



Statistical Learning and Uncommon Soil Microbiota Explain Biogeochemical Responses after Wildfire

 Alexander S. Honeyman,^a Timothy S. Fegel,^b Henry F. Peel,^a Nicole A. Masters,^a David C. Vuono,^a William Kleiber,^c Charles C. Rhoades,^b  John R. Spear^{a,d}

^aCivil and Environmental Engineering, Colorado School of Mines, Golden, Colorado, USA

^bRocky Mountain Research Station, USDA Forest Service, Fort Collins, Colorado, USA

^cApplied Mathematics, University of Colorado, Boulder, Colorado, USA

^dQuantitative Biosciences and Engineering, Colorado School of Mines, Golden, Colorado, USA

ABSTRACT Wildfires are a perennial event globally, and the biogeochemical underpinnings of soil responses at relevant spatial and temporal scales are unclear. Soil biogeochemical processes regulate plant growth and nutrient losses that affect water quality, yet the response of soil after variable intensity fire is difficult to explain and predict. To address this issue, we examined two wildfires in Colorado, United States, across the first and second postfire years and leveraged statistical learning (SL) to predict and explain biogeochemical responses. We found that SL predicts biogeochemical responses in soil after wildfire with surprising accuracy. Of the 13 biogeochemical analytes analyzed in this study, 9 are best explained with a hybrid microbiome + biogeochemical SL model. Biogeochemical-only models best explain 3 features, and 1 feature is explained equally well with the hybrid and biogeochemical-only models. In some cases, microbiome-only SL models are also effective (such as predicting NH_4^+). Whenever a microbiome component is employed, selected features always involve uncommon soil microbiota (i.e., the “rare biosphere” [existing at <1% mean relative abundance]). Here, we demonstrate that SL paired with DNA sequence and biogeochemical data predicts environmental features in postfire soils, although this approach could likely be applied to any biogeochemical system.

IMPORTANCE Soil biogeochemical processes are critical to plant growth and water quality and are substantially disturbed by wildfire. However, soil responses to fire are difficult to predict. To address this issue, we developed a large environmental data set that tracks postfire changes in soil and used statistical learning (SL) to build models that exploit complex data to make predictions about biogeochemical responses. Here, we show that SL depends upon uncommon microbiota in soil (the “rare biosphere”) to make surprisingly accurate predictions about soil biogeochemical responses to wildfire. Using SL to explain variation in a natively chaotic environmental system is mechanism independent. Likely, the approach that we describe for combining SL with microbiome and biogeochemical parameters has practical applications across a range of issues in the environmental sciences where predicting responses would be useful.

KEYWORDS biogeochemistry, microbial ecology, soil, statistical learning, wildfire

Wildfire is a natural event, but wildfire extent, frequency, and severity are changing as a consequence of climate, fuel profiles, and sources of ignition (1–3). While wildfire effects on ecosystems are typically assessed based on their macroscopic effects (e.g., burned trees and basin water quality impacts), the importance of the soil microbial world to ecosystem biogeochemical function is receiving increasing attention (4, 5). The severity of wildfire substantially affects soil ecosystems inclusive of the microbiome and

Editor Isaac Cann, University of Illinois at Urbana-Champaign

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

Address correspondence to Alexander S. Honeyman, ahoneyman@mines.edu, or John R. Spear, jspear@mines.edu.

The authors declare no conflict of interest.

Received 23 February 2022

Accepted 16 May 2022

Published 15 June 2022

associated organic and inorganic chemical environments (6, 7). Many factors, including spatial heterogeneity (in soil types, vegetation, topography, etc.), soil burn severity (SBS [rated as high, moderate, low, and no burn]), and time since fire create soil complexity (8–12), so a general response to wildfire is implausible (13, 14). However, improved predictions of postfire microbial and biogeochemical responses will advance understanding of ecosystem disturbance and recovery. These predictions should be quantitative rather than descriptive and will rely upon the truism that the microbial and biogeochemical worlds are intertwined (15–20). There are postfire legacy effects in soil (9, 21–24), and it would be valuable to predict these long-term effects with quantitative precision. However, to date, quantitative predictions about biogeochemical responses in disturbed soil remain elusive.

Generation of vast amounts of data from environmental DNA and RNA has become relatively straightforward (25–27). Nonetheless, explaining a complicated system such as soil remains challenging—especially so in relating the microbiome to biogeochemistry at field scale. One response to these challenges is to systemically probe mechanisms with experimental designs that either rule in or rule out a specific behavior as important. Such experiments, however, potentially introduce confirmation bias; when we look for specific behavior, we tend to find it (even if it has unknown relevance at the spatial scales that fires burn on). Further, it is unusual to uncover new, unexpected, relationships in data when we are not explicitly searching for or testing them in the first place. Rather, as suggested by Nurse in 2021, we may need to first consider “the flow of information through living systems,” which in turn could help us “make better sense of the flood of biological data” (28).

Soils are rich in data: 1 g contains on the order of 1 billion microbial cells (29), each of which has a genome with millions of nucleic acid base pairs. Conservatively, there is on the order of 1 petabyte (PB) of genetic information in a single gram of soil. Statistical learning (SL) and machine learning (ML) are increasingly being applied to gain inference from and to predict complicated behavior and patterns from large data sets (30). Advancements in statistical theory (31) and the development of more efficient algorithms (32) have helped. The biomedical sciences (33–35), among others, are increasingly capitalizing on SL and ML techniques for a variety of health applications. Yet, despite the enormous potential of SL and ML (36), they have not been widely adopted in designed field-based environmental research (37, 38). As such, the environmental sciences remain ripe for a wider adoption of data science and statistics (38) as our ability to generate large environmental data sets—hundreds to thousands of observations entailing thousands to millions of features (variables), i.e., “Big Data”—is outstripping our capacity to make sense of the inherently complicated environment at scales of importance.

To be used effectively, the dimensions of data sets need to be created at the outset with solid experimental design and SL/ML techniques in mind as the evaluative framework. Some ML algorithms are data-hungry—requiring far more samples (n) than can be reliably collected at field scale. SL, on the other hand, does not require a large n , but modern SL machineries are required when the number of features (p) far exceeds n (the system is overdetermined). Environmental biogeochemistry is primed for SL approaches given that p is generally far larger than n .

In this application of SL to fire-disturbed soils, we tracked soil biogeochemical recovery at two fires in Colorado, United States—the 416 Fire outside Durango, CO, and the Decker Fire outside Salida, CO—with 47 unique spatial plots (resampled over 3 years and three seasons) representing a gradient of SBS classes and both organic (Org) and mineral (Min) horizons over multiple depths (see Fig. S1 and S3 in the supplemental material). To do this, we acquired a large microbial and biogeochemical data set by sequencing 542 microbiome samples and quantifying the biogeochemistry of 178 samples from the same soils. There are 18 biogeochemical features that were measured in every soil sample (Tables S1 to S4). Thirteen of these features represent non-redundant information and were used to train models (Table S5). We then paired these

13 independent biogeochemical features with tens of thousands of microbiome features—quantified by the relative abundance of small subunit (SSU) rRNA genes as amplicon sequence variants (ASVs) via high-throughput DNA sequencing. Thus, there are on the order of 15 million data points in our study. These data represent great variance (Fig. 1) and encompass key SBS, spatial heterogeneity, time since fire, and soil horizon variables. Our aim was to address two questions with these data. (i) Is it possible to train statistical models that well explain biogeochemical responses after wildfire? (ii) What microbial and biogeochemical features are identified as skillful to those predictions?

We find that sparsity-inducing linear statistical models (i.e., models that select which features are important) are capable of predicting concentrations of biogeochemical analytes in soil. We discovered which microbiological and biogeochemical features are important by shrinking the vast majority of coefficients to exactly zero using the least absolute shrinkage and selection operator (lasso) penalty applied to multiple linear regression (equation 1) (31):

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (1)$$

Equation 1 minimizes the squared error of predictions, while the coefficients β are subject to a “penalty” for their inclusion in the model. This minimization occurs over the target value (y_i) and feature space (x_i) for all samples ($i = 1$ to n) using all features for each sample ($j = 1$ to p). We estimate the weight λ by minimizing equation 1 for a sequence of possible values of λ , and choose the one for which cross-validation predictions on held-out testing data are the most accurate. For the optimal λ , we identify the coefficients β that minimize equation 1. The result is a set of coefficients β that construct the best possible linear model for predictions.

Although the conventional wisdom is that the abundant microbiota are the main drivers of ecological response (39), our research has led us to conclude that the microbial features useful for the prediction of responses tend to be members of the “rare biosphere” and are uncommon (taxa binned at any phylogenetic rank present at <1% mean relative abundance across all samples in the data set). This suggests that abundant microbiota, though likely still major drivers in ecosystems, may be less predictive than previously understood for biogeochemical responses at field scale—at least in this complex postwildfire soil environment.

The analysis presented in this article is intended to showcase the usefulness of SL in environmental biogeochemistry while simultaneously inviting further analysis of our data set by the scientific community (see the “Data availability” section below).

RESULTS

Summary of soil biogeochemistry. Soil undergoes obvious, rapid, transformation after fire (see Fig. S3 in the supplemental material), and there is great variance in biogeochemical profiles postfire (Fig. 1). We observe that soil biogeochemistry is dependent on soil burn severity (SBS), soil horizon, and time since fire and the most change in soil occurs in the first year after burning (Tables S1 to S5). For example, in Org horizons from high-SBS plots at the 416 Fire, mean water-extractable dissolved organic carbon (DOC), dissolved total nitrogen (DTN), and ammonium (NH_4^+) decrease by 93%, 77%, and 44% from immediately postfire to 1 year later (2018 to 2019), respectively (Tables S1 and S2). Heat from fire has a limited distance that it can transfer in soil, and this distance will differ depending on the severity with which soil burned and the nature of the soil (e.g., bulk density and water content); thus, different patches of soil and different soil horizons within those patches will respond differently to wildfire. Embracing the variability of soil across principal compositional features (e.g., SBS, space, time, and soil horizon) in our study was critical to establishing a representative data set.

Summary of the soil microbiome. Several phyla of microbiota are common in both burned and unburned soils (Fig. 1)—though these groups of microorganisms are

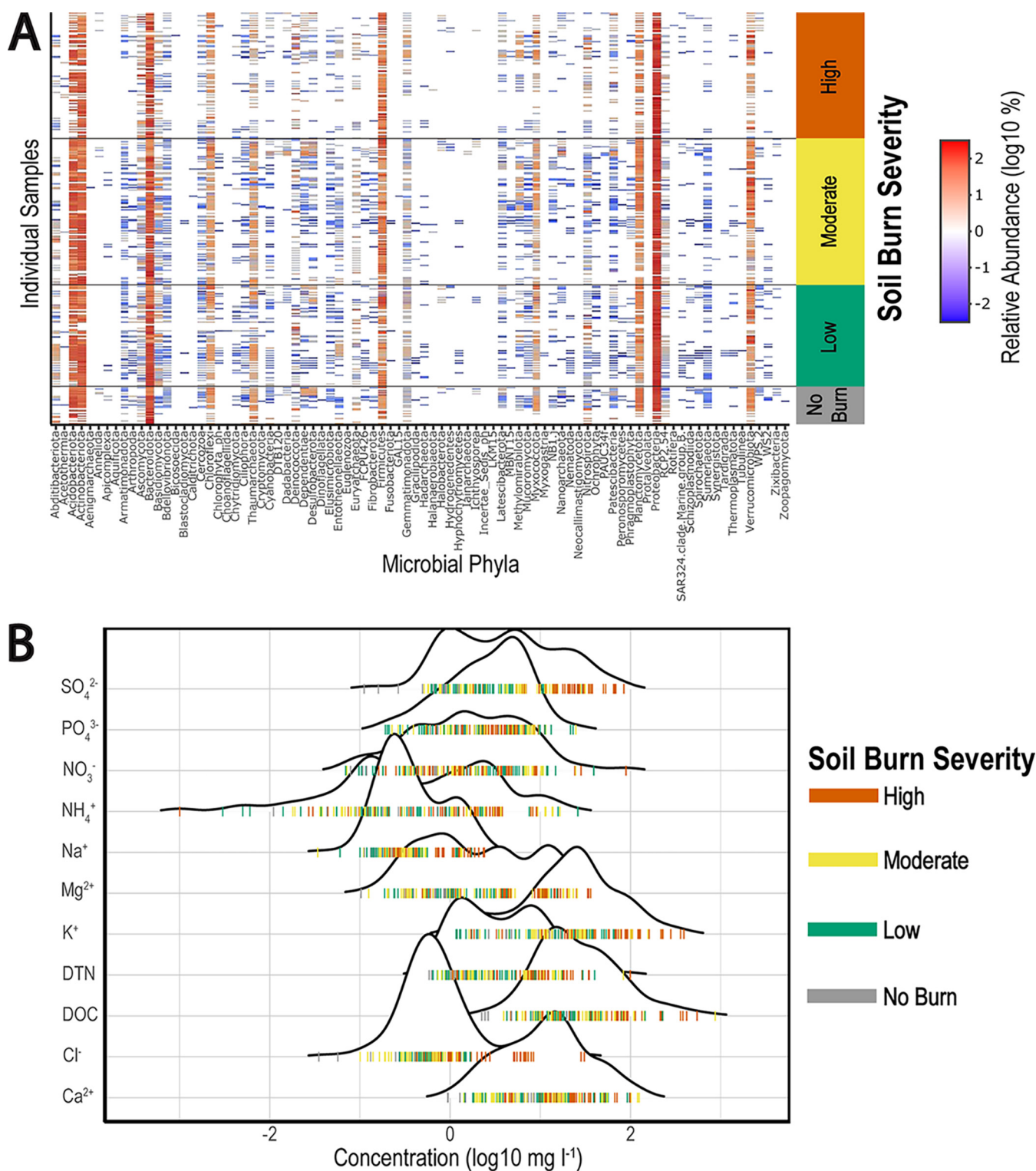


FIG 1 The total sample set ($n_{\text{microbiome}} = 542$ and $n_{\text{biogeochemistry}} = 178$) is highly variable across soil burn severity (SBS), space, time since fire, and soil horizon. SBS, specifically, is insufficient to clearly differentiate soil samples. Due to the large number of features measured from the microbiome, there are on the order of 15 million data points in our study. (A) Relative abundance of all detected phyla—the highest, and simplest, level of microbial taxonomy—arranged vertically by SBS across all soil samples. The color of individual bars denotes the percentage of relative abundance (color key at far right). Contaminant, chloroplast, and mitochondrial sequences were removed from the data set. Although Arthropoda (and other nonmicrobial Eukarya) are not part of the microbiome, these DNA sequences represent real information from soil and were retained in the data set. These nonmicrobial reads are few in number; for example, Arthropoda sequences have a mean relative abundance of only 0.08%. A phylum with no detected abundance is plotted as white. A similar graphic depicting all amplicon sequence variants (ASVs) that were measured and incorporated into the study would be hundreds of times wider. In general, there are several phyla that are common in soil, and heuristically, these are identified by dark red bands that traverse vertically up the entire

(Continued on next page)

TABLE 1 Predicting leachable dissolved organic carbon and dissolved total nitrogen from burned soil^a

Prediction	Model type	RMSE (log ₁₀ mg L ⁻¹)		No. of selected features		Fraction from rare biosphere
		Training	Testing	Biogeochemistry	Microbiome	
DOC	Biogeochemistry only	0.27	0.29	5	0	NA ^b
	Hybrid microbiome + biogeochemistry	0.21	0.23	4	24.5	0.77
	Microbiome only	0.33	0.40	0	35.5	0.80
DTN	Biogeochemistry only	0.26	0.30	7	0	NA
	Hybrid microbiome + biogeochemistry	0.19	0.22	6	38	0.81
	Microbiome only	0.32	0.40	0	53	0.82

^aDissolved organic carbon (DOC) and dissolved total nitrogen (DTN) concentrations were log₁₀ transformed prior to modeling. Values are median model scores across 10 different training and testing data set splits of 75% and 25% of total data, respectively. Reported numbers of features are the numbers of features selected by SL as important. A feature count of one-half is a result of calculating medians. Ten-fold cross-validation was also employed on each of the training sets. Different types of microbiome and biogeochemical data sets are compared. For both DOC and DTN, the hybrid biogeochemical + genus microbiome data set performs best. Models are evaluated with the RMSE of residuals (log₁₀ space in milligrams per liter). For both model types using microbiome data, the vast majority of important genera were rare (<1% mean relative abundance across all samples in the data set).

^bNA, not applicable.

not unexpected for any soil habitat: *Acidobacteriota*, *Actinobacteriota*, *Bacteroidota*, *Thaumarchaeota*, *Firmicutes*, *Proteobacteria*, and *Verrucomicrobiota*. There are also dozens of rare phyla present in subsets of our samples and, similarly, tens of thousands of rare amplicon sequence variants (ASVs). Despite the difficulty with which DNA can be extracted from burned soils with a dramatic reduction in microbial abundance—that likely also have high concentrations of PCR inhibitors—we were able to successfully amplify DNA from each SBS and reliably remove contaminating DNA sequence reads (Fig. S5). SBS, space, time since fire, and soil horizon all impact microbial diversity as well as the amount of abundant versus rare microbiota—as quantified by observed unique ASVs and Hill numbers (40, 41) (Fig. S6). Soil is not sterilized after wildfire, or at least repopulates quickly. In the first postfire sampling (weeks after burning), high SBS Org horizon soils from both the 416 and Decker Fires contain $>1 \times 10^5$, $>1 \times 10^4$, and $>1 \times 10^4$ copies of bacterial, archaeal, and eukaryal marker genes, respectively, per gram of soil, as determined by absolute quantification via digital PCR (dPCR) (Fig. S4). Within 1 year of burning, microbial populations (from all three domains of life) in these high-SBS Org horizons increase by at least 1 order of magnitude. These observed microbial recoveries are likely attributed to colonization from residual Org horizon, upward movement from mineral and rhizosphere soils, and precipitation (42)/dust inputs.

Statistical learning. Our goal is to explain the complex variation represented in Fig. 1 in a way that is useful. Trained linear models with a lasso penalty term are able to successfully predict biogeochemistry concentrations—as evaluated by the root mean square error (RMSE) of residuals (a smaller RMSE means a model is more accurate) (Table 1 and Fig. 2 and 3). Hybrid biogeochemistry + microbiome models are the most accurate for 9 of 13 measured biogeochemical analytes in fire-impacted soil: DTN, DOC, acid-neutralizing capacity (ANC), PO_4^{3-} , SO_4^{2-} , NH_4^+ , Na^+ , Cl^- , and pH (Tables S7 and S8). For NO_3^- , Ca^{2+} , and Mg^{2+} , the biogeochemical-only models are most accurate (Table S6). K^+ is predicted equally well with biogeochemical-only and hybrid models (Tables S6 to S8). Microbiome-only models never had the best RMSE for an analyte—though predictions are still accurate beyond expectation (Tables S9 and S10). In a direct comparison of biogeochemical-only models and microbiome-only models, the biogeochemical models are more accurate for 10 of the 13 analytes; however, for the remaining 3 analytes, PO_4^{3-} is explained equally well by both model types, and the microbiome model is better than the biogeochemical model at predicting NH_4^+ and pH. We ran >150 models that predict biogeochemical

FIG 1 Legend (Continued)

figure. Though highly abundant, these phyla are less helpful for making predictions about soil due to their ubiquity. Rare events (i.e., bands of any color surrounded by white) are far more unique and diagnostic in predicting biogeochemical responses. (B) Measured water-extractable biogeochemical concentrations vary over several orders of magnitude in the study. Each band below the distributions is one measured extract from soil and is color coded by SBS.

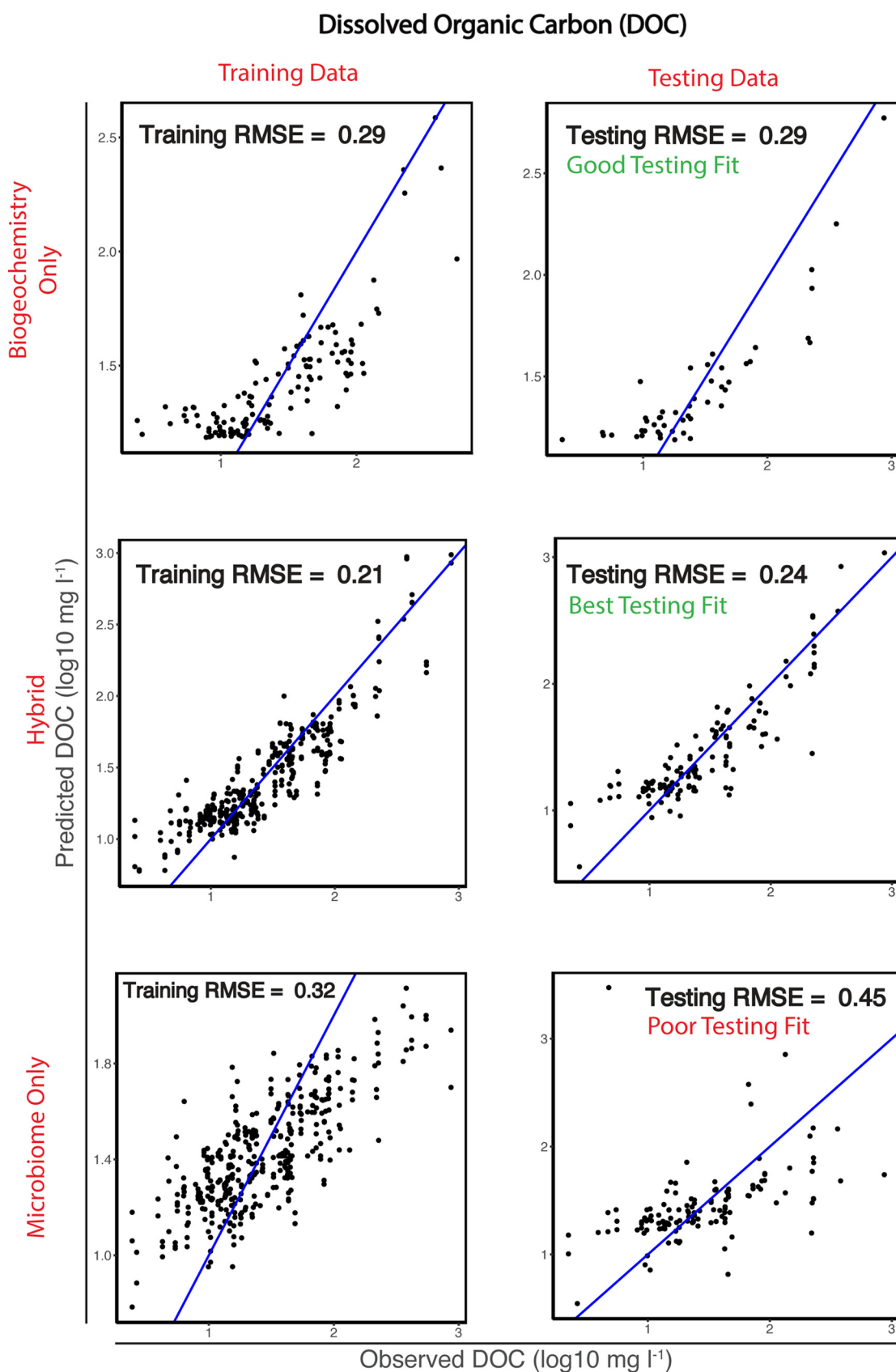


FIG 2 SL predictions of leachable dissolved organic carbon (DOC) from soil after wildfire. Target values were $\log_{10}$ transformed prior to modeling. Blue lines are a reference 1:1 line for the predicted versus observed values ($\log_{10}$ milligrams per liter), and

(Continued on next page)

responses (Tables S6 to S10); here, we provide two in-depth examples of the application of SL to microbiome + biogeochemistry data sets: (i) DOC and DTN requiring hybrid (microbiome + biogeochemistry) models and (ii) NH_4^+ , an in-depth extension of the microbiome-only model motivated by our finding that the microbiome is better at predicting NH_4^+ than biogeochemistry. These two examples provide a snapshot of the predictability of the variation in Fig. 1, but analogous in-depth explanations for any feature in our data set are possible.

(i) Predicting DOC and DTN with hybrid SL models. Hybrid microbiome-biogeochemistry data sets make more accurate predictions than biogeochemistry or the microbiome, alone. We use the RMSE of residuals in addition to simple predicted value versus observed value scatterplots to evaluate the accuracy of statistical models. Here, we report median $\log_{10}$ space testing data RMSEs of 0.23 mg L^{-1} and 0.22 mg L^{-1} for dissolved organic carbon (DOC) and dissolved total nitrogen (DTN), respectively, from a hybrid biogeochemistry and microbiome feature space (Table 1). Given the broad range of DOC and DTN observed in this study (Table S5), these accuracies are notable. We mitigated the effect of outlier predictions on model RMSEs through cross-validation and report the median performance of 10 different training/testing data splits (Table 1). We also visualized prediction versus observation scatterplots to facilitate intuitive assessment of model accuracy (Fig. 2 and 3). Cross-validation determined which biogeochemical and microbiological features are important for making predictions (Fig. 4). Examination of coefficients selected during SL shows that, for hybrid data set models, fewer biogeochemical terms are used than in biogeochemistry-only models; microbiome data are favored and assist in higher-accuracy predictions of biogeochemical responses. Further, selection of useful microbiological features for making predictions shows that the rare biosphere is more useful than the abundant biosphere (Fig. 5).

(ii) Predicting NH_4^+ with only the microbiome. To further probe whether or not the soil microbiome alone is useful for explaining biogeochemical responses (with no biogeochemical or SBS information), we conducted a case study of NH_4^+ —a biogeochemical analyte that typically increases after wildfire but is quickly depleted in the following months and years (Tables S2 and S4) (43–45). First, we tested for the effect of SBS on NH_4^+ concentrations across all soil samples (one-way analysis of variance [ANOVA], $P = 1.67 \times 10^{-4}$). Since NH_4^+ varies significantly by SBS (an important general measure of fire impact on soil), we explicitly tested whether or not the median NH_4^+ concentration for each SBS class is predictive of biogeochemical response (Fig. 6).

Despite the significant differences in NH_4^+ among SBS classes, SBS has poor accuracy as a predictor of NH_4^+ concentration. This finding is unsurprising as there are only four discrete levels of burn severity from which to make predictions (high, moderate, low, and no burn); nonetheless, it explicitly motivates the need for SL models that exploit nuance to make more accurate predictions. Next, since elevated NH_4^+ postfire (relative to unburned soil [Tables S2 and S4]) provides an energetic nitrogen source that chemolithoautotrophic ammonia-oxidizing archaea (AOA) (46–50) and ammonia-oxidizing bacteria (AOB) (51–53) may exploit—and this is supported by concomitant elevations in nitrate (NO_3^-) in the same soils—we examined the relationship between NH_4^+ concentration and AOA/AOB (54). To test this relationship, we regressed NH_4^+ on AOA/AOB relative abundances and found no correlation (Fig. 7). Further, elevated NH_4^+ and NO_3^- concentrations in burned versus unburned soils immediately postfire are rapidly depleted in the first and second postfire years (44% and 74% losses in NH_4^+ and NO_3^- , respectively, in high-SBS Org horizons at the 416 Fire in the first year), which suggests that impacts on nitrogen nutrients from geophysical/hydrological processes

FIG 2 Legend (Continued)

the visual spread of data about the blue line is an important measure of model accuracy. Median RMSEs (Table 1) are a fair assessment of the aggregate accuracy of a model type, but single-model RMSEs (as are necessary here for plotting) are sensitive to outlier predictions. Three model types are compared: biogeochemistry alone, a hybrid biogeochemistry + genus microbiome data set, and the genus microbiome alone. Independent training and testing data sets evaluate the accuracy of learned models. Note that axis scales are different for each scatterplot. The hybrid model performs better than either the biogeochemistry or microbiome model alone.

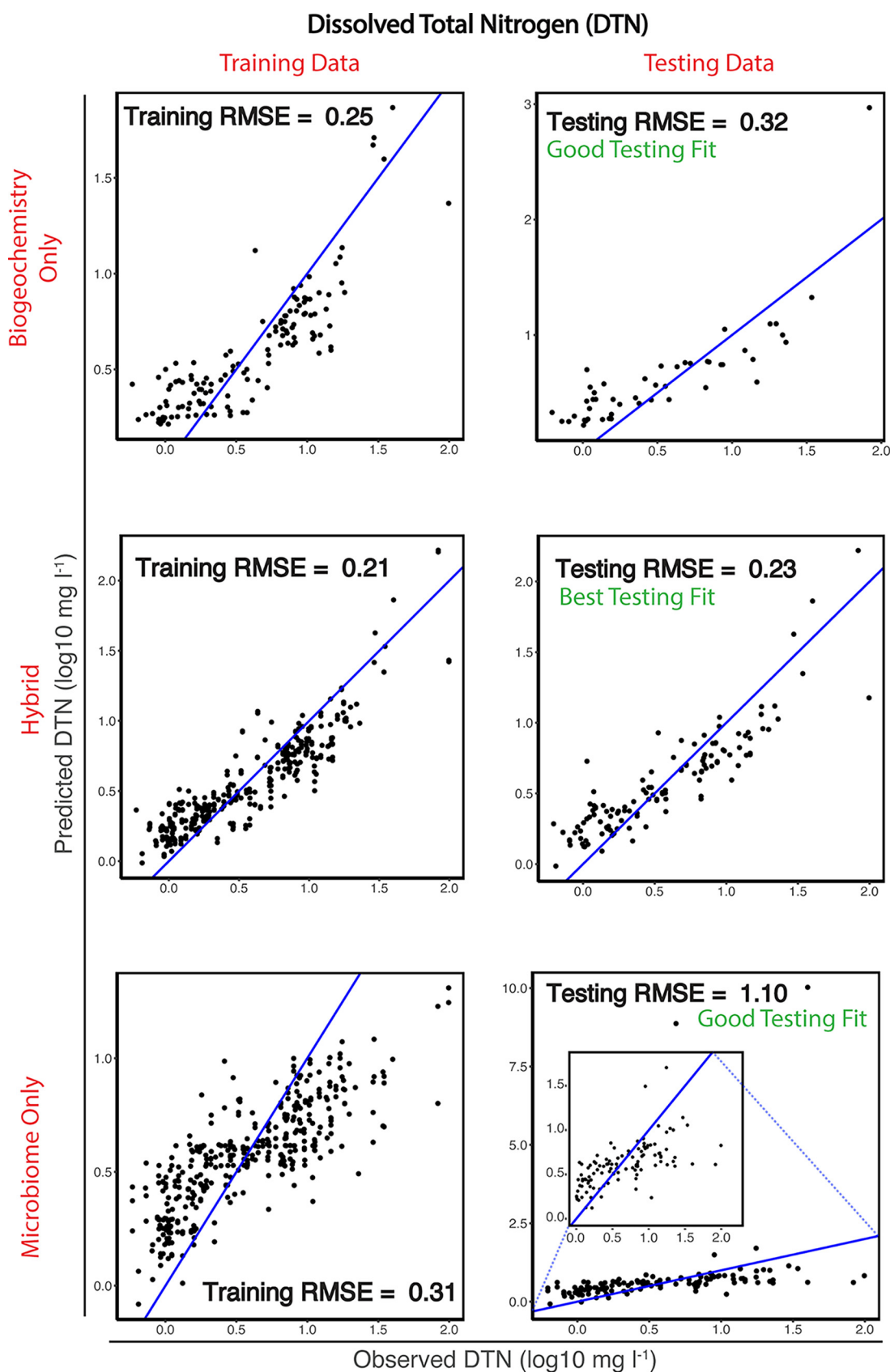


FIG 3 SL predictions of leachable dissolved total nitrogen (DTN) from soil after wildfire. Target values were $\log_{10}$ transformed prior to modeling. Blue lines are a reference 1:1 line for the predicted versus observed values ($\log_{10}$ milligrams per liter), and

(Continued on next page)

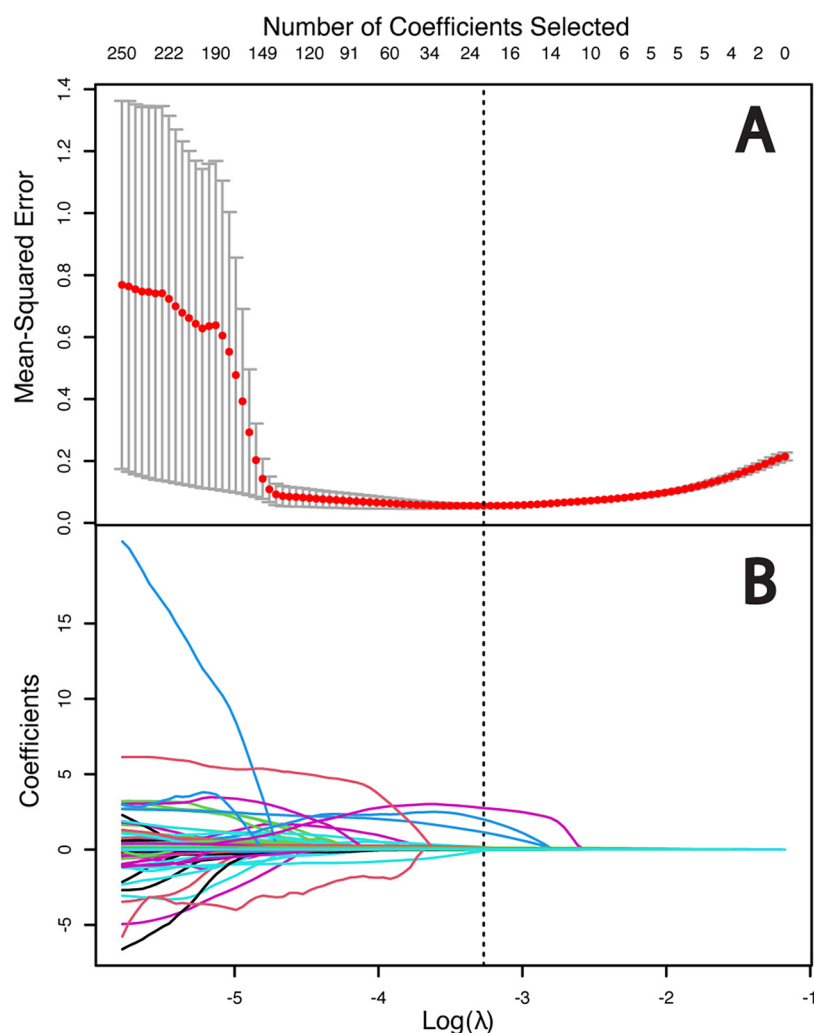


FIG 4 Statistical learning (via lasso-penalized linear regression) of the best features—and their coefficients—for predicting leachable dissolved total nitrogen (DTN) from soil after wildfire with a hybrid SL model. The objective was to learn a useful set of parameters while discarding unhelpful ones; both too many and too few features included in the final model result in an undesirable accuracy. (A) Learning the optimal penalty weight (λ on x axis) that coefficients must overcome to enter the model (dotted line) (equation 1). This is done by identifying the λ that minimizes the mean model error (red points) with 10-fold cross-validation (gray error bars are $\pm$ standard error). (B) There is an optimal number of selected coefficients (each colored line represents a feature) in the hybrid model that produces the best model fit (dotted line), and this combination is an ensemble of biogeochemical and genus microbiome terms. The coefficients of features shrink to exactly 0 (unimportant) with increasing λ . The optimal value of λ (dotted line) is large enough to remove noise from the data set—overcoming the statistical nuisance of variance inflation—but small enough to retain enough complexity necessary to make accurate predictions.

are substantial. Neither SBS nor AOA/AOB alone explains NH_4^+ response immediately postfire and in the postfire years; far more information is likely necessary.

Finally, we tested whether or not the microbiome alone and SL accurately predict NH_4^+ response (Fig. 7). The microbiome was examined by extending our lasso-penalized linear

FIG 3 Legend (Continued)

the visual spread of data about the blue line is an important measure of model accuracy. Median RMSEs (Table 1) are a fair assessment of the aggregate accuracy of a model type, but single-model RMSEs (as are necessary here for plotting) are sensitive to outlier predictions. Three model types are compared: biogeochemistry alone, a hybrid biogeochemistry + genus microbiome data set, and the genus microbiome alone. Independent training and testing data sets evaluate the accuracy of learned models. Note that axis scales are different for each scatterplot. There are two outlier points that lengthen the y axis in the microbiome-only testing data scatterplot; the inset scatterplot is a magnification of data around the 1:1 reference line. The hybrid model performs better than either the biogeochemistry or microbiome model alone.

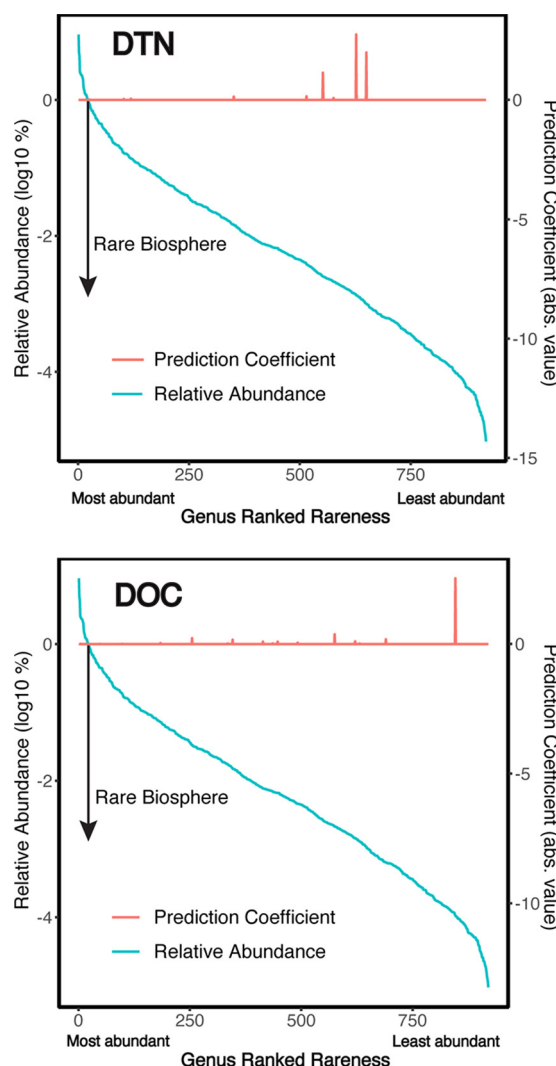


FIG 5 Contributions of rare genera of soil microorganisms to effective hybrid model predictions of water-extractable dissolved total nitrogen (DTN) and dissolved organic carbon (DOC). With decreasing abundance of genera of microorganisms (blue curve), their predictive utility, as measured by their coefficients in the learned linear model (red peaks), becomes greater. A prediction coefficient value of zero means that the microorganism was not included in the learned model at all. The black arrow begins at the threshold for rare microbiota on the left y axis (relative abundance) and indicates the direction of increasing rareness: any organism with a $\log_{10}$ relative abundance of $<0\%$ is rare ($<1\%$ mean relative abundance across all samples in the data set). The vast majority of nonzero coefficients (important features) come from the rare biosphere, though some abundant microbiota still have extremely small prediction coefficients on this figure.

statistical model by first conducting principal-component analysis (PCA) on the data. PCA is a dimension reduction technique that learns a relatively small set of new features (principal components [PCs]) that are linear combinations of all features in the data set. If only a few PCs are necessary to explain data, then the system is not complicated. We regressed NH_4^+ on PCs with lasso-penalized linear regression (trained on 75% of samples and tested on the remaining 25%) to determine how many PCs are important. Surprisingly, dozens of PCs were selected, and predictions on testing data were accurate, indicating that considering only the dominant microbiota in explaining responses is insufficient. Paradoxically, although the microbiome can predict enriched NH_4^+ concentrations in soil postfire, this process is not thought to be biological; NH_4^+ derives from combusted vegetation and Org horizons (43–45). However, in this case study of SL with principal-component regression, we found that the microbiome provides the sufficient complexity required to predict NH_4^+ . This suggests that even abiotic analyte pools such as NH_4^+ after wildfire can be

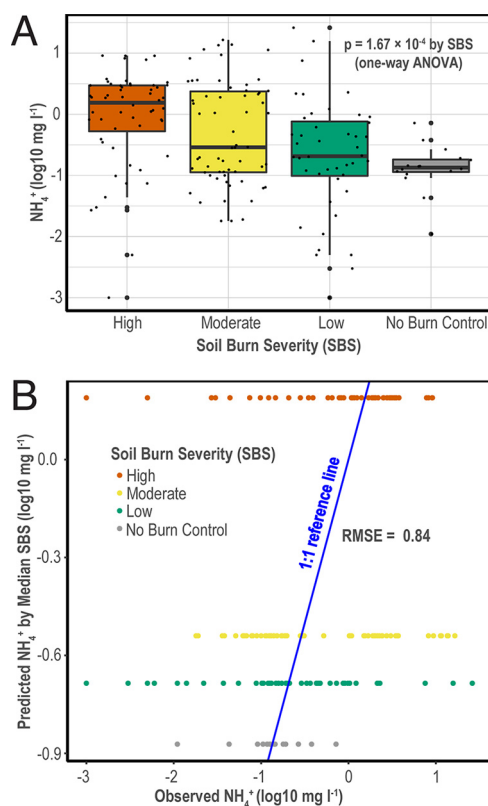


FIG 6 Although NH_4^+ varies significantly by SBS, using SBS to make explicit predictions about NH_4^+ response results in poor accuracy. (A) Box and whisker plot of NH_4^+ concentration for each SBS. These variations are significant (one-way ANOVA, $P = 1.67 \times 10^{-4}$). (B) Using the median NH_4^+ by SBS to predict NH_4^+ response. In a successful prediction model, data points would follow the 1:1 prediction line. Although on average, there is significant variation in NH_4^+ that is explained by SBS (A), using SBS alone to predict NH_4^+ response results in poor accuracy (as shown by horizontal scatter in panel B). SBS is important in a descriptive sense to understanding generalized biogeochemical responses after wildfire, but a multidimensional statistical model is required to accurately predict explicit NH_4^+ concentrations.

explained by microbial tenants (55) of an ecosystem—independent of whether or not those microbiota actually produce or consume the analyte. These tenants may be responding to other environmental changes postfire (e.g., increases in pH and acid-neutralizing capacity [ANC]), but their aggregate responses to these changes still aid in the prediction of NH_4^+ through a complex web of covariances and interdependencies. It is possible to use the microbiome to predict NH_4^+ concentrations both immediately postfire and in the postfire years, and this SL approach provides far more accurate predictions than the descriptive use of either SBS or microbiota that have been shown to actively oxidize NH_4^+ (AOA/AOB) in soils (54).

DISCUSSION

Soils are critical to understand due to their essential roles in maintaining ecosystem productivity (56, 57). Fire is a natural part of many ecosystems, but it is unclear how soils and their microbiomes will respond to shorter fire return intervals and combinations of disturbances and climatic stress in burned landscapes (1, 3, 58). Here, we conducted a study that examines a large multivariate data set from wildfire-impacted soil and show that biogeochemical responses can be explained with linear SL models. To our knowledge, there are no data sets of similar construction on which we could further test and validate our models; however, we have extensively evaluated predictions on held-out testing data from two different fires/forests.

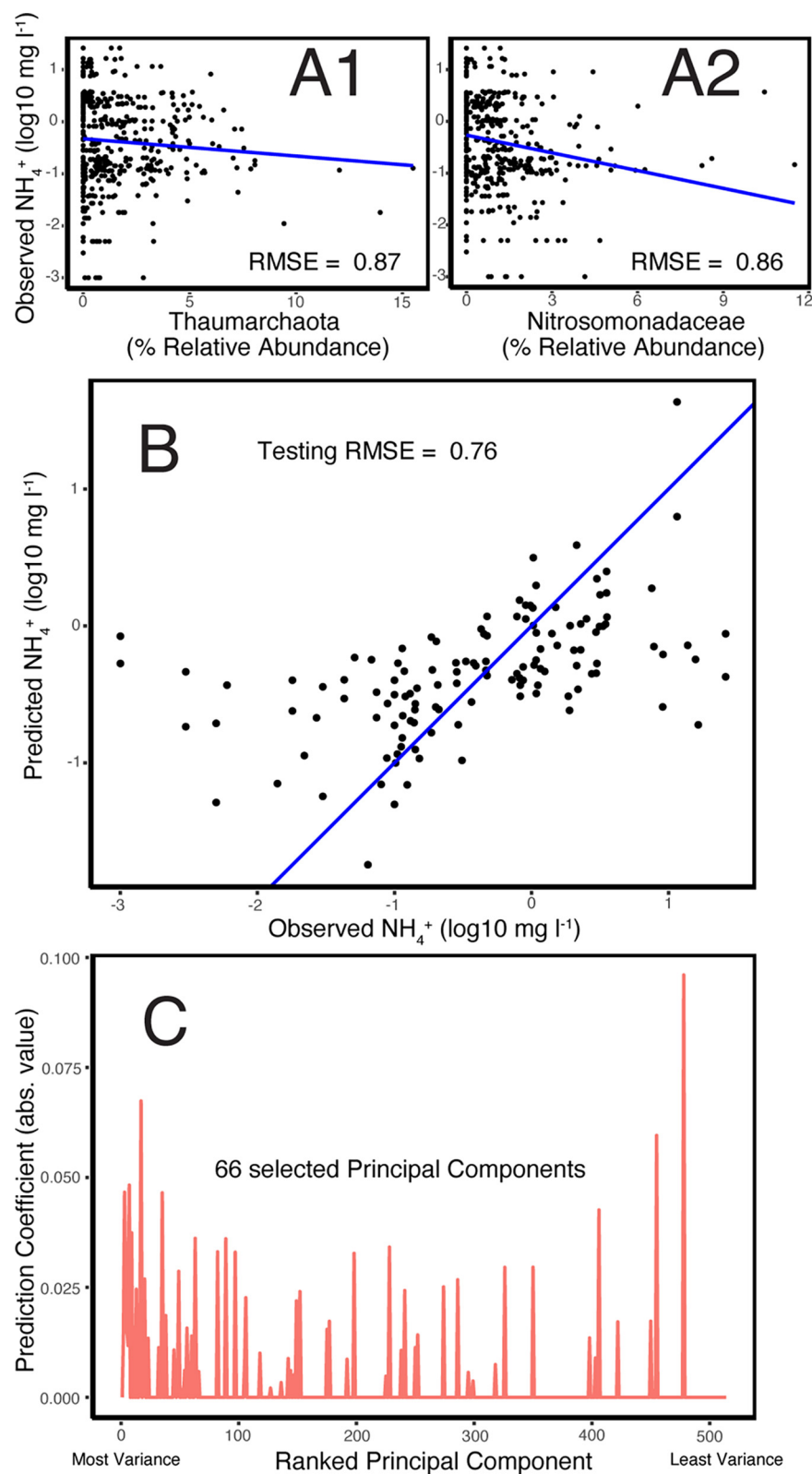


FIG 7 Predicting and explaining soil ammonium (NH_4^+) in the postfire years using only the microbiome. (A1) There is no relationship between NH_4^+ and the abundance of canonical ammonia-oxidizing archaea (AOA) from phylum *Thaumarchaeota* in soil (a poor linear fit [blue line]); 100% of the *Thaumarchaeota* sequenced in our study were of class *Nitrososphaera*. (A2) The linear fit (blue line) between ammonia-oxidizing bacteria (AOB) from family *Nitrosomonadaceae* and NH_4^+ is also poor. (B) However, NH_4^+ response is well explained by PCA regression. The blue line is a 1:1

(Continued on next page)

We learned what aspects of the soil biogeochemical world are important to explaining the variance in soil at field scale. We find that hybrid microbiome + biogeochemical models are the most accurate (Table 1). Further, most microbial features selected by SL as useful for making accurate predictions of biogeochemical responses are uncommon ($<1\%$ mean relative abundance at any phylogenetic rank) (Fig. 5; see Tables S7 to S10 in the supplemental material). We can use SL to exploit complexities to make predictions about specific soil responses in this burned environment. There is likely a far greater density of information in soil biogeochemistry and the soil microbiome that can be leveraged for applied inference and prediction than previously recognized.

For particularly important ecosystem features such as NH_4^+ (4, 45), it is worth comparing the predictive power of physical descriptions (such as SBS) as well as NH_4^+ -associated microbiota (AOA/AOB) to the accuracy of SL models. Although NH_4^+ varies significantly as a function of SBS, we find that SBS is not predictive of NH_4^+ response (Fig. 6). Members of the AOA phylum *Thaumarchaeota* and AOB family *Nitrosomonadaceae* are common in our data set (Fig. 1 and 7), and their ability to use ammonia for chemolithotrophy is important to global nitrogen cycling in both marine and terrestrial systems (46–54, 59–61). Nonetheless, AOA/AOB relative abundances do not predict NH_4^+ response. The presence of AOA and AOB in the soils we analyzed is a case in point of the potential of SL for understanding the biogeochemistry of a system (Fig. 7).

Environmental scientists acknowledge that environmental systems are enormously complex, yet we often aim to distill that complexity to a handful of what we think are principal, controlling features. This dissonance is especially acute as the scale of the system increases. One might expect a correlation, even if weak, between the relative abundances of AOA/AOB and NH_4^+ since these microbiota can exploit NH_4^+ for energy and/or nitrogen (62), but we did not observe this relationship (Fig. 7). Though genomes/transcriptomes of AOA/AOB in our study would be necessary to fully explain their activity, a possible alternative view supported by our SL analysis is that cost-efficient, high-throughput microbial community profiling with the SSU rRNA gene may be sufficient if the goal is strictly to predict responses in soil (a large n from high-throughput sequencing may compensate for the loss in large p offered by multiomic methods). Gene-specific investigations of postfire soils found that neither *nifH* nor *amoA* (the genes that encode the enzymes for nitrogen fixation and ammonia oxidation, respectively) were consistently expressed or even present (5, 63, 64). A functional model of postfire soil may not predict DTN or NH_4^+ responses if the enzymes responsible for their generation/transformation are not consistently quantifiable yet DTN and NH_4^+ postfire vary greatly on a continuous scale (Tables S1 to S5).

As one possible solution, we show it is feasible to leverage the composition of the total microbiome cost-efficiently with hundreds of SSU rRNA gene amplicons to explain DTN and NH_4^+ (and other) responses in the first and second postfire years. A critical distinction is that this does not mean that the microbial community is necessarily responsible for generating or consuming NH_4^+ postfire: in fact, elevated NH_4^+ is a typical product of organic matter oxidizing in wildfire (43, 44), and postfire erosion and plant uptake are additional sinks for NH_4^+ (65, 66). We show that independent of how

FIG 7 Legend (Continued)

reference for predicted (using the PC regression model) versus observed NH_4^+ . The model was trained on 75% of samples, and the performance of the model on test samples is plotted here (the remaining 25% of data). The spread of data about blue lines is critical to assessing model accuracy; RMSE reports the error of the model in $\log_{10}$ space, and so small differences in RMSE are substantial differences in the accuracies of models. (C) Surprisingly, many PCs were selected during PCA regression, suggesting that microbiome complexities (and not just abundant microbiota) are critical to accurately predicting NH_4^+ response in the postfire years. NH_4^+ can be explained by an integrative approach to unpacking the microbiome; the presence of the usual suspects of NH_4^+ oxidation (AOA/AOB) does not explain the response of NH_4^+ with the same skill as SL. PC regression overestimates NH_4^+ at low concentrations and tends to underestimate it at high concentrations. Low concentrations may be hard to predict in general due to instrument detection limits, but high concentrations might be better predicted with the development of more nuanced statistical models that allow for nonlinear biogeochemical relationships.

NH_4^+ is mechanistically regulated in postfire soil, it is possible to use the impressive data density of DNA to predict responses in the first two postfire years without microbial genomes or transcriptomes (though multiomic methods are certainly valuable and necessary to a detailed molecular description of environmental processes). The regulation of NH_4^+ and other analytes postfire is likely dependent, in ways still unknown, on the complex milieu of geophysical processes and macroscopic organisms as well as competition, symbioses, and horizontal gene transfer in the microbial world—even if redox reactions themselves are carried out by a small set of core genes (67).

The power of SL, coupled with an appropriately large data set, is the ability to explore complex systems in a heuristic fashion to possibly discern patterns and relationships that were previously unknown or unknowable in the absence of the SL techniques. SL is unaware of the meaningfulness of features to domain scientists, and therein lies an opportunity to generate new hypotheses and discover unexpected phenomena (such as the predictive power of the microbial rare biosphere) (Fig. 5). Features possibly unrelated to microbial activity immediately after fire (e.g., NH_4^+) can still be explained with surprising accuracy by quantifying the response of the microbial community—even if these responses are primarily governed by other harsh fire/post-fire conditions (e.g., excessive heat and increases in soil pH). We hypothesize that the composition of microbial communities in soil records the life history and potential of an environment, and this record is valuable for making predictions about innumerable biogeochemical responses in temporal forward trajectories.

The rare biosphere is critical to making successful predictions (Table 1 and Fig. 5), while SBS and AOA/AOB relative abundances, by themselves, are unhelpful (Fig. 6 and 7). The microbial rare biosphere (such as that we characterized shortly after wildfire) is poorly understood, yet likely serves as a nearly inexhaustible source of genomic innovation, as described by Sogin and colleagues in 2006 (68). We suggest that the rare biosphere and the inherent complexities of diverse microbial communities in general possess the information necessary for accurate biogeochemical predictions and explanations as learned by statistical models (Table 1). Microbial communities and their diversity/traits are important to the resilience and trajectory of ecosystems (69, 70), but quantifying these attributes to explain variation in soil is an ongoing challenge (71). A new approach to describing complex systems may be necessary in environmental science—one that exploits, rather than constrains, the variability of natural systems. Such an approach may yield true predictive capability in the environment while simultaneously elucidating impactful, unexpected, mechanisms to test experimentally. The feasibility of explicit, accurate, predictions for soil responses suggests that the potential, trajectory, and success of postfire restoration interventions could be quantitatively evaluated in an SL framework—possibly enabling the optimization of restoration strategies for impact and cost-effectiveness in environmental policy. Additional research should investigate the feasibility of using microbiome and biogeochemical data to forecast future responses in environmental systems—including more precise quantification of the reliability and biases of statistical models.

Leveraging the complexities of environmental systems as an asset (72) with SL, ML, or artificial intelligence (AI), in general, likely has numerous practical applications. As with post-wildfire soils, agricultural soils (73), biological wastewater treatment systems (74), and oceans (75) all have critical biogeochemical cycles that require greater understanding. Global microbial and biogeochemical systems (67, 76) offer a near inexhaustible supply of data density—possibly enabling the principles of biomedical “personalized medicine” to be extended to environmental science for the precision management of soil, water, and air.

MATERIALS AND METHODS

Wildfire sampling. Two wildfires in Colorado, United States, were studied with recurrent sampling of 47 unique plots over the first and second postfire years (collections from one fall, one spring, and three summers); in total, these data are 178 bulk soil collections and 542 microbiome samples. The 416 Fire, ~13 miles north of Durango, CO, in the San Juan National Forest (ignition on 1 June 2018; 54,130-acre final footprint) and the Decker Fire, south of Salida, CO, in the Rio Grande and San Isabel National Forests (detected on 8 September 2019; 8,910-acre final footprint) both burned in mosaic patterns with various SBS (77, 78). Both fires burned in montane western forests at an ~8,000-ft elevation with mixed

tree communities comprised of conifers (e.g., Ponderosa pine and Douglas fir) and hardwoods (e.g., quaking aspen and Gambel oak). The 416 Fire was sampled three times: 27 June 2018, 13 June 2019, and 18 June 2020. The Decker Fire was also sampled three times: 7 November 2019, 22 May 2020, and 29 July 2020.

Regions of the 416 and Decker fires that represented a transect of SBS classes were targeted for sampling. Forty-seven unique spatial plots were established (17 at the 416 Fire and 30 at the Decker Fire). Both Org horizon and Min horizon bulk soil samples were collected at the 416 Fire, and only Org horizon samples were collected at the Decker Fire. Sampling of only Org horizon samples at the Decker Fire enabled greater spatial representation for statistical models (more sites were sampled) at the expense of resolution in soil horizons. Biological triplicate soil samples were taken for DNA from each bulk soil collection, and the remainder bulk soil (singlicate) was saved for biogeochemistry. A detailed description of field sampling methods is provided in the supplemental material.

Biogeochemistry. Bulk soil samples were air-dried, protected from light, in the humidity-controlled laboratory at 22°C for up to 14 days until mass loss from evaporation ceased. Triplicate leachate tubes were made from each sample with a 1:10 mass ratio of soil to Milli-Q purified water (>18.2 M Ω -cm resistivity) and gently mixed for 2 h on a tabletop shaker. A bell jar vacuum filtration system (Kimble; DWK Life Sciences, NJ, USA) was used to filter leachates through two different filter types: (i) 0.7- μ m-pore glass fiber filters without binders (MilliporeSigma; AP4004705) for dissolved organic carbon and dissolved total nitrogen (DOC/DTN) analyses and (ii) 0.45- μ m-pore polyvinylidene difluoride (PVDF) membrane filters (MilliporeSigma; Durapore HVLP04700) for ion chromatography (IC) analyses. The third leachate was saved, unfiltered, at 4°C for acid-neutralizing capacity (ANC) and pH determination. DOC/DTN and IC filtrates were stored at -20°C. For every 20 samples filtered, raw Milli-Q water and a filtrate blank for each filter type were processed the same as all other samples. Laboratory analytical methods for the characterization of filtrates are reported in the supplemental information.

DNA extraction and sequencing. (i) Extraction. DNA from samples stored in Zymo ZR BashingBead lysis tubes was extracted according to the manufacturer's instructions for the ZymoBIOMICS DNA miniprep kit, including the optional PCR inhibitor removal step (polyphenolics, humic/fulvic acids, and melanin), and then stored at -80°C.

(ii) Sequencing. Singlicate 50- μ L PCRs amplified the V4-5 hypervariable regions of the SSU rRNA gene with universal primers (*Bacteria*, *Archaea*, and *Eukarya*) (79). Primers and reaction concentrations are reported in the supplemental material. Sequencing was performed with MiSeq V2 paired-end 2 $\times$ 250 chemistry (Illumina, Inc., San Diego, CA, USA) at the Duke Center for Genomic and Computational Biology or the University of Colorado Anschutz Medical Campus Genomics and Microarray Core.

Digital PCR. Ninety DNA samples were examined for absolute bacterial, archaeal, and microbial eukaryotic abundance with a QIAcuity One 5-plex digital PCR (dPCR) instrument (Qiagen, Inc., USA, Germantown, MD, USA). dPCR samples represented both the 416 Fire across SBS, space, time, and soil horizon and the Decker Fire across SBS, space, and time. A QIAcuity 8.5k 96-well nanoplate and the QIAcuity EG PCR kit were used for amplification and imaging according to the manufacturer's instructions. The primer, reaction, thermocycling, and imaging/quantification details are reported in the supplemental material.

Bioinformatics and statistical learning. (i) Bioinformatics. ASVs were called and taxonomy was assigned by Dada2 (26, 80, 81). ASV, taxonomy, and associated sample metadata were organized as phyloseq objects (82). Contaminating DNA reads were identified and removed by Decontam (83). Relative abundances of taxa were computed by *ampvis2* in R (84). Details for each bioinformatic step are presented in the supplemental material.

(ii) Statistical learning. All SL and statistics were conducted in R (85). Training data sets (75% of data) and testing data sets (the remaining 25%) were randomized from the total data set ($n_{\text{microbiome}} = 542$ and $n_{\text{biogeochemistry}} = 178$), and the process of splitting data into various feature and target matrices was automated with in-house code. Within our code base, any combination of biogeochemical or microbiome data—at any user-defined phylogenetic rank—can be coerced into training, testing, feature, and target matrices immediately compatible with either of the two following statistical models.

(a) lasso-penalized linear regression. The penalty weight λ was chosen with 10-fold cross-validation on training data with the *glmnet* R package (32). Root mean square error (RMSE) was reported for both training and testing data.

(b) lasso-penalized principal-component regression. Principal components were computed on the full sample set in R with *prcomp* (options to center and scale data were set to TRUE). lasso-penalized multiple linear regression was trained on principal components with samples from the training data set, and cross-validation was conducted exactly as described above. The testing data set was used for plotting and model evaluation.

(c) Randomized training and testing data splits. For DOC and DTN predictions, model sensitivity to training/testing data randomization was assessed by training on 10 different random splits of data; each of the random splits still received 10-fold cross-validation on training data. Median results from the 10 different data splits are reported for DOC and DTN in Table 1; plotted points (Fig. 2 and 3) are from the train/test split that produced a hybrid model testing the RMSE closest to the median score. The ammonium (NH₄⁺) case study (Fig. 7) and all other model scores reported in the supplemental material use this train/test data split. Testing the variation in NH₄⁺ concentrations across SBS was conducted via one-way ANOVA in R with the *aov()* function (85).

Data availability. Our data are FAIR (findable, accessible, interpretable, and reusable) (86). All raw DNA reads were deposited at the NCBI SRA under BioProject no. [PRJNA767436](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA767436). All biogeochemical data, postprocessed DNA sequence data, metadata, and code are available on Zenodo (<https://doi.org/10.5281/zenodo.5539890>). Our FAIR data and code are sufficient to reproduce all of our analyses. Our code

is broken down into one preprocessing file, one analysis file, and one file used to make Earth maps—all as RMarkdown files (.Rmd) with comments and explanations for every step. SL modeling of any analyte in our study can be done within our analysis code, and figures analogous to those we present in this article can be automatically produced. Interested parties can explore their own questions relevant to our data set. Our full analysis pipeline, including figure generation, is reproducible in one hands-off run of code in ~1 h on a 2018 MacBook Pro.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 10.6 MB.

ACKNOWLEDGMENTS

We are grateful to Laura Leonard and Michael Remke (Mountain Studies Institute, Durango, Colorado, USA) for assistance with field work, to Blake Stamps for molecular biology technical advice, to James Ranville and Shaun Bevers for helpful discussions about geochemistry, and to Kalen Rasmussen and Michael Vega for assistance with DNA sequencing preparation. Henry and Lin Ballard, Eric Bader, and Chief Michael Schmitt (Sunshine Fire, Boulder County, Colorado, USA), Maya MacHamer (Fourmile Watershed Coalition, Boulder County, Colorado, USA), and Jared Whitmer (Columbine Ranger District, San Juan National Forest, USA) generously lent their time to help us identify and gain access to active wildfire. The Columbine Ranger District of the San Juan National Forest and the Salida Ranger District of the San Isabel National Forest graciously permitted our access to burn areas.

This research was supported by the National Science Foundation Graduate Research Fellowship Program (to A.S.H., Fellow ID 2019258966) and NASA Exobiology research grant 80NSSC19K0479 (to J.R.S.).

We declare no conflict of interest.

Conceptualization, A.S.H.; methodology, A.S.H. and J.R.S.; field investigation, A.S.H., H.F.P., and J.R.S.; biogeochemistry investigation, T.S.F., A.S.H., and C.C.R.; molecular biology investigation, A.S.H., H.F.P., N.A.M., and J.R.S.; software, A.S.H.; formal analysis, A.S.H., D.C.V., and W.K.; resources, J.R.S. and C.C.R.; writing — original draft, A.S.H.; writing — review & editing, all authors; supervision, J.R.S.; funding acquisition, A.S.H. and J.R.S.

REFERENCES

1. Abatzoglou JT, Williams AP. 2016. Impact of anthropogenic climate change on wildfire across western US forests. *Proc Natl Acad Sci U S A* 113:11770–11775. <https://doi.org/10.1073/pnas.1607171113>.
2. McLauchlan KK, Higuera PE, Miesel J, Rogers BM, Schweitzer J, Shuman JK, Tepley AJ, Varner JM, Veblen TT, Adalsteinsson SA, Balch JK, Baker P, Battlori E, Bigio E, Brando P, Cattau M, Chipman ML, Coen J, Crandall R, Daniels L, Enright N, Gross WS, Harvey BJ, Hatten JA, Hermann S, Hewitt RE, Kobziar LN, Landesmann JB, Loranthy MM, Maezumí SY, Mearns L, Moritz M, Myers JA, Pausas JG, Pellegrini AFA, Platt WJ, Roozeboom J, Safford H, Santos F, Scheller RM, Sherrieff RL, Smith KG, Smith MD, Watts AC. 2020. Fire as a fundamental ecological process: research advances and frontiers. *J Ecol* 108:2047–2069. <https://doi.org/10.1111/1365-2745.13403>.
3. Higuera PE, Shuman BN, Wolf KD. 2021. Rocky Mountain subalpine forests now burning more than any time in recent millennia. *Proc Natl Acad Sci U S A* 118:e2103135118. <https://doi.org/10.1073/pnas.2103135118>.
4. Rhoades CC, McCutchan JH, Cooper LA, Clow D, Detmer TM, Briggs JS, Stednick JD, Veblen TT, Ertz RM, Likens GE, Lewis WM. 2013. Biogeochemistry of beetle-killed forests: explaining a weak nitrate response. *Proc Natl Acad Sci U S A* 110:1756–1760. <https://doi.org/10.1073/pnas.1221029110>.
5. Nelson AR, Narrowe AB, Rhoades CC, Fegle TS, Daly RA, Roth HK, Chu RK, Amundson KK, Geonczy SE, Emerson JB, Young RB, Steindorff AS, Mondo SJ, Grigoriev IV, Salamov A, Borch T, Wilkins MJ. 2021. Playing with FIRE: a genome resolved view of the soil microbiome responses to high severity forest wildfire. *bioRxiv* 2021.08.17.456416. <https://doi.org/10.1101/2021.08.17.456416>.
6. Whitman T, Whitman E, Woollet J, Flannigan MD, Thompson DK, Parisien M-A. 2019. Soil bacterial and fungal response to wildfires in the Canadian boreal forest across a burn severity gradient. *Soil Biol Biochem* 138: 107571. <https://doi.org/10.1016/j.soilbio.2019.107571>.
7. de Miera LES, Pinto R, Gutierrez-Gonzalez JJ, Calvo L, Ansola G. 2020. Wildfire effects on diversity and composition in soil bacterial communities. *Sci Total Environ* 726:138636. <https://doi.org/10.1016/j.scitotenv.2020.138636>.
8. Qin Q, Liu Y. 2021. Changes in microbial communities at different soil depths through the first rainy season following severe wildfire in North China artificial *Pinus tabulaeformis* forest. *J Environ Manage* 280:111865. <https://doi.org/10.1016/j.jenvman.2020.111865>.
9. Xiang X, Shi Y, Yang J, Kong J, Lin X, Zhang H, Zeng J, Chu H. 2015. Rapid recovery of soil bacterial communities after wildfire in a Chinese boreal forest. *Sci Rep* 4:3829. <https://doi.org/10.1038/srep03829>.
10. Yang T, Tedersoo L, Lin X, Fitzpatrick MC, Jia Y, Liu X, Ni Y, Shi Y, Lu P, Zhu J, Chu H. 2020. Distinct fungal successional trajectories following wildfire between soil horizons in a cold-temperate forest. *New Phytol* 227: 572–587. <https://doi.org/10.1111/nph.16531>.
11. Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 15:579–590. <https://doi.org/10.1038/nrmicro.2017.87>.
12. Lombao A, Barreiro A, Fontúrbel MT, Martín A, Carballas T, Díaz-Raviña M. 2020. Key factors controlling microbial community responses after a fire: importance of severity and recurrence. *Sci Total Environ* 741:140363. <https://doi.org/10.1016/j.scitotenv.2020.140363>.
13. Brown SP, Veatch AM, Horton JL, Ford E, Jumpponen A, Baird R. 2019. Context dependent fungal and bacterial soil community shifts in response to recent wildfires in the Southern Appalachian Mountains. *For Ecol Manage* 451:117520. <https://doi.org/10.1016/j.foreco.2019.117520>.

14. Woollet J, Whitman T. 2020. Pyrogenic organic matter effects on soil bacterial community composition. *Soil Biol Biochem* 141:107678. <https://doi.org/10.1016/j.soilbio.2019.107678>.
15. Spear JR, Walker JJ, McCollom TM, Pace NR. 2005. Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci U S A* 102:2555–2560. <https://doi.org/10.1073/pnas.0409574102>.
16. Nealson KH. 2005. Hydrogen and energy flow as “sensed” by molecular genetics. *Proc Natl Acad Sci U S A* 102:3889–3890. <https://doi.org/10.1073/pnas.0500211102>.
17. Pérez-Izquierdo L, Clemmensen KE, Strengbom J, Granath G, Wardle DA, Nilsson M, Lindahl BD. 2021. Crown-fire severity is more important than ground-fire severity in determining soil fungal community development in the boreal forest. *J Ecol* 109:504–518. <https://doi.org/10.1111/1365-2745.13529>.
18. Holden SR, Rogers BM, Treseder KK, Randerson JT. 2016. Fire severity influences the response of soil microbes to a boreal forest fire. *Environ Res Lett* 11:e035004. <https://doi.org/10.1088/1748-9326/11/3/035004>.
19. Smith NR, Kishchuk BE, Mohn WW. 2008. Effects of wildfire and harvest disturbances on forest soil bacterial communities. *Appl Environ Microbiol* 74:216–224. <https://doi.org/10.1128/AEM.01355-07>.
20. 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. <https://doi.org/10.1038/ismej.2013.11>.
21. Taş N, Prestat E, McFarland JW, Wickland KP, Knight R, Berhe AA, Jorgenson T, Waldrop MP, Jansson JK. 2014. Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest. *ISME J* 8:1904–1919. <https://doi.org/10.1038/ismej.2014.36>.
22. Murphy JD, Johnson DW, Miller WW, Walker RF, Carroll EF, Blank RR. 2006. Wildfire effects on soil nutrients and leaching in a Tahoe Basin watershed. *J Environ Qual* 35:479–489. <https://doi.org/10.2134/jeq2005.0144>.
23. Certini G. 2005. Effects of fire on properties of forest soils: a review. *Oecologia* 143:1–10. <https://doi.org/10.1007/s00442-004-1788-8>.
24. González-Pérez JA, González-Vila FJ, Almendros G, Knicker H. 2004. The effect of fire on soil organic matter—a review. *Environ Int* 30:855–870. <https://doi.org/10.1016/j.envint.2004.02.003>.
25. Pace NR. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276:734–740. <https://doi.org/10.1126/science.276.5313.734>.
26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
27. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Silva RD, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwards CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Jansson S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe C, Keim P, Kelley ST, Knights D, Koester I, Kosciorek T, Kreps J, Langille MGI, Lee J, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852–857. <https://doi.org/10.1038/s41587-019-0209-9>.
28. Nurse P. 2021. Biology must generate ideas as well as data. *Nature* 597:305–305. <https://doi.org/10.1038/d41586-021-02480-z>.
29. Bickel S, Or D. 2020. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nat Commun* 11:116. <https://doi.org/10.1038/s41467-019-13966-w>.
30. Lopez AM, Wells A, Fendorf S. 2021. Soil and aquifer properties combine as predictors of groundwater uranium concentrations within the Central Valley, California. *Environ Sci Technol* 55:352–361. <https://doi.org/10.1021/acs.est.0c05591>.
31. Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 58:267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
32. Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
33. Rhee S-Y, Taylor J, Wadhwa G, Ben-Hur A, Brutlag DL, Shafer RW. 2006. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc Natl Acad Sci U S A* 103:17355–17360. <https://doi.org/10.1073/pnas.0607274103>.
34. Guan L, Tian X, Gombas S, Zemek AJ, Krishnan G, Scott R, Narasimhan B, Tibshirani RJ, Pham TD. 2017. Big data modeling to predict platelet usage and minimize wastage in a tertiary care system. *Proc Natl Acad Sci U S A* 114:11368–11373. <https://doi.org/10.1073/pnas.1714097114>.
35. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciorek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, McKay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579:567–574. <https://doi.org/10.1038/s41586-020-2095-1>.
36. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. 2019. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A* 116:22071–22080. <https://doi.org/10.1073/pnas.1900654116>.
37. Miller TH, Gallidabino MD, MacRae JI, Hogstrand C, Bury NR, Barron LP, Snape JR, Owen SF. 2018. Machine learning for environmental toxicology: a call for integration and innovation. *Environ Sci Technol* 52:12953–12955. <https://doi.org/10.1021/acs.est.8b05382>.
38. Jeganathan P, Holmes SP. 2021. A statistical perspective on the challenges in molecular microbial biology. *JABES* 26:131–160. <https://doi.org/10.1007/s13253-021-00447-1>.
39. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. *Science* 359:320–325. <https://doi.org/10.1126/science.aap9516>.
40. Vuono DC, Benecke J, Henkel J, Navidi WC, Cath TY, Munakata-Marr J, Spear JR, Drewes JE. 2015. Disturbance and temporal partitioning of the activated sludge metacommunity. *ISME J* 9:425–435. <https://doi.org/10.1038/ismej.2014.139>.
41. Hill MO. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–432. <https://doi.org/10.2307/1934352>.
42. Honeyman AS, Day ML, Spear JR. 2018. Regional fresh snowfall microbiology and chemistry are driven by geography in storm-tracked events, Colorado, USA. *PeerJ* 6:e5961. <https://doi.org/10.7717/peerj.5961>.
43. Wan S, Hui D, Luo Y. 2001. Fire effects on nitrogen pools and dynamics in terrestrial ecosystems: a meta-analysis. *Ecol Appl* 11:1349–1365. [https://doi.org/10.1890/1051-0761\(2001\)011\[1349:FEONPA\]2.0.CO;2](https://doi.org/10.1890/1051-0761(2001)011[1349:FEONPA]2.0.CO;2).
44. Rhoades CC, Formwalt PJ, Paschke MW, Shanklin A, Jonas JL. 2015. Recovery of small pile burn scars in conifer forests of the Colorado Front Range. *For Ecol Manage* 347:180–187. <https://doi.org/10.1016/j.foreco.2015.03.026>.
45. Rhoades CC, Chow AT, Covino TP, Fegell TS, Pierson DN, Rhea AE. 2019. The legacy of a severe wildfire on stream nitrogen and carbon in headwater catchments. *Ecosystems* 22:643–657. <https://doi.org/10.1007/s10021-018-0293-6>.
46. Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM, DeLong EF. 2006. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* 4:e95. <https://doi.org/10.1371/journal.pbio.0040095>.
47. Kerou M, Offre P, Valledor L, Abby SS, Melcher M, Nagler M, Weckwerth W, Schleper C. 2016. Proteomics and comparative genomics of Nitrososphaera viennensis reveal the core genome and adaptations of archaeal ammonia oxidizers. *Proc Natl Acad Sci U S A* 113:E7937–E7946. <https://doi.org/10.1073/pnas.1601212113>.
48. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6:245–252. <https://doi.org/10.1038/nrmicro1852>.
49. Brochier-Armanet C, Gribaldo S, Forterre P. 2012. Spotlight on the Thaumarchaeota. *ISME J* 6:227–230. <https://doi.org/10.1038/ismej.2011.145>.
50. Reyes C, Hodgskiss LH, Kerou M, Pribasnik T, Abby SS, Bayer B, Kraemer SM, Schleper C. 2020. Genome wide transcriptomic analysis of the soil ammonia oxidizing archaeon Nitrososphaera viennensis upon exposure to copper limitation. *ISME J* 14:2659–2674. <https://doi.org/10.1038/s41396-020-0715-2>.
51. Schleper C. 2010. Ammonia oxidation: different niches for bacteria and archaea? *ISME J* 4:1092–1094. <https://doi.org/10.1038/ismej.2010.111>.
52. Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A, Kirkegaard RH, von Bergen M, Rattei T, Bendinger B, Nielsen PH, Wagner M. 2015. Complete nitrification by Nitrospira bacteria. *Nature* 528:504–509. <https://doi.org/10.1038/nature16461>.
53. Baskaran V, Patil PK, Antony ML, Avunje S, Nagaraju VT, Ghate SD, Nathamuni S, Dineshkumar N, Alavandi SV, Vijayan KK. 2020. Microbial community profiling of ammonia and nitrite oxidizing bacterial enrichments from brackishwater ecosystems for mitigating nitrogen species. *Sci Rep* 10:5201. <https://doi.org/10.1038/s41598-020-62183-9>.
54. Huang L, Chakrabarti S, Cooper J, Perez A, John SM, Daroub SH, Martens-Habbena W. 2021. Ammonia-oxidizing archaea are integral to nitrogen

- cycling in a highly fertile agricultural soil. *ISME Commun* 1:19. <https://doi.org/10.1038/s43705-021-00020-4>.
55. Petryshyn VA, Junkins EN, Stamps BW, Bailey JV, Stevenson BS, Spear JR, Corsetti FA. 2021. Builders, tenants, and squatters: the origins of genetic material in modern stromatolites. *Geobiology* 19:261–277. <https://doi.org/10.1111/gbi.12429>.
 56. Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Paruelo J, Raskin RG, Sutton P, van den Belt M. 1997. The value of the world's ecosystem services and natural capital. *Nature* 387:253–260. <https://doi.org/10.1038/387253a0>.
 57. Hill BH, Kolka RK, McCormick FH, Starry MA. 2014. A synoptic survey of ecosystem services from headwater catchments in the United States. *Ecosyst Serv* 7:106–115. <https://doi.org/10.1016/j.ecoser.2013.12.004>.
 58. Westerling AL, Hidalgo HG, Cayan DR, Swetnam TW. 2006. Warming and earlier spring increase western U.S. forest wildfire activity. *Science* 313:940–943. <https://doi.org/10.1126/science.1128834>.
 59. Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci U S A* 102:14683–14688. <https://doi.org/10.1073/pnas.0506625102>.
 60. Alves RJE, Minh BQ, Ulrich T, von Haeseler A, Schleper C. 2018. Unifying the global phylogeny and environmental distribution of ammonia-oxidizing archaea based on amoA genes. *Nat Commun* 9:1517. <https://doi.org/10.1038/s41467-018-03861-1>.
 61. Tournay M, Stieglmeier M, Spang A, Könneke M, Schintlmeister A, Ulrich T, Engel M, Schlöter M, Wagner M, Richter A, Schleper C. 2011. Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci U S A* 108:8420–8425. <https://doi.org/10.1073/pnas.1013488108>.
 62. Stahl DA, de la Torre JR. 2012. Physiology and diversity of ammonia-oxidizing Archaea. *Annu Rev Microbiol* 66:83–101. <https://doi.org/10.1146/annurev-micro-092611-150128>.
 63. Yeager CM, Northup DE, Grow CC, Barns SM, Kuske CR. 2005. Changes in nitrogen-fixing and ammonia-oxidizing bacterial communities in soil of a mixed conifer forest after wildfire. *Appl Environ Microbiol* 71:2713–2722. <https://doi.org/10.1128/AEM.71.5.2713-2722.2005>.
 64. Cobo-Díaz JF, Fernández-González AJ, Villadas PJ, Robles AB, Toro N, Fernández-López M. 2015. Metagenomic assessment of the potential microbial nitrogen pathways in the rhizosphere of a Mediterranean forest after a wildfire. *Microb Ecol* 69:895–904. <https://doi.org/10.1007/s00248-015-0586-7>.
 65. Gimeno-García E, Andreu V, Rubio JL. 2000. Changes in organic matter, nitrogen, phosphorus and cations in soil as a result of fire and water erosion in a Mediterranean landscape. *Eur J Soil Sci* 51:201–210. <https://doi.org/10.1046/j.1365-2389.2000.00310.x>.
 66. Coskun D, Britto DT, Shi W, Kronzucker HJ. 2017. How plant root exudates shape the nitrogen cycle. *Trends Plant Sci* 22:661–673. <https://doi.org/10.1016/j.tplants.2017.05.004>.
 67. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. *Science* 320:1034–1039. <https://doi.org/10.1126/science.1153213>.
 68. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc Natl Acad Sci U S A* 103:12115–12120. <https://doi.org/10.1073/pnas.0605127103>.
 69. 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* 3:417. <https://doi.org/10.3389/fmicb.2012.00417>.
 70. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. <https://doi.org/10.1126/science.aac9323>.
 71. Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N. 2013. Temporal variability in soil microbial communities across land-use types. *ISME J* 7:1641–1650. <https://doi.org/10.1038/ismej.2013.50>.
 72. Ibarra JT, Cockle K, Altamirano T, van der Hoek Y, Simard S, Bonacic C, Martin K. 2020. Nurturing resilient forest biodiversity: nest webs as complex adaptive systems. *Ecol Soc* 25:27. <https://doi.org/10.5751/ES-11590-250227>.
 73. Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 2020. Plant-microbiome interactions: from community assembly to plant health. *Nat Rev Microbiol* 18:607–621. <https://doi.org/10.1038/s41579-020-0412-1>.
 74. Newhart KB, Marks CA, Rauch-Williams T, Cath TY, Hering AS. 2020. Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control. *J Water Process Eng* 37:101389. <https://doi.org/10.1016/j.jwpe.2020.101389>.
 75. Duhamel S, Diaz JM, Adams JC, Djaoudi K, Steck V, Waggoner EM. 2021. Phosphorus as an integral component of global marine biogeochemistry. *Nat Geosci* 14:359–368. <https://doi.org/10.1038/s41561-021-00755-8>.
 76. Green JL, Bohannan BJM, Whitaker RJ. 2008. Microbial biogeography: from taxonomy to traits. *Science* 320:1039–1043. <https://doi.org/10.1126/science.1153475>.
 77. Hansen L. 2019. 416 Fire burned area emergency response (BAER) executive summary. San Juan National Forest, Durango, CO. https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fseprd611411.pdf.
 78. Great Basin National Incident Management Team. 2019. Decker Fire update Thursday, October 24, 2019 as of 8:00 a.m. Salida, CO. https://inciweb.nwcg.gov/photos/CORGF/2019-09-10-2357-Decker-Fire/related_files/pict20190924-101559-0.pdf.
 79. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 18:1403–1414. <https://doi.org/10.1111/1462-2920.13023>.
 80. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W. 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J Biotechnol* 261:169–176. <https://doi.org/10.1016/j.jbiotec.2017.06.1198>.
 81. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267. <https://doi.org/10.1128/AEM.00062-07>.
 82. McMurdie PJ, Holmes S. 2011. PHYLOSEQ: a Bioconductor package for handling and analysis of high-throughput phylogenetic sequence data, p 235–246. *In* Altman RB, Dunker AK, Hunter L, Murray TA, Klein TE (ed), Pacific Symposium on Biocomputing 2012. World Scientific, Hackensack, NJ.
 83. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. <https://doi.org/10.1186/s40168-018-0605-2>.
 84. Andersen KS, Kirkegaard RH, Karst SM, Albertsen M. 2018. ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* <https://doi.org/10.1101/299537>.
 85. R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
 86. Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, Parsons M, Robinson E, Wyborn L. 2019. Make scientific data FAIR. *Nature* 570:27–29. <https://doi.org/10.1038/d41586-019-01720-7>.