3. A 16S rRNA clone library and T-RFLP study of three soil microbial communities that were each proximate to active coal seam vents in China also reported a proportionally large number of Crenarcheaota among detected archaeal clones (8), suggesting that these may be common inhabitants of soils impacted by long-term fires.

Supplemental References

1. Caporaso, J. G., C. L. Lauber, W. a Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. a Gilbert, G. Smith, and R. Knight. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621–1624.
2. Rasmussen, R. 2001. Quantification on the LightCycler. Rapid Cycle Real-Time PCR: Methods and Applications:p21–34. Springer Berlin Heidelberg

3. Ribbe, M., D. Gadkari, and O. Meyer. 1997. N2 Fixation by Streptomyces thermoautotrophicus Involves a Molybdenum-Dinitrogenase and a Manganese-Superoxide Oxidoreductase That Couple N2Reduction to the Oxidation of Superoxide Produced from O2by a Molybdenum-CO Dehydrogenase. Journal of Biological Chemistry 272:26627–26633.
4. Wang, X., S. I. Elshahawi, K. A. Shaaban, L. Fang, L. V. Ponomareva, Y. Zhang, G. C. Copley, J. C. Hower, C. G. Zhan, M. K. Kharel, and J. S. Thorson. 2014a. Ruthmycin, a new Tetracyclic Polyketide from Streptomyces sp. RM-4-15. Organic Letters 16:456–459.
5. Wang, X., K. A. Shaaban, S. I. Elshahawi, L. V Ponomareva, M. Sunkara, G. C. Copley, J. C. Hower, A. J. Morris, M. K. Kharel, and J. S. Thorson. 2014b. Mullinamides A and B, new cyclopeptides produced by the Ruth Mullins coal mine fire isolate Streptomyces sp. RM-27-46. The Journal of antibiotics 67:571–5.
6. Blumer-Schuette, S. E., S. D. Brown, K. B. Sander, E. A. Bayer, I. Kataeva, J. V. Zurawski, J. M. Conway, M. W. W. Adams, and R. M. Kelly. 2014. Thermophilic lignocellulose deconstruction. FEMS Microbiology Reviews 38:393–448.
7. Garg, N., W. Tang, Y. Goto, S. K. Nair, and W. a. van der Donk. 2012. Lantibiotics from Geobacillus thermodenitrificans. Proceedings of the National Academy of Sciences of the United States of America 109:5241–5246.
8. Zhang, T., J. Xu, J. Zeng, and K. Lou. 2013. Diversity of prokaryotes associated with soils around coal-fire gas vents in MaNasi county of Xinjiang, China. Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology 103:23–36.

Appendix B

Supplemental tables

Table B.1. Primers used in this study.

Primer name	sequence (5' - 3')	Target	target site	Product size (bp)	Tm	Reference
515F	GTGCCAGCMGCCGCGGTAA	16S V4	515-534	291	69.5	Caporaso et al., ISME J. 2012
806R	GGACTACHVGGGTWTCTAAT		787-806		45.1	

Table B.2. Mean and standard deviation of phylogenetic diversity and richness across technical sequencing replicates.

Three replicate DNA extractions, amplifications and sequencing reactions were performed per soil, and, after calculating the technical variability, these sequences were pooled into one aggregate set of sequences to achieve deep coverage of the community within each soil.

SampleID	*PD_mean*	*PD_sd*	*Richness_mean*	*Richness_sd*
C01	393.96	16.22	4073.67	55.77
C02	392.48	9.42	3805.00	48.50
C03	403.12	15.25	4498.67	39.72
C04	374.95	6.51	4420.33	89.51
C05	405.05	14.17	4389.33	109.25
C06	332.89	13.26	3718.67	117.33
C07	371.50	7.80	4253.00	67.01
C08	525.93	5.37	6011.67	191.04
C09	312.71	32.40	2328.33	352.23
C10	267.32	27.06	2128.00	225.08
C11	343.84	12.26	3886.67	81.56
C12	249.92	29.65	2106.67	280.73
C13	316.18	58.27	2471.00	816.28
C14	307.29	16.47	2688.67	232.20
C15	330.40	38.06	3011.67	435.15
C16	356.85	12.24	3546.33	83.93
C17	506.13	19.77	5724.00	179.43
C18	392.64	13.98	4210.67	105.61

Table B.3.

(A) Percent variation explained for PCoA axes 1 and 2 for weighted and unweighted UniFrac, Sorensen-dice, and Bray-Curtis distances/dissimilarities. Nonnormalized Weighted UniFrac was chosen because it was most informative in explaining the variance along the first two axes. (B) Pairwise resemblance correlations calculated with Mantel and PROTEST. All $p < 0.001$ for all tests.

A.

	PCoA1	*PCoA2*
Weighted UniFrac	77.1	12.7
Normalized Weighted Unifrac	74.6	10.9
Unweighted UniFrac	18.3	13.6
Sorensen-dice	20.1	15.2
Bray-Curtis	23.9	13.7

B.

Dist1	**Dist2**	**Mantel_R**
weighted_UniFrac	unweighted_UniFrac	0.63
weighted_UniFrac	normalized_weighted_UniFrac	0.96
weighted_UniFrac	BrayCurtis	0.72
weighted_UniFrac	Sorenson	0.68
unweighted_UniFrac	normalized_weighted_UniFrac	0.61
unweighted_UniFrac	BrayCurtis	0.81
unweighted_UniFrac	Sorensen	0.94
normalized_weighted_UniFrac	BrayCurtis	0.70
normalized_weighted_UniFrac	Sorensen	0.69
BrayCurtis	Sorensen	0.85

Table B.4. Explanatory value of soil contextual data to changes in Centralia soil community structure along PCoA axes for all soils.

Factors significant at $p < 0.10$ are in bold.

	PCoA1	PCoA2	R2	P value	
% explanation	77.1	12.7			
Soil Temperature	**0.968**	**-0.252**	**0.787**	**0.002**	**
NO_3N (ppm)	**0.226**	**-0.974**	**0.290**	**0.067**	.
pH	**0.185**	**0.983**	**0.649**	**0.008**	**
K (ppm)	-0.813	0.582	0.006	0.946	
Mg (ppm)	-0.148	0.989	0.123	0.374	
Organic matter	0.812	-0.583	0.002	0.984	
NH_4N (ppm)	**0.194**	**-0.981**	**0.287**	**0.088**	.
SulfateSulfur (ppm)	0.121	-0.993	0.116	0.372	
Ca (ppm)	**0.182**	**0.983**	**0.529**	**0.022**	*
Fe (ppm)	**0.253**	**-0.967**	**0.271**	**0.094**	.
Fire history	-0.605	0.797	0.253	0.169	
As (ppm)	-0.014	-1.000	0.124	0.404	
P (ppm)	0.435	-0.900	0.093	0.462	
Soil Moisture (%)	**0.263**	**-0.965**	**0.405**	**0.035**	*
Significant codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1					
Number of permutations: 999					

Table B.5. Explanatory value of soil contextual data to changes in Centralia soil community structure along PCoA axes for fire-affected soils.

Factors significant at $p < 0.10$ are in bold.

	PCoA1	PCoA2	R2	P value	
% explanation	70.9	22.0			
SoilTemperature_to10cm	**0.765**	**-0.644**	**0.578**	**0.088**	**.**
NO3N_ppm	-0.002	-1.000	0.328	0.236	
pH	**0.490**	**0.872**	**0.823**	**0.002**	******
K_ppm	0.282	-0.959	0.232	0.429	
Mg_ppm	**0.767**	**0.641**	**0.604**	**0.058**	**.**
OrganicMatter_500	0.407	-0.913	0.218	0.498	
NH4N_ppm	-0.021	-1.000	0.342	0.155	
SulfateSulfur_ppm	-0.216	-0.976	0.118	0.759	
Ca_ppm	**0.613**	**0.790**	**0.694**	**0.015**	*****
Fe_ppm	0.044	-0.999	0.355	0.204	
As_ppm	-0.492	-0.871	0.388	0.228	
P_ppm	0.142	-0.990	0.238	0.453	
SoilMoisture_Per	-0.023	-1.000	0.460	0.143	
Fire_history	0.742	-0.670	0.136	0.637	
Significant codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1					
Number of permutations: 999					

Table B.6. Explanatory value of soil contextual data to changes in Centralia soil community structure along the contrained PCoA axes for fire-affeted soils, after removing the influence of temperature.

Factors significant at $p < 0.10$ are in bold.

	CAP_A1	CAP_A2	R2	P value	
% explanation	64.2	25.9			
SoilTemperature_to10cm	1.000	0.000	0.000	1.000	
NO3N_ppm	-0.973	-0.233	0.354	0.285	
pH	**0.771**	**0.637**	**0.729**	**0.014**	*
K_ppm	-0.416	-0.909	0.093	0.730	
Mg_ppm	0.641	0.767	0.370	0.247	
OrganicMatter_500	0.070	-0.997	0.128	0.613	
NH4N_ppm	-0.962	-0.273	0.367	0.240	
SulfateSulfur_ppm	-0.988	0.154	0.234	0.446	
Ca_ppm	**0.652**	**0.759**	**0.551**	**0.092**	.
Fe_ppm	-0.862	-0.508	0.396	0.355	
As_ppm	-0.948	-0.317	0.378	0.216	
P_ppm	-0.132	-0.991	0.287	0.350	
SoilMoisture_Per	-0.813	-0.583	0.419	0.203	
Fire_history	0.636	-0.771	0.276	0.375	

Significant codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1

Number of permutations: 999

Table B.7. Parameters and fits of neutral models.

Model parameter	*all*	*Fire-affected*	*Recovered*
m	0.04	0.08	0.10
m.ci	0.00	0.00	0.00
m.mle	0.04	0.08	0.10
maxLL	-5838.12	1187.68	-2735.42
binoLL	475.69	1162.47	-143.93
poisLL	475.67	1162.46	-143.94
Rsqr	0.45	0.12	0.53
Rsqr.bino	-1.19	-0.86	-0.47
Rsqr.pois	-1.19	-0.86	-0.47
RMSE	0.20	0.26	0.21
RMSE.bino	0.39	0.38	0.37
RMSE.pois	0.39	0.38	0.37
AIC	-11672.24	2379.36	-5466.85
BIC	-11655.75	2394.86	-5451.16
AIC.bino	955.38	2328.94	-283.86
BIC.bino	971.88	2344.43	-268.17
AIC.pois	955.35	2328.92	-283.88
BIC.pois	971.84	2344.42	-268.19
N	321000.00	321000.00	321000.00
Samples	18.00	9.00	7.00
Richness	28220.00	17097.00	18866.00
Detect	0.00	0.00	0.00
%AbovePred	0.14	0.12	0.13
%BelowPred	0.10	0.07	0.12

Table B.8. Welch's t-tests comparing the mean relative abundances of phyla across fire-affected and recovered soils.

Bold values are significant at $p < 0.05$.

Phylum	T-statistic	DF	p-value
Crenarchaeota	**2.80**	**8.36**	**0.02**
Euryarchaeota	-0.47	11.86	0.65
[Parvarchaeota]	**-3.31**	**11.34**	**0.01**
Unidentified Bacteria	**2.33**	**8.22**	**0.05**
AD3	-1.58	7.28	0.16
Acidobacteria	-1.74	13.64	0.10
Actinobacteria	-0.22	13.12	0.83
Armatimonadetes	-0.58	13.21	0.57
Bacteroidetes	**-4.00**	**9.73**	**0.00**
Chlamydiae	-1.68	10.73	0.12
Chlorobi	-0.43	10.96	0.67
Chloroflexi	**2.82**	**9.67**	**0.02**
Cyanobacteria	1.85	8.07	0.10
Elusimicrobia	**-3.45**	**8.01**	**0.01**
FCPU426	-0.79	11.28	0.45
Firmicutes	0.60	10.97	0.56
Gemmatimonadetes	**-2.24**	**12.33**	**0.04**
Nitrospirae	0.04	12.47	0.97
OD1	-1.28	10.05	0.23
OP11	-1.82	7.56	0.11
Planctomycetes	**-3.33**	**11.61**	**0.01**
Proteobacteria	**-2.42**	**12.89**	**0.03**
SBR1093	2.02	8.00	0.08
Spirochaetes	**-2.43**	**6.68**	**0.05**
TM6	**-2.48**	**7.47**	**0.04**
Tenericutes	0.14	10.06	0.89
Verrucomicrobia	**-3.78**	**10.92**	**0.00**
WPS-2	0.41	10.37	0.69
WS3	-2.26	6.59	0.06
Below_0.01	-0.27	8.39	0.79

Appendix C

Supplemental figures

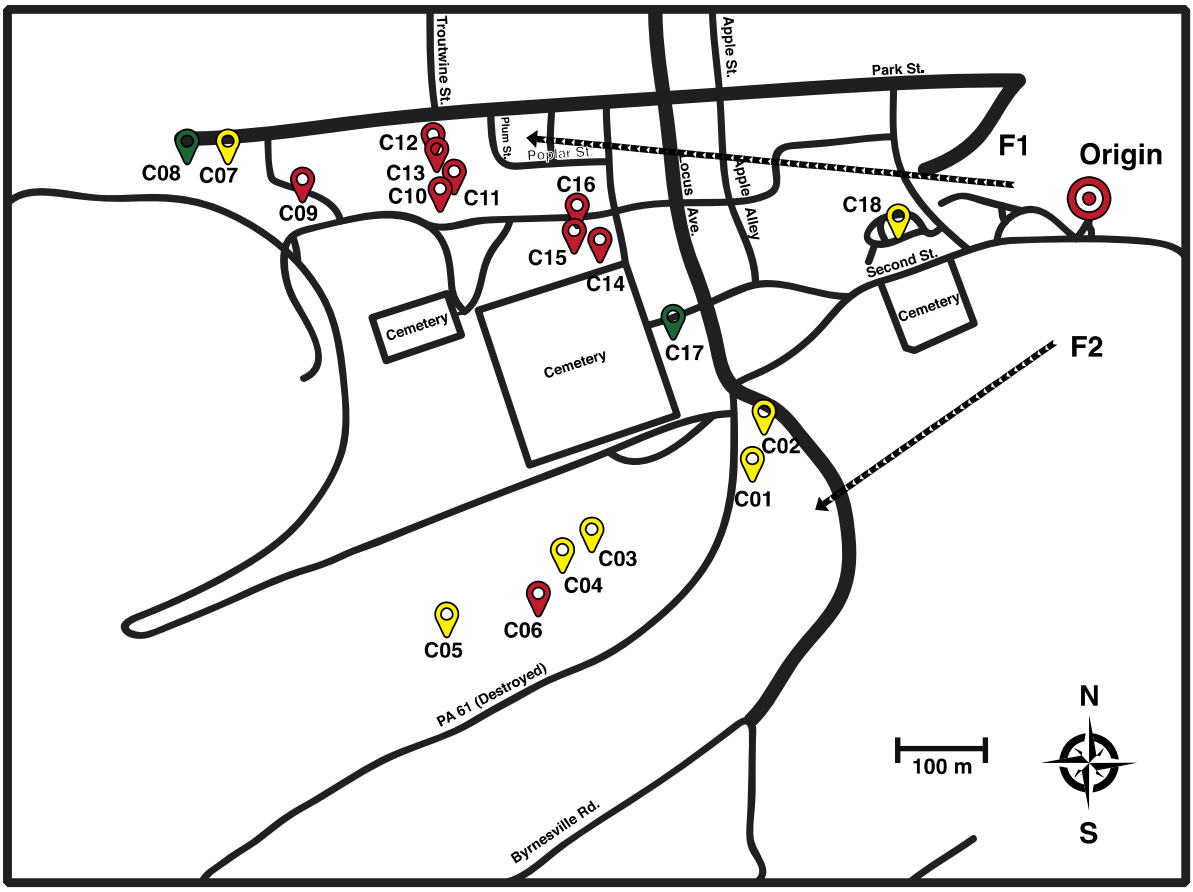

Figure C.1. Soil sampling sites at Centralia mine fire.

In total, 18 surface soil samples (5.08 cm x 20 cm PVC core) were collected along two fire fronts in Centralia, on 15/16 October 2014. Sampling sites encompass a gradient of historical fire activity (red flags: Fire-affected in 2014 (temperature > 21°C); yellow flags: recovered in temperature, post-fire; and green flags: reference soils).

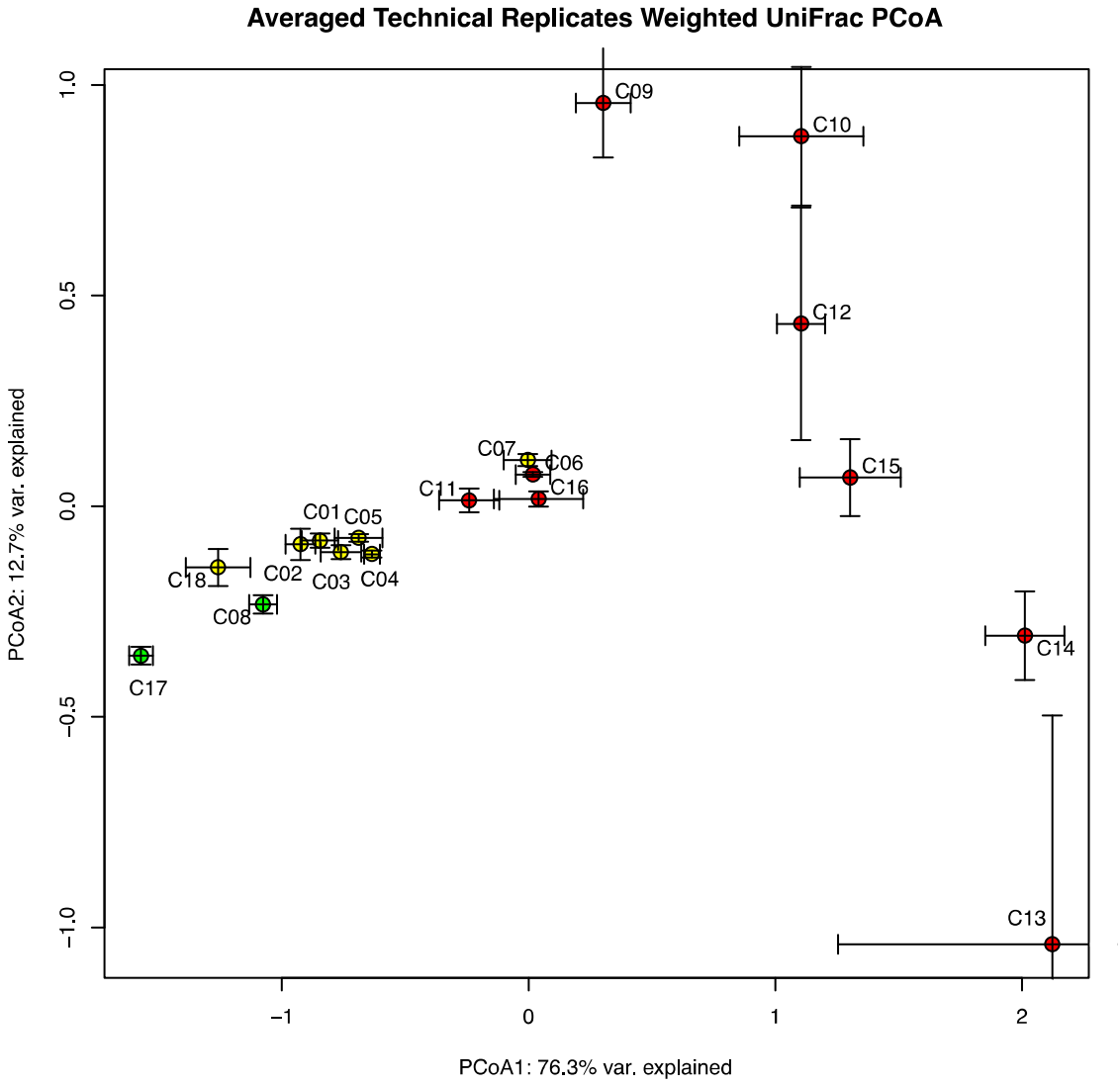

Figure C.2. PCoA showing the variability among technical replicates.

Three replicate DNA extractions, amplifications and sequencing reactions were performed per soil, and these sequences were subsequently pooled into one aggregate set of sequences to achieve deep coverage of the community within each soil. Error bars are standard deviation around the mean weighted UniFrac distance among technical replicates, each subsampled to an even 53,000 sequences per replicate.

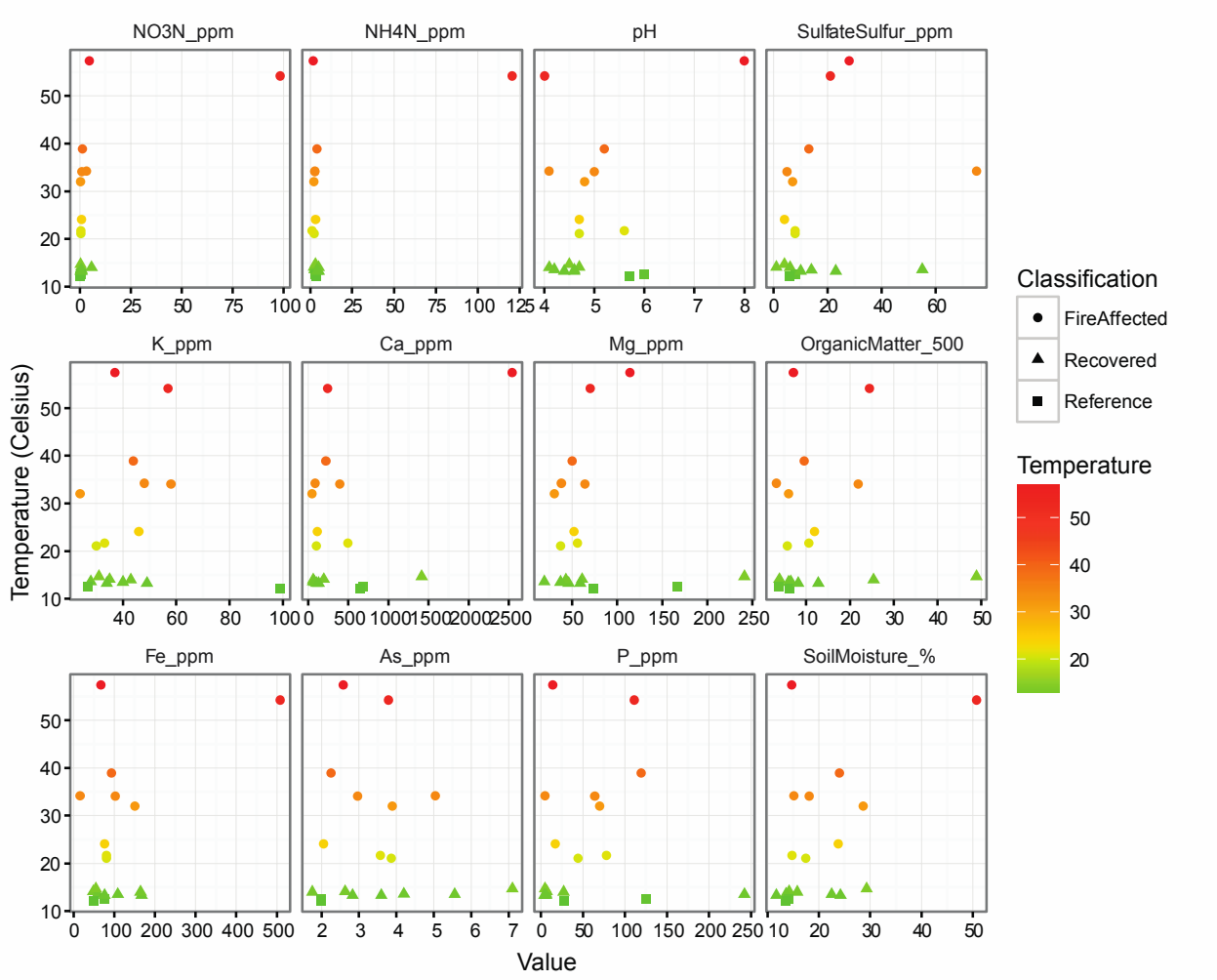

Figure C.3. Soil physical and chemical data plotted against temperature.

Color gradient shows the soil temperature, and symbols show soil fire classification in October 2014 as fire-affected, recovered, or reference.

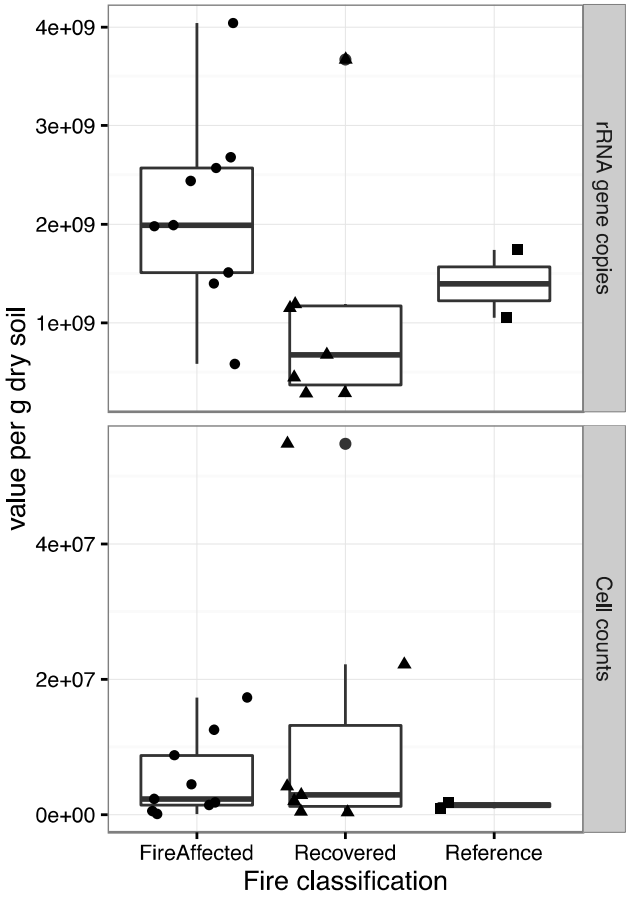

Figure C.4. Community size measurements.

Quantification of (**A**) 16S rRNA copies per gram of dry soil and (**B**) cell counts per gram of dry soil in fire-affected, recovered, and reference soils. 16S rRNA copies were assessed using quantitative PCR, and cell counts were assessed using cell separation from soil, staining and microscope imaging. There were no statistical differences in values across fire classification for either measurement (all pairwise $p > 0.09$ with a student's t-test).

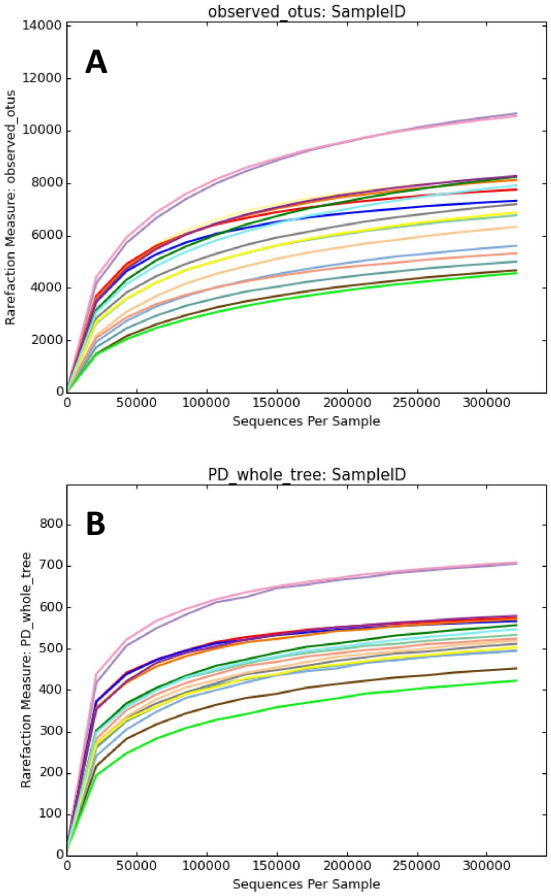

Figure C.5. Rarefaction curves.

Centralia 16S rRNA amplicon sequencing effort assessed by subsampling/rarefaction of (**A**) richness and (**B**) Faith's phylogenetic diversity with increasing total number of sequences.

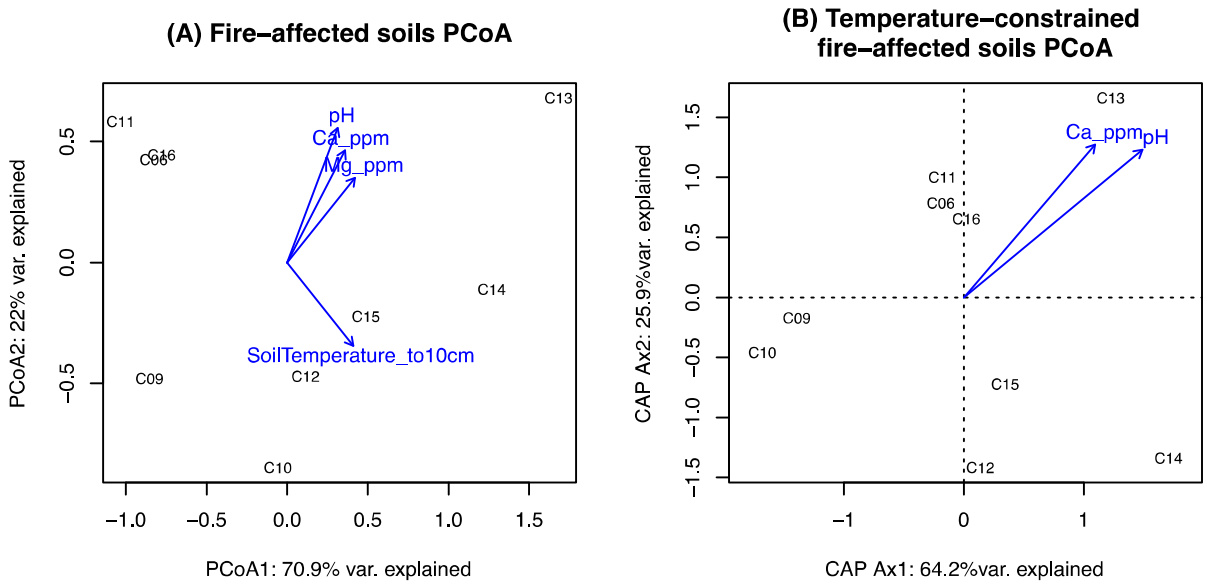

Figure C.6. Divergence in fire-affected soils is not well explained by temperature.

(A) Principal coordinate analysis (PCoA) based on weighted UniFrac distances of phylogenetic bacterial and archaeal community structure in fire-affected soils. The strength of statistically significant ($p < 0.10$) explanatory variables are shown with blue arrows. **(B)** Constrained analysis (CAP) based on weighted UniFrac distances, where the explanatory value of temperature is removed from the analysis to understand the influence of the remaining explanatory variables.

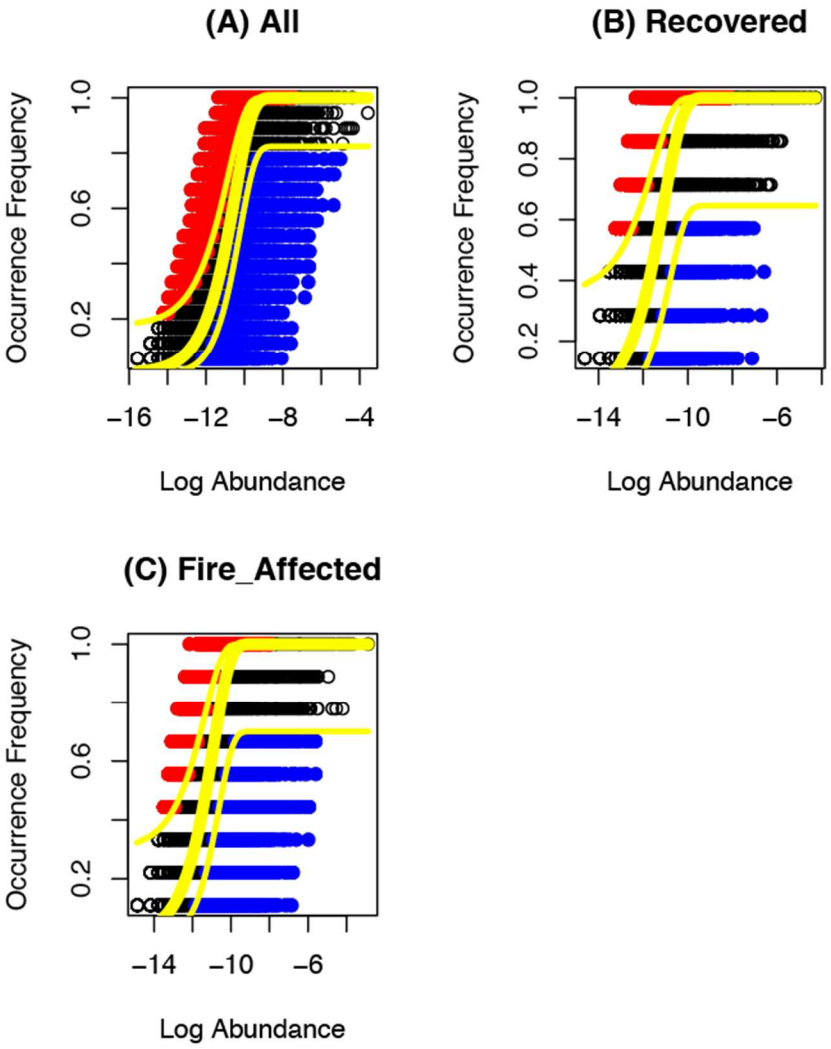

Figure C.7. Neutral models of community assembly.

(**A**) the total community ("All", n= 18), **(B)** recovered soils ("Recovered" n=7), and **(C)** fire-affected soils ("Fire_Affected", n=9). Red symbols show OTUs that had higher abundance than their prediction, and blue symbols show OTUs that had lower abundance than their prediction. The thick yellow line is the neutral model prediction, and the thin yellow lines show a 95% confidence interval around the prediction.

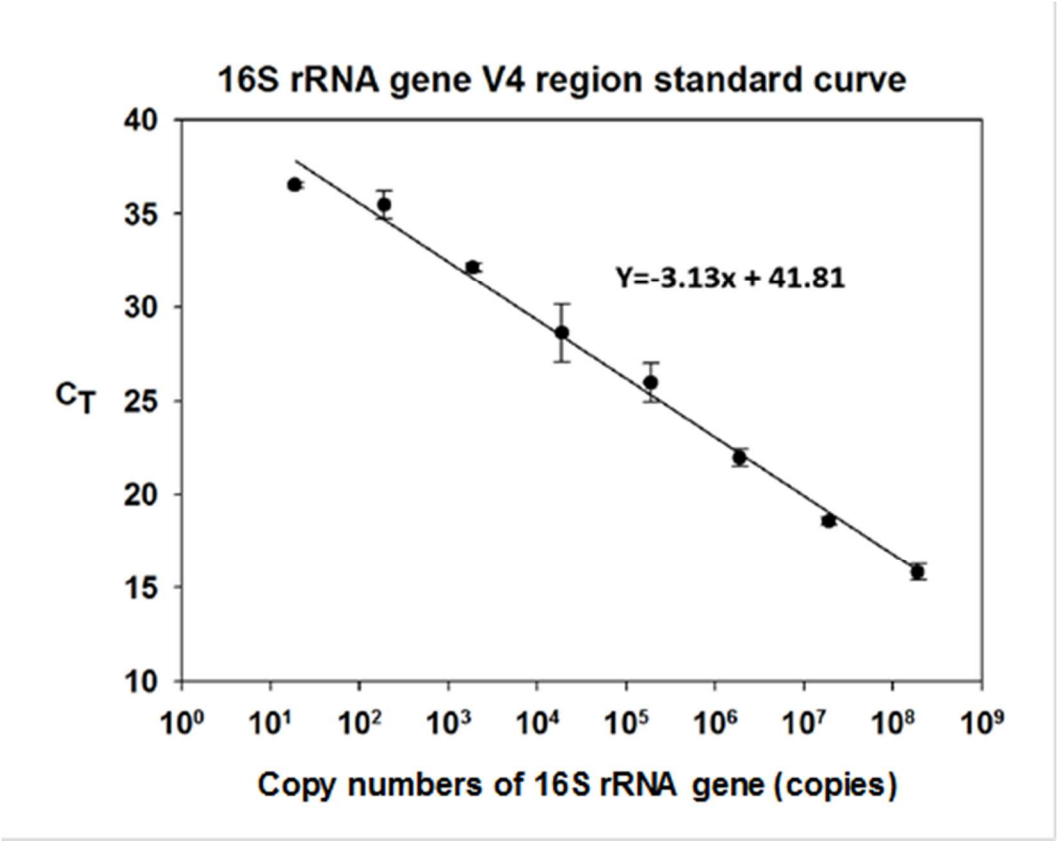

Figure C.8. qPCR standard curve.

Quantitative PCR standard curve for the amount of *E.coli* 16S rRNA gene copies (cloned into plasmids) versus C_T values. The solid line is the regression ($R^2 = 0.988$). The error bars are the standard deviations obtained in three independent experiments.

REFERENCES

REFERENCES

1. Allen MR, Barros VR, Broome J, Cramer W, Christ R, Church JA, *et al.* (2014). IPCC Fifth Assessment Synthesis Report-Climate Change 2014 Synthesis Report. *IPCC Fifth Assess Synth Report-Climate Chang 2014 Synth Rep* pages: 167.

2. Vitousek PM, Mooney HA, Lubchenco J, Melillo JM. (2008). Human domination of Earth's ecosystems. In: *Urban Ecology: An International Perspective on the Interaction Between Humans and Nature*. pp 3–13.

3. Bender EEA, Case TJT, Gilpin ME. (1984). Perturbation Experiments in Community Ecology: Theory and Practice. *Ecology* **65**: 1–13.

4. Thrush SF, Hewitt JE, Dayton PK, Coco G, Lohrer AM, Norkko A, *et al.* (2009). Forecasting the limits of resilience: integrating empirical research with theory. *Proc R Soc B Biol Sci* **276**: 3209–3217.

5. Ruberto L, Dias R, Lo Balbo A, Vazquez SC, Hernandez EA, Mac Cormack WP. (2009). Influence of nutrients addition and bioaugmentation on the hydrocarbon biodegradation of a chronically contaminated Antarctic soil. *J Appl Microbiol* **106**: 1101–1110.

6. Desai C, Pathak H, Madamwar D. (2010). Advances in molecular and '-omics' technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. *Bioresour Technol* **101**: 1558–1569.

7. Ma Y, Rajkumar M, Zhang C, Freitas H. (2016). Beneficial role of bacterial endophytes in heavy metal phytoremediation. *J Environ Manage* **174**: 14–25.

8. Fuentes S, Barra B, Gregory Caporaso J, Seeger M. (2015). From rare to dominant: A fine-tuned soil bacterial bloom during petroleum hydrocarbon bioremediation. *Appl Environ Microbiol* **82**: 888–896.

9. Shade A, Peter H, Allison SD, Baho D, Berga M, Buergmann H, *et al.* (2012). Fundamentals of microbial community resistance and resilience. *Front Microbiol* **3**: 417.

10. Vellend M. (2010). Conceptual synthesis in community ecology. *Q Rev Biol* **85**: 183–206.

11. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, *et al.* (2013). Patterns and Processes of Microbial Community Assembly. *Microbiol Mol Biol Rev* **77**: 342–356.

12. Dini-Andreote F, Stegen JC, van Elsas JD, Salles JF. (2015). Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial

succession. *Proc Natl Acad Sci* **112**: E1326–E1332.

13. Evans S, Martiny JB, Allison SD. (2016). Effects of dispersal and selection on stochastic assembly in microbial communities. *ISME J* 1–10.

14. Vellend M, Srivastava DS, Anderson KM, Brown CD, Jankowski JE, Kleynhans EJ, *et al.* (2014). Assessing the relative importance of neutral stochasticity in ecological communities. *Oikos* **123**: 1420–1430.

15. Tucker CM, Shoemaker LG, Davies KF, Nemergut DR, Melbourne BA. (2016). Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity. *Oikos* **125**: 778–789.

16. Ferrenberg S, O'Neill SP, Knelman JE, Todd B, Duggan S, Bradley D, *et al.* (2013). Changes in assembly processes in soil bacterial communities following a wildfire disturbance. *Isme J* **7**: 1102–1111.

17. Melody S, Johnston F. (2015). Coal mine fires and human health: What do we know? *Int J Coal Geol* **152**: 1:14.

18. Nolter M a, Vice DH. (2004). Looking back at the Centralia coal fire: a synopsis of its present status. *Int J Coal Geol* **59**: 99–106.

19. Elick JM. (2011). Mapping the coal fire at Centralia, Pa using thermal infrared imagery. *Int J Coal Geol* **87**: 197–203.

20. Janzen C, Tobin-Janzen T. (2008). Microbial Communities in Fire-Affected Soils. In: *Microbiology of Extreme Soils*. Springer, pp 299–316.

21. Tobin-Janzen T, Shade A, Marshall L, Torres K, Beblo C, Janzen C, *et al.* (2005). Nitrogen Changes and Domain Bacteria Ribotype Diversity in Soils Overlying the Centralia, Pennsylvania Underground Coal Mine Fire. *Soil Sci* **170**: 191–201.

22. Portillo MC, Leff JW, Lauber CL, Fierer N. (2013). Cell size distributions of soil bacterial and archaeal taxa. *Appl Environ Microbiol* **79**: 7610–7617.

23. Robertson GP, Coleman DC, Bledsoe C (eds). (1999). Standard Soil Methods for Long-Term Ecological Research. Oxford University Press: Cary, NC, USA.

24. Edelstein AD, Tsuchida M a, Amodaj N, Pinkard H, Vale RD, Stuurman N. (2014). Advanced methods of microscope control using μManager software. *J Biol Methods* **1**: 10.

25. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, *et al.* (2012). Fiji: an open source platform for biological image analysis. *Nat Methods* **9**: 676–682.

26. Schneider C a, Rasband WS, Eliceiri KW. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**: 671–675.

27. Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–1624.

28. Rice P, Longden I, Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

29. Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996–8.

30. Edgar RC, Flyvbjerg H. (2014). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**: 3476–3482.

31. Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, *et al.* (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**: e545.

32. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nature* **7**: 335–336.

33. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–267.

34. Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

35. Price MN, Dehal PS, Arkin AP. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.

36. Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

37. Lozupone CA, Hamady M, Kelley ST, Knight R. (2007). Quantitative and qualitative diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* **73**: 1576–1585.

38. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**: 169–172.

39. Oksanen AJ, Blanchet FG, Kindt R, Minchin PR, Hara RBO, Simpson GL, *et al.* (2011).

vegan : community ecology package. *R Packag version 115-1*. http://cran.r-project.org/, http://vegan.r-forge.r-project.org.

40. Sloan WT, Woodcock S, Lunn M, Head IM, Curtis TP. (2007). Modeling taxa-abundance distributions in microbial communities using environmental sequence data. In: Vol. 53. *Microbial Ecology*. pp 443–455.

41. Burns AR, Zac Stephens W, Stagaman K, Wong S, Rawls JF, Guillemin K, *et al.* (2015). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *Isme J* 1–10.

42. Wickham H. (2009). ggplot2: elegant graphics for data analysis. Springer: New York.

43. Li Y, Wen H, Chen L, Yin T. (2014). Succession of bacterial community structure and diversity in soil along a chronosequence of reclamation and re-vegetation on coal mine spoils in China. *PLoS One* **9**. e-pub ahead of print, doi: 10.1371/journal.pone.0115024.

44. Quadros PD de, Zhalnina K, Davis-Richardson AG, Drew JC, Menezes FB, Camargo FA d. O, *et al.* (2016). Coal mining practices reduce the microbial biomass, richness and diversity of soil. *Appl Soil Ecol* **98**: 195–203.

45. Lennon JT, Jones SE. (2011). Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* **9**: 119–130.

46. Hubert C, Loy a., Nickel M, Arnosti C, Baranyi C, Bruchert V, *et al.* (2009). A Constant Flux of Diverse Thermophilic Bacteria into the Cold Arctic Seabed. *Science (80-)* **325**: 1541–1544.

47. MCBEE RH, MCBEE VH. (1956). The incidence of thermorphilic bacteria in arctic soils and waters. *J Bacteriol* **71**: 182–187.

48. Portillo MC, Santana M, Gonzalez JM. (2012). Presence and potential role of thermophilic bacteria in temperate terrestrial environments. *Naturwissenschaften* **99**: 43–53.

49. Buerger S, Spoering A, Gavrish E, Leslin C, Ling L, Epstein SS. (2012). Microbial scout hypothesis, stochastic exit from dormancy, and the nature of slow growers. *Appl Environ Microbiol* **78**: 3221–3228.

50. Fukami T. (2015). Historical contingency in community assembly : integrating niches, species pools, and priority effects. *Annu Rev Ecol Evol Syst* **46**: 1–23.

51. O'Brien SL, Gibbons SM, Owens SM, Hampton-Marcell J, Johnston ER, Jastrow JD, *et al.* (2016). Spatial scale drives patterns in soil bacterial diversity. *Environ Microbiol* **18**: 2039–2051.

CHAPTER 3: Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient

Work presented in this chapter has been published as

Sorensen JW, Dunivin TK, Tobin TC and Shade A. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nature Microbiology* **4**, 55-61(2019)

Abstract

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 in particular to microbes inhabiting complex and highly variable environments like 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 overlying the ongoing Centralia, Pennsylvania (USA) coal seam fire. We found that hot soils harbored distinct communities with small genomes and small cell sizes relative to ambient soils. Hot soils notably lacked genes encoding known two-component regulatory systems and antimicrobial production and resistance. 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.

Main

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, causing surface soil temperatures to reach as high as > 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 with ambient temperature.

We recently assessed 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 from fire ("fire-affected"), previously hot but now recovered to ambient temperatures ("recovered") and never impacted by the fire ("reference"). Fire-affected soils had starkly different community structure from ambient soils. These hot soils also had overlapping 16S rRNA gene compositions but differences in which taxa were most abundant. However, after the fire advanced, soils reasonably recovered towards reference community structure. This suggested a considerable capacity of soil microbiomes for resilience, even after exposure to a severe and unanticipated stressor, and prompted us to ask what microbial attributes underlay the community changes in fire-affected soils.

From twelve metagenomes (six fire-affected, five recovered, and one reference; **Table E.1**), we calculated average genome sizes inclusive of chromosomes and plasmids. Average genome sizes were negatively and strongly correlated with temperature (**Figure 3.1A**, 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 (**Table E.2**). We used three additional methods to assess changes in genome size with soil temperature and found them all to be in agreement (**Figure F.1**). Though unmeasured variables might provide additional information, only temperature was explanatory out of thirteen measured soil variables (**Table E.3**). To the best of our knowledge, this is the first report of decreases in average genome size across an *in situ* temperature gradient that spans physiological requirements from mesophiles to thermophiles.

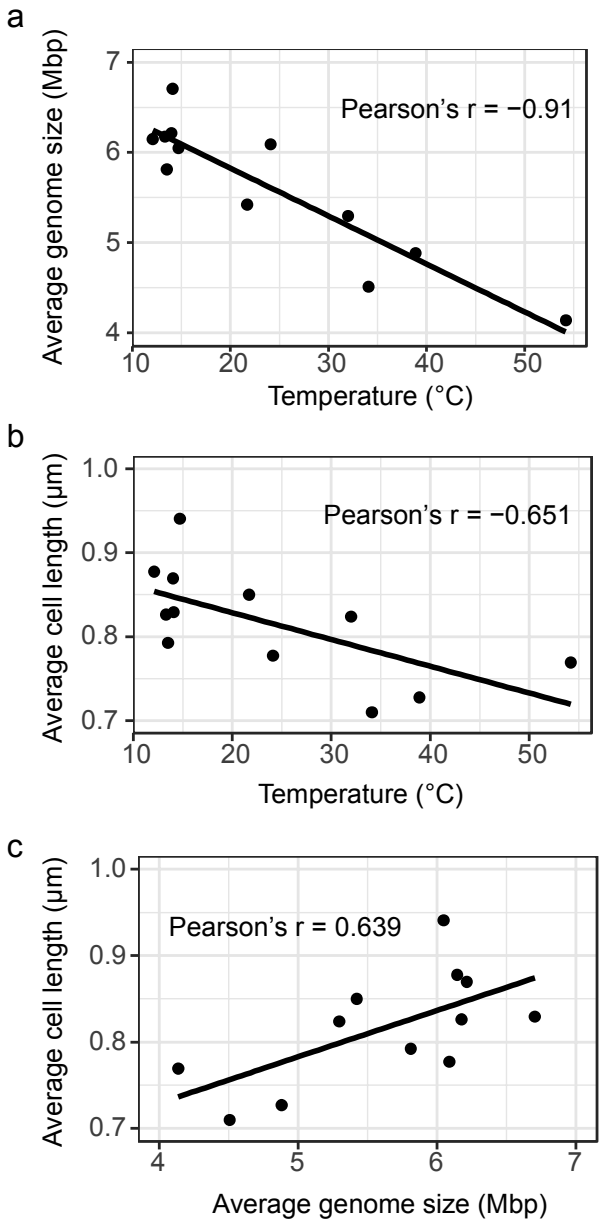

Figure 3.1. Changes in average genome size and cell sizes with temperature.

Changes in average genome and cell sizes across the soil temperature gradient in Centralia. (**A**) Average genome size in each metagenome was calculated using MicrobeCensus and plotted against site temperature (Pearson's correlation $p = 4.095 \times 10^{-5}$). (**B**) Average cell length was measured from 44-910 cells from 3-9 replicate fields for each soil and plotted against soil temperature (Pearson's correlation $p = 0.022$). (**C**) Average genome size had a direct

Figure 3.1. (cont'd)

relationship with average cell size (Pearson's correlation p = 0.025). All Pearson's correlations were two-sided and had n=12 soils.

We next compared average genome sizes estimated from Centralia metagenomes to those from 22 public soil metagenomes (**Figure 3.2A**, **Table E.4**). Generally, hot Centralia soils had small genomes relative to other soils, while 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 , **Figure 3.2B**). Genome sizes from RefSoil were not different across several soil types (**Figure 3.2C**), suggesting a minimal influence of soil type on genome size. While the average genome size in hot Centralia soils is not as small as the soil oligotroph Candidatus *Udaeobacter* (2.81 Mbp (6)), it is significantly smaller than directly-comparable ambient Centralia soils and small relative to other soils (**Figure 3.2A**). 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.

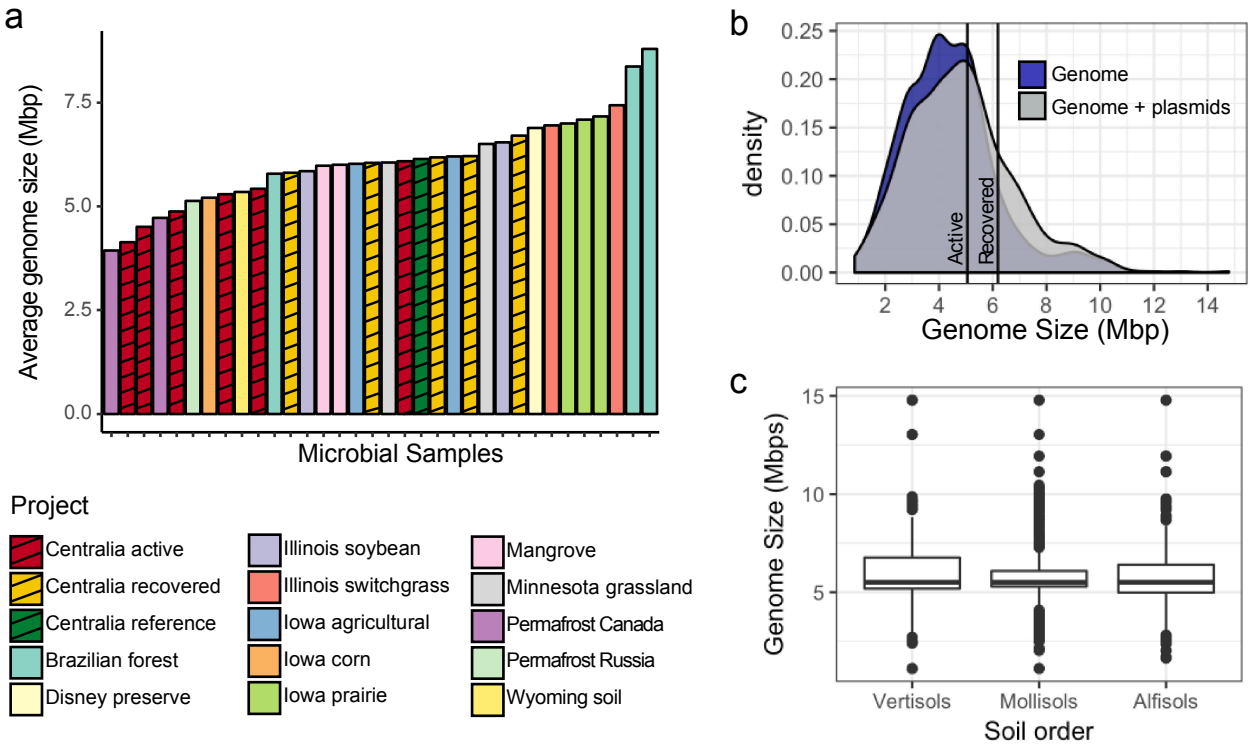

Figure 3.2. Comparison of Centralia genome sizes to other soils.

Comparison of Centralia genome sizes to other soils. (**A**) Comparison to publicly available metagenomes of similar coverage and quality from the database MG-RAST. Average genome size in soil metagenomes, estimated using MicrobeCensus. Samples are ordered by average genome size and colored by sample location. (**B**) Distribution of genome size from cultivable soil microorganisms (RefSoil) with and without plasmids. The mean genome size of Centralia active 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. Midlines of each boxplot corresponds 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.

It was hypothesized by Sabath and colleagues (2013) that small cells may be selected to minimize cellular maintenance costs at high temperatures and that small cells indirectly select for small genomes (3). We re-analyzed microscope images from soil cell counts in Centralia (10) to extract size information. Average cell sizes were also negatively correlated with temperature (**Figure 3.1B**; Pearson's R = -0.65, p =0.021, n = 12 soils, **Table E.5**). Accordingly, cell size correlated with genome size **(Figure 3.1C**; 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 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 (e.g., 3,14). As we previously reported (10) and as confirmed by this work using phylogenetic inference of genome size (**Figure F.1B**), there were stark changes in community structure between fire-affected and ambient soils (**Figure F.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; **Figure F.1C, Table E.6**), which represent some of the most abundant taxa, we asked if small MAGs typical of hot Centralia soils were relatives of thermophiles or lineages that have characteristically small genomes (**Figure 3.3**, **Figure F.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 not described as thermotolerant (**Figure 3.3A**). Taxonomy could not be assigned to 51 (out of 104) MAGs beyond the phylum level, and two Bacteria were unable to be assigned beyond the domain level, suggesting previously undescribed taxa (**Figure 3.3B**, **Table E.6**). For some phyla,

Centralia MAGs trended small relative to the median genome sizes of isolate references (e.g., Acidobacteria and Actinobacteria; **Figure 3.3B**), although there were exceptions (e.g, 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.

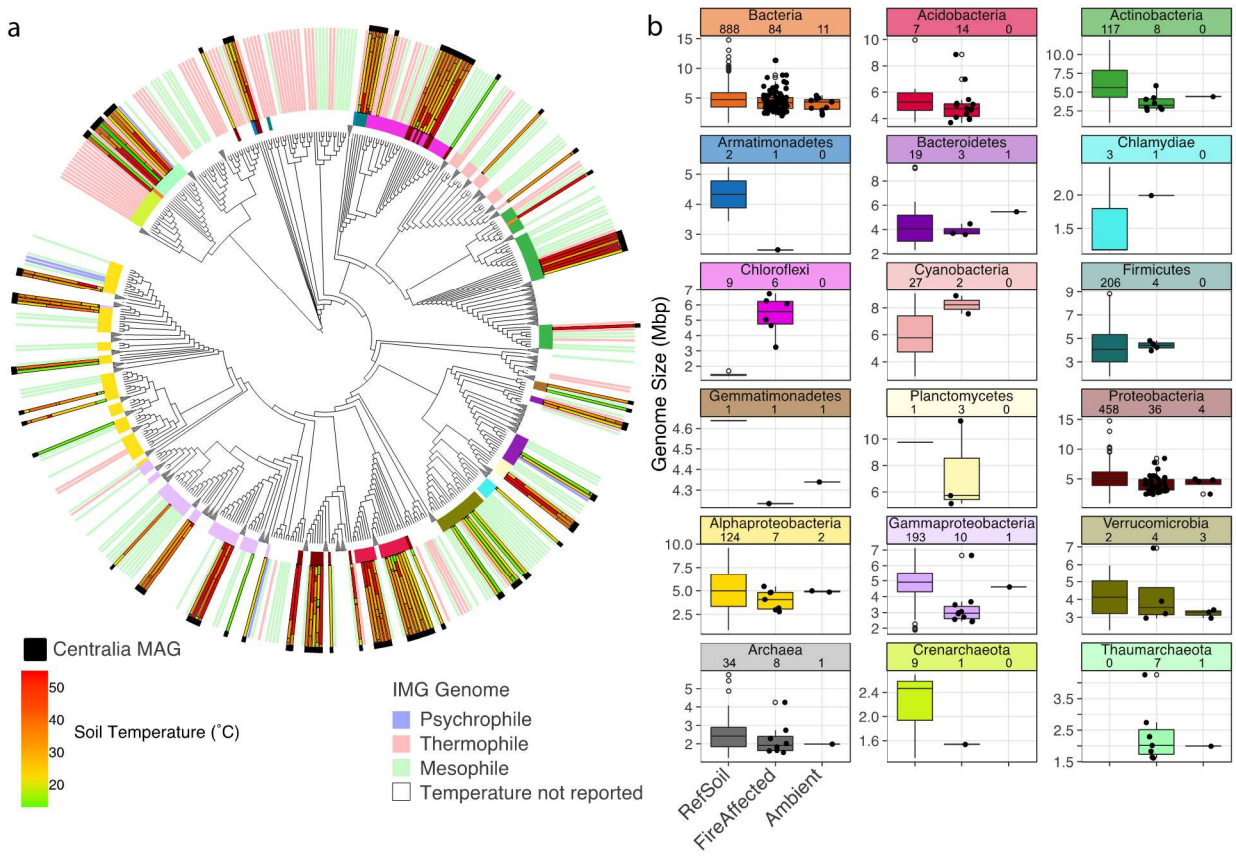

Figure 3.3. Distribution and diversity of Centralia MAGs in comparison to IMG and RefSoil database.

Temperature distributions and diversity of Centralia MAGs compared to reference soil genomes from IMG and the RefSoil database. (**A**) Microbial reference phylogeny based on single-copy (aka "marker") genes (45) that was expanded to include Centralia MAGs. For clarity, large clades that did not contain MAGs are collapsed. The inner color ring shows phylum-level taxonomy, matched to phyla in panel B. The outer color ring shows the temperature reported for IMG reference lineages (muted) and the distribution and measured soil temperatures for Centralia MAGs (bright, black flags). (**B**) Genome sizes of RefSoil isolates compared to 104 of the highest-quality Centralia MAGs from fire-affected and ambient soils (taxonomy assigned by MiGA (44)). Sample sizes indicated in the panel headers are the total number of RefSoil

Figure 3.3. (cont'd)

genomes or Centralia MAGs detected within each lineage. Note differences in y-axis ranges. Because the many of the highest-quality MAGs assembled from hot soils, **Figure 3.3B** does not provide robust MAG comparisons across Centralia fire impact categories. Midlines of each boxplot corresponds to the median values. The top and bottom of each boxplots represent the 75^{th} and 25^{th} percentiles respectively. The upper and lower whiskers extend to the furthest values that are not outliers. Sample numbers for each box plot are indicated in the panel headers and refer to either genomes in the RefSoil database or MAGs detected in this study.

We used metagenome annotations from the KEGG module (KM) database to determine changes in functional genes with increasing temperature. KMs are groups of KEGG Orthologs (KOs) that represent complexes, functional sets, metabolic pathways, or signatures. Eighty-one percent of KOs detected in Centralia metagenomes were detected in all 12 soils, and many patterns with temperature were attributable to changes in normalized KO abundance rather than in KO detection. In total, 284 (out of 541 detected; 52.50%) were correlated with temperature (**Figure 3.4, Table E.8, Supplementary Results**).

Twenty-seven KMs were positively correlated with temperature (Pearson's $R > 0.656$, false discovery rate < 0.05; **Figure 3.4A, Table E.8**). Anaerobic processes, including dissimilatory sulfate reduction (M00596), dissimilatory nitrate reduction (M00530) and denitrification (M00529), were enriched in hot soils (**Figure 3.4A**, *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 likely causes inundated and anaerobic microhabitats. Prior work in Centralia indicated an importance of these metabolisms: sulfur, sulfate, nitrate and ammonium 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 KMs had similar dynamics to several archaeal proteins (**Figure 3.4A,** *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 sulfate reducers (19) and has nine soil reference genomes that average 2.26 Mbp (**Figure 3.3B**). Together, these data suggest that pathways enriched in small genomes from hot soils encode functions attuned to the Centralia habitat.

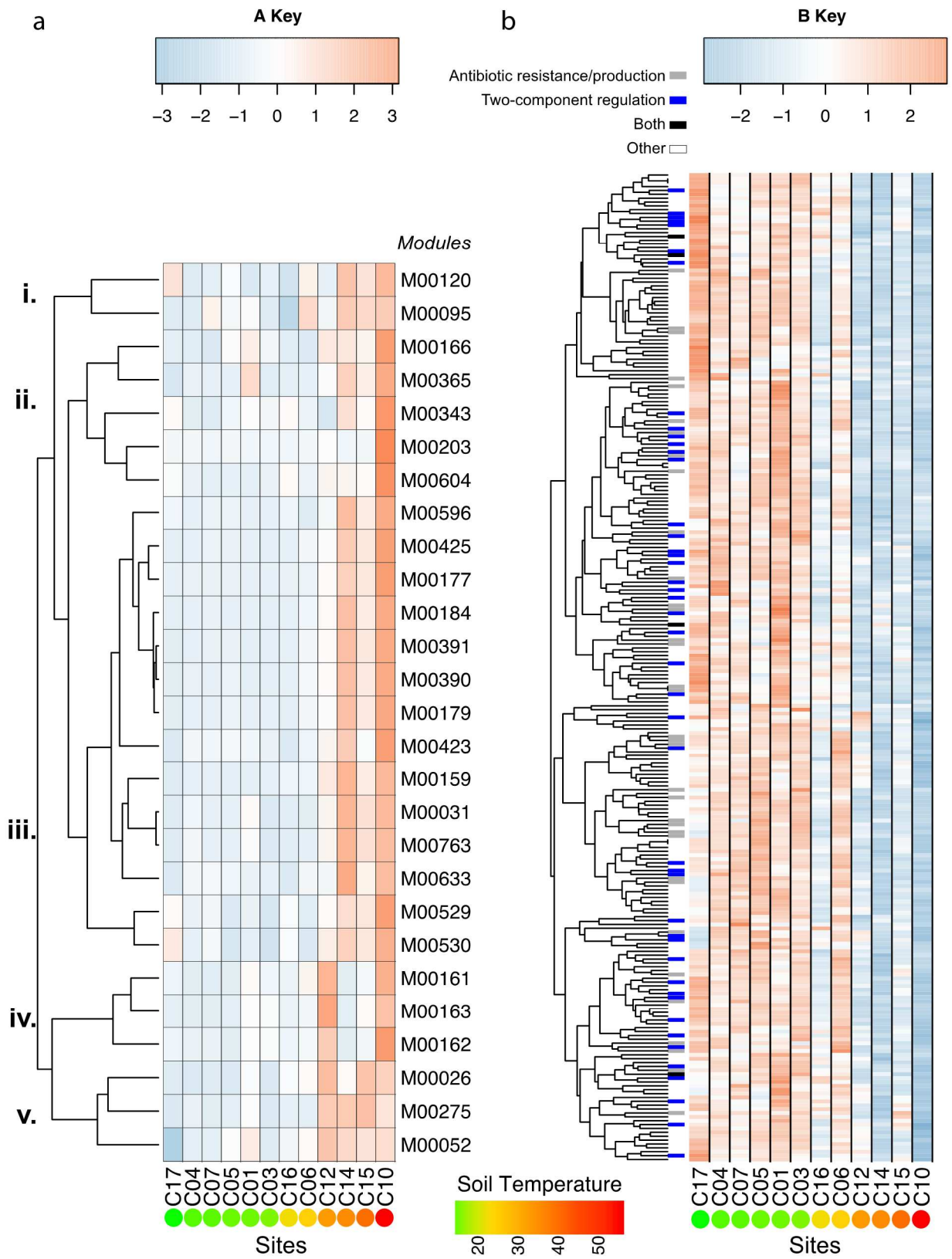

Figure 3.4. KEGG modules correlated with temperature.

(Figure 3.4. cont'd)

Modules (rows) are centered and standardized across Centralia metagenomes (columns), with warm colors showing relative enrichment and cool colors showing relative depletion. Modules with significant relationships with temperature are shown. Sites are arranged by increasing temperature from left to right. (**A**) 27 KEGG modules were positively correlated with temperature (Pearson's R range = 0.646 to 0.933, n = 12 soils). (**B**) 257 KEGG modules were negatively correlated with temperature (Pearson's R range = -0.642 to -0.925, n = 12 soils). A third of the KEGG modules negatively correlated with temperature were either two-component regulatory systems (blue dendrogram tips), antimicrobial resistance or production (gray tips), or both (black tips). Note differences in color gradient ranges across panels A and B. Row dendrograms show the hierarchical clustering of KEGG modules by response patterns to the temperature gradient. Numbers (e.g. i, ii) denote clusters of KEGG modules with similar response patterns to the temperature gradient. Full information on correlation statistics for each KEGG module is listed in **Table E.8**.

Temperature was negatively correlated with 257 KMs (47.5%, Pearson's R < -0.6, false discovery rate < 0.05; **Figure 3.4B**, **Table E.8**). In general, depleted KMs were detected across ambient soils. Of note were antimicrobial resistance and production and two component regulatory systems, which comprised 32.7% of KMs negatively correlated with temperature (84 out of 257, **Figure 3.4B**). This trend was striking, but some KMs belonging to these categories had no relationships with temperature and these KM categories were always detected in fire-affected soils.

Thirty-nine modules for antimicrobial production and resistance were negatively correlated with temperature, which agrees with our prior 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 encoding resistance to described antimicrobials. Like most databases, KEGG is biased towards genomes and genes from fast-growing mesophiles and may miss annotation of under-described thermophile antimicrobial genes. However, 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), suggesting that there is a smaller pool of potential competitors inhabiting the hot soils.

Forty-nine detected two-component regulatory system 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 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 (**Figure F.3**), providing evidence that the soils experience similar environmental stability in temperature, albeit at different ranges. This suggests that wild small genomes are not necessarily conditional on stable environments (7) and invites investigation of whether two-component regulatory systems are consistently less prevalent among thermophiles.

Our cultivation-independent field study supports cultivation-dependent studies that suggest higher temperatures support growth of bacteria and archaea with small genomes (3). Surprisingly, it also suggests that microbial populations inhabiting complex environments, like soils, may generally reflect similar overarching traits in genome size as 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) via abiotic environmental filtering, not the evolutionary process of natural selection towards streamlining. Though taxa enriched in hot soils characteristically had smaller genomes and cells, there is no evidence for contemporary genome streamlining in

Centralia. Rather, we suspect that these thermotolerant cells were resuscitated from the vast dormant pool in soil. This is supported by three lines of evidence. First, there was turnover in community membership across hot and ambient Centralia soils (10), providing evidence against contemporary streamlining within local lineages. Second, many of the lineages that we detected in high abundance in certain hot sites were also detected in low abundance in other sites, including ambient sites (**Figure 3.3A** and (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, suggesting that thermophilic lineages are widespread but typically inactive (16,34,35). Therefore, we posit that Centralia small genomes 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 microbes. 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 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.

Materials and 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 (MoBio, Solana Beach, CA, USA) according the manufacturer's instructions. To briefly summarize the published methods we used (5), after the four cycles of freeze thawing, 10 mL of a Phenol-chloroform-isoamyl alcohol mixture (25:24:1) was added to each sample, mixed and centrifuged at 7,500 g for 10 minutes. After precipitation, DNA was pelleted via centrifugation at 7,500 g for 15 min. The pelleted DNA was resuspended in 100 μL of TE buffer (10mM Tris-HCl, 1mM EDTA•Na_2. The resulting manually extracted DNA was then purified using the MoBio DNEasy PowerSoil kit per the manufacturer's instructions, omitting the 10 min vortexing step after adding 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 base pairs. Samples had between 19Gbp and 50Gbp of sequence data. Additional methodology details are provided in Supporting Materials.

Quality control, assembly and annotation

Adapters were removed and quality trimmed at values less 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 40bp after trimming, or an average quality score less than 8. Reads matching 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 kmer size ranging from 31-121 with a k-step of 10. Coverage of resulting contigs was estimated using seal to map all reads onto the contigs.

To use all sequencing data, we worked with assembled and unassembled reads processed by Integrated Microbial Genomes (IMG) using their standard annotation pipeline (38). After comparing several annotation methods (Supplementary Discussion), we chose to use the KEGG Orthology database (39) for analyzing the Centralia data due to its inherent structure and ability to integrate metabolic pathways. KEGG Ortholog (KO) abundances were relativized to the median abundance in each site of a set of 36 single copy genes published previously (40) (**Table E.7**). One single copy gene (K01519) was an outlier in 7 out of 12 samples as assessed by Grubb's test for outliers and removed. We analyzed patterns in KEGG Modules (KMs) (39), a set of manually defined functional units made up of multiple KOs. KM abundances were calculated based on the median abundance of their constituent KOs that were present in the metagenomes. KMs were included in analysis if 50% or more of their constituent KOs were identified in the dataset. Approximately one third of the open reading frames per sample were able to be annotated with KEGG (**Table E.1**). As a caveat to the study, unannotated open reading frames can result from erroneous reads and mis-assemblies but also could be previously undescribed and or divergent genes 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 average genome size by calculating the percent of sampled reads that match to a set of single copy genes (41). We also

used three additional methods to calculate average genome size (**Figure F.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 (**Table E.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 KOs, previously used by He et al. 2015 (40) to estimate average genome sizes. Odds ratios were determined at each site by comparing their abundance within a site to their average abundance across all 12 sites. One KO (K01519, Triosephosphate isomerase) was an outlier in seven out of twelve metagenomes, as determined by grubbs test, and was removed.

We used previously published 16S rRNA gene sequencing data (3) to estimate average genome size. 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 size were identified using the Tukey method and omitted from calculation of the mean phylum genome size (13). Phyla present in Centralia metagenomes but without representative genomes in IMG were combined at the Domain level, and a mean Domain genome size was calculated in the same manner. Each site's weighted mean genome size site was calculated based on the relative abundance of the phyla at each site.

Quality filtered metagenome reads were downloaded from JGI GOLD database. Paired-end reads from all 12 soils were assembled together using MEGAHIT (v1.0.2) (6) with a kmer range from 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%. Resulting SAM files were converted to sorted

BAM files using SAMTools (v1.3). Contigs larger than 2,500bp were binned into metagenome-assembled genomes (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 greater 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 inverse of its completeness (6). The average MAG size at each site was calculated by taking the mean of size of all MAGs detected in a site.

Average Cell Size

To calculate cell size, we re-analyzed microscope images previously used 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 cell size range and deviations (**Table E.5**) were consistent with those previously reported (42).

Construction of metagenome-assembled genomes (MAGs), taxonomic assignments, and visualization

Assembled contigs from quality filtered reads were binned into MAGs using MetaBAT (43) (v0.26.3) with the "--veryspecific" flag. Detailed description of assembly and binning procedures can be found in supplemental. Completeness and contamination were estimated for each MAG using CheckM (v1.0.5). MAG's we assigned taxonomy using the Microbial Genome Atlas (MiGA) NCBI Prokaryote project (44). Highest quality MAGs with greater 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 inverse of its completeness (6). The average MAG size at each site was calculated by taking the mean of size of all MAGs detected in a site.

The CheckM (45) genome tree was extended to include Centalia high-quality MAGs. The Interactive Tree of Life (iTOL) (46) was used for visualization (https://itol.embl.de/tree/3520411744356315278585 34#). Temperature range and taxonomy for each genome in the tree was collected from JGI IMG. MAGs were classified as fire-affected or ambient based on in which group of samples they had a higher coverage, and 95% of MAGs had at least 10x greater coverage in one soil category as compared to the other.

Comparisons with other soil metagenomes and genomes

All metagenome data sets 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, public = true. The resulting list of metagenome data sets were ordered by number of base pairs (bp). Metagenomic data sets with the most bp 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 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, data sets with the greatest bp were selected. This search yielded 22 data sets from 12 locations and five countries (**Table E.4**). Sequences from MG-RAST data sets were quality checked using FastQC (v0.11.3, (47) and quality controlled using the FASTX toolkit (fastq_quality_filter, "-Q33 -q 30 -p 50").

Average genome size for each dataset was calculated from the quality filtered DNA sequences using MicrobeCensus with default parameters.

The RefSoil database of soil genomes (48) was used to estimate 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 for calculating two-sided Pearson's correlations (49). The outliers package (50) was used for identifying outlying KOs. The ggplot2 package was used for visualization (51). Heat maps were created with heatmap2 from the gplots package (52).

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.

Code availability

All analysis workflows are available on GitHub (ShadeLab/PAPER_Sorensen_NatMicro_2018).

APPENDICES

APPENDIX D

Supplemental results

Supplementary Results

Comparison of metagenome annotation methods on results

We first compared both assembled and unassembled data using the Cluster of Orthologous Groups (COG) (7), Pfam (8), KEGG Orthology (9-11), and Enzyme IMG (12) databases to investigate whether any of these databases provided more complete annotation or resulted in different overarching community patterns. We found that COG, pfam and KO annotated between 29 and 42% of the genes present at each site, while the Enzyme database annotated only 17.15% to 20.80%. Bray Curtis distance matrices calculated from the Centralia gene tables of each of these databases were all correlated (Mantel test all $R > 0.738$, $p < 0.001$), but the Enzyme database consistently had the lowest correlation with other databases.

Additional calculations of average genome size in Centralia

For each metagenome, we assessed the abundance of 36 single copy genes (1) that were annotated to KEGG Orthologs (KO) (2). Twenty-nine single-copy gene KOs had odds ratios positively correlated with temperature ($p < 0.04$, Pearson's $R > 0.59$, **Figure F.1A**, **Table E.7**). None of the single-copy genes had correlations with metagenome sizes (all $p > 0.15$), affirming that this method is robust to differences in metagenome size. There were increases in single copy gene abundance with temperature, despite that the metagenomes had similar sequencing efforts. Thus, the odds ratios of single copy genes and estimates of genome size support a reduction in average genome size with increased soil temperature.

We also calculated an average genome size for each soil based on the 16S rRNA gene phylum-level composition of the community (3), which allowed us to also assess whether the

changes in genome size could be attributed to replacement of members along the thermal gradient (community turnover). In agreement with the above estimates of genome size changes, this phylogenetic inference analysis revealed a negative correlation between average genome size and temperature (**Figure F.1B**, Pearson's r = -0.860, p <0.001, n = 12 amplicon datasets). This suggests that the shift in average genome size was, at least in part, due to compositional changes favoring taxa with smaller genomes.

We also estimated the sizes of metagenome-assembled genomes (MAGs) that were observed along the temperature gradient. We assembled 104 MAGs from Centralia that had < 5% contamination and > 90% completeness (**Table E.6**). There was an inverse trend in the average sizes of these MAGs with temperature (R = -0.63, p = 0.03, n = 12 metagenomes, **Figure F.1C**). An analysis limitation is that these MAGs represent a subset of the most community members that were prevalent such that their genomes could be well-assembled from metagenomes, and so we do not know how representative these genome sizes are of their community. We expect that this trend is conservative because we did not weight by MAG abundance at each site to be cautious about abundance normalization. However, together with our other independent assessments of genome size, this provides additional support for the prevalence of small genomes in hot Centralia soils.

Patterns of enriched KEGG Modules in hot soils

For KMs positively correlated with temperature, all were enriched in the hottest soil (C10; 54.2 ºC). Most temperature-correlated KMs in soils C06 and C16 (21.7°C and 24.1°C, respectively) had relatively low abundances that were comparable to KM levels in sites with ambient temperatures. Broadly, the response patterns of the positively correlated KMs fell into

five clusters. Cluster *i* had a relatively linear response with increasing temperature, with no particular modules of note. Cluster *ii* also had a linear response to the temperature gradient, but contained KMs that were especially enriched in the hottest site C10. Cluster *ii* contained modules related to archaeal proteasome (M00343) and isoprenoid biosynthesis (M00365). In addition, cluster *ii* included two carbon transport systems: glucose/arabinose transporters (M00203) and trehalose transporters (M00604). Cluster *iii* contained 14 modules that were consistently enriched in the three hottest sites, C14(34.1ºC), C15(38.9 ºC) and C10(54.2 ºC), suggesting a threshold for these enriched modules with temperatures > 30ºC (**Figure 3.4A**). Site C15 generally had lower representation of these KMs than C14 and C10. Archaeal ribosome (M00179), archaeal RNA polymerase (M00184) and archaeal exosome (M00390) respectively were more abundant in C10 (the hottest site, 54.2 ºC) than in C14 or C15. The KM for dissimilatory sulfate reduction (M00596) was also present in cluster *iii*. This clustering of dissimilatory sulfate reduction with archaeal proteins points to an enrichment in sulfate reducing archaea in hot soils, and is also supported by an increase in Crenarchaeota in fire-affected soils (3), an archaeal phylum including known sulfate reducers (4). In addition to sulfate reduction, the KMs for dissimilatory nitrate reduction (M00530) and denitrification (M00529) were also part of cluster *iii.* (**Figure 3.4A**). Clusters *iv* and *v* both had KMs enriched in C12. Cluster *iv* shares enrichment in KMs between C12 and C10 soils and includes three KMs related to photosynthesis (M00161, M00162, M00163). Cluster *v* includes KMs generally shared across all soils > 30 ºC, and included primary metabolisms (e.g. histidine biosynthesis).

Supplementary References

1. He, S. *et al.* Patterns in wetland microbial community composition and functional gene repertoire associated with methane emissions. *MBio* **6,** e00066-15 (2015).

2. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27,** 29–34 (1999).

3. 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).

4. 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).

5. 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. Microbiology* 229–235 (2006).

6. Li, D., Liu, C.-M. C.-M., Luo, R., Sadakane, K. & Lam, T.-W. T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijin graph. *Bioinformatics* **31,** 1674–1676 (2015).

7. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** (2003).

8. Punta, M. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **40,** 290–301 (2012).

9. 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).

10. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–D205 (2014).

11. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucl. Acids Res.* **28,** 27–30 (2000).

12. Markowitz, V. M. *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42,** D560–D567 (2014).

13. Tukey, J. W. *Exploratory Data Analysis*. *Analysis* **2,** (1977).

APPENDIX E

Supplemental tables

Table E.1. Sequence summary information for Centralia metagenomes.

	Sequencing Depth (Gbases)	Raw Reads	Quality Reads	Aligned Reads	Assembled Length	Number Contigs	Percent Genes Annotated with KO	Site Temperature (°C)
Cen01	23	1.59E+08	1.55E+08	1.18E+08	2.44E+09	5.05E+06	29.56	14.1
Cen03	26	1.77E+08	1.74E+08	1.21E+08	2.85E+09	6.33E+06	30.85	14.7
Cen04	25	1.71E+08	1.68E+08	1.08E+08	2.50E+09	5.98E+06	30.9	13.3
Cen05	25	1.70E+08	1.67E+08	1.14E+08	2.77E+09	6.32E+06	30.34	14.0
Cen06	22	1.51E+08	1.49E+08	1.10E+08	2.22E+09	4.63E+06	30.63	24.1*
Cen07	21	1.41E+08	1.39E+08	9.64E+07	2.26E+09	4.84E+06	31.32	13.5
Cen10	36	2.43E+08	2.38E+08	2.22E+08	1.17E+09	1.89E+06	35.57	54.2*
Cen12	24	1.64E+08	1.61E+08	1.51E+08	1.20E+09	1.59E+06	33.04	32.0*
Cen14	24	1.59E+08	1.57E+08	1.40E+08	1.42E+09	2.23E+06	34.21	34.1
Cen15	20	1.32E+08	1.28E+08	1.09E+08	1.14E+09	1.84E+06	34.44	38.9*
Cen16	51	3.40E+08	3.30E+08	2.93E+08	4.06E+09	6.61E+06	32.43	21.7
Cen17	24	1.61E+08	1.55E+08	7.87E+07	1.90E+09	5.17E+06	31.14	12.1

Table E.2. Two-sided Pearson's correlations of Eukaryotic-specific ribosomal KEGG Orthologs and plasmid pfam categories with temperature.

KO/pfam	Description	Pearsons_R	p_value
K02868	Large subunit L11	-0.321	0.309
K02997	Small subunit S9	-0.126	0.695
K02870	Large subunit L12	-0.163	0.613
K02865	Large subunit L10	-0.423	0.17
K02901	Large subunit L27	-0.144	0.656
K02932	Large subunit L5	-0.25	0.432
K02981	Small subunit S2	-0.303	0.338
K02953	Small subunit S13	-0.339	0.281
K02949	Small subunit S11	-0.331	0.294
K02891	Large subunit L22	-0.332	0.291
K02900	Large subunit L27	-0.246	0.441
K02920	Large subunit L37	0.303	0.339
K02964	Small subunit S18	-0.303	0.338
K02969	Small subunit S20	-0.206	0.521
K02985	Small subunit S3	-0.337	0.284
K02993	Small subunit S7	-0.186	0.562
K02893	Large subunit L23	0.065	0.841
K02905	Large subunit L29	Not Detected	Not Detected
K02923	Large subunit L38	-0.384	0.218
pfam01446	Rep_1	0.104	0.747
pfam01719	Rep_2	Not Detected	Not Detected
pfam01051	Rep_3	-0.393	0.207
pfam05732	RepL	-0.276	0.385
pfam07042	TrfA protein	0.456	0.136
pfam04796	RepA_C	0.75	0.005
pfam02486	Rep_trans	0.16	0.62
pfam01402	RHH_1	0.368	0.239
pfam01815	Rop protein	Not Detected	Not Detected
pfam03428	RP-C	-0.518	0.084
pfam10134	RPA	-0.032	0.921
pfam06970	RepA_N	Not Detected	Not Detected
pfam06504	RepC	0.065	0.842
pfam03090	Replicase	-0.491	0.105

Table E.3. Two-sided Pearson's correlations of soil environmental variables with average genome size.

	Pearson's_R	Test_Statistic	FDR_AdjustedP_value
SoilTemperature_to10cm	-0.910	-6.920	0.001
OrganicMatter_500	-0.131	-0.418	0.926
NO3N_ppm	-0.591	-2.317	0.113
NH4N_ppm	-0.592	-2.325	0.113
pH	0.030	0.096	0.926
SulfateSulfur_ppm	-0.390	-1.338	0.391
K_ppm	-0.110	-0.350	0.926
Ca_ppm	0.072	0.229	0.926
Mg_ppm	0.121	0.385	0.926
Fe_ppm	-0.607	-2.415	0.113
P_ppm	-0.447	-1.582	0.314
As_ppm	-0.039	-0.125	0.926
SoilMoisture_Per	-0.590	-2.310	0.113

Table E.4. MG-RAST metadata for soil metagenomes used in this study.

Project Name	Sample Location	Country	Sample Shortname	Project ID	Sample Name	Gbp
ARMO	Rondonia	Brazil	Brazilian forest	mgp3731	mgm4546395.3	13.27
ARMO	Rondonia	Brazil	Brazilian forest	mgp3731	mgm4536139.3	9.04
ARMO	Rondonia	Brazil	Brazilian forest	mgp3731	mgm4535554.3	9.69
Axel Heiberg Permafrost: Part 4A	Central Axel Heiberg Island	Canada	Permafrost Canada	mgp252	mgm4523023.3	6.52
Axel Heiberg Permafrost: Part 4A	Central Axel Heiberg Island	Canada	Permafrost Canada	mgp252	mgm4523145.3	5.52
CedarCreek minsoil June2013	Bethel, MN	USA	Minnesota grassland	mgp5588	mgm4541646.3	10.65
CedarCreek minsoil June2013	Bethel, MN	USA	Minnesota grassland	mgp5588	mgm4541645.3	9.77
Fermi-syntheticlongreads	Fermi National Accelerator Laboratory	USA	Illinois switchgrass	mgp14596	mgm4653791.3	7.95
Fermi-syntheticlongreads	Fermi National Accelerator Laboratory	USA	Illinois switchgrass	mgp14596	mgm4653788.3	7.14
GED prairie unassembled	Iowa	USA	Iowa prairie	mgp6377	mgm4539575.3	18.79
GED prairie unassembled	Iowa	USA	Iowa prairie	mgp6377	mgm4539572.3	17.58
GED prairie unassembled	Iowa	USA	Iowa prairie	mgp6377	mgm4539576.3	17.43
GP corn unassembled	Iowa	USA	Iowa corn	mgp6368	mgm4539523.3	8.12
Hofmockel Soil Aggregate COB KBASE	Boone County, IA	USA	Iowa agricultural	mgp2592	mgm4509400.3	24.98

Table E.4. (cont'd)

Hofmockel Soil Aggregate COB KBASE	Boone County, IA	USA	Iowa agricultural	mgp2592	mgm4509401.3	7.86
ISA-SMC-2011	Auburn, IL	USA	Illinois soybean	mgp2076	mgm4502542.3	12.54
ISA-SMC-2011	Auburn, IL	USA	Illinois soybean	mgp2076	mgm4502540.3	10.60
Mining of new genes and pathways from soil of mangrove forest	Matang Mangrove Forest	Malaysia	Mangrove	mgp11628	mgm4603402.3	24.38
Mining of new genes and pathways from soil of mangrove forest	Matang Mangrove Forest	Malaysia	Mangrove	mgp11628	mgm4603270.3	24.54
NEON	Disney Wilderness Preserve, FL	USA	Disney preserve	mgp13948	mgm4664918.3	11.20
Permafrost sediments, North-East Siberia, Kolyma lowland	Kolyma river lowland	Russia	Permafrost Russia	mgp7176	mgm4546813.3	19.20
Ungulate Exclosure 2015	Wyoming	USA	Wyoming soil	mgp15600	mgm4670120.3	6.41

Table E.5. Cell size measurements from microscope images, quantified with FIJI software.

Soil	Average_Area	SD_Area	Average_Length	SD_Length	Average_Minor	SD_Minor	Number_Cells
Cen01	0.345	0.245	0.829	0.395	0.502	0.130	327
Cen03	0.450	0.349	0.941	0.409	0.557	0.221	225
Cen04	0.371	0.328	0.826	0.439	0.495	0.218	910
Cen05	0.376	0.306	0.869	0.473	0.507	0.171	434
Cen06	0.355	0.257	0.777	0.315	0.527	0.180	431
Cen07	0.347	0.269	0.793	0.378	0.511	0.163	581
Cen10	0.323	0.215	0.769	0.299	0.498	0.160	390
Cen12	0.370	0.280	0.824	0.385	0.517	0.187	217
Cen14	0.298	0.207	0.710	0.278	0.487	0.164	515
Cen15	0.290	0.201	0.727	0.341	0.464	0.162	44
Cen16	0.430	0.383	0.850	0.385	0.578	0.215	841
Cen17	0.385	0.313	0.878	0.335	0.524	0.156	455

Table E.6. Completeness, contamination, and taxonomy of Metagenome Assembled Genomes (MAGs).

MAG	Completeness	Contamination	MiGA.Taxonomy
METABAT_VerySpecific.86	100	0	f__Acidobacteriaceae
METABAT_VerySpecific.338	100	0.99	p__Proteobacteria
METABAT_VerySpecific.126	100	0.68	c__Spartobacteria
METABAT_VerySpecific.189	99.62	4	p__Proteobacteria
METABAT_VerySpecific.119	99.51	0.97	p__Thaumarchaeota
METABAT_VerySpecific.132	99.5	0.28	c__Alphaproteobacteria
METABAT_VerySpecific.561	99.44	2.31	p__Proteobacteria
METABAT_VerySpecific.135	99.15	1.28	c__Actinobacteria
METABAT_VerySpecific.57	99.12	1.75	f__Solibacteraceae
METABAT_VerySpecific.244	99.06	0.06	f__Beijerinckiaceae
METABAT_VerySpecific.384	98.99	1.36	f__Intrasporangiaeae
METABAT_VerySpecific.343	98.9	2.2	c__Gemmatimonadetes
METABAT_VerySpecific.180	98.65	2.2	p__Verrucomicrobia
METABAT_VerySpecific.78	98.54	0.97	p__Thaumarchaeota
METABAT_VerySpecific.138	98.41	3.17	p__Proteobacteria
METABAT_VerySpecific.36	98.25	1.1	c__Solibacteres
METABAT_VerySpecific.134	98.24	2.61	p__Acidobacteria
METABAT_VerySpecific.41	98.21	2.15	f__Hyphomicrobiaceae
METABAT_VerySpecific.396	98.18	1.82	c__Anaerolineae
METABAT_VerySpecific.166	98.08	4.27	c__Solibacteres
METABAT_VerySpecific.167	98.06	2.91	f__Nitrosophaeraceae
METABAT_VerySpecific.71	98.03	0.74	o__Chitinophagales
METABAT_VerySpecific.115	98.02	2.38	p__Chloroflexi
METABAT_VerySpecific.209	97.82	1.71	c__Acidobacteriia
METABAT_VerySpecific.334	97.8	1.37	p__Bacteroidetes
METABAT_VerySpecific.176	97.73	0.97	p__Thaumarchaeota
METABAT_VerySpecific.65	97.69	0.99	p__Chloroflexi
METABAT_VerySpecific.377	97.69	1.98	p__Proteobacteria
METABAT_VerySpecific.339	97.66	1.37	p__Bacteroidetes
METABAT_VerySpecific.231	97.48	2.52	p__Proteobacteria
METABAT_VerySpecific.52	97.4	4.03	p__Proteobacteria
METABAT_VerySpecific.306	97.3	0.72	c__Spartobacteria
METABAT_VerySpecific.258	97.29	1.31	o__Xanthomonadales
METABAT_VerySpecific.412	97.29	3.94	f__Isosphaeraceae
METABAT_VerySpecific.434	97.23	3.8	o__Nostocales
METABAT_VerySpecific.152	97.13	3.34	c__Alphaproteobacteria
METABAT_VerySpecific.140	97.08	1.01	c__Alphaproteobacteria

Table E.6. (cont'd)

METABAT_VerySpecific.42	97.07	2.23	p__Actinobacteria
METABAT_VerySpecific.82	97.01	3.85	p__Actinobacteria
METABAT_VerySpecific.56	97.01	2.14	p__Actinobacteria
METABAT_VerySpecific.79	96.84	3.73	c__Gammaproteobacteria
METABAT_VerySpecific.331	96.7	2.2	c__Gemmatimonadetes
METABAT_VerySpecific.325	96.7	2.38	p__Chloroflexi
METABAT_VerySpecific.106	96.7	4.16	p__Proteobacteria
METABAT_VerySpecific.385	96.66	0.58	c__Alphaproteobacteria
METABAT_VerySpecific.154	96.64	0.84	p__Proteobacteria
METABAT_VerySpecific.443	96.62	0	c__Spartobacteria
METABAT_VerySpecific.199	96.59	0.57	p__Planctomycetes
METABAT_VerySpecific.294	96.59	1.68	p__Proteobacteria
METABAT_VerySpecific.137	96.58	2.23	p__Actinobacteria
METABAT_VerySpecific.73	96.52	2.03	o__Rhizobiales
METABAT_VerySpecific.388	96.51	2.33	p__Planctomycetes
METABAT_VerySpecific.89	96.36	0.89	p__Proteobacteria
METABAT_VerySpecific.457	96.32	2.94	p__Crenarchaeota
METABAT_VerySpecific.175	96.25	1.12	c__Alphaproteobacteria
METABAT_VerySpecific.38	96.15	2.94	c__Actinobacteria
METABAT_VerySpecific.97	95.99	4.63	p__Firmicutes
METABAT_VerySpecific.96	95.94	0	c__Solibacteres
METABAT_VerySpecific.399	95.8	3.83	p__Proteobacteria
METABAT_VerySpecific.59	95.7	1.68	p__Proteobacteria
METABAT_VerySpecific.53	95.63	0.97	p__Thaumarchaeota
METABAT_VerySpecific.32	95.07	0.12	c__Gammaproteobacteria
METABAT_VerySpecific.117	95.06	1.55	c__Gammaproteobacteria
METABAT_VerySpecific.18	94.69	1.26	c__Gammaproteobacteria
METABAT_VerySpecific.278	94.54	2.52	p__Proteobacteria
METABAT_VerySpecific.536	94.47	2.14	c__Gammaproteobacteria
METABAT_VerySpecific.45	94.44	2.96	p__Firmicutes
METABAT_VerySpecific.491	94.44	0.43	c__Solibacteres
METABAT_VerySpecific.520	94.14	1.85	p__Armatimonadetes
METABAT_VerySpecific.23	94.13	1.71	p__Actinobacteria
METABAT_VerySpecific.33	94.07	3.36	p__Proteobacteria
METABAT_VerySpecific.596	94.06	3.07	o__Oscillatoriales
METABAT_VerySpecific.505	93.94	2.55	p__Proteobacteria
METABAT_VerySpecific.174	93.91	4.4	c__Solibacteres
METABAT_VerySpecific.129	93.82	1.4	f__Rhodanobacteraceae

Table E.6. (cont'd)

METABAT_VerySpecific.223	93.78	1.94	f__Nitrosophaeraceae
METABAT_VerySpecific.67	93.73	0.99	p__Proteobacteria
METABAT_VerySpecific.208	93.18	4	p__Chloroflexi
METABAT_VerySpecific.445	93.18	1.16	c__Gammaproteobacteria
METABAT_VerySpecific.243	93.07	0.68	c_Chlamydiia
METABAT_VerySpecific.47	93	0.93	p__Firmicutes
METABAT_VerySpecific.155	92.87	0.93	p__Proteobacteria
METABAT_VerySpecific.593	92.74	4.16	p__Chloroflexi
METABAT_VerySpecific.342	92.72	4.82	c__Actinobacteria
METABAT_VerySpecific.449	92.4	1.36	c__Gammaproteobacteria
METABAT_VerySpecific.6	92.2	1.8	p__Acidobacteria
METABAT_VerySpecific.91	92.19	1.71	d__Bacteria
METABAT_VerySpecific.675	92.08	0.99	p__Proteobacteria
METABAT_VerySpecific.233	92.08	0	p__Firmicutes
METABAT_VerySpecific.577	92.06	1.56	d__Bacteria
METABAT_VerySpecific.68	91.89	1.12	f__Acidiferrobacteraceae
METABAT_VerySpecific.164	91.67	0.79	c__Alphaproteobacteria
METABAT_VerySpecific.554	91.47	2.36	c__Spartobacteria
METABAT_VerySpecific.427	91.45	0	c__Acidobacteriia
METABAT_VerySpecific.507	91.36	1	c_Chitinophagia
METABAT_VerySpecific.109	91.26	0.97	p__Thaumarchaeota
METABAT_VerySpecific.483	91.22	4.89	p__Verrucomicrobia
METABAT_VerySpecific.187	91.01	0.87	c__Solibacteres
METABAT_VerySpecific.122	90.6	2.28	p__Proteobacteria
METABAT_VerySpecific.692	90.59	2.73	p__Verrucomicrobia
METABAT_VerySpecific.580	90.58	1.78	c__Acidobacteriia
METABAT_VerySpecific.27	90.57	0.85	p__Acidobacteria
METABAT_VerySpecific.333	90.29	0.07	c__Nitrososphaeria
METABAT_VerySpecific.3	90.22	3.39	c__Alphaproteobacteria

Table E.7. Two-sided Pearson's correlations of single-copy KEGG Ortholog odds ratios with temperature.

KEGG	Test Statistic T	Pearson's r	FDR Adjusted p-value
K00773	2.502	0.621	3.89E-02
K01409	2.180	0.568	5.92E-02
K01889	4.385	0.811	1.89E-03
K01890	-2.364	-0.599	4.49E-02
K01937	-1.718	-0.477	1.20E-01
K02428	5.214	0.855	6.15E-04
K02519	-1.865	-0.508	9.72E-02
K02864	6.485	0.899	1.57E-04
K02867	6.863	0.908	1.41E-04
K02874	8.547	0.938	7.88E-05
K02876	6.307	0.894	1.67E-04
K02881	6.339	0.895	1.67E-04
K02886	4.662	0.828	1.28E-03
K02890	7.058	0.913	1.41E-04
K02906	5.893	0.881	2.74E-04
K02926	2.361	0.598	4.49E-02
K02931	6.994	0.911	1.41E-04
K02933	8.036	0.931	8.15E-05
K02946	5.791	0.878	3.00E-04
K02948	9.381	0.948	5.13E-05
K02950	6.542	0.900	1.57E-04
K02952	6.442	0.898	1.57E-04
K02956	6.807	0.907	1.41E-04
K02959	2.726	0.653	2.74E-02
K02961	6.606	0.902	1.57E-04
K02965	7.171	0.915	1.41E-04
K02967	4.877	0.839	9.67E-04
K02982	5.421	0.864	4.79E-04
K02988	6.565	0.901	1.57E-04
K02992	8.233	0.934	8.15E-05
K02994	9.828	0.952	5.13E-05
K02996	6.990	0.911	1.41E-04
K03106	1.680	0.469	1.24E-01
K03470	6.948	0.910	1.41E-04
K03596	-2.426	-0.609	4.28E-02

Table E.8. Significant two-sided Pearson's correlations of KEGG Modules with temperature.

Module	Module Description	Completeness	Pearson's r	FDR adjusted p-value
M00432	Leucine biosynthesis, 2-oxoisovalerate => 2-oxoisocaproate	1	-0.925	1.32E-03
M00183	RNA polymerase, bacteria	1	-0.922	1.32E-03
M00709	Macrolide resistance, MacAB-TolC transporter	1	-0.917	1.32E-03
M00477	EvgS-EvgA (acid and drug tolerance) two-component regulatory system	1	-0.916	1.32E-03
M00729	Fluoroquinolone resistance, gyrase-protecting protein Qnr	0.667	-0.915	1.32E-03
M00453	QseC-QseB (quorum sensing) two-component regulatory system	1	-0.912	1.32E-03
M00499	HydH-HydG (metal tolerance) two-component regulatory system	1	-0.912	1.32E-03
M00086	beta-Oxidation, acyl-CoA synthesis	0.5	-0.912	1.32E-03
M00082	Fatty acid biosynthesis, initiation	0.571	-0.91	1.32E-03
M00565	Trehalose biosynthesis, D-glucose 1P => trehalose	1	-0.909	1.32E-03
M00258	Putative ABC transport system	1	-0.906	1.43E-03
M00096	C5 isoprenoid biosynthesis, non-mevalonate pathway	0.9	-0.905	1.43E-03
M00501	PilS-PilR (type 4 fimbriae synthesis) two-component regulatory system	1	-0.901	1.62E-03
M00446	RstB-RstA two-component regulatory system	1	-0.901	1.62E-03
M00017	Methionine biosynthesis, apartate => homoserine => methionine	0.846	-0.899	1.66E-03
M00037	Melatonin biosynthesis, tryptophan => serotonin => melatonin	0.5	-0.899	1.66E-03
M00134	Polyamine biosynthesis, arginine => ornithine => putrescine	1	-0.895	1.90E-03
M00028	Ornithine biosynthesis, glutamate => ornithine	0.714	-0.891	2.13E-03
M00050	Guanine ribonucleotide biosynthesis IMP => GDP,GTP	0.833	-0.888	2.30E-03
M00042	Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline	0.5	-0.885	2.54E-03
M00509	WspE-WspRF (chemosensory) two-component regulatory system	1	-0.883	2.60E-03
M00649	Multidrug resistance, efflux pump AdeABC	1	-0.882	2.60E-03
M00251	Teichoic acid transport system	1	-0.881	2.60E-03

Table E.8. (cont'd)

M00135	GABA biosynthesis, eukaryotes, putrescine => GABA	0.6	-0.881	2.60E-03
M00539	Cumate degradation, p-cumate => 2-oxopent-4-enoate + 2-methylpropanoate	1	-0.881	2.60E-03
M00454	KdpD-KdpE (potassium transport) two-component regulatory system	1	-0.88	2.60E-03
M00394	RNA degradosome	1	-0.88	2.60E-03
M00655	AdeS-AdeR two-component regulatory system	1	-0.878	2.70E-03
M00046	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate	0.667	-0.877	2.70E-03
M00170	C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type	0.5	-0.876	2.70E-03
M00627	beta-Lactam resistance, Bla system	0.75	-0.873	2.91E-03
M00221	Putative simple sugar transport system	1	-0.873	2.91E-03
M00247	Putative ABC transport system	1	-0.872	2.91E-03
M00722	Cationic antimicrobial peptide (CAMP) resistance, phosphoethanolamine transferase PmrC	1	-0.87	2.93E-03
M00457	TctE-TctD (tricarboxylic acid transport) two-component regulatory system	1	-0.87	2.93E-03
M00045	Histidine degradation, histidine => N-formiminoglutamate => glutamate	1	-0.869	3.00E-03
M00115	NAD biosynthesis, aspartate => NAD	0.857	-0.867	3.08E-03
M00013	Malonate semialdehyde pathway, propanoyl-CoA => acetyl-CoA	0.667	-0.866	3.08E-03
M00498	NtrY-NtrX (nitrogen regulation) two-component regulatory system	1	-0.866	3.08E-03
M00605	Glucose/mannose transport system	1	-0.866	3.08E-03
M00006	Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P	0.667	-0.863	3.25E-03
M00126	Tetrahydrofolate biosynthesis, GTP => THF	0.565	-0.862	3.29E-03
M00298	Multidrug/hemolysin transport system	1	-0.861	3.40E-03
M00049	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	0.556	-0.857	3.57E-03
M00156	Cytochrome c oxidase, cbb3-type	1	-0.857	3.57E-03

Table E.8. (cont'd)

M00012	Glyoxylate cycle	0.667	-0.857	3.57E-03
M00502	GlrK-GlrR (amino sugar metabolism) two-component regulatory system	1	-0.856	3.57E-03
M00475	BarA-UvrY (central carbon metabolism) two-component regulatory system	1	-0.855	3.57E-03
M00631	D-Galacturonate degradation (bacteria), D-galacturonate => pyruvate + D-glyceraldehyde 3P	1	-0.853	3.77E-03
M00641	Multidrug resistance, efflux pump MexEF-OprN	1	-0.853	3.77E-03
M00169	CAM (Crassulacean acid metabolism), light	1	-0.851	3.94E-03
M00015	Proline biosynthesis, glutamate => proline	0.75	-0.85	3.94E-03
M00639	Multidrug resistance, efflux pump MexCD-OprJ	1	-0.85	3.94E-03
M00210	Phospholipid transport system	1	-0.85	3.94E-03
M00662	Hk1-Rrp1 (glycerol uptake and utilization) two-component regulatory system	0.5	-0.848	4.15E-03
M00467	SasA-RpaAB (circadian timing mediating) two-component regulatory system	1	-0.846	4.22E-03
M00140	C1-unit interconversion, prokaryotes	1	-0.845	4.28E-03
M00250	Lipopolysaccharide transport system	1	-0.844	4.31E-03
M00359	Aminoacyl-tRNA biosynthesis, eukaryotes	0.955	-0.843	4.41E-03
M00669	gamma-Hexachlorocyclohexane transport system	0.75	-0.842	4.46E-03
M00670	Mce transport system	0.75	-0.842	4.46E-03
M00451	BasS-BasR (antimicrobial peptide resistance) two-component regulatory system	1	-0.841	4.46E-03
M00129	Ascorbate biosynthesis, animals, glucose-1P => ascorbate	0.714	-0.841	4.46E-03
M00551	Benzoate degradation, benzoate => catechol / methylbenzoate => methylcatechol	1	-0.84	4.48E-03
M00230	Glutamate/aspartate transport system	1	-0.838	4.58E-03
M00497	GlnL-GlnG (nitrogen regulation) two-component regulatory system	1	-0.837	4.58E-03
M00549	Nucleotide sugar biosynthesis, glucose => UDP-glucose	0.833	-0.837	4.58E-03

Table E.8. (cont'd)

M00172	C4-dicarboxylic acid cycle, NADP - malic enzyme type	0.75	-0.837	4.58E-03
M00027	GABA (gamma-Aminobutyrate) shunt	0.857	-0.837	4.58E-03
M00525	Lysine biosynthesis, acetyl-DAP pathway, aspartate => lysine	1	-0.836	4.64E-03
M00526	Lysine biosynthesis, DAP dehydrogenase pathway, aspartate => lysine	0.9	-0.836	4.64E-03
M00220	Rhamnose transport system	1	-0.833	4.85E-03
M00699	Multidrug resistance, efflux pump AmeABC	0.5	-0.833	4.86E-03
M00307	Pyruvate oxidation, pyruvate => acetyl-CoA	1	-0.831	5.07E-03
M00473	UhpB-UhpA (hexose phosphates uptake) two-component regulatory system	1	-0.83	5.15E-03
M00216	Multiple sugar transport system	1	-0.828	5.25E-03
M00168	CAM (Crassulacean acid metabolism), dark	0.5	-0.828	5.28E-03
M00527	Lysine biosynthesis, DAP aminotransferase pathway, aspartate => lysine	0.909	-0.827	5.33E-03
M00002	Glycolysis, core module involving three-carbon compounds	0.917	-0.826	5.36E-03
M00572	Pimeloyl-ACP biosynthesis, BioC-BioH pathway, malonyl-ACP => pimeloyl-ACP	0.8	-0.826	5.36E-03
M00238	D-Methionine transport system	1	-0.826	5.37E-03
M00204	Trehalose/maltose transport system	1	-0.823	5.66E-03
M00628	beta-Lactam resistance, AmpC system	1	-0.823	5.66E-03
M00157	F-type ATPase, prokaryotes and chloroplasts	1	-0.822	5.71E-03
M00503	PgtB-PgtA (phosphoglycerate transport) two-component regulatory system	1	-0.822	5.71E-03
M00189	Molybdate transport system	1	-0.821	5.76E-03
M00455	TorS-TorR (TMAO respiration) two-component regulatory system	1	-0.82	5.83E-03
M00097	beta-Carotene biosynthesis, GGAP => beta-carotene	0.833	-0.819	6.04E-03
M00754	Nisin resistance, phage shock protein homolog LiaH	1	-0.816	6.34E-03
M00193	Putative spermidine/putrescine transport system	1	-0.816	6.34E-03
M00083	Fatty acid biosynthesis, elongation	0.8	-0.814	6.67E-03

Table E.8. (cont'd)

M00377	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)	1	-0.813	6.67E-03
M00718	Multidrug resistance, efflux pump MexAB-OprM	1	-0.813	6.67E-03
M00119	Pantothenate biosynthesis, valine/L-aspartate => pantothenate	1	-0.813	6.67E-03
M00459	VicK-VicR (cell wall metabolism) two-component regulatory system	1	-0.812	6.77E-03
M00485	KinABCDE-Spo0FA (sporulation control) two-component regulatory system	1	-0.811	6.77E-03
M00595	Thiosulfate oxidation by SOX complex, thiosulfate => sulfate	1	-0.811	6.79E-03
M00445	EnvZ-OmpR (osmotic stress response) two-component regulatory system	1	-0.811	6.79E-03
M00504	DctB-DctD (C4-dicarboxylate transport) two-component regulatory system	1	-0.809	6.91E-03
M00255	Lipoprotein-releasing system	1	-0.809	6.91E-03
M00040	Tyrosine biosynthesis, prephanate => pretyrosine => tyrosine	0.6	-0.807	7.19E-03
M00328	Hemophore/metalloprotease transport system	1	-0.807	7.19E-03
M00546	Purine degradation, xanthine => urea	0.95	-0.805	7.26E-03
M00044	Tyrosine degradation, tyrosine => homogentisate	0.833	-0.805	7.26E-03
M00668	Tetracycline resistance, TetA transporter	0.5	-0.803	7.53E-03
M00036	Leucine degradation, leucine => acetoacetate + acetyl-CoA	1	-0.803	7.53E-03
M00208	Glycine betaine/proline transport system	1	-0.802	7.60E-03
M00004	Pentose phosphate pathway (Pentose phosphate cycle)	0.867	-0.802	7.64E-03
M00478	DegS-DegU (multicellular behavior control) two-component regulatory system	1	-0.801	7.64E-03
M00593	Inositol transport system	1	-0.801	7.68E-03
M00644	Vanadium resistance, efflux pump MexGHI-OpmD	1	-0.801	7.68E-03
M00212	Ribose transport system	1	-0.8	7.68E-03
M00760	Erythromycin resistance, macrolide 2-phosphotransferase I MphA	0.5	-0.798	8.04E-03
M00324	Dipeptide transport system	1	-0.798	8.04E-03
M00743	Aminoglycoside resistance, protease HtpX	1	-0.797	8.18E-03

Table E.8. (cont'd)

M00227	Glutamine transport system	1	-0.796	8.22E-03
M00009	Citrate cycle (TCA cycle, Krebs cycle)	0.769	-0.794	8.38E-03
M00127	Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P	0.7	-0.794	8.38E-03
M00658	VanS-VanR (actinomycete type vancomycin resistance) two-component regulatory system	1	-0.794	8.41E-03
M00519	YesM-YesN two-component regulatory system	1	-0.793	8.41E-03
M00063	CMP-KDO biosynthesis	1	-0.793	8.41E-03
M00136	GABA biosynthesis, prokaryotes, putrescine => GABA	1	-0.793	8.44E-03
M00672	Penicillin biosynthesis, aminoadipate + cycteine + valine => penicillin	1	-0.791	8.55E-03
M00505	KinB-AlgB (alginate production) two-component regulatory system	1	-0.791	8.59E-03
M00360	Aminoacyl-tRNA biosynthesis, prokaryotes	1	-0.79	8.68E-03
M00555	Betaine biosynthesis, choline => betaine	1	-0.79	8.68E-03
M00277	PTS system, N-acetylgalactosamine-specific II component	1	-0.786	9.28E-03
M00656	VanS-VanR (VanB type vancomycin resistance) two-component regulatory system	1	-0.785	9.46E-03
M00350	Capsaicin biosynthesis, L-Phenylalanine => Capsaicin	0.6	-0.784	9.70E-03
M00022	Shikimate pathway, phosphoenolpyruvate + erythrose-4P => chorismate	0.8	-0.783	9.74E-03
M00020	Serine biosynthesis, glycerate-3P => serine	1	-0.783	9.74E-03
M00642	Multidrug resistance, efflux pump MexJK-OprM	1	-0.782	9.81E-03
M00300	Putrescine transport system	1	-0.781	1.01E-02
M00339	RaxAB-RaxC type I secretion system	1	-0.78	1.02E-02
M00260	DNA polymerase III complex, bacteria	1	-0.78	1.02E-02
M00766	Streptomycin resistance, deactivating enzyme StrAB	1	-0.779	1.02E-02
M00077	Chondroitin sulfate degradation	1	-0.779	1.02E-02
M00237	Branched-chain amino acid transport system	1	-0.779	1.02E-02
M00361	Nucleotide sugar biosynthesis, eukaryotes	0.857	-0.777	1.04E-02
M00025	Tyrosine biosynthesis, chorismate => tyrosine	0.857	-0.775	1.08E-02

Table E.8. (cont'd)

M00532	Photorespiration	0.8	-0.773	1.12E-02
M00335	Sec (secretion) system	1	-0.773	1.13E-02
M00010	Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	1	-0.77	1.18E-02
M00024	Phenylalanine biosynthesis, chorismate => phenylalanine	0.857	-0.768	1.19E-02
M00728	Cationic antimicrobial peptide (CAMP) resistance, envelope protein folding and degrading factors DegP and DsbA	1	-0.768	1.19E-02
M00589	Putative lysine transport system	1	-0.768	1.19E-02
M00093	Phosphatidylethanolamine (PE) biosynthesis, PA => PS => PE	1	-0.768	1.19E-02
M00149	Succinate dehydrogenase, prokaryotes	1	-0.768	1.19E-02
M00232	General L-amino acid transport system	1	-0.767	1.20E-02
M00318	Iron/zinc/copper transport system	1	-0.765	1.23E-02
M00579	Phosphate acetyltransferase-acetate kinase pathway, acetyl-CoA => acetate	1	-0.765	1.23E-02
M00474	RcsC-RcsD-RcsB (capsule synthesis) two-component regulatory system	1	-0.764	1.24E-02
M00456	ArcB-ArcA (anoxic redox control) two-component regulatory system	1	-0.764	1.24E-02
M00306	PTS system, fructose-specific II-like component	1	-0.763	1.26E-02
M00524	FixL-FixJ (nitrogen fixation) two-component regulatory system	1	-0.762	1.27E-02
M00538	Toluene degradation, toluene => benzoate	1	-0.762	1.27E-02
M00124	Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P	1	-0.759	1.32E-02
M00078	Heparan sulfate degradation	0.667	-0.759	1.32E-02
M00200	Putative sorbitol/mannitol transport system	1	-0.758	1.33E-02
M00713	Fluoroquinolone resistance, efflux pump LfrA	0.5	-0.758	1.33E-02
M00714	Multidrug resistance, efflux pump QacA	0.5	-0.758	1.33E-02
M00253	Sodium transport system	1	-0.756	1.36E-02
M00240	Iron complex transport system	1	-0.756	1.36E-02
M00368	Ethylene biosynthesis, methionine => ethylene	0.667	-0.755	1.39E-02
M00493	AlgZ-AlgR (alginate production) two-component regulatory system	1	-0.754	1.39E-02

Table E.8. (cont'd)

M00034	Methionine salvage pathway	0.842	-0.754	1.39E-02
M00439	Oligopeptide transport system	1	-0.753	1.40E-02
M00035	Methionine degradation	0.857	-0.752	1.41E-02
M00542	EHEC/EPEC pathogenicity signature, T3SS and effectors	0.765	-0.752	1.41E-02
M00299	Spermidine/putrescine transport system	1	-0.752	1.41E-02
M00014	Glucuronate pathway (uronate pathway)	0.667	-0.752	1.42E-02
M00330	Adhesin protein transport system	1	-0.748	1.48E-02
M00500	AtoS-AtoC (cPHB biosynthesis) two-component regulatory system	1	-0.747	1.51E-02
M00531	Assimilatory nitrate reduction, nitrate => ammonia	0.667	-0.745	1.55E-02
M00740	Methylaspartate cycle	0.667	-0.745	1.55E-02
M00570	Isoleucine biosynthesis, threonine => 2-oxobutanoate => isoleucine	1	-0.745	1.55E-02
M00447	CpxA-CpxR (envelope stress response) two-component regulatory system	1	-0.744	1.57E-02
M00511	PleC-PleD (cell fate control) two-component regulatory system	1	-0.742	1.61E-02
M00704	Tetracycline resistance, efflux pump Tet38	0.5	-0.742	1.61E-02
M00697	Multidrug resistance, efflux pump MdtEF-TolC	0.5	-0.742	1.61E-02
M00654	ParS-ParR (polymyxin-adaptive resistance) two-component regulatory system	1	-0.741	1.63E-02
M00101	Cholesterol biosynthesis, squalene 2,3-epoxide => cholesterol	0.727	-0.74	1.66E-02
M00468	SaeS-SaeR (staphylococcal virulence regulation) two-component regulatory system	1	-0.738	1.69E-02
M00256	Cell division transport system	1	-0.736	1.74E-02
M00323	Urea transport system	1	-0.735	1.75E-02
M00259	Heme transport system	1	-0.734	1.77E-02
M00622	Nicotinate degradation, nicotinate => fumarate	1	-0.734	1.78E-02
M00167	Reductive pentose phosphate cycle, glyceraldehyde-3P => ribulose-5P	0.625	-0.733	1.79E-02
M00326	RTX toxin transport system	1	-0.732	1.81E-02
M00618	Acetogen	1	-0.732	1.81E-02

Table E.8. (cont'd)

M00016	Lysine biosynthesis, succinyl-DAP pathway, aspartate => lysine	0.929	-0.732	1.81E-02
M00575	Pertussis pathogenicity signature 2, T1SS	1	-0.731	1.82E-02
M00535	Isoleucine biosynthesis, pyruvate => 2-oxobutanoate	1	-0.731	1.82E-02
M00533	Homoprotocatechuate degradation, homoprotocatechuate => 2-oxohept-3-enedioate	1	-0.729	1.85E-02
M00003	Gluconeogenesis, oxaloacetate => fructose-6P	1	-0.729	1.85E-02
M00254	ABC-2 type transport system	1	-0.729	1.85E-02
M00742	Aminoglycoside resistance, protease FtsH	0.833	-0.728	1.85E-02
M00448	CssS-CssR (secretion stress response) two-component regulatory system	1	-0.728	1.85E-02
M00741	Propanoyl-CoA metabolism, propanoyl-CoA => succinyl-CoA	0.846	-0.728	1.85E-02
M00100	Sphingosine degradation	1	-0.728	1.85E-02
M00023	Tryptophan biosynthesis, chorismate => tryptophan	0.688	-0.728	1.85E-02
M00488	DcuS-DcuR (C4-dicarboxylate metabolism) two-component regulatory system	1	-0.727	1.86E-02
M00517	RpfC-RpfG (cell-to-cell signaling) two-component regulatory system	1	-0.726	1.86E-02
M00215	D-Xylose transport system	1	-0.726	1.86E-02
M00663	SsrA-SsrB two-component regulatory system	1	-0.725	1.89E-02
M00648	Multidrug resistance, efflux pump MdtABC	1	-0.724	1.89E-02
M00338	Cysteine biosynthesis, homocysteine + serine => cysteine	1	-0.724	1.89E-02
M00325	alpha-Hemolysin/cyclolysin transport system	1	-0.724	1.90E-02
M00613	Anoxygenic photosynthesis in green nonsulfur bacteria	1	-0.724	1.90E-02
M00607	Glycerol transport system	1	-0.723	1.91E-02
M00471	NarX-NarL (nitrate respiration) two-component regulatory system	1	-0.722	1.93E-02
M00356	Methanogenesis, methanol => methane	1	-0.721	1.97E-02
M00060	Lipopolysaccharide biosynthesis, KDO2-lipid A	1	-0.72	1.98E-02
M00513	LuxQN/CqsS-LuxU-LuxO (quorum sensing) two-component regulatory system	1	-0.719	2.00E-02

Table E.8. (cont'd)

M00566	Dipeptide transport system, Firmicutes	1	-0.718	2.02E-02
M00761	Undecaprenylphosphate alpha-L-Ara4N biosynthesis, UDP-GlcA => Undecaprenyl phosphate alpha-L-Ara4N	1	-0.716	2.07E-02
M00660	Xanthomonas spp. pathogenicity signature, T3SS and effectors	1	-0.716	2.07E-02
M00011	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	0.735	-0.716	2.07E-02
M00222	Phosphate transport system	1	-0.716	2.07E-02
M00778	Type II polyketide backbone biosynthesis, acyl-CoA + malonyl-CoA => polyketide	0.545	-0.714	2.09E-02
M00591	Putative xylitol transport system	1	-0.714	2.09E-02
M00019	Valine/isoleucine biosynthesis, pyruvate => valine / 2-oxobutanoate => isoleucine	1	-0.714	2.09E-02
M00118	Glutathione biosynthesis, glutamate => glutathione	0.5	-0.714	2.09E-02
M00449	CreC-CreB (phosphate regulation) two-component regulatory system	1	-0.713	2.09E-02
M00332	Type III secretion system	1	-0.711	2.16E-02
M00236	Putative polar amino acid transport system	1	-0.708	2.25E-02
M00244	Putative zinc/manganese transport system	1	-0.706	2.29E-02
M00727	Cationic antimicrobial peptide (CAMP) resistance, N-acetylmuramoyl-L-alanine amidase AmiA and AmiC	1	-0.706	2.29E-02
M00213	L-Arabinose transport system	1	-0.704	2.33E-02
M00779	Dihydrokalafungin biosynthesis, octaketide => dihydrokalafungin	0.75	-0.698	2.54E-02
M00214	Methyl-galactoside transport system	1	-0.697	2.58E-02
M00652	Vancomycin resistance, D-Ala-D-Ser type	0.6	-0.695	2.65E-02
M00664	Nodulation	1	-0.694	2.69E-02
M00696	Multidrug resistance, efflux pump AcrEF-TolC	1	-0.692	2.73E-02
M00362	Nucleotide sugar biosynthesis, prokaryotes	1	-0.691	2.76E-02
M00008	Entner-Doudoroff pathway, glucose-6P => glyceraldehyde-3P + pyruvate	1	-0.685	2.98E-02
M00568	Catechol ortho-cleavage, catechol => 3-oxoadipate	1	-0.681	3.11E-02

Table E.8. (cont'd)

M00476	ComP-ComA (competence) two-component regulatory system	1	-0.677	3.25E-02
M00480	VraS-VraR (cell-wall peptidoglycan synthesis) two-component regulatory system	1	-0.673	3.41E-02
M00235	Arginine/ornithine transport system	1	-0.673	3.43E-02
M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	0.867	-0.67	3.54E-02
M00506	CheA-CheYBV (chemotaxis) two-component regulatory system	1	-0.669	3.59E-02
M00320	Lipopolysaccharide export system	1	-0.668	3.59E-02
M00701	Multidrug resistance, efflux pump EmrAB	1	-0.668	3.59E-02
M00514	TtrS-TtrR (tetrathionate respiration) two-component regulatory system	1	-0.668	3.60E-02
M00617	Methanogen	0.557	-0.664	3.77E-02
M00150	Fumarate reductase, prokaryotes	1	-0.664	3.77E-02
M00674	Clavaminate biosynthesis, arginine + glyceraldehyde-3P => clavaminate	1	-0.663	3.77E-02
M00194	Maltose/maltodextrin transport system	1	-0.661	3.90E-02
M00319	Manganese/zinc/iron transport system	1	-0.66	3.91E-02
M00581	Biotin transport system	1	-0.653	4.24E-02
M00673	Cephamycin C biosynthesis, aminoadipate + cycteine + valine => cephamycin C	1	-0.651	4.34E-02
M00645	Multidrug resistance, efflux pump SmeABC	1	-0.648	4.50E-02
M00721	Cationic antimicrobial peptide (CAMP) resistance, arnBCADTEF operon	1	-0.648	4.50E-02
M00584	Acetoin utilization transport system	0.75	-0.647	4.50E-02
M00487	CitS-CitT (magnesium-citrate transport) two-component regulatory system	1	-0.647	4.51E-02
M00121	Heme biosynthesis, glutamate => protoheme/siroheme	0.826	-0.642	4.74E-02
M00095	C5 isoprenoid biosynthesis, mevalonate pathway	0.875	0.646	4.53E-02
M00163	Photosystem I	1	0.657	4.06E-02
M00161	Photosystem II	1	0.68	3.13E-02

Table E.8. (cont'd)

M00052	Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP	0.571	0.692	2.74E-02
M00120	Coenzyme A biosynthesis, pantothenate => CoA	0.909	0.693	2.70E-02
M00343	Archaeal proteasome	1	0.713	2.09E-02
M00162	Cytochrome b6f complex	0.875	0.714	2.09E-02
M00203	Glucose/arabinose transport system	1	0.737	1.72E-02
M00633	Semi-phosphorylative Entner-Doudoroff pathway, gluconate/galactonate => glycerate-3P	0.8	0.75	1.46E-02
M00275	PTS system, cellobiose-specific II component	1	0.76	1.30E-02
M00365	C10-C20 isoprenoid biosynthesis, archaea	1	0.794	8.38E-03
M00166	Reductive pentose phosphate cycle, ribulose-5P => glyceraldehyde-3P	0.857	0.806	7.26E-03
M00530	Dissimilatory nitrate reduction, nitrate => ammonia	1	0.807	7.19E-03
M00031	Lysine biosynthesis, mediated by LysW, 2-aminoadipate => lysine	1	0.829	5.20E-03
M00423	Molybdate/tungstate transport system	1	0.831	5.07E-03
M00763	Ornithine biosynthesis, mediated by LysW, glutamate => ornithine	0.833	0.833	4.85E-03
M00026	Histidine biosynthesis, PRPP => histidine	0.789	0.84	4.49E-03
M00596	Dissimilatory sulfate reduction, sulfate => H2S	1	0.846	4.22E-03
M00604	Trehalose transport system	1	0.86	3.40E-03
M00529	Denitrification, nitrate => nitrogen	0.917	0.872	2.92E-03
M00159	V/A-type ATPase, prokaryotes	1	0.893	2.04E-03
M00179	Ribosome, archaea	0.985	0.908	1.32E-03
M00184	RNA polymerase, archaea	0.875	0.91	1.32E-03
M00390	Exosome, archaea	1	0.916	1.32E-03
M00391	Exosome, eukaryotes	0.5	0.917	1.32E-03
M00177	Ribosome, eukaryotes	0.671	0.93	1.17E-03
M00425	H/ACA ribonucleoprotein complex	0.5	0.933	1.13E-03

Table E.9. Permanent finished genomes per phylum in Integrated Microbial Genomes database used in Figure F.2.

Phylum	Genomes
Chlorobi	21
Crenarchaeota	224
Acidobacteria	68
Bacteroidetes	1900
Proteobacteria	23638
Planctomycetes	104
Bacteria	48637
Cyanobacteria	391
Chloroflexi	154
Verrucomicrobia	96
Nitrospirae	49
Actinobacteria	6075
Armatimonadetes	13
candidate division TM6	1
Gemmatimonadetes	26
Chlamydiae	270
Elusimicrobia	42
Candidatus Parcubacteria	60
Firmicutes	14186
candidate division WPS-2	8
Euryarchaeota	651
Candidatus Parvarchaeota	7
OP11	0
Spirochaetes	746
Fusobacteria	132
Candidatus Omnitrophica	60
BRC1	2
WS1	1
Tenericutes	336
candidate division GAL15	3
Candidatus Saccharibacteria	47
Unclassified	0
FBP	0
Fibrobacteres	30
Archaea	882
NC10	0

Table E.9. (cont'd)

Candidatus Aerophobetes	4
Deferribacteres	51
Deinococcus-Thermus	74
Candidatus Fervidibacteria	13
Aquificae	36
Lentisphaerae	8
cadidate division SR1	0

APPENDIX F

Supplemental figures

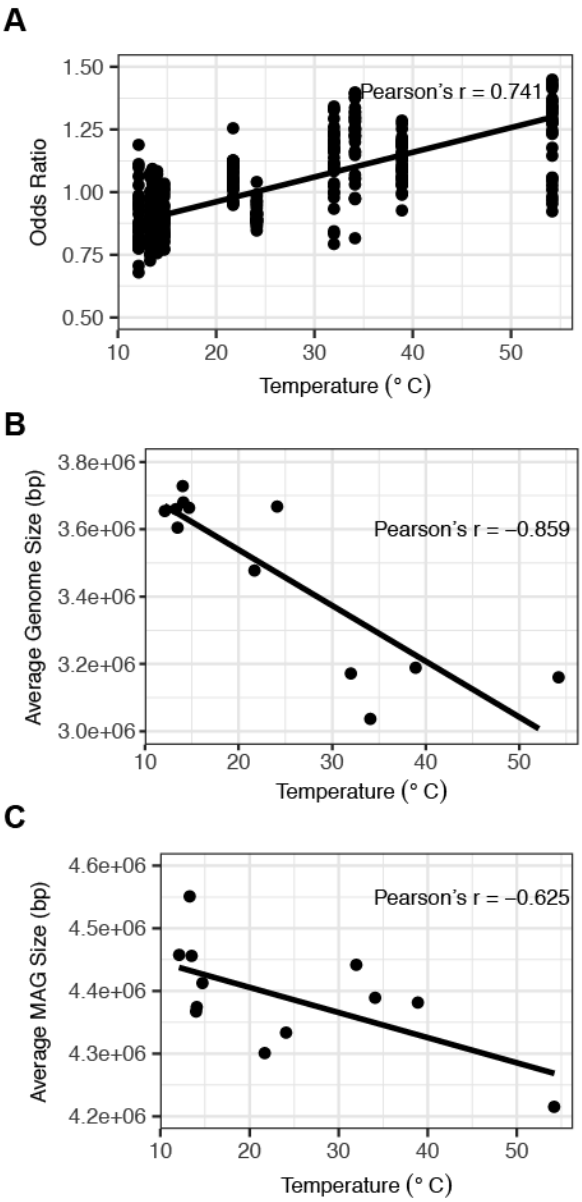

Figure F.1. Complementary methods used to assess changes in average genome size across the soil temperature gradient in Centralia.

(A) Odds ratios were calculated for 35 single-copy gene KEGG Orthologs in each site and plotted against site temperature. Reported two-sided Pearson's correlation is between all single copy gene odds ratios and temperature($p = 2.2x10^{-16}$). (B) Average genome size in each site was

Figure F.1. (cont'd)

calculated based on phylum level abundances from 16S rRNA gene amplicon data, using weighted average genome sizes of each phylum present in JGI IMG (accessed 19 June 2017, two-sided Pearson's correlation $p = 0.0003$). (C) Average MAG size at each site was calculated based on presence/absence of 104 MAGs (two-sided Pearson's correlation $p = 0.029$,). For all Pearson's correlations, n= 12 soils.

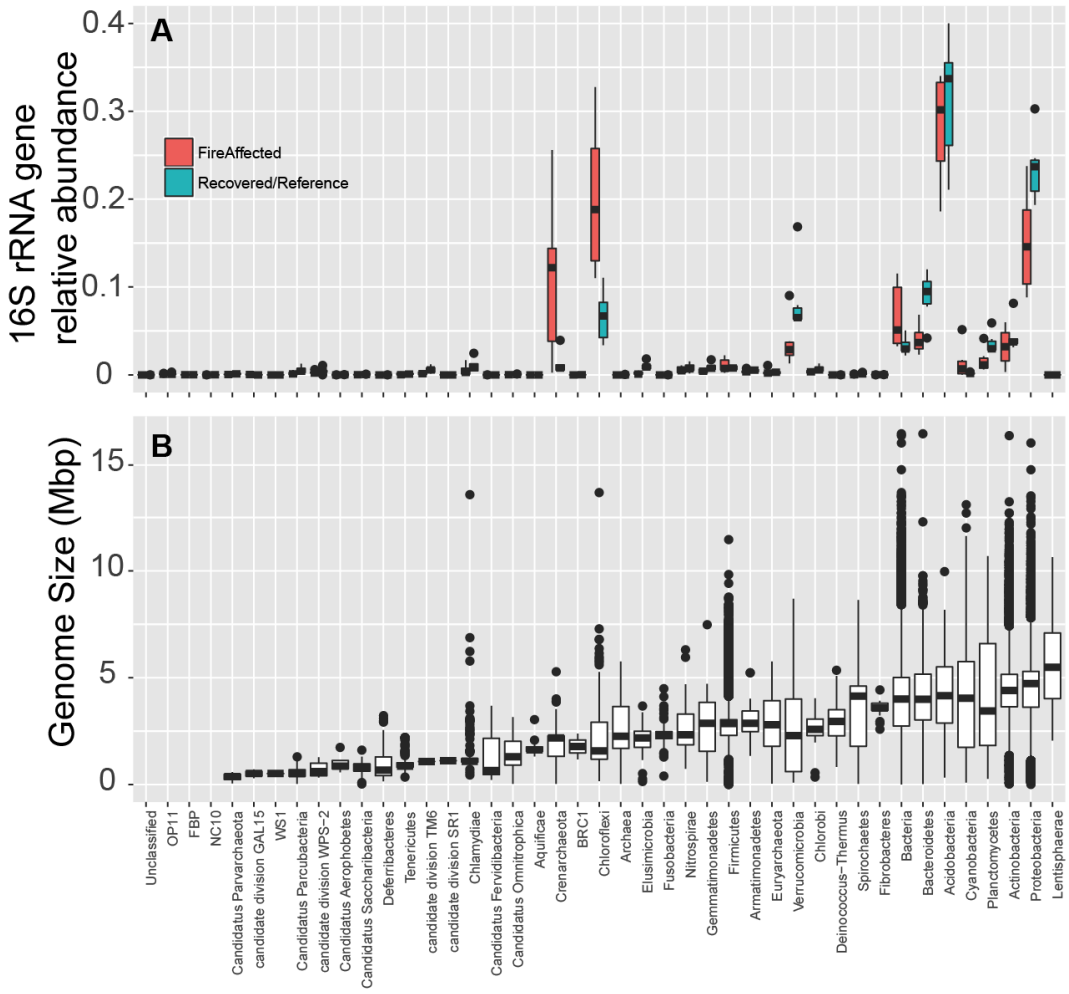

Figure F.2. Community structure in Centralia.

(A) Relative abundance of phyla in fire-affected (red, n=6 soils) and recovered/reference (blue, n=6 soils) sites based on 16S rRNA gene amplicon sequences. Taxonomic assignments were with the RDP classifier against the greengenes database (B) Sizes of permanent draft and finished genomes in IMG from phyla detected in Centralia. Midlines of each boxplot correspond to 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. Number of genomes per boxplot is described in **Table E.9**.

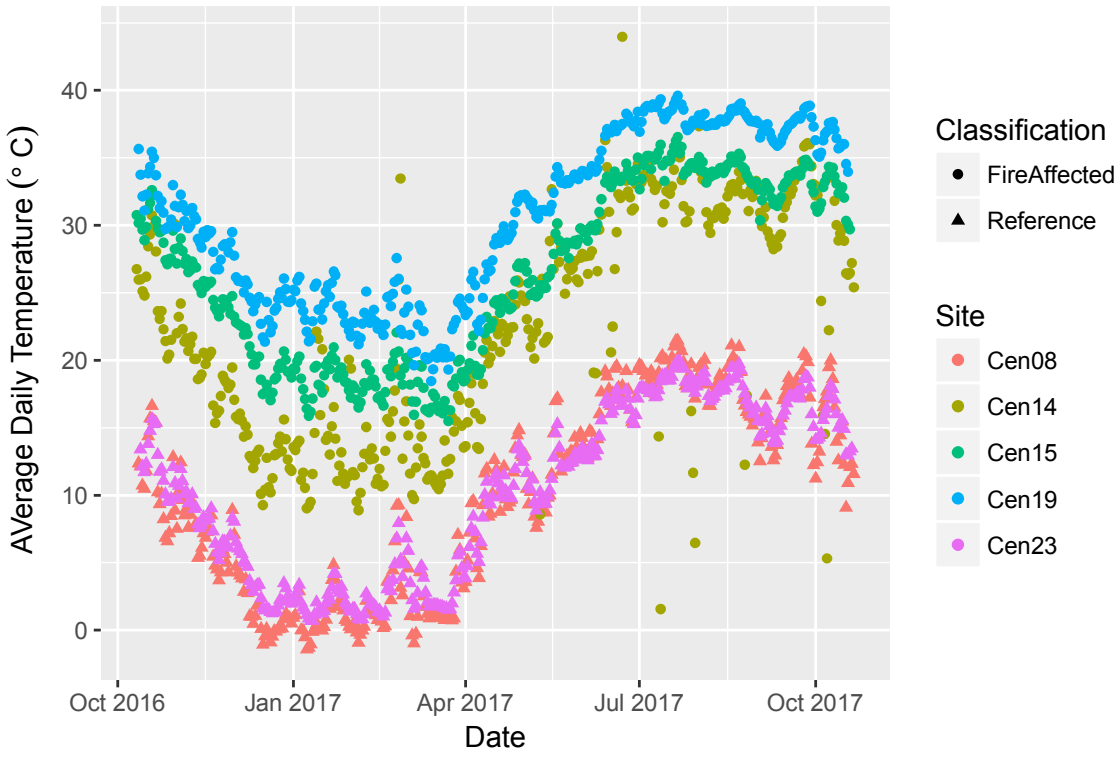

Figure F.3. Annual temperature fluctuations at three fire-affected and two ambient Centralia sites

Annual temperature fluctuations at three fire-affected (circles) and two ambient (triangles) Centralia sites, measured using *in situ* temperature loggers (HOBOs) that were buried 5 - 10 cm below the surface. Temperature loggers were deployed after the soils were collected for this study.

REFERENCES

REFERENCES

1. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science.* 351, 6280 (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, (2016).

5. Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci.* 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, (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. Microbial communities in fire-affected soils. in *Microbiology of Extreme Soils* 299–316 (Springer, 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, (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. Torre, R. De, Dodsworth, J. A. & Hungate, B. Measuring Nitrification , Denitrification , and Related Biomarkers in Terrestrial Geothermal Ecosystems. 486, (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.* fiy016, (2018).

21. Gao, Z. M. *et al.* Symbiotic Adaptation Drives Genome Streamlining of the Cyanobacterial Sponge Symbiont 'Candidatus Synechococcus spongiarum'. *MBio* 5, 1–11 (2014).

22. Grote, J. *et al.* Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *MBio* 3, 1–13 (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 1. 956, 951–956 (1999).

25. Hoch, J. A. Two-component and phosphorelay signal transduction. 165–170 (2000).

26. Whitworth, D. E. & Cock, Æ. P. J. A. Evolution of prokaryotic two-component systems : insights from comparative genomics. 459–466 (2009). doi:10.1007/s00726-009-0259-2

27. Ranea, J. A. G., Grant, A., Thornton, J. M. & Orengo, C. A. Microeconomic principles explain an optimal genome size in bacteria. 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 (80-.).* 326, 1263–1268 (2009).

31. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology* 10, 13–26 (2012).

32. Kussell, E. & Leibler, S. S. Phenotypic diversity, population growth , and information in fluctuating environments. *Science (80-.).* 309, 2075 (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, 1153–65 (2014).

36. Giovannoni, S. J. *et al.* Genetics: Genome streamlining in a cosmopolitan oceanic bacterium. *Science (80-.).* 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. Microbiology* 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, (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. Microbial Genomes Atlas: Standardizing genomic and metagenomic analyses for Archaea and Bacteria.

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).

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. (2017).

50. Komsta, L. outliers: Tests for outliers. *R package version 0.14* . http://CRAN.R-project.org/package=outliers (2011). doi:doi:10.1201/9780203910894.ch6

51. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer-Verlag, 2009).

52. Warnes, G. R. *et al.* gplots: Various R Programming Tools for Plotting Data. (2016).

53. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* (2017). doi:10.1038/nature24621

54. Lan, Y., Wang, Q., Cole, J. R. & Rosen, G. L. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS One* 7, (2012).

55. DeSantis, T. Z. *et al.* Greengenes: a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *App Env. Microb* 72, 5069–5072 (2006).

CHAPTER 4: Dormancy dynamics and dispersal contribute to soil microbiome resilience

This work is currently in review and revision at Philosophical Transactions B as
Sorensen, JW and Shade A. Dormancy dynamics and dispersal contribute to soil microbiome resilience

Abstract

In disturbance ecology, stability is composed of resistance to change and resilience towards recovery after the disturbance subsides. Two key microbial mechanisms that can support microbiome stability include dormancy and dispersal. Specifically, microbial populations that are sensitive to disturbance can be re-seeded by local dormant pools of viable and reactivated cells, or by immigrants dispersed from regional metacommunities. However, it is difficult to quantify the contributions of these mechanisms to stability without, first, distinguishing the active from inactive membership, and, second, distinguishing the populations recovered by local resuscitation from those recovered by dispersed immigrants. Here, we investigate the contributions of dormancy dynamics (activation and inactivation), and dispersal to soil microbial community resistance and resilience. We designed a replicated, 45-week time-series experiment to quantify the responses of the active soil microbial community to a thermal press disturbance, including unwarmed control mesocosms, disturbed mesocosms without dispersal, and disturbed mesocosms with dispersal after the release of the stressor. Communities changed in structure within one week of warming. Though the disturbed mesocosms did not fully recover within 29 weeks, resuscitation of thermotolerant taxa was key for community transition during the press, and both resuscitation of opportunistic taxa and immigration contributed to community resilience. Also, mesocosms with dispersal were more resilient than mesocosms without. This work advances the mechanistic understanding of how microbiomes respond to disturbances in their environment.

Introduction

Ongoing changes to Earth's climate are projected to alter disturbance regimes and to pervasively expose ecosystems to stressors like elevated atmospheric greenhouse gases and increased temperatures (1). Microbial communities, or *microbiomes*, provide vital ecosystem functions and are key players in determining ecosystem responses to environmental changes (2, 3). Understanding the mechanisms that underpin microbiome responses to environmental disturbances will support efforts to predict, and, potentially, manage, microbiomes for stable functions within their ecosystems.

In disturbance ecology, stability refers to consistent properties in the face of a stressor (4). Here, we apply terms from disturbance ecology as they have been adopted in microbial ecology (5–7). Stability includes components of both resistance and resilience. Resistance is the capacity of a system to withstand change in the face of a stressor, and its inverse is sensitivity. Resilience is the extent to which a system recovers following a disturbance, and is often expressed as a rate of change over time. Secondary succession is the process of community reassembly after a disturbance, and it can lead to either a state of recovery or an alternative stable state. Recovery is when a system fully returns to either its pre-disturbance state or is indistinguishable from a comparative control, and this term can be applied both to the state of the stressor and to the responsive community. Similarly, an alternative stable state is when the system does not return but rather assumes a different state. Together, resistance and resilience are the major quantifiable components of stability, and they can be calculated from community measurements of alpha diversity, beta diversity, or function (6, 8).

There are two related microbial mechanisms that support population persistence in the face of disturbance, and therefore contribute to community resistance, resilience, and recovery. One

mechanism is microbial dispersal, as successful immigrants can support resilience and recovery of sensitive populations. Across an interconnected landscape, microbial metacommunities are linked via dispersal, and so immigrants originate from the regional species pool (9–12). A second important but less-considered mechanism is microbial dormancy dynamics (13, 14). Dormancy dynamics include initiation and resuscitation. Initiation into dormancy can support local survival of populations sensitive to the disturbance, and therefore support community resistance by stabilizing community structure. Resuscitation from dormancy can support resilience and recovery by re-seeding sensitive populations from the local dormant pool. Thus, while both dispersal and resuscitation can support microbiome stability, dispersed immigrants originate regionally while resuscitated members originate locally. After a disturbance, if sensitive populations are not repopulated via immigration or resuscitation, they will become locally extinct and contribute to necromass (aka relic DNA, (15)).

We designed a replicated time-series experiment to quantify the contributions of dormancy dynamics and dispersal to the response of a soil microbiome to a thermal press disturbance. We targeted a soil microbiome because terrestrial microbiomes are front-line responders to climate change and sequesters of carbon (2, 3), and therefore an important constituent to understand for predicting ecosystem outcomes to environmental change. Also, soils harbor the highest known microbial diversity (16–18) and present a maximum challenge in deciphering microbiome responses to disturbance. Furthermore, a majority of the microbial cells or richness in soil is dormant (13, 19), reportedly as high as 80%, representing a considerable pool of microbial functional potential. Finally, across heterogeneous soils, an average of 40% of the microbiome DNA was necromass that existed extracellularly (15). This suggests that DNA-based methods of determining microbiome dynamics include both inactive and necromass

reservoirs, and that there is need for increased precision to move forward to quantify mechanisms underpinning microbiome stability.

The mesocosm experiment reported here follows prior field work in Centralia, Pennsylvania (20–24). Centralia is the site of an underground coal seam fire that ignited in 1962 and advances 5-7 my^{-1} along the coal seams (25, 26). The coal seams are highly variable in depth, but average 70 m below the surface (25), so as the fire advances underground it warms the overlying surface soils from ambient to mesothermal to thermal conditions. After the fire advances, previously warmed soils cool to ambient temperatures. In the field, we observed that previously warmed soils recovered towards reference soils in bacterial and archaeal community structure, with the exception of a slightly increased selection for Acidobacteria in the recovered soils (attributable to lower soil pH after coal combustion, (20)). However, during fire impact, there was high divergence among soil communities, and we hypothesized that differences in dormancy dynamics (e.g., different members resuscitating and initiating priority effects during the stress) may explain the divergences. We also hypothesized that resuscitation would shift community structure during the thermal disturbance, but that resuscitation and dispersal would together support resilience after the disturbance subsided. Therefore, in this experiment, we aimed to control dispersal, and also to quantify activity dynamics and determine their consistency and test our hypotheses.

Materials and Methods

Soil collection, mesocosm design, and soil sampling

Eight kg of soil was collected in Whirlpack bags from the top ten centimeters of a reference site in Centralia, PA (site C08, 40 48.084N 076 20.765W) on March 31st, 2018. The

site is temperate with the following chemical-physical properties: Organic Matter 4.8%; Nitrate 7.9 ppm; Ammonium 20.5 ppm; pH 5; Sulfur 19 ppm; Potassium 69 ppm; Calcium 490 ppm, Magnesium 59 ppm; Iron 110 ppm, and Phosphorus 395 ppm. The ambient soil temperature when collected was 4°C. The sample was stored at 4°C until the experiment was initiated. Soil was sieved through a 4mm mesh, homogenized, and ~300 g were dispensed into 15 autoclaved quart-sized glass canning jars that were used as mesocosms (Ball). The homogenized soil sample intentionally was used in all 15 mesocosms to assess the reproducibility of community temporal dynamics starting from the same soil source. Percent soil moisture was determined using by massing and drying. Each mesocosm was massed weekly to assess evaporation and any loss of water mass was replaced with sterile water to maintain percent soil moisture throughout the experiment. Sterile metal canning lids were secured loosely to prevent anaerobiosis. All set-up and manipulation of the mesocosms was performed in a Biosafety Level 2 cabinet (ThermoScientific 1300 Series A2) and we used aseptic technique.

Mesocosms first were acclimated at 14°C to mimic the ambient soil temperature at the typical time of fall soil collection and to coordinate with our previous field study (20). Acclimation proceeded for four weeks in a cooling incubator (Fischer Scientific Isotemp), and then soils were divided into three treatment groups (**Figure 4.1**). Six unwarmed control mesocosms ("Control") were maintained at 14°C for the duration of the experiment. Nine warmed mesocosms ("Disturbance") were subjected to a 12-week disturbance regime to simulate a press thermal disturbance. First, the temperature was gradually increased to 60°C, by 3°C to 3.5°C daily increments over two weeks. Second, the temperature was maintained at 60°C for 8 weeks. Sixty degrees was chosen because it was close to the observed maximum thermal temperature that we have measured in surface soils impacted by the Centralia coal seam fire (20).

Next, the temperature was gradually decreased to 14°C, by 3°C to 3.5°C daily increments over two weeks. Finally, the mesocosms were maintained at 14°C for four weeks until the penultimate sampling. From the nine disturbed mesocosms, four were randomly selected for the dispersal treatment ("Disturbance + Immigration"). These four disturbed mesocosms received a dispersal event one week after the temperature was recovered to 14°C after the thermal disturbance. Each was inoculated with 0.5 mL of a 10% weight by volume soil slurry made from a composite soil sample from the six unwarmed control mesocosms, and then gently mixed with a sterile spatula. Using qPCR data from control mesocosms at week 16, we estimate that approximately $6.37x10^6$ cells were dispersed into each Disturbance + Immigration mesocosm. We used soil from the control mesocosms to simulate dispersal from similar, adjacent soils to repopulate disturbed communities, as expected in the field. Finally, all mesocosms were left undisturbed at 14°C for another 25 weeks prior to the final 45-week sampling. During the final 25-week incubation, percent moisture was not monitored.

Mesocosms were non-destructively sampled after 4, 5, 6, 10, 14, 15, 16, 20, and 45 weeks of incubation. At each time point, approximately 15 g soil was removed from a mesocosm, of which ~13 g was flash-frozen in liquid nitrogen for RNA preservation and stored at -80°C until RNA/DNA co-extraction.

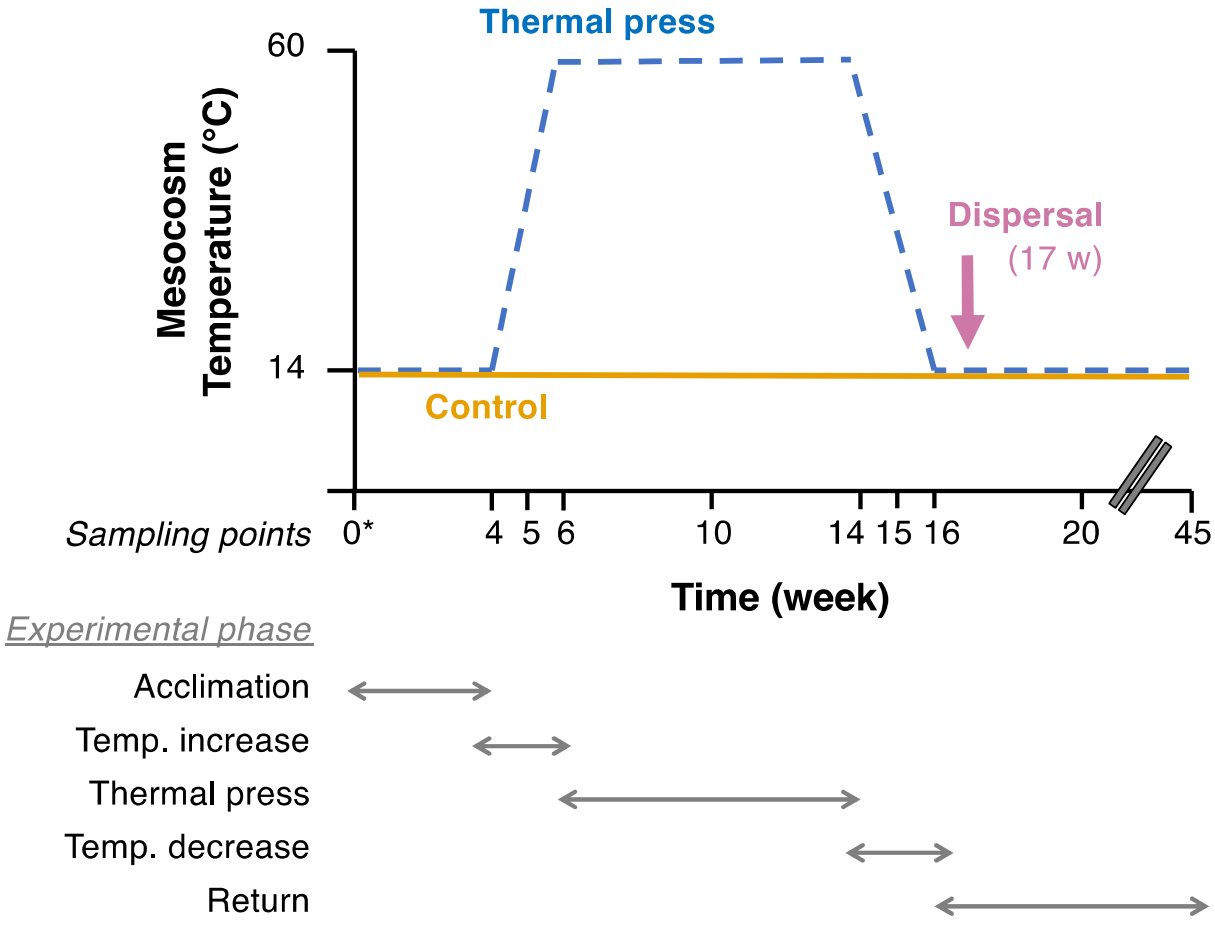

Figure 4.1. Experimental design of the mesocosm study.

At time 0 (indicated by the asterisk), reference temperate soil (0-20 cm depth from surface) was homogenized and divided among fifteen 1 L glass mesocosms that were maintained at ambient moisture through the experiment. Nondestructive sampling of each mesocosm proceeded from week 4 onward as indicated by the x-axis. Unwarmed Control mesocosms (solid gold line, n = 6) were maintained at 14°C, which was ambient soil temperature at the time of collection. Disturbed mesocosms (dashed blue line, n = 9, including Disturbance and Disturbance + Immigration groups) were acclimated for four weeks at 14°C, increased to 60°C over two weeks, maintained at 60°C as a thermal press disturbance for eight weeks, then decreased back to 14°C

Figure 4.1. (cont'd)

over two weeks, and finally maintained for a total of 45 weeks. Four of the disturbance mesocosms received homogenized soil slurry from Control mesocosms as a dispersal event at week 17, after the thermal press was released (Disturbance + Immigration treatment; see methods). Note the break in the x-axis time scale between weeks 20 and 45.

RNA/DNA co-extraction

To obtain RNA and DNA from the same cell pool, we minimally modified a manual coextraction protocol originally published by (27). For each sample, 0.5 g of flash-frozen soil was added to Qiagen PowerBead Tubes containing 0.70 mm garnet beads. Next, 500 uL of a 5% CTAB/Phosphate buffer and 500 uL of phenol:chloroform:isoamyl alcohol were added to each PowerBead tube. Cells were then lysed using a Model 607 MiniBeadBeater-16 (BioSpec Products Inc.) for 30 seconds, followed by a 10 min centrifugation at 10,000 x g and 4°C. The top aqueous layer was transferred to a fresh tube and 500 uL chloroform:isoamyl alcohol was added. The tubes were inverted several times to form an emulsion before a five minute centrifugation at 16,000 x g and 4°C. The top aqueous layer was transferred to a clean 1.5 mL centrifuge tube. Nucleic acids were precipitated by adding two volumes of a 30% PEG6000 1.6M NaCL solution, inverting several times to mix, and incubating on ice for two hours. After incubation, nucleic acids were pelleted by a 20 min centrifugation at 16,000 x g and 4°C. The supernatant was removed from each tube and one mL of ice-cold ethanol was added to the pelleted nucleic acids. Tubes were centrifuged for 15 min at 16,000 x g and 4°C, and the ethanol supernatant was removed. Pelleted nucleic acids were left to air dry before resuspending in 30 uL of sterile DEPC-treated water.

To purify the RNA, co-extracted nucleic acids were diluted 1:100 before treatment with Ambion Turbo DNA-free DNase kit, using the robust treatment option in the manufacturer's instructions. Extracted nucleic acids were mixed with 0.1 volumes of the 10X Turbo DNase Buffer and three uL of TURBO Dnase enzyme (six units total) and incubated at 37°C for 30 min. After incubation, 0.2 volumes of DNase inactivation reagent was added and incubated for five minutes at room temperature before a five min centrifugation at 2,000 x g and room temperature.

The treated supernatant was removed and used as the template for reverse transcription. RNA purity was assessed by PCR (see below for details) and showed no amplification. Reverse transcription was performed with random hexamers using the SuperScript III First-Strand Synthesis System for RT-PCR(Invitrogen) per manufacturer's instructions.

PCR of cDNA and no-RT controls was performed using the Earth Microbiome Project 16S rRNA gene V4 primers(515F 5'-GTGCCAGCMGCCGCGGTAA-3', 806R 5'-GGACTACHVGGGTWTCTAAT-3') (16, 28). Temperature cycling was as follows: 94°C for four minutes followed by 30 cycles of 94°C for 45 seconds, 50°C for 60 seconds and 72°C for 90 seconds followed by a final elongation step at 72°C for 10 minutes. Products were visualized using gel electrophoresis.

16S rRNA and 16S rRNA gene sequencing and processing

Here, for simplicity we use "microbiome" to refer to the bacterial and archaeal community members captured by amplifying and Illumina sequencing of the 16S ribosomal RNA and DNA (rRNA gene). Library preparation and sequencing was performed by the Michigan State University Genomics Core Research Facility. A single library was prepped using the method in Kozich et al (2013) (29). PCR products were normalized using Invitrogen SequalPrep DNA Normalization Plates. This library was loaded onto 4 separate Illumina MiSeq V2 Standard flow cells and sequenced using 250bp paired end format with a MiSeq V2 500 cycle reagent cartridge. Base calling was performed by the Illumina Real Time Analysis (RTA) V1.18.54.

All samples were first checked for any contaminating primer sequences using cutadapt(30), before being processed together using the USEARCH pipeline (31, 32). Briefly,

paired end reads were merged using -fastq_mergepairs and then dereplicated using -fastx_uniques. Reads were clustered *de novo* at 97% identity and then the original merged reads were mapped to the representative sequences of each cluster. Each OTU was classified using SINTAX(33) and with the Silva database (version 123, (34)).

Designating Total and Active Communities

Each RNA and DNA sample was rarefied to 50,000 reads in R using the vegan package version 2.5-4 (35) discarding any samples which did not contain sufficient reads (**Figure I.1**). Samples for which either the RNA or DNA did not have 50,000 reads were omitted from the analysis presented here (12 out of 135 in total). The Total community was defined as the community recovered in the DNA reads. The Active community was defined per sample, using the DNA read numbers of those taxa that had 16S rRNA:rRNA gene ratio was >1 in each sample(36). Consequently, while every sample was initially rarefied to 50,000 reads, each sample's active community varied slightly in total reads. Finally, we did not include taxa that had undefined rRNA:rRNA gene ratios ("phantoms") in the analysis (**Figure I.2**, see discussion in supplementary materials).

Quantitative PCR (qPCR)

qPCR was performed on the V4 region of the 16S rRNA gene and conducted in a BioRad CFX qPCR machine using the Absolute QPCR Mix, SYBR Green, no ROX (Thermo Scientific). Each reaction contained 12.5ul of the 2X Absolute QPCR Mix, 1.25 ul each of 10uM primers 515F and 806R, 3uL of template DNA and 2uL of PCR grade water. Temperature cycling conditions were as follows: 15 minutes at 95°C, followed by 39 cycles of 94°C for 45 seconds,

50°C for 60 seconds, and 72°C for 90 seconds, followed by a final elongation step at 72°C for 10 minutes. Fluorescence was measured in each well at the end of every cycle. Extracted gDNA from *E. coli* MG1655 was used for the standard curve, and was run in triplicate with every plate. Samples were run in duplicate across different plates and those that amplified after the lowest point of the standard curve (27 copies per reaction) were treated as zeroes. No template controls were included in every qPCR plate and they never amplified. Amplification specificity was assessed by melt curve (60°C to 95°C, 0.5°C increments).

Calculating resistance and resilience of community structure

We calculated resistance and resilience as described in Shade and Peter 2012 (6) and Orwin and Wardle 2004 (8). These are unitless metrics that have a theoretical range from -1 to 1. Resistance of the active community structure at week 10 was calculated for every disturbed mesocosm using Equation 1:

Eq. 1

$$RS = 1 - \frac{2*|y_c - y_d|}{y_c + |y_c - y_d|}$$

, where y_c is the mean Bray Curtis similarity for Control mesocosms at week 10 compared to week 4 (pre-disturbance), and y_d is the individually calculated Bray Curtis similarity of each disturbed mesocosm at week 10 to week 4. Resilience of the active community in each disturbed mesocosm was calculated for the observed secondary succession (week 16 to 45) as well as the initial (week 16 to 20) and the long-term (week 20 to 45) secondary succession using Equation 2.

Eq 2.

$$RL = \frac{2 * |y_{c,s} - y_{d,s}|}{(|y_{c,s} - y_{d,s}| + |y_{c,e} - y_{d,e}|)} - 1$$

, where *s* is the start of the secondary succession and *e* is the end, $y_{c,s}$ is the mean Bray Curtis similarity of Control mesocosms at week *S* to week 4 (pre-disturbance), $y_{d,s}$ is the Bray Curtis similarity of each disturbed mesocosm at week S to week 4 (pre-disturbance), $y_{c,e}$ is the mean Bray Curtis similarity of Control mesocosms at week *e* to week 4, and $y_{d,e}$ is the Bray Curtis similarity of each disturbed mesocosms at week *e* to week 4.

Ecological statistics

Ecological analyses were performed in R (37). The adonis and anosim function in the vegan package was used to perform PERMANOVAs (38) and ANOSIM respectively, to assess disturbance and immigration effects on community composition, and the betadisper function was used to quantify beta dispersion (39) with Tukey's Honestly Significant Difference post-hoc test across Control, Disturbance, and Disturbance + Immigration treatments. Pairwise tests for alpha diversity (Richness and Pielou's Evenness), community size (i.e. 16S rRNA gene copies per gram of soil), and resilience values were performed using the Kruskal-Wallis test, with Dunn's post-hoc correction for multiple comparisons when needed to assess differences between control, disturbance, and immigration treatments. Principal coordinates analysis was used for ordination of pairwise sample differences based on Bray-Curtis dissimilarity. Procrustes superimposition (PROTEST) was performed using the procrustes function in the vegan package to compare community structure trajectories in direction and extent of change and a false discovery rate adjustment was used for multiple tests. Data visualizations were performed using ggplot2 (40). Heatmaps were made using the heatmap.2 function in the gplots package (41).

To understand potential roles of dormancy initiation and resuscitation in driving community resistance and resilience, we distinguished between taxa that changed in their activity from taxa that changed in their detection over the course of the disturbance. Taxa that fell below detection (there was no rRNA gene detected in a particular sample) were coded differently for the heatmap than taxa that became inactive (rRNA:rRNA gene shifted from > 1 to < 1). For the heatmap, we used the Active community for the input data, but coded taxa that fell below detection in the Total community as NAs to distinguish them from inactive taxa, which were coded as 0. Notably, taxa that fell below detection in the Total community could have been either active, inactive, or locally extinct. To conservatively attribute activity dynamics, we restricted the heatmap visualization only to the taxa that were among the 50 most abundant in Active samples over the course of the experiment.

Responsive taxa were those that changed in activity over secondary succession (between weeks 16, 20, and 45) by their 16S rRNA:rRNA gene ratio, either from < 1 to > 1 or > 1 to < 1. Immigrant taxa were undetected in all disturbed mesocosms at week 16, but detected in Control mesocosms at Week 16 and Disturbance + Immigration mesocosms at either week 20 or week 45 while remaining undetected in the Disturbance mesocosms. Contributions of responsive and immigrant taxa to beta diversity were calculated as the Bray-Curtis dissimilarity attributed to the responsive taxa subset and divided by the total Bray-Curtis dissimilarity, both calculated from the Total (DNA) community, as done previously to assess the contributions of conditionally rare taxa (42) and the contributions of core taxa (43) to beta diversity

Data availability and code

Sequence workflows, OTU tables, and statistical workflows to reproduce the analyses described here are available on GitHub (https://github.com/ShadeLab/PAPER_Sorensen_InPrep_Mesocosms). All raw sequence data are deposited in the NCBI Short Read Archive under BioProject PRJNA559185.

Results

Sequencing summary

In total, we sequenced 135 pairs of samples (cDNA and DNA) across nine timepoints and 15 mesocosms. We rarefied all samples to 50,000 reads, and removed those samples with fewer than 50,000 reads. This resulted in the removal of 12 samples and left 53 unwarmed Control, 36 Disturbance, and 34 Disturbance + Immigration pairs of samples. After rarefaction, sample richness ranged from 84 to 4,108, with 16,854 total OTUs observed, inclusive of both DNA and RNA datasets.

Overarching responses to the thermal press disturbance

Total community richness responded consistently and as expected to the thermal press disturbance. There was a notable bottle effect of maintaining field soil in mesocosms, indicated by the gradual decrease in richness over time in the unwarmed Control treatment (**Figure 4.2AB**). In the Disturbance treatment, there was a modest but statistically supported decrease in richness one week after warming from 14°C to 37°C (week 5 all Disturbance v. Control comparison, Kruskal-Wallis test, $p = 0.003$), and then a more substantial decrease after warming to 60°C at week 6 (Kruskal-Wallis test, $p = 0.002$). Disturbance community size decreased over

weeks four to seven and then maintained at a median of 1.03 x 10^7 rRNA gene copies per g soil (**Figure 4.3**). Control communities decreased until week seven (bottle effect) and then increased rapidly by week ten and generally stabilized at median of 2.98 X 10^8 16S rRNA gene copies/g soil (**Figure 4.3A**). Together, these results show that the warming treatment acted as an environmental filter, resulting either in death or population decreases past the limits of detection for taxa that were otherwise fit in unwarmed conditions. Furthermore, there was a weak increase in richness after the dispersal event in the Disturbance + Immigration treatment, relative to the Disturbance treatment (Kruskal – Wallis test p= 0.088 at week 20, and p = 0.168 at week 45), and this increase was also observed for community size, which approaches that of the unwarmed control (Kruskal – Wallis test Control vs Disturbance + Immigration p=0.11, Control vs Disturbance p=0.0004, Disturbance vs Disturbance + Immigration p=0.013) (**Figure 4.3B**). This suggests that the dispersal treatment was effective in promoting the process of recovery in richness and community size. Importantly, Disturbance and Disturbance + Immigration mesocosms were not significantly different in either richness nor community size prior to the immigration event (**Table H.1** and **Table H.2**) However, disturbed mesocosms did not completely recover richness to the level of the ambient Controls, even by week 45 (**Figure 4.2B**). Evenness followed the same overarching patterns as richness (**Figure 4.2CD**).

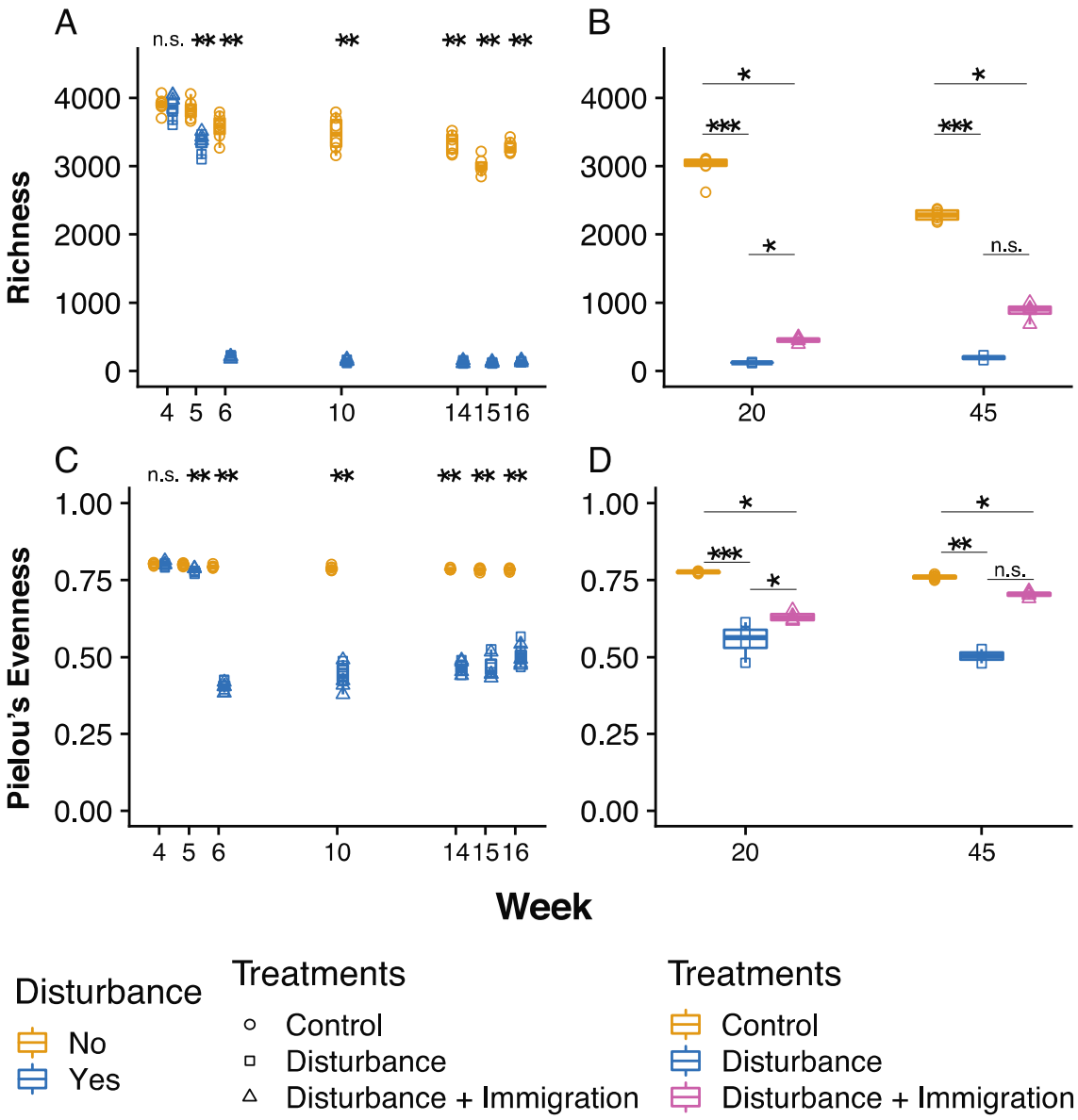

Figure 4.2. Changes in alpha diversity over the disturbance experiment.

Alpha diversity was assessed using operational taxonomic units clustered at 97% sequence identity, after 16S rRNA gene sequencing and rarefaction to 50,000 sequences per sample. (**A**) Changes in the observed no. OTUs (richness) in Control (gold, circles) and Disturbance (blue, squares and triangles) mesocosms over the thermal press (weeks 4-16). (**B**) Changes in richness in Control (gold circles), Disturbance (blue squares), and Disturbance + Immigration (pink triangles) mesocosms over the recovery period, weeks 20-45. The Disturbance + Immigration

Figure 4.2. (cont'd)

mesocosms received a dispersal event at week 17. (**C**) Changes in evenness over weeks 4-16. (**D**) Changes in evenness over weeks 20-45. Asterisks indicate significant differences by a Kruskal Wallis test (n.s = not significant; * $p<0.1$, ** $p<0.01$, *** $p<0.001$, with a Dunn correction for multiple comparisons in **B** and **D**).

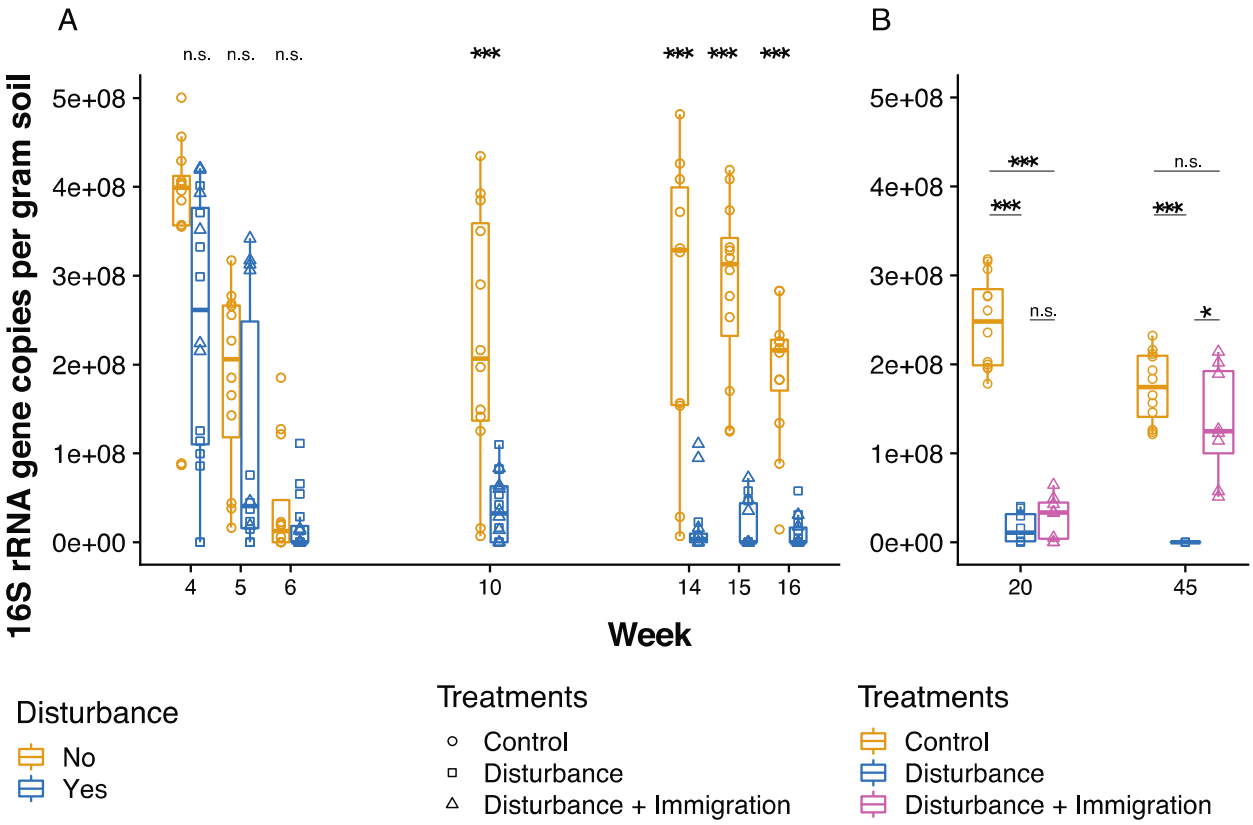

Figure 4.3. Changes in community size over the disturbance experiment.

Community size was estimated using qPCR of the 16S rRNA gene and standardized per gram of soil from which nucleic acids were extracted. (**A**) Changes in the 16S rRNA gene copies in Control (gold, circles) and disturbed (blue, squares and triangles) mesocosms over the thermal press (weeks 4-16). (**B**) Changes in the 16S rRNA gene copies in Control, Disturbance (blue squares) and Disturbance + Immigration (pink triangles) mesocosms over the recovery period, weeks 20-45. The Disturbance + Immigration mesocosms received a dispersal event at week 17. Asterisks indicate significant differences by a Kruskal Wallis test (n.s. = not significant, * $p<0.1$, ** $p<0.01$, *** $p<0.001$, with a Dunn correction for multiple comparisons in **B**).

We compared community structure across treatments for the Total community dataset, rRNA gene; 14,159 OTUs) and the Active dataset (rRNA:rRNA gene > 1; 6,693 = OTUs). There were clear and consistent shifts in beta diversity in the disturbed mesocosms (n=9, inclusive of Disturbance and Disturbance + Immigration), as well as high reproducibility among replicates in community structure within treatments as shown by the overlap of symbols per treatment and timepoint in the ordination (**Figure 4.4**). As compared to the Controls, the disturbed mesocosms had increased betadispersion (variability in community structure) starting at week 6 onward, with the exception of week 10 (**Figure 4.5**). Over the experiment, disturbed mesocosms had distinct community structures compared to Control (disturbed v. Control PERMANOVA PsuedoF = 63.87, Rsqr = 0.345, p=0.001 for Total communities, and PsuedoF=35.97, Rsqr=0.229, p=0.001 for Active communities, all timepoints). Control communities were relatively stable over the study, while disturbed communities changed directionally, and were significantly different from Control communities after a single week of warming (week 5 Control vs Disturbed PERMANOVA PsuedoF = 3.06, Rsqr= 0.218, p=0.001 for Total community and PsuedoF= 2.88, Rsqr=0.208, p=0.001 for Active community, **Table H.3**). Disturbed communities continued to shift with temperature during the course of the experiment, and then shifted slightly back towards the Control after the stressor was released and Disturbance and Disturbance + Immigration communities had similar structures during the press (**Table H.4**). Though no disturbed mesocosms fully recovered to overlap with the Control communities, the Disturbance + Immigration mesocosms were more similar to the Control than the Disturbance mesocosms without dispersal (**Figures 4.2B, 4.3B, 4.4**) . Across all treatments, Total communities and Active communities were synchronous in their temporal trajectories (Mantel R =0.943, p = 0.001 on 999 permutations; Protest Sum of Squares =0.238, R= 0.873, p=0.001), but there was higher

betadispersion in the disturbed treatments for the Active communities (Comparing Total v. Active for disturbed mesocosms, Kruskal Wallis $p=0.029$). This suggests that there was Active community variability masked by the contributions of dead and dormant taxa to the Total community.

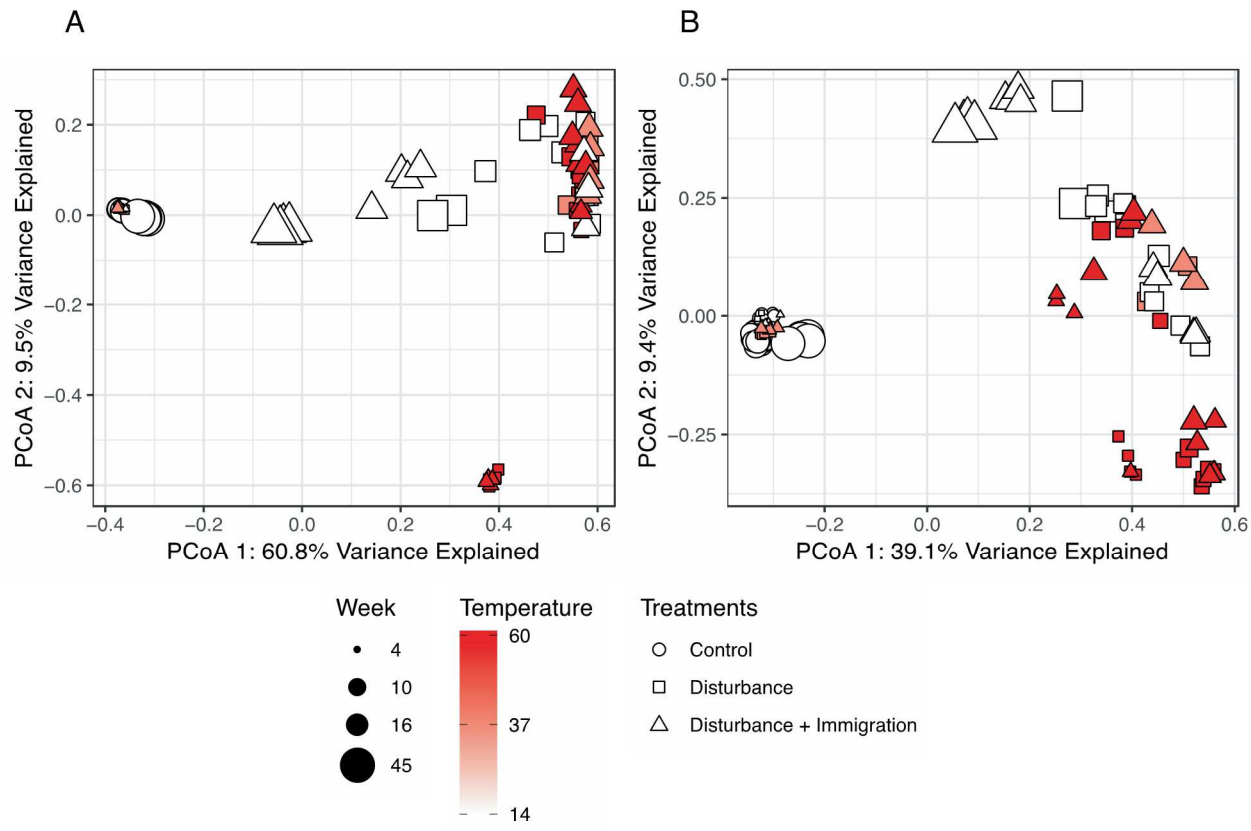

Figure 4.4. Changes in beta diversity over the disturbance experiment.

Pairwise differences in community structure was quantified using pairwise Bray-Curtis dissimilarity and then ordinated using Principal Coordinates Analysis (PCoA). Time is shown by symbol size, and mesocosm temperature is indicated by heat colors, with the brightest red indicating the warmest time point. Control mesocosms are circles, Disturbance are squares, and Disturbance + Immigration are triangles. (**A**) PCoA of the Total community, assessed using sequencing of the 16S rRNA gene. (**B**) PCoA of the Active community, including only OTUs that had 16S rRNA:rRNA gene > 1.

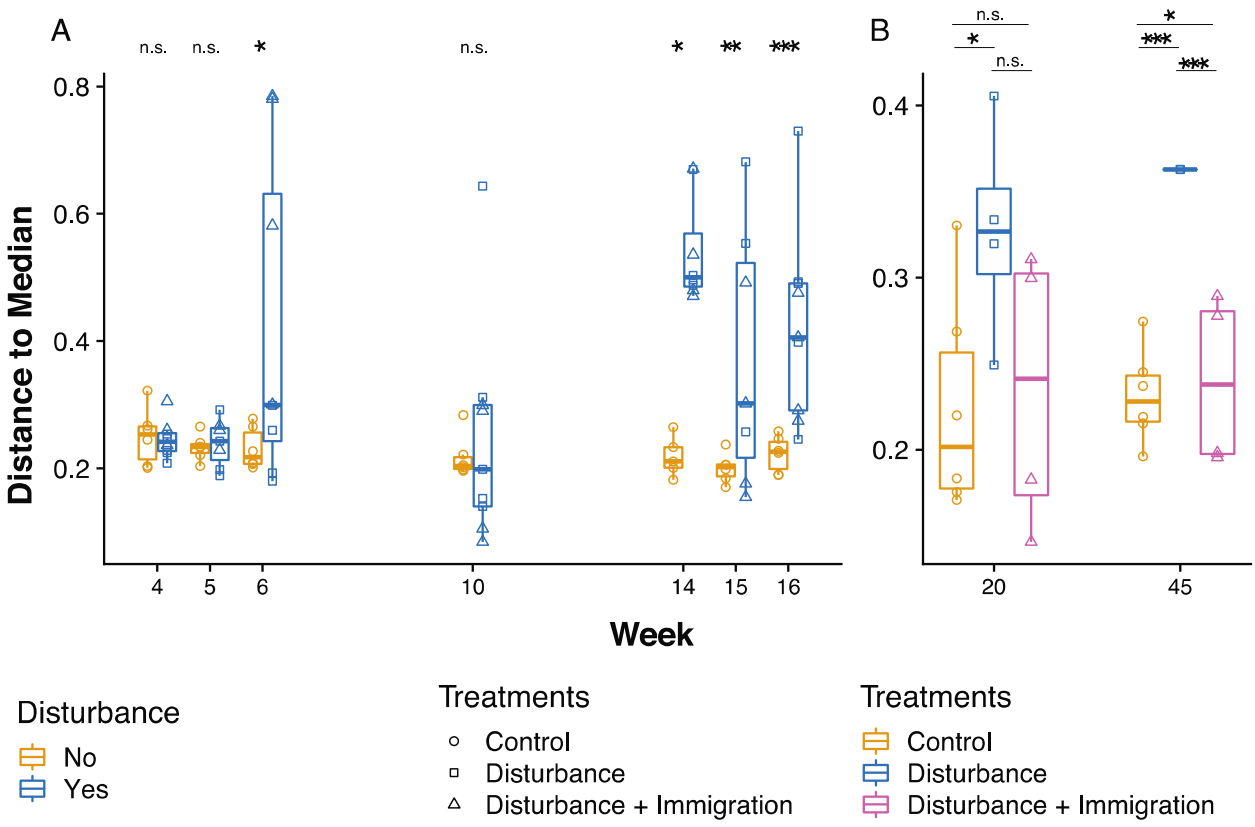

Figure 4.5. Changes in beta dispersion over the disturbance experiment.

Beta dispersion, an indicator of variability in community structure, was quantified using the distance to the median in ordination space (**Figure 4.4.**), which was constructed based on Bray-Curtis dissimilarity. (**A**) Changes in beta dispersion in Control (gold, circles) and Disturbance (blue, squares and triangles) mesocosms over the thermal press (weeks 4-16). (**B**) Changes in beta dispersion in Control, Disturbance (blue squares), and Disturbance + Immigration (pink triangles) mesocosms over the recovery period, weeks 20-45. The Disturbance + Immigration mesocosms received a dispersal event at week 17. Asterisks indicate significant differences with a Tukey's Honestly Significant Difference post-hoc test (n.s. = not significant, * $p<0.1$, ** $p<0.01$, *** $p<0.001$). Note differences in y-axis ranges between **A** and **B**.

Replicate disturbed mesocosms (again, inclusive of Disturbance and Disturbance + Immigration) had highly reproducible responses during the press. They had high overlap in membership and overall synchronous trajectories (i.e. changes in community structure through time), even after the immigration event at week 16 (33 of 36 PROTEST all $R > 0.89$ and false-discovery rate adjusted p-values < 0.05).

Resistance and resilience

For the Active community, we calculated resistance and resilience of the disturbed mesocosms relative to the Control using community divergence from the first sampling time (Week 4, end of acclimatization period) as the reference (**Figure 4.6A**). Even in the Control communities, there was an initial drop in similarity between weeks 4 and 5, which we attribute to incomplete acclimatization and a bottle effect. However, after that, the Control communities remain relatively stable with no additional divergence, while the disturbed communities decreased to their maximum divergence at week 10 (60°C).

Disturbance + Immigration communities converge slightly after the dispersal event. Overall resistance was low (**Figure 4.6B**), and resilience reached its maximum, 0.41, in the immigration treatment between weeks 16 (the time point at which the thermal press was released) and the final week 45, but ranged from a minimum of 0.04 between week 16 and 20 in the Disturbance without immigration treatment (**Figure 4.6C-E**). Immigration enhanced resilience from week 16 to week 20 (Kruskal Wallis p value 0.034) and from week 16 to week 45 (Kruskal Wallis p value 0.083), but not from week 20 to 45, possibly because of insufficient power (Kruskal Wallis p value 0.180). Notably, there were only two Disturbance mesocosm replicates (out of five) that met the rarefaction threshold for week 45.

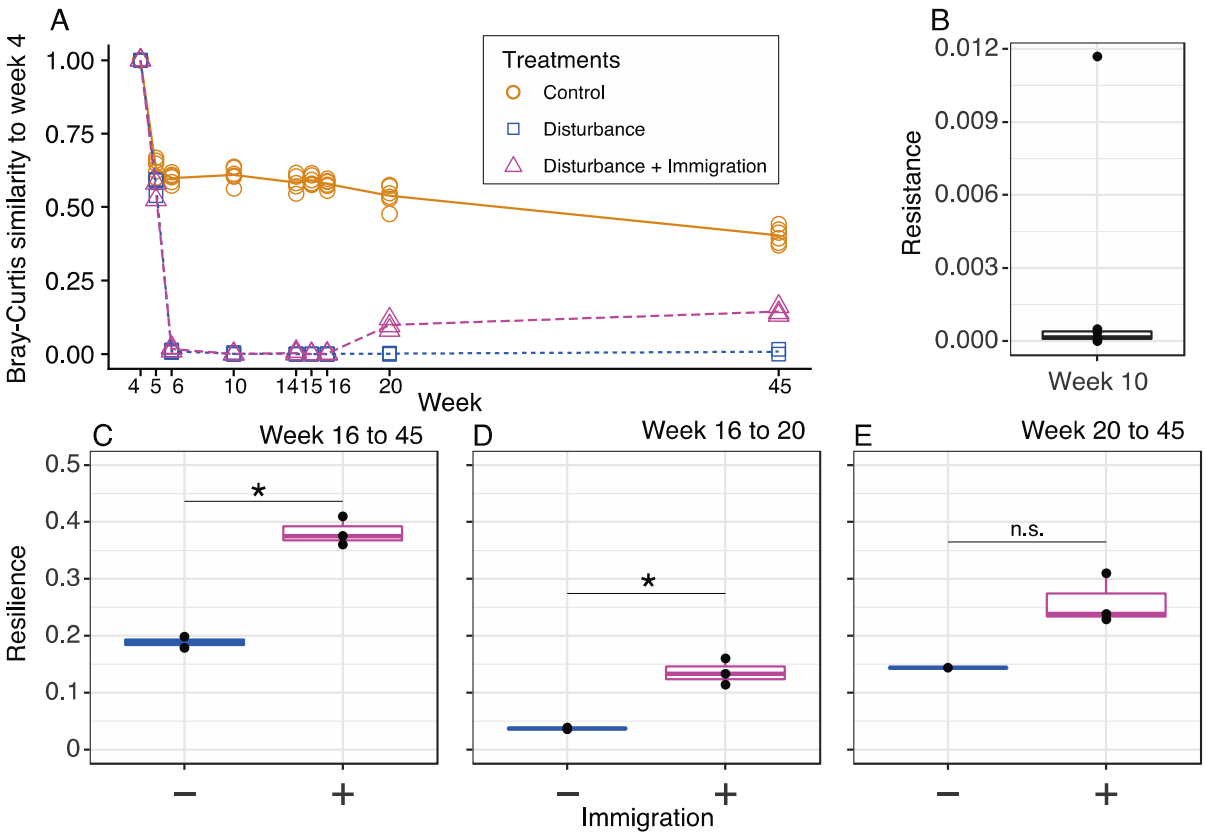

Figure 4.6. Resistance and resilience of soil mesocosm communities to a thermal press. (**A**) Temporal series of community divergence from pre-disturbance community (week 4) in Control (gold solid line), Disturbance (blue short dashed line), and Disturbance + Immigration (pink long dashed line) to calculate resistance and resilience. (**B**) Resistance of disturbed mesocosms at week 10, the time point of maximum community change after the thermal press begins. (**C-E**) Resilience of disturbed mesocosms without (-) and with (+) immigration, calculated after the thermal press is released (week 16) for the (**C**) full recovery to week 45, (**D**) initial recovery to week 20, and also for (**E**) long-term recovery from weeks 20 to 45. Asterisks indicate significant differences by a Kruskal Wallis test (n.s. = not significant, * $p<0.1$).

We wanted to assess the relative contributions of taxa that activate or inactivate after the disturbance subsides to the overall beta diversity (weeks 16-45). We also wanted to assess the relative contributions of taxa that colonized after dispersal. We calculated the relative contribution of activity dynamics by identifying taxa that switched between an active and inactive state during secondary succession. We found that these dynamically active taxa contributed 11.7% to 58.9% (median 28.6%) of the observed beta diversity, while immigrants contributed 7.9% to 26.3% (median 14.7%) of the observed beta diversity during the same time period**.**

Activity dynamics of abundant taxa

We investigated the activity dynamics of the top 50 most abundant taxa within the Active communities, and distinguished taxa that became inactive (rRNA:rRNA gene < 1, white cells in **Figure 4.7A**) from taxa that fell below detection (rRNA gene = 0, black cells in **Figure 4.7A**, see Methods for details). Within this set of 50, we detected no purely resistant taxa that were consistently active throughout the experiment. This finding agrees with the analyses showing low resistance (**Figure 4.6B**) and substantial shifts in the disturbed communities (**Figure 4.5**). We detected 17 taxa that were sensitive to the disturbance (**Figure 4.7B**). Sensitive taxa were active prior to the warming but became inactive or dropped below detection during the warming, and then did not reactivate. We also detected 19 transition taxa that were inactive prior to the warming, active during the warming, and then became inactive after the stressor was released. Because there was no external dispersal into the system, these thermotolerant taxa were likely in the dormant pool of the soil. We could divide these responses generally into early and late transition taxa. There were 6 early transition taxa that became active during week 5 or 6 of the

experiment, but then became inactive at weeks 10 and 14. There were also 13 late transition taxa that remained inactive during weeks 5 and 6 but became active during weeks 10 and 14.

Among the top 50 Active taxa, we did not detect purely resilient taxa that were active prior to the warming, became inactive during the warming, but then reactivated after the return to ambient temperature. This suggests that dormancy strategies responsive to warming were not a substantial contributor to member preservation, nor to eventual re-seeding. Instead, opportunists and immigrants facilitated resilience in the mesocosms. The opportunists were defined as inactive or below detection prior to and during the warming, but then activated after the temperature returned, likely due to resuscitation, and there were five taxa in this category. Eight immigrants were generally active prior to the warming, dropped to below detection or became inactive during the warming, and then in the end, were active again only in the Disturbance + Immigration treatment (and not in the Disturbance mesocosms without immigration).

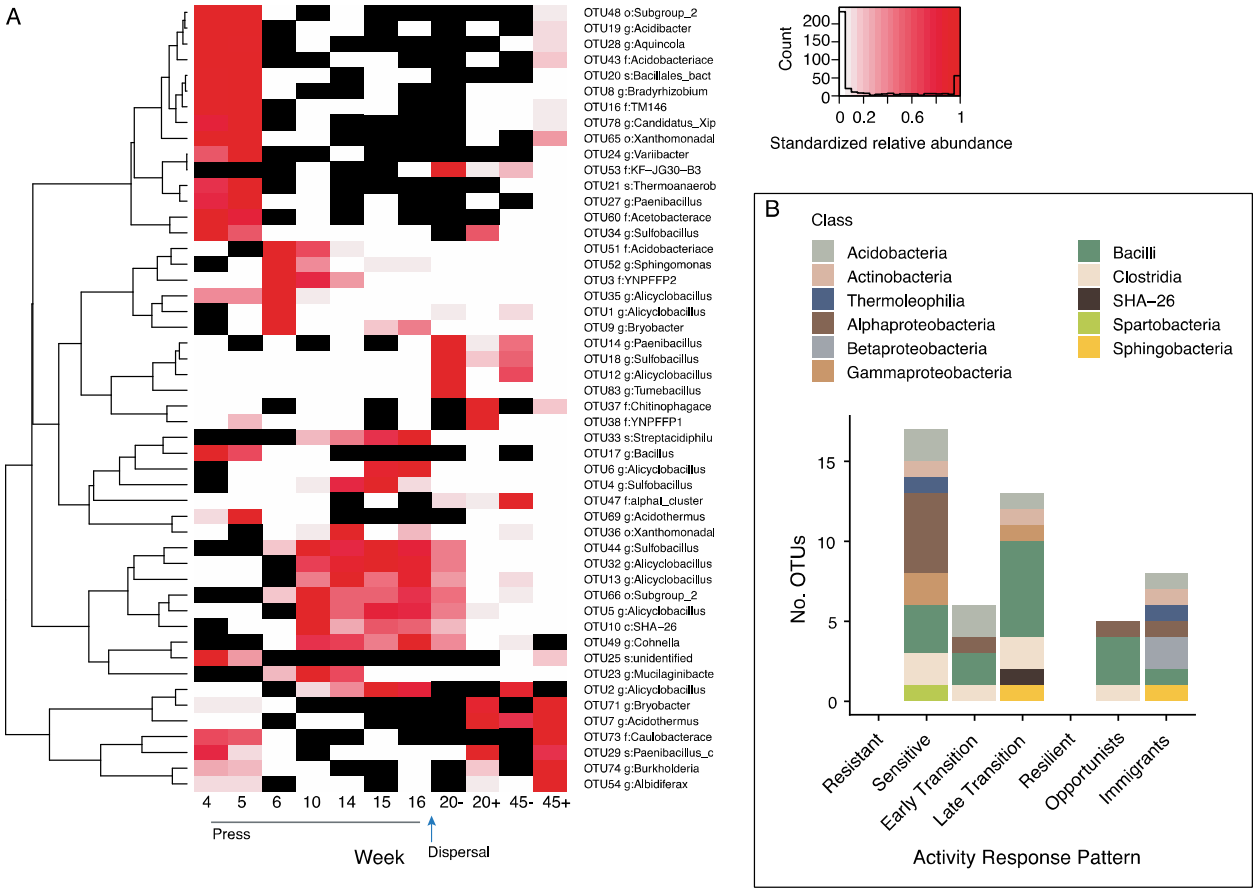

Figure 4.7. Activity dynamics of abundant taxa in response to the press disturbance.

(**A**) Heatmap and dendrogram of abundant taxa reveal common patterns of detection and activity. Black cells are taxa that were undetected (coded as NA) in the 16S rRNA gene (DNA) community, and white cells are taxa that were detected in the DNA but had 16S rRNA:rRNA gene < 1 (inactive, coded as 0). The heat gradient indicates each taxon's abundance relative to its maximum observed in disturbance treated mesocosms during the experiment. Immigration is indicated for weeks 20 and 45 by minus (no) and plus (yes) signs. (**B**) Summary of activity response patterns to the disturbance of the top 50 taxa, including resistant, sensitive, early and late transition, resilient, opportunist, and immigrant taxa. Definitions of each of these categories of taxa are found in the main text.

Discussion

Our results show that both dispersal and local dormancy dynamics, including activation and inactivation, can contribute to overarching patterns of community resilience. The dispersal event simulated in this experiment posed an optimistic scenario: well-mixed, control soils were mixed into disturbed soils to maximize the volume of the disturbed soil that came into contact with the inoculum. Regardless, by all metrics (beta diversity, alpha diversity, community size), immigration was impactful. These data directly show that dispersal can augment resilience towards recovery, supporting our hypothesis. Given that the influences of dispersal on community assembly has been investigated previously (often indirectly for bacterial and archaeal microbiomes, as inferred from the contributions of stochastic or neutral processes e.g., (20, 44–47)), this result is in agreement with the consensus of the literature that dispersal and dispersal limitation can matter for assembly (48–50).

A new result is that local resuscitation also contributes to microbiome community transitions during disturbance, and to resilience after the stress is released. Among the most abundant taxa, there were near equal numbers of taxa that contributed to resilience via resuscitation and to resilience via immigration. While, the influence of resuscitation on resilience was not as impactful as that of dispersal (**Figure 4.6**), changes in activity dynamics contributed 28.9% to the observed beta diversity during secondary succession. Therefore, both mechanisms – local resuscitation and regional immigration – contribute to microbiome stability, but potentially to different extents. The microbial dormant pool is important for maintaining microbial diversity (51) and has evolutionary implications for traits that persist within inactive populations (52). To make more explicit the role of dormancy dynamics for community disturbance responses (e.g., (53)), the phenomenon of the "storage effect" underpins modern coexistence theory (54) and

refers to the ability of competing species to coexist when their growth and activities are separately partitioned over time, typically in dynamic environments (55). Given the severity of the thermal stressor in Centralia and in this experiment, our results suggest that the soil microbial dormant pool is deep, in that it contains functionality for distinctive conditions, like thermal stress, that are not within the expected range of environmental variability. Our finding support other studies which have found thermophiles in unexpected environments such are arctic sediments and temperate soils (56–58).

Alternatively, it could be that, rather than local resuscitation, extremely rare but active taxa that were below the limits of detection grew rapidly and repopulated to become among the most active and abundant taxa. These data cannot rule out this possibility, and, if true, it would suggest an interesting role for release of rare taxa from competition (via death or inactivation of the competitors sensitive to the warming) in driving post-disturbance assembly. However, given that no resistant taxa were detected that could withstand the wide temperature range in the experiment, conditional rarity may be a less common scenario than opportunistic resuscitation.

Another goal of the experiment was to understand the reproducibility of member resuscitation given the press disturbance, and from the same soil. Because we observed high divergence in the hot soil communities in Centralia that was not attributable to any measured environmental variable, including temperature (20), we hypothesized that stochastic resuscitation could initiate priority effects (e.g., (10)), leading to divergent hot communities. However, we did not see the strongest differences in beta dispersion between Control and disturbed mesocosms until the press was subsiding (Weeks 15 and 16 in **Figure 4.5**). This, along with the overall strongly-correlated trajectories of disturbed community structures, suggest that the disturbance responses were consistent across disturbed mesocosms and do not support our hypothesis that

priority effects (initiated by different resuscitating membership) determines community structure during the press. Therefore, we interpret that resuscitation in response to the thermal stress was largely deterministic, and that observed divergences among hot soil communities in the field may be instead attributed to either differences local edaphic factors that were unmeasured, different structures of the underlying dormant pools, or stochasticity in regional dispersal during secondary succession.

Moving forward, there are several insights gleaned from this experiment. For soil, measuring dispersal in the field is difficult, given the various means by which microorganisms may arrive to a locality, including wind, ground water, and invertebrate vectors. Therefore, controlled experimentation is needed to quantify the contributions of dispersal to secondary succession. However, measuring activity dynamics and estimating the dormant pool of microbes in field samples, while imperfect, is possible (19, 36, 59, 60). Because our experiment suggests a role of resuscitation in determining the community that thrives during the disturbance, and also an influence of resuscitation for secondary succession towards recovery, we recommend to collect member activity data. More generally, routine characterization of the dormant pool of soil microbes, including its stability, diversity, and functions, can provide insights into the roles of these inactive taxa for disturbance responses.

Microbiome stability is a progression along a trajectory, including a pre-disturbance community with a variance around a mean structure or a routine seasonal dynamic, a transition to an ephemeral community structure during the disturbance, and finally, after the disturbance is released, secondary succession towards either recovery or an alternative stable state. Longitudinal series of microbiome structure inclusive of all stages of this trajectory can be informative. Characterizing the full disturbance trajectory will allow for quantification of the

different and potentially changing mechanisms that support stability (e.g., resuscitation, conditional rarity, immigration), and will facilitate prediction given new stressors. In our experiment, one week of stress was sufficient to observe community sensitivity (by week 5, the control and the disturbance treatments were statistically different), but 29 weeks after the stress was released was not sufficient to observe complete recovery, though it seems that recovery is possible given the trajectory toward the controls. We expect that this time frame of response may be typical for many soils (61) and it can be used to inform future studies.

To conclude, this experiment shows both dispersal and dormancy dynamics can contribute to soil microbiome resilience in response to a press stress. Specifically, resuscitation of thermotolerant members contributed to microbiome transition during press, and then immigration provided a substantial boost to recovery beyond what was achieved with resuscitated opportunists. Because activity responses to the disturbance were consistent, these results suggest that predictive insights into microbiome resilience can be advanced more generally. We expect that accounting for mechanisms of local resuscitation and regional dispersal together will advance quantitative understanding of environmental microbiome stability.

APPENDICES

APPENDIX G

Supplemental results

Supplemental Results

Relationships between taxon activity and abundance

The conventional thought is that relative abundance is the outcome of growth and therefore an indicator of fitness, and so high relative abundance is indicative of recent or current activity in the environment. However, we detected a weak, but statistically supported, inverse (log10) relationship between OTU 16S rRNA:rRNA gene ratio and relative abundance for those taxa with an rRNA:rRNA gene ratio >1 (**Figure I.2A**, Pearson's $R = -.14$, $p < 0.0001$). This result is in agreement with other studies that have suggested that rare taxa may have high activity levels relative to their abundance in the community (42–46). We present it here to be transparent that there are likely additional active but rare members that contribute to stability that have not been considered in our analyses.

The inverse relationship between activity and abundance could not include taxa that had RNA but no DNA detected (aka "phantom taxa", (44)) because they have an undefined 16S rRNA:rRNA gene ratio. We make clear that, to be conservative, phantom taxa (that have RNA but no DNA detected) were not included in the analyses, and that rare taxa that had high activity ratios were not included in the description of activity response patterns among the top 50 most abundant taxa. On balance, phantom taxa contributed proportionally few rRNA reads and few unique OTUs to the dataset (**Figure I.2 B and C**). However, there were a few exceptions, including five samples that had >10% rRNA reads and > 50% of richness attributed to phantom taxa. Four of these were from the Disturbance mesocosms at week 14 (peak-thermal press), and one sample was from week 16, at the end of the press. These samples also had relatively low richness and community size (**Figure 4.2** and **4.3**). We speculate that, by reducing community

size and likely also total microbial biomass, the disturbance indirectly provoked relatively higher contributions by phantom taxa and conditionally rare taxa (47).

APPENDIX H

Supplemental tables

Table H.1. Kruskal Wallis tests for Richness between Disturbance and Disturbance + Immigration mesocosms during the press.

Week	KW rank sum statistic	p value
4	5.00	0.025
5	1.13	0.289
6	5.33	0.021
10	0.96	0.327
14	0.02	0.885
15	2.00	0.157
16	1.50	0.221

Table H.2. Kruskal Wallis tests for on community size between Disturbance and Disturbance + Immigration treatments during press.

Week	KW rank sum statistic	p value
4	0.59	0.441
5	0.05	0.821
6	3.38	0.066
10	0.90	0.342
14	0.72	0.396
15	4.21	0.040
16	0.55	0.456

Table H.3. ANOSIM tests on influence of disturbance on community structure.

Week	ANOSIM R	P value
4	0.17	0.055
5	0.57	0.001
6	1.00	0.002
10	1.00	0.002
14	1.00	0.001
15	1.00	0.002
16	1.00	0.001
20	1.00	0.001
45	0.64	0.003

Table H.4. ANOSIM results of community structure differences between Disturbance and Disturbance + Immigration mesocosms during the press.

Week	ANOSIM R	p value
4	0.54	0.038
5	0.15	0.222
6	-0.06	0.515
10	-0.05	0.63
14	0.07	0.449
15	0.20	0.196
16	0.04	0.359

APPENDIX I

Supplemental figures

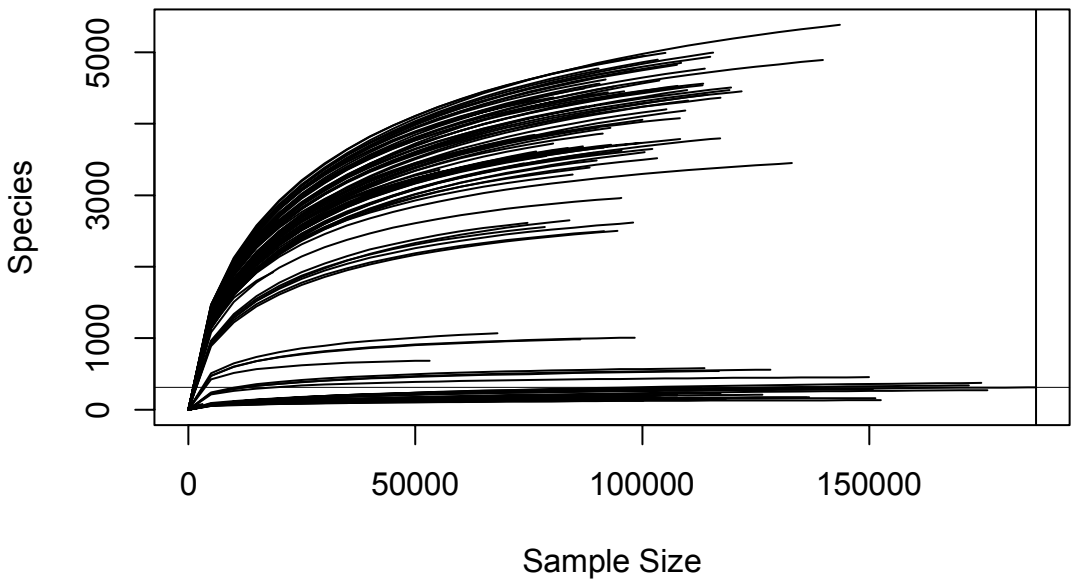

Figure I.1. **Rarefaction curves for soil mesocosm microbial communities.**

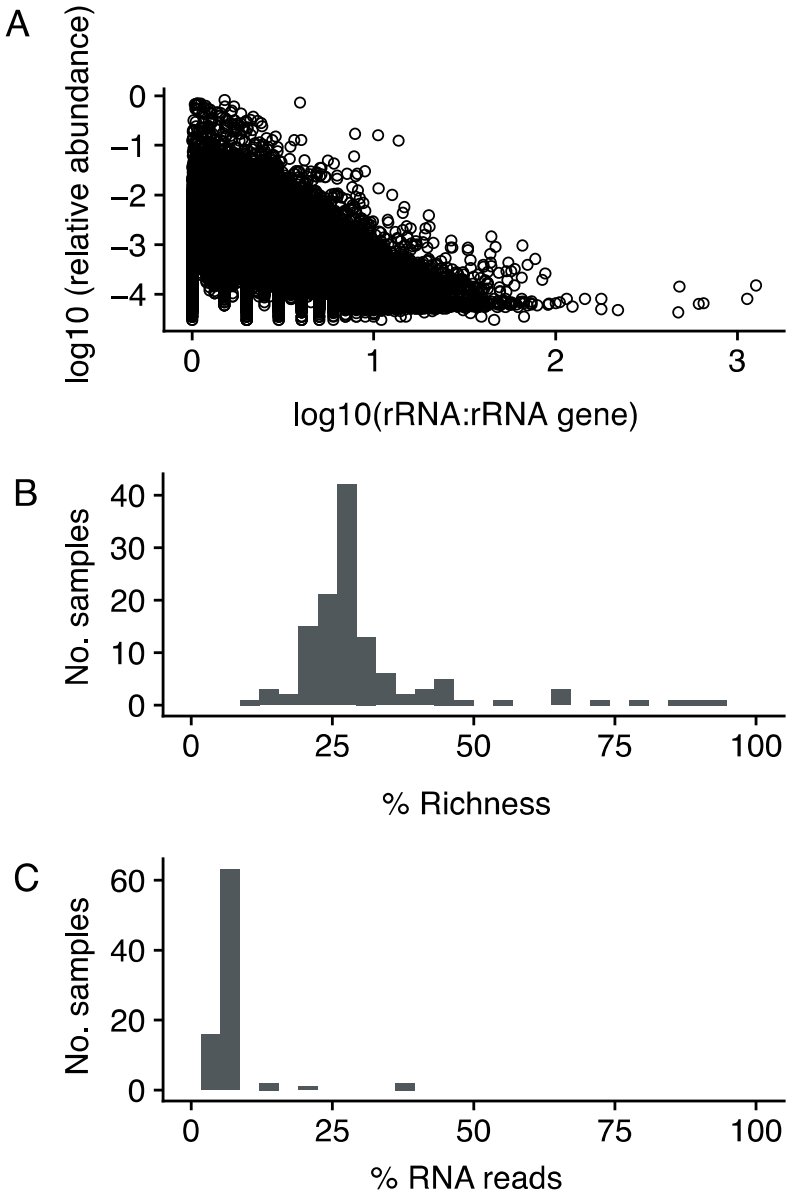

Figure I.2. Taxon activity and abundance relationships.

(A) Log10 relative abundance and log10 rRNA:rRNA gene ratio were inversely correlated. Each point is a different OTU detected in the dataset that had 16S rRNA:rRNA gene greater than or equal to 1. (B) Distribution of percent sample richness (No. OTUs detected, inclusive of DNA and RNA datasets) that were phantom taxa (16S rRNA detected but not 16S rRNA gene). (C) Distribution of percent RNA reads attributed to phantom taxa.

REFERENCES

REFERENCES

1. IPCC. 2014. Climate Change 2014Climate Change 2014: Synthesis Report.

2. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT, Crowther TW, Danovaro R, Foreman CM, Huisman J, Hutchins DA, Jansson JK, Karl DM, Koskella B, Mark Welch DB, Martiny JBH, Moran MA, Orphan VJ, Reay DS, Remais J V., Rich VI, Singh BK, Stein LY, Stewart FJ, Sullivan MB, van Oppen MJH, Weaver SC, Webb EA, Webster NS. 2019. Scientists' warning to humanity: microorganisms and climate change. Nat Rev Microbiol.

3. Singh BK, Bardgett RD, Smith P, Reay DS. 2010. Microorganisms and climate change: Terrestrial feedbacks and mitigation options. Nat Rev Microbiol.

4. Pimm SL. 1984. The complexity and stability of ecosystems. Nature 307:321–326.

5. Allison SD, Martiny JBH. 2008. Resistance, resilience, and redundancy in microbial communities. Proc Natl Acad Sci U S A 105:11512–11519.

6. Shade A, Peter H, Allison SD, Baho D, Berga M, Buergmann H, Huber DH, Langenheder S, Lennon JT, Martiny JBH, Matulich KL, Schmidt TM, Handelsman J. 2012. Fundamentals of microbial community resistance and resilience. Front Microbiol 3:417.

7. Kearns PJ, Shade A. 2018. Trait-based patterns of microbial dynamics in dormancy potential and heterotrophic strategy: case studies of resource-based and post-press succession. ISME J.

8. Orwin KH, Wardle DA. 2004. New indices for quantifying the resistance and resilience of soil biota to exogenous disturbances. Soil Biol Biochem 36:1907–1912.

9. Leibold MA, Holyoak M, Mouquet N, Amarasekare P, Chase JM, Hoopes MF, Holt RD, Shurin JB, Law R, Tilman D, Loreau M, Gonzalez A. 2004. The metacommunity concept: A framework for multi-scale community ecology. Ecol Lett 7:601–613.

10. Fukami T. 2015. Historical contingency in community assembly : integrating niches, species pools, and priority effects. Annu Rev Ecol Evol Syst 46:1–23.

11. Langenheder S, Berga M, Östman Ö, Székely AJ. 2012. Temporal variation of β-diversity and assembly mechanisms in a bacterial metacommunity. ISME J 6:1107–1114.

12. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and Processes of Microbial Community Assembly. Microbiol Mol Biol Rev 77:342–356.

13. Lennon JT, Jones SE. 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol 9:119–130.

14. Hawkes C V., Keitt TH. 2015. Resilience vs. historical contingency in microbial responses to environmental change. Ecol Lett 18:612–625.

15. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol 2:16242.

16. Thompson LRLRLR, Sanders JGJG, McDonald D, Amir A, Ladau J, Locey KJKJ, Prill RJRJ, Tripathi A, Gibbons SMSM, Ackermann G, Navas-Molina JAJA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JTJT, Mirarab S, Zech Xu Z, Jiang L, Haroon MFMFMF, Kanbar J, Zhu Q, Jin Song SS, Kosciolek T, Bokulich NANA, Lefler J, Brislawn CJCJ, Humphrey G, Owens SMSM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JAJA, Clauset A, Stevens RLRL, Shade A, Pollard KSKS, Goodwin KDKD, Jansson JKJK, Gilbert JAJA, Knight R, Rivera JLA, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll J, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart H-P, Guyton K, Haig S-J, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler A V., Krezalek M, Kueneman J, Lamendella R, Landon EM, Lane-deGraaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguin-Orlando A, Seldin L, Seyler LM, Shakhsheer

B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M, Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Tas¸ N, Taylor DL, Thomas T, Timling I, Turner BL, Urich T, Ursell LK, van der Lelie D, Van Treuren W, van Zwieten L, Vargas-Robles D, Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551.

17. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci Early Edit:1–6.

18. Louca S, Mazel F, Doebeli M, Parfrey LW. 2019. A census-based estimate of earth's bacterial and archaeal diversity. PLoS Biol.

19. Blagodatskaya E, Kuzyakov Y. 2013. Active microorganisms in soil: Critical review of estimation criteria and approaches. Soil Biol Biochem 67:192–211.

20. Lee S-H, Sorensen JW, Grady KL, Tobin TC, Shade A. 2017. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. ISME J 11:1447–1459.

21. Sorensen JW, Dunivin TK, Tobin TC, Shade A. 2018. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. Nat Microbiol.

22. Dunivin TK, Shade A. 2018. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. FEMS Microbiol Ecol fiy016.

23. Kearns PJ, Shade A. 2017. Trait-based patterns of microbiome succession in dormancy and heterotrophic strategy. PeerJ Prepr 5.

24. Tobin-Janzen T, Shade A, Marshall L, Torres K, Beblo C, Janzen C, Lenig J, Martinez A, Ressler D. 2005. Nitrogen Changes and Domain Bacteria Ribotype Diversity in Soils Overlying the Centralia, Pennsylvania Underground Coal Mine Fire. Soil Sci 170:191–201.

25. Nolter M a, Vice DH. 2004. Looking back at the Centralia coal fire: a synopsis of its present status. Int J Coal Geol 59:99–106.

26. Elick JM. 2011. Mapping the coal fire at Centralia, Pa using thermal infrared imagery. Int J Coal Geol 87:197–203.

27. Griffiths R, Whiteley A, O'Donnell A. 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of Ribosomal DNA- and rRNA-Based Microbial Community Composition. Appl Environ Microbiol 66:5488–5491.

28. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108:4516.

29. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–20.

30. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10-12.

31. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

32. Edgar RC, Flyvbjerg H. 2014. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482.

33. Edgar RC. 2016. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv.

34. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41.

35. Oksanen J, Kindt R, Legendre P, O'Hara B. 2006. The vegan Package for Community Ecology1.8-3.

36. Bowsher AW, Kearns PJ, Shade A. 2019. 16S rRNA/rRNA Gene Ratios and Cell Activity Staining Reveal Consistent Patterns of Microbial Activity in Plant-Associated Soil. mSystems 4:e00003-19.

37. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria.

38. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46.

39. Anderson MJ. 2005. Distance□Based Tests for Homogeneity of Multivariate Dispersions. Biometrics 62:245–253.

40. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Book, Springer, New York.

41. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. 2016. gplots: Various R Programming Tools for Plotting Data.

42. Shade A, Jones SESE, Caporaso JG, Handelsman JJ, Knight R, Fierer N, Gilbert JAAJA, Gregory Caporaso J, Handelsman JJ, Knight R, Fierer N, Gilbert JAAJA. 2014. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. MBio 5:e01371-14.

43. Grady KL, Sorensen JW, Stopnisek N, Guittar J, Shade A. 2019. Assembly and seasonality of core phyllosphere microbiota on perennial biofuel crops. Nat Commun 19.

44. Ferrenberg S, O'Neill SP, Knelman JE, Todd B, Duggan S, Bradley D, Robinson T, Schmidt SK, Townsend AR, Williams MW, Cleveland CC, Melbourne BA, Jiang L, Nemergut DR. 2013. Changes in assembly processes in soil bacterial communities following a wildfire disturbance. Isme J 7:1102–1111.

45. Dini-Andreote F, Stegen JC, van Elsas JD, Salles JF. 2015. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. Proc Natl Acad Sci 112:E1326–E1332.

46. Burns AR, Zac Stephens W, Stagaman K, Wong S, Rawls JF, Guillemin K, Bohannan BJ. 2015. Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. Isme J 1–10.

47. Zhou J, Liu W, Deng Y, Jiang YH, Xue K, He Z, Van Nostrand JD, Wu L, Yang Y, Wang A. 2013. Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. MBio.

48. Evans S, Martiny JB, Allison SD. 2016. Effects of dispersal and selection on stochastic assembly in microbial communities. ISME J 1–10.

49. Günther S, Faust K, Schumann J, Harms H, Raes J, Müller S. 2016. Species-sorting and mass-transfer paradigms control managed natural metacommunities. Environ Microbiol 18:4862–4877.

50. Nemergut DR, Knelman JE, Ferrenberg S, Bilinski T, Melbourne B, Jiang L, Violle C, Darcy JL, Prest T, Schmidt SK, Townsend AR. 2016. Decreases in average bacterial community rRNA operon copy number during succession. ISME J 10:1147–1156.

51. Jones SE, Lennon JT. 2010. Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci U S A 107:5881–5886.

52. Shoemaker WR, Lennon JT. 2018. Evolution with a seed bank: The population genetic consequences of microbial dormancy. Evol Appl 11:60–75.

53. Miller AD, Chesson P. 2009. Coexistence in disturbance-prone communities: how a resistance-resilience trade-off generates coexistence via the storage effect. Am Nat 173:E30–E43.

54. Warner RR, Chesson PL. 1985. Coexistence mediated by recruitment fluctuations - a field guide to the storage effect. Am Nat 125:769–787.

55. Barabás G, D'Andrea R, Stump SM. 2018. Chesson's coexistence theory. Ecol Monogr.

56. Hubert C, Loy A, Nickel M, Arnosti C, Baranyi C, Brüchert V, Ferdelman T, Finster K, Christensen FM, de Rezende JR, Vandieken V, Jørgensen BB, Bruchert V, Ferdelman T, Finster K, Christensen FM, Rosa de Rezende J, Vandieken V, Jorgensen BB. 2009. A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. Science (80-) 325:1541–1544.

57. Portillo MC, Santana M, Gonzalez JM. 2012. Presence and potential role of thermophilic bacteria in temperate terrestrial environments. Naturwissenschaften 99:43–53.

58. Marchant R, Franzetti A, Pavlostathis SG, Tas DO, Erdbrugger I, Unyayar A, Mazmanci M a., Banat IM. 2008. Thermophilic bacteria in cool temperate soils: Are they metabolically active or continually added by global atmospheric transport? Appl Microbiol Biotechnol 78:841–852.

59. Blazewicz SJ, Barnard RL, Daly R a, Firestone MK. 2013. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J 7:2061–8.

60. Dlott G, Maul JE, Buyer J, Yarwood S. 2015. Microbial rRNA: RDNA gene ratios may be unexpectedly low due to extracellular DNA preservation in soils. J Microbiol Methods 115:112–120.

61. Shade A, Gregory Caporaso J, Handelsman J, Knight R, Fierer N. 2013. A meta-analysis of changes in bacterial and archaeal communities with time. ISME J 7:1493–1506.

CHAPTER 5: Conclusions and Future Directions

Summary

The work presented in this dissertation used the coal fire in Centralia, PA as a model disturbance to answer questions about the disturbance ecology of soil microbial communities. Chapter 2 broadly looked at changes in microbial community diversity in response to and in recovery from temperature disturbance. Fire affected soil microbial communities harbored fewer microbial taxa and were more divergent in the community structure than either reference soils or recovered soils. Using the framework of Vellend(1) and Nemergut(2) to investigate this divergence in community structure, little support was found for the community assembly processes of drift, dispersal, or selection driving this observed divergence. We hypothesized that stochastic resuscitations of local dormant microbes initiated priority effects in the soils, and thereby causing the observed divergence. Further, despite this increased divergence in community structure during disturbance, soils that had recovered in temperature from the disturbance also showed clear signs of recovery of community. We proposed a conceptual model for the soil microbial community response to the coal fire wherein community structure was hypothesized to be driven by priority effects during the disturbance, and by weak environmental filtering post disturbance.

In Chapter 3, the traits and functional potential of the microbial communities within the fire affected soils was investigated using shotgun metagenomics. We found that the average genome size of the soil microbial communities had a strong negative correlation with the temperature of the soil at the time of collection. Using fluorescence microscopy of soil microbial cell suspensions revealed that there was also a negative correlation between average cell size (length) and soil temperature at the time of collection. The changes in genome size were in part attributable to shifts in community structure, and not contemporary genome streamlining. These

microbial genomes tended to have fewer two-component regulatory systems and fewer antimicrobial resistance and production mechanisms. This work provided culture independent support for the relationship between cell size, genome size, and temperature that had largely been observed in isolate based studies.

Finally in Chapter 4, we made use of a soil warming mesocosm experiment in order to test our hypothesis from Chapter 2, that stochastic resuscitations from dormancy initiate priority effects and drive divergence in community structure across disturbed sites. We used homogenized soil from a reference site in Centralia, PA to create replicate mesocosms and subjected them to warming for a period of 12 weeks. While we found no evidence that supported our hypothesis of priority effects, we were able to assess the importance of dispersal for recovery from disturbance. A subset of disturbed mesocosms received a dispersal event and these mesocosms showed much higher resilience than their no dispersal counterparts. These results reveal the importance of dispersal for recovery from disturbance while suggesting resuscitations from a dormant seedbank may play a larger during the disturbance itself.

Together these works offer insights into the disturbance ecology of soil microbial communities in response to elevated temperature. They demonstrate the benefit of apply the community assembly synthesis of Vellend(1) and Nemergut(2), specifically for understanding community responses to and recovery from disturbance (Chapter 2), particularly the importance of dispersal (Chapter 4). Further they provide support that some generalizable relationships discovered using large isolate collections extend to environmental systems as well (Chapter 3).

Future Directions

These studies offer a jumping off point for future research on disturbance ecology and the microbiology of thermal terrestrial systems. Dispersal was shown to be particular important for resilience of microbial community structure post disturbance. However, due to the design of our experiment we were unable to assess it's importance for initial disturbance response. It is tempting to conclude that dispersal must be important for disturbance response since we observed a much lower richness of microbes in our warmed mesocosms as compared to our fire affected field sites. However, bottle effects are common in mesocosm experiments, and our decision to maintain warmed mesocosms in an aerobic environment at a constant percent moisture differs from our sampled fire affected sites, which tended to have higher moisture content and were actively venting high levels of CO_2. Assessing the importance of dispersal for disturbance response could be with the use of reciprocal transplant experiments, where soil cores from a reference site are placed into a dialysis bag(thereby limiting dispersal into the core) and moved into a fire-affected site, and vice versa. Similar experiments have been performed to look at the role of community structure vs environmental conditions on ecosystem processes(3, 4), but their extension into investigations on dispersal's role in determining disturbance response could be valuable.

One process which we were unable to investigate in these studies was diversification. Due to their large population sizes and capacity for horizontal gene transfer, diversification could play a particularly large role in the assembly of microbial communities, particularly in cases where communities remain isolated from each other due to dispersal barriers. However, these processes are difficult to measure in the environment. Some progress has been made, combining assembly of shotgun metagenomes, genome binning, and single cell genomics has allowed for

insights into genetic diversity of wild populations in lakes and sediments(5, 6). However, due to the complexity and vast diversity of microbes present in soils, using these techniques in those systems will remain difficult.

Another avenue for future work is in the characterization of the dormant seedbank. An unfortunate drawback to the methods employed to look at active and dormant communities in this dissertation is that the designation is made on the per taxa level. That is, the 16S rRNA:rRNA gene ratio method results in classifying a taxon as either active or dormant. However, microbes exhibit phenotypic diversity and this extends to cells' activity rates as well. In order to predict how a microbial community may respond to a disturbance, it will be beneficial to know the relative size and composition of the active and dormant community. Recently, advances in flow cytometry and different labeling methods have led the field to be able to make an active/dormant classification on a per cell basis, instead of on a per taxon basis. Bioorthogonal non-canonical amino acid tagging (BONCAT) is a technique used to label translationally active cells from environmental samples(7, 8). Microbes are extracted from an environmental sample and then incubated with homopropargylglycine (HPG), a methionine analog, which is incorporated into new proteins. A fluorescent dye is added that conjugates with HPG containing proteins, thereby labeling translationally active cells. The translationally active cells can then be separated from inactive or dormant cells using fluorescence assisted cell sorting (FACS), and sequenced using traditional high-throughput sequencing methods. Using this cell specific technique could allow for identifying different dormancy strategies, such as responsive vs spontaneous initiation into dormancy, based on the relative abundance of taxa in the dormant vs active fractions.

REFERENCES

REFERENCES

1. Vellend M. 2010. Conceptual synthesis in community ecology. Q Rev Biol 85:183–206.

2. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and Processes of Microbial Community Assembly. Microbiol Mol Biol Rev 77:342–356.

3. Balser TC, Firestone MK. 2005. Linking microbial community composition and soil processes in a California annual grassland and mixed-conifer forest. Biogeochemistry 73:395–415.

4. Reed HE, Martiny JBH. 2007. Testing the functional significance of microbial composition in natural communities. FEMS Microbiol Ecol 62:161–170.

5. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10:1–13.

6. Starnawski P, Bataillon T, Ettema TJG, Jochum LM, Schreiber L, Chen X, Lever MA, Polz MF, Jørgensen BB, Schramm A, Kjeldsen KU. 2017. Microbial community assembly and evolution in subseafloor sediment. Proc Natl Acad Sci U S A 114:2940–2945.

7. Hatzenpichler R, Connon SA, Goudeau D, Malmstrom RR, Woyke T, Orphan VJ. 2016. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal - bacterial consortia. Proc Natl Acad Sci U S A 113:E4069–E4078.

8. Couradeau E, Sasse J, Goudeau D, Nath N, Hazen TC, Bowen BP, Chakraborty R, Malmstrom RR, Northen TR. 2019. Probing the active fraction of soil microbiomes using BONCAT-FACS. Nat Commun 10.