ISME

www.nature.com/ismecomms

ARTICLE OPEN

Check for updates

Ecological insights into soil health according to the genomic traits and environment-wide associations of bacteria in agricultural soils

Roland C. Wilhelm[1]✉, Joseph P. Amsili[1], Kirsten S. M. Kurtz[1], Harold M. van Es[1] and Daniel H. Buckley[1]

Soil microbiomes are sensitive to current and previous soil conditions, and bacterial 'bioindicators' of biological, physical, and chemical soil properties have considerable potential for soil health assessment. However, the lack of ecological or physiological information for most soil microorganisms limits our ability to interpret the associations of bioindicators and, thus, their utility for guiding management. We identified bioindicators of tillage intensity and twelve soil properties used to rate soil health using a 16S rRNA gene-based survey of farmland across North America. We then inferred the genomic traits of bioindicators and evaluated their environment-wide associations (EWAS) with respect to agricultural management practice, disturbance, and plant associations with 89 studies from agroecosystems. Most bioindicators were either positively correlated with biological properties (e.g., organic matter) or negatively correlated with physical and chemical properties. Higher soil health ratings corresponded with smaller genome size and higher coding density, while lower ratings corresponded with larger genomes and higher *rrn* copy number. Community-weighted genome size explained most variation in health ratings. EWAS linked prominent bioindicators with the impacts of environmental disturbances. Our findings provide ecological insights into bioindicators of soil properties relevant to soil health management, illustrating the tight coupling of microbiome and soil function.

ISME Communications; https://doi.org/10.1038/s43705-022-00209-1

INTRODUCTION

Managing soil health promotes the long-term fertility and ecological integrity of agricultural lands [1, 2]. Soil health encompasses a range of soil properties that contribute value to agroecosystems, including nutrient and water cycling, biodiversity, plant pathogen suppression, and pollution mitigation. Soil health is monitored using biological, physical, and chemical indicators that correspond with these functions [3–5]. Ideally, indicators should be directly linked to soil function, interpretable, and exhibit a dynamic response to management practices [6–10]. The soil microbiome has considerable potential to serve in this capacity. Microbial communities are highly sensitive to management practices [11–14], including those that shape properties that determine soil health in agricultural systems [15–20]. The broad ecological and functional diversity of bacteria in soil provides rich information about soil conditions, which was recently used to predict soil health status [21]. However, our ability to interpret the responses of bacterial 'bioindicators' is limited by our sparse understanding of the ecology and function of most bacteria in soil. Bridging this gap between soil microbial ecology and soil health will improve the use of microbiome data in soil health monitoring.

Ecological insight into soil microbiome structure and function can be derived by leveraging the large amounts of DNA sequencing data available in public repositories. One form of ecological inference can be derived from genomic data, whereby microbial traits can be estimated from representative genomes that are close relatives of taxa observed in phylogenetic gene marker surveys [22]. Genomic traits, such as genome size, codon usage bias, and *rrn* copy number, can be used to derive ecological information from trends in soil microbiome composition [23, 24] based on the evolutionary tradeoffs between growth, survival, and reproduction shaping these traits [25–27]. Genomic traits form the basis of several life-history frameworks that group bacteria by ecological strategies (e.g., 'generalist' vs. 'specialist') [28]; adaptive tradeoffs between growth rate, yield, and stress tolerance [26, 29, 30]; or metabolic dependency (e.g., 'prototrophic' vs. 'auxotrophic') [31]. These frameworks have been used to interpret microbiome trends associated with agricultural management practices, such as tillage intensity and nutrient management [32, 33].

While promising, the genomic inference of ecological traits has notable limitations. For example, many of the most active and abundant microorganisms in agricultural soils lack representative genomes from which traits might be predicted [34–37]. Ecological information can still be derived for these non-cultivated organisms by profiling their phylogenetic gene markers across the growing number of publicly available amplicon sequencing projects [38, 39]. An 'environment-wide association survey' (EWAS) approach follows the principle of reverse ecology, where information is inferred from changes in the abundance and distribution of genes across sites [40], in our case the 16S rRNA phylogenetic marker gene across environmental conditions. Traditional approaches assign a trait using curated databases [41, 42], which tend to exclude uncultured or poorly characterized taxa. This is problematic since unclassified taxa are often indicative

[1]School of Integrative Plant Sciences, Bradfield Hall, Cornell University, Ithaca, NY 14853, USA. ✉email: rwilhelm@cornell.edu

Received: 5 September 2022 Revised: 12 December 2022 Accepted: 20 December 2022
Published online: 09 January 2023

of soil properties relevant to soil health management [21, 37, 43, 44]. In contrast, EWAS requires no prior knowledge, given the capacity to obtain information for any organism with a phylogenetic gene marker present in sequencing databases [45–48]. An EWAS approach is primarily limited by the poor quality of metadata reported for most sequencing projects [49] and a historical lack of standardization in sequencing workflows. These drawbacks are partially compensated for by the sheer volume of available sequencing projects and renewed efforts to systematize data publishing will improve the efficacy of EWAS over time [50].

Our study identified and characterized bacterial bioindicators of soil properties used in soil health assessment using a large amplicon sequencing survey of farmland across North America. Our first objective was to utilize 16S rRNA gene sequencing data to identify bioindicators that correlate with twelve biological (e.g., organic matter), physical, and chemical soil properties used in soil health assessment. We focused on profiling specific bioindicator species given the relatively minor differences observed in diversity metrics reported for our dataset [21]. Our second objective was to evaluate trends in bioindicators using (i) inferred genomic traits and (ii) a 16S rRNA gene-based EWAS to understand the ecological basis for their associations with soil health. For (i), we tested whether trends in community-weighted genomic traits corresponded to variation in soil health ratings. For (ii), we explored the environment-wide associations (EWAS) of key bioindicators using a database comprised of agricultural microbiomes (derived from 89 prior studies) that included diverse metadata grouped by study factors into broad (management practice, disturbance, and plant association) and specific categories (fertilization, land-use, tillage, drought etc.). This combined approach yielded ecological information about the most abundant bioindicators of soil health and provided new perspectives on the relationships between the soil microbiome and properties related to healthy soil function.

METHODS

Soil health and bacterial community data collection

Our primary dataset consisted of 778 soil samples sourced from farmland across the USA, representing diverse cropping systems, as part of a soil health initiative led by Cornell University and the USDA Natural Resources Conservation Service. Soils originated from 191 unique locations that differed in agricultural management practices and soil health ratings. This dataset was used in a separate study to test the accuracy of microbiome-based machine learning for predicting soil health [21]. Our study aims to identify bioindicators and explore the underlying ecological basis for their association with soil health ratings, which have yet to be examined.

The soil properties of each sample were collected using the Comprehensive Assessment of Soil Health (CASH) framework (Table S1), which uses *biological* (soil organic matter, respiration, ACE protein, and active carbon, also known as ‘permanganate oxidizable organic carbon’), *chemical* (pH, phosphorus, potassium, and minor elements), and *physical ratings* (aggregate stability, available water capacity, soil texture, and surface and sub-surface hardness) to assess soil health [7]. Tillage data was collected for most soils ($n = 599$) and was coded as ‘till’ vs. ‘no till.’ Surface and sub-surface hardness ratings were inverted so that more compacted soils corresponded with higher ratings (opposite of CASH framework); these ratings were present for a subset of samples ($n = 309$ and 292, respectively). Measurements for each soil property were transformed using a scoring function to create a normalized rating that accounts for differences in soil texture [7]. A total health score was then calculated from the unweighted mean of all twelve ratings. Perspectives on the nature of soil health assessment and health indicators continues to evolve [10]. The soil properties in the CASH framework have been used extensively to assess the impacts of soil management practices on soil function [7].

Total DNA was extracted from soils to determine bacterial community composition and was also used to estimate microbial biomass [51]. DNA was extracted using the DNeasy PowerSoil Kit, as per manufacturers recommendation (QIAGEN, Germantown, MD, USA). DNA concentration was quantified in triplicate using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA). Bacterial community composition was determined through amplicon sequencing of the V4 region of the 16S rRNA gene using Illumina MiSeq (2 × 250 paired-end) and dual-indexed barcoded primers (515f/806r; sequences provided in Table S2) as previously described [21, 52]. Demultiplexing, filtering and trimming, and chimera removal were performed with *QIIME2* (v. 2020.2) [53] using default parameters and trimming left and right by 5 bp. Operational taxonomic units (OTUs) were defined as amplicon sequence variants and assigned taxonomic classifications using *QIIME2* with dependencies on *DADA2* [38] and the *Silva* database (nr_v132) [54], respectively. Raw sequencing data was archived at the National Centre for Biotechnology Information (BioProject: PRJEB35975).

Identifying bacterial ‘bioindicators’ of soil health ratings

Bacterial OTUs indicative of soil health were determined by Spearman rank correlations using the ‘rcorr’ function in the R package *Hmisc* (v. 1.34.0) [55]. Prior to correlation analyses, OTUs occurring at low frequency (fewer than 10 samples), and at low relative abundance (<0.01% of average read depth) were removed and data was normalized by sequencing depth and reported as counts per thousand reads. *p*-values were adjusted according to the Benjamini and Hochberg false discovery rate [56]. Weak correlations ($r < |0.3|$ and $p_{\mathrm{adj}} > 0.05$) were removed [57, 58]. Indicator species analyses was used to identify bioindicators for tillage intensity using the “multipatt” function in the R package *indicspecies* (v. 1.7.12) [59]. All analyses can be reproduced with scripts included in the Supplementary Data package.

Analysis of community-weighted genomic traits

OTUs were assigned genomic trait values using representative genomes present in public databases. Genomic traits summarized by IMG-ER [60] were downloaded (March 15th, 2020) for all isolate ($n = 68{,}600$), single-cell amplified ($n = 3400$), and metagenome-assembled genomes ($n = 8800$). Genomic traits were selected based on prior evidence of their correlation with life-history strategies [25–27] and availability in the IMG-ER portal, namely: genome size, coding density (total length of coding regions/genome size), *rrn* operon copy number, CRISPR arrays, and biosynthetic gene clusters (BGCs). Gene abundances were normalized by genome size. OTUs were assigned a trait value iteratively based on taxonomic classification. Unclassified OTUs at the rank genus were progressively matched to averaged trait values at higher taxonomic ranks. Most OTUs were assigned a trait value (20,148/21,463; 94%) and the majority were assigned at their lowest classified taxonomic rank (58%). The community-weighted average trait values were calculated for whole bacterial communities using the weighted mean based on the relative abundance of each OTU in the community. In addition, community-weighted *rrn* abundance was re-calculated using the *rrn*DB (v. 5.6) [61], yielding results consistent with those derived from IMG-ER data.

Environment-wide association survey (EWAS)

The EWAS of bioindicators were determined from trends in the relative abundance of identical OTUs in other 16S rRNA gene amplicon datasets from agricultural and related terrestrial environments (full details in Supplementary Methods). In brief, we compiled 89 studies totaling 14,780 individual 16S rRNA gene amplicon libraries, termed the ‘AgroEcoDB.’ Amplicon libraries were downloaded from the Short Read Archive (May 15th, 2020) for BioProjects with taxonomic IDs for “soil metagenome” (taxID: 410658), “compost metagenome” (702656), “decomposition metagenome” (1897463), “fertilizer metagenome” (1765030), “manure metagenome” (1792145), “rhizosphere metagenome” (939928), and “wood decay metagenome” (1593443). The final database was filtered from an initial 729 BioProjects to 89 based on the following criteria: (i) common overlap of the V4 region of the 16S rRNA gene, (ii) experimental manipulations, when used, were typical of agricultural management, and (iii) contained at least 15 samples with well-curated metadata. Sequences were quality filtered and assigned to OTUs using the identical methods applied to our primary soil health amplicon data. Common OTUs were those with exact sequence matches (i.e., based on amplicon sequence variant IDs) after ensuring all sequences had the exact same length prior to processing with DADA2.

Indicator species analyses was then used to calculate an indicator value for all OTUs in the AgroEcoDB based on each individual study factor (‘EWAS indicators’; $p_{\mathrm{adj}} < 0.05$). Study factors were categorized by management categories (e.g., inorganic vs. organic fertilizer and other broad strategies, like crop rotation), disturbance (tillage, drought etc.), plant association (bulk vs. rhizosphere soil), biome (grassland vs. cropland) and other minor categories (decomposition, soil depth etc.; see Table S3). The indicator value of EWAS indicators were scored as positive or negative based on whether the study factor was positively (reduced tillage, OM management, etc.) or negatively associated with soil health (Table S3). Our subsequent analyses provided a test of these assumptions. EWAS indicator values were averaged and assigned to their corresponding OTUs in the soil health dataset. Assigned indicator values were used to calculate community-weighted averages grouped by categories (i.e., management, disturbance, plant-association, biome, etc.) in the same way as genomic trait values.

Statistical analyses

Statistics were performed using R (v. 4.0.3) [62] with dependency on the following packages: *reshape2* (v. 1.4.4), *ggplot2* (v. 3.3.2), *plyr* (v. 1.8.6) [63–65], and *phyloseq* (v. 1.34.0) [66]. Permutational multivariate analysis of variance (PERMANOVA) was performed on Bray-Curtis dissimilarity using the R package *vegan* (v. 2.5.7) with 999 permutations. PERMANOVA was repeated with 50 permutations of factor order to obtain average R^2 values. The relative importance of community-weighted traits and EWAS for explaining variation in community composition was compared with *relaimpo* (v. 2.2.6) [67]. Co-occurrence networks were constructed for bacterial taxa (aggregated by genus) based on whether two genera shared a common bioindicator status for each of the twelve soil health ratings. Edges were weighted by the number of OTUs co-occurring between nodes (i.e., each genus) and bioindicators with negative and positive correlations with ratings were visualized separately. Co-occurrence networks were visualized using Gephi (v.0.9.2) [68] with network topography determined by the Yifan Hu 'proportional' force-directed graphing algorithm (relative strength = 2) [69].

RESULTS

Relationships among soil health ratings

Biological ratings of soil health were highly interrelated and positively correlated with total health score and with aggregate stability (Fig. S1). Total health score was negatively correlated with surface and sub-surface hardness ratings (where a higher rating indicates greater compaction) and sand content. DNA yield was significantly positively correlated with total health score ($r = 0.51$; $p < 0.001$) but was heavily influenced by clay content, likely due to the absorptive effects of clay on DNA extraction reagents (Fig. S2).

Bioindicators of soil properties and health ratings

We evaluated whether variance in OTU relative abundance was correlated with each of twelve health ratings and with total health score to identify bacterial bioindicators ($r > |0.3|$ and $p_{\text{adj}} < 0.05$). A subset of OTUs (8.7%; 1874/21,463) were identified as correlated with one or more health ratings ($\mu = 1.5$ ratings per OTU; max = 5). These 'bioindicators' were taxonomically diverse (348 different classifications at rank genus) with most belonging to candidate groups or unclassified genera (62%). Approximately twice as many unclassified or candidate genera (1.9-fold) were present in the bioindicator set (215/348) compared to the overall dataset (430/943). Correlations of bioindicators with biological ratings (i.e., organic matter quantity and composition) were primarily positive, while correlations with physical or chemical ratings were largely negative (Fig. 1). The majority of bioindicator OTUs were correlated in a consistent direction with one or more health ratings (96%; 1798/1874). Many genera (46%) contained a diversity of bioindicator OTUs that differed in their relationship to soil health ratings.

The main bioindicators of high biological health status were OTUs classified to Candidatus *Udaeobacter* (*Verrucomicrobia*) and *Illumatobacteraceae* (*Actinobacteria*), as well as unclassified groups of *Chloroflexi* (order *KD4-96*), *Alphaproteobacteria* (*Xanthobacteraceae*) and *Actinobacteria* (class *MB-A2-108*; full list in Table S4). The most consistent bioindicators of low physical, chemical and total health scores were OTUs classified as *Sphingomonas* (*Alphaproteobacteria*), and unclassified groups of *Chloroflexi* (order *JG30-KF-CM45*), *Archaea* (Ca. *Nitrososphaeraceae*), and *Acidobacteria* (genus *RB41*). Many highly abundant taxa (occurring at 1–5% of total read counts) were differentially abundant in tilled soils ($n_{\text{OTU}} = 292$) as compared to no till systems ($n_{\text{OTU}} = 18$; Table S4). The predominant bioindicators of tillage were members of *Alphaproteobacteria* (*Sphingomonadaceae*, *Rhizobiaceae* and *Caulobacteraceae*), *Acidobacteria* (*Pyrinomonadaceae*), *Verrucomicrobia* (genus *Chthoniobacter*) and *Actinobacteria* (genus *Terrabacter*). The main bioindicators of untilled fields coincided with the previously mentioned bioindicators of high biological health ratings, as well as *Actinobacteria* classified to *Gaiella* and unclassified groups of *Solirubrobacterales* (Table S4).

Genomic traits linked to soil health and tillage

We evaluated whether community-weighted genomic traits (see Methods) explained variance in soil health ratings. Several inferred genomic traits correlated with soil health ratings. Community-weighted genome size, CRISPR array frequency, and number of BGCs were all negatively correlated with total health scores (Fig. 2A). The relationships among traits in community-weighted data broadly reflected the existing correlations among genomic traits (Mantel statistic $r = 0.64$; $p = 0.01$). However, the relationships among genome size, BGCs, and *rrn* copies in community-weighted data exhibited opposite trends from those observed in genomic databases (Fig. 2B).

Community-weighted CRISPR array frequency exhibited some of the strongest correlations with health ratings, being negatively correlated to water capacity, OM, and total health score (Fig. 3), and positively correlated with sand content ($r = 0.44$; $p < 0.001$). Community-weighted coding density was positively correlated with total health score and ratings of OM quality (active carbon and ACE protein; Fig. 3), and negatively correlated with DNA yield, a proxy for microbial biomass. The bacterial bioindicators of DNA yield with the lowest coding density were classified to *Chloroflexi* (*Ktedonobacterales*: HSB_OF53-F07 and JG30a-KF-32; $\mu_{density} = 76.2$) and a family of *Planctomycetes* (*Gemmataceae*; $\mu_{density} = 79.1$).

Community-weighted *rrn* copy number was not correlated with total health score ($r = 0.003$), but it was significantly higher in tilled vs. untilled soils (Wilcoxon, $p < 0.001$; Fig. 4), and was correlated with surface and sub-surface hardness ratings (Fig. 3). Variance in community-weighted *rrn* copy number was driven by the abundances of *Georgenia* ($\mu_{rrn} = 5.7$; *Actino.*), *Bacillaceae* ($\mu_{rrn} = 5.5$; *Firmicutes*), and *Planococcaceae* ($\mu_{rrn} = 5.4$; *Firmicutes*), which were all favored by tillage. Community-weighted genome size and BGC number were also higher in tilled soils relative to untilled soils (Fig. 4), and this result is consistent with their negative correlation to total health score. Community-weighted genome size and BGCs were primarily correlated with biological health ratings, unlike *rrn* copy number which was exclusively correlated with physical or chemical health ratings (Fig. 3).

Environment-wide associations of bioindicators of soil health

The majority of OTUs in the soil health dataset were present in the AgroEcoDB ($n_{\text{OTU}} = 17{,}818/21{,}573$ of OTUs at 100% identity), representing a total of 96.9% of all sequences. A total of 8760 OTUs found in both datasets were identified as significant EWAS indicators ($p_{\text{adj}} < 0.05$) of one or more study factors in the AgroEcoDB. The indicator values of EWAS indicators were used to calculate community-weighted averages of broad categories and sub-categories to assess general correlations between the EWAS of bacteria and soil health. Community-weighted EWAS, inferred from amplicon data, explained more variation in bacterial community composition than community-weighted genomic traits, inferred from genomic databases (Table 1A). In contrast, community-weighted genome size explained more variation in total health score than any of the EWAS categories or sub-categories (Table 1B).

We examined the EWAS of the most abundant bioindicator taxa driving the relationship between community-weighted genome size and active carbon rating (Fig. 5). Active carbon rating is a

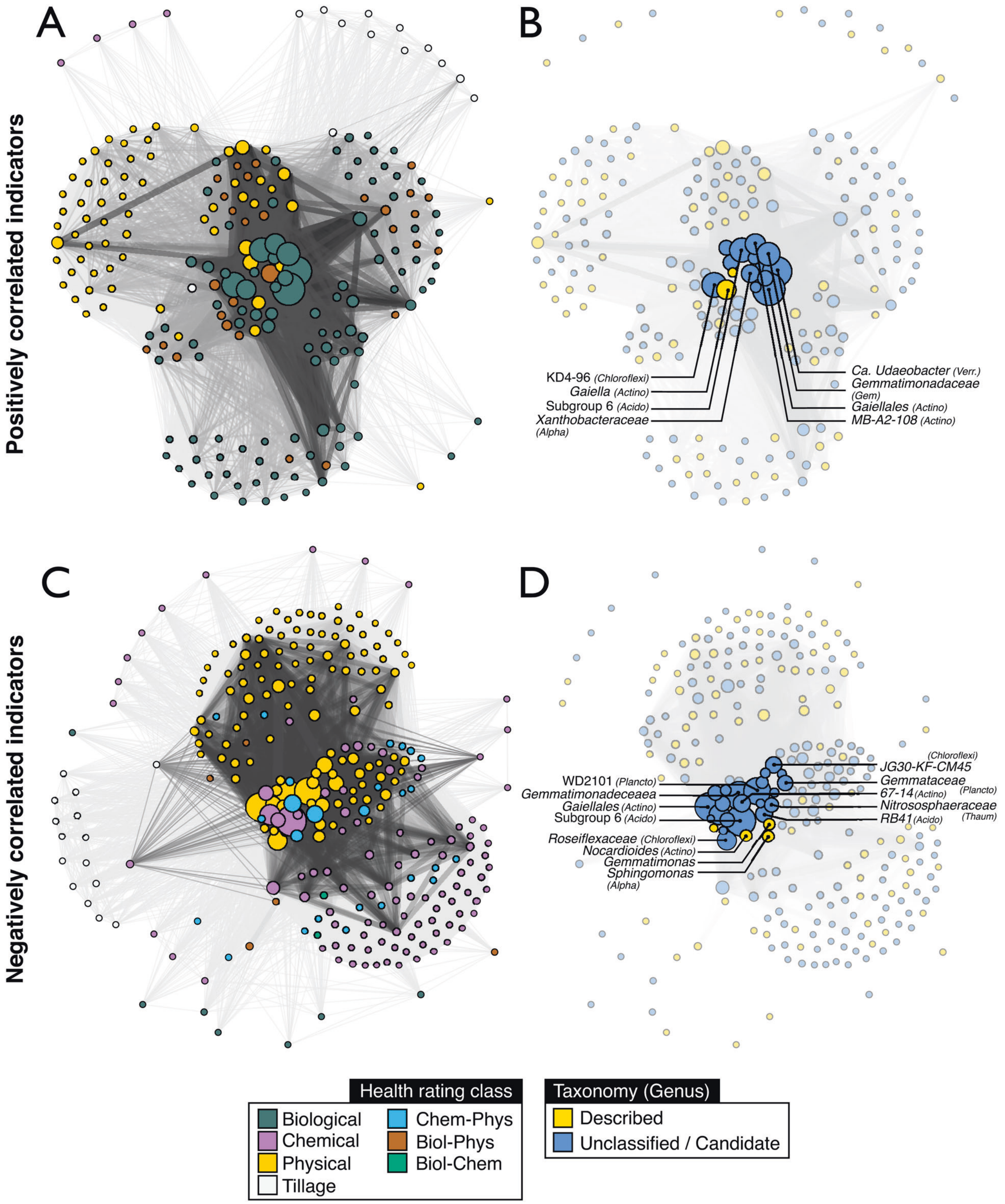

Fig. 1 An overview of the general relationship between bioindicators of biological, physical, and chemical classes of soil properties. Among all correlated bioindicators of soil health properties shown in these co-occurrence network diagrams, the majority of positive correlations where with biological ratings (**A** and **B**), while the majority of negative correlations where with physical and chemical ratings (**C** and **D**). Networks were divided based on whether indicators exhibited positive (**A** and **B**) or negative correlations (**C** and **D**) with health ratings. In (**A**) and (**C**), nodes are colored according to health rating class and, in (**B**) and (**D**), according to whether a taxon is represented by a described species. The relationship among bioindicators was visualized in a network to highlight the prevalence of key taxa (see labeling of nodes in **B** and **D**); differences in the relationships of bioindicators with soil health classes (in **A** and **C**); and the high number of uncultured/unclassified bioindicators (in **B** and **D**). Nodes represent bioindicator OTUs aggregated to their lowest resolved taxonomic rank and scaled by the total number of OTUs. Edges represent co-occurrence of indicator OTUs for one or more of the same health rating. Edge weights are scaled by the number of co-occurring OTUs common between nodes. In (**A**) and (**C**), nodes were colored based on majority rules according to the number of OTUs representing a given health class. Classes were hyphenated when no majority was achieved.

measure of oxidizable soil C and it is a strong predictor of total soil health (Fig. S1). The bioindicators of active carbon with the largest estimated genome size were classified as *Chthoniobacter* (μ_{size} = 7.8 Mb; $\mu_{rel.abund.}$ = 0.5% of total counts), *Geodermatophilaceae* (4.8 Mb; 1.0%), and *Sphingomonas* (4.2 Mb; 1.7%), and those with the smallest were: *Gaiella* (1.5 Mb; 2.1%), Ca. *KD4-96* (2.3 Mb; 4.9%), and Ca. *Udaeobacter* (2.7 Mb; 3.5%). Of these representative taxa, those with larger genomes had consistently higher relative abundance in soils having low active carbon ratings relative to those taxa with smaller genomes (Fig. 6). In addition, bioindicators with larger genomes were associated with tilled soils (Fig. 6). The EWAS revealed that all of these taxa were associated with bulk soil rather than rhizosphere soil (Fig. 7A). The EWAS also revealed the relationship between genome size and disturbance (Fig. 7B). These trends were driven by disturbances related to tillage (Fig. S3) and watering regimes (Fig. S4). The EWAS did not reveal any consistent

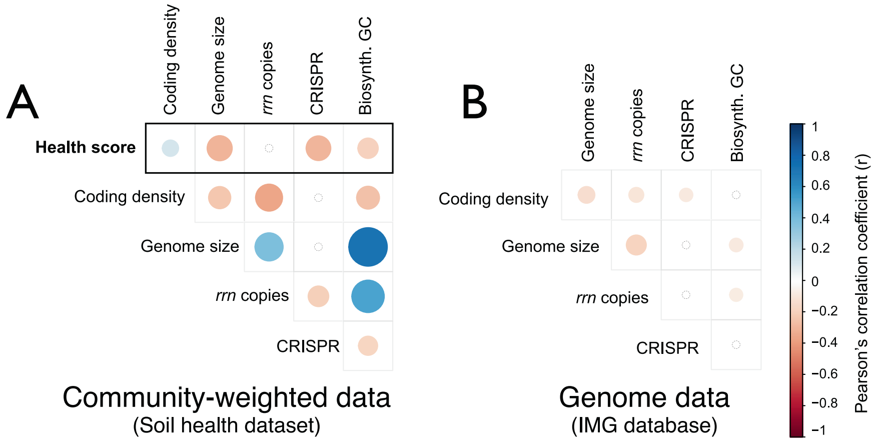

Fig. 2 Correlations among genomic traits and also total health score. In (**A**), correlations were based on average trait scores weighted by the relative abundance of taxa-specific traits values (i.e., community-weighted data) in the soil health data. In (**B**), the same calculation was made from the genomic database used to assign trait values to taxa. This side-by-side comparison illustrates that the relationships among traits in community-weighted data partially reflected the existing relationships observed in the genomic data (Mantel statistic $r = 0.64$; $p = 0.01$). The strength of each Pearson's correlation corresponds with color intensity as indicated by the scale provided. Circle area corresponds to the inverse of p value with non-significant values indicated by a small, colorless circle.

			Biological				Physical				Chemical			
Community-weighted trait	Health score	DNA	Active carbon	ACE protein	Org. matter	Respir.	Water cap.	Agg. stabil.	Surf. hard.	Sub. hard.	pH	P	K	Minor element
Genome size	**-0.306**	-0.108	**-0.444**	**-0.302**	**-0.218**	**-0.195**	-0.016	**-0.141**	-0.172	-0.006	**-0.222**	0.062	0.091	**-0.29**
CRISPR	**-0.3**	-0.043	**-0.192**	**-0.165**	**-0.322**	-0.103	**-0.337**	**-0.149**	-0.175	-0.146	**0.291**	**-0.223**	-0.109	-0.002
BGCs	**-0.199**	-0.079	**-0.31**	**-0.316**	**-0.165**	-0.078	-0.034	-0.003	-0.02	-0.001	**-0.128**	-0.035	**0.121**	-0.098
rrn	0.003	-0.066	0.001	-0.026	0.034	-0.011	0.096	**-0.146**	**-0.316**	**-0.316**	0.118	**-0.153**	0.078	0.04
Coding density	**0.126**	**-0.152**	**0.16**	0.158	0.065	-0.084	0.073	-0.073	-0.002	0.141	**0.139**	0.034	**0.156**	**-0.138**

Fig. 3 A summary of correlations between soil health ratings and community-weighted traits. Community-weighted genome size explains the most variance in overall health ratings. All Pearson's $r > |0.3|$ are shaded blue and all significant correlations are shown in bold.

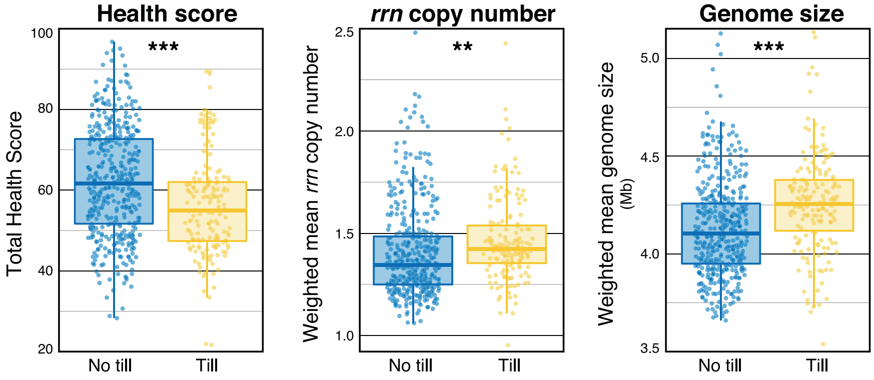

Fig. 4 The relationship between tillage, soil health, and community weighted genomic traits. Tillage was significantly different according to variation in overall soil health rating, community-weighted *rrn* copy number, and community-weighted genome size. Significant differences are denoted with asterisk based on *t*-tests (** $p < 0.01$; *** $p < 0.001$).

Table 1. Community-weighted traits and environment-wide associations (shaded) explain variation in (A) community composition according to PERMANOVA based on Bray-Curtis dissimilarity and (B) total health score according to relative importance values.

A Variation in Community Composition According to PERMANOVA

	Community-weighted trait / environment-wide association	R^2
Category	Disturbance	0.045
	Management	0.039
	Genome size	0.035
	Plant association	0.03
	Coding density	0.019
	rrn copy number	0.017
Sub-category	Organic matter quality	0.031
	Tillage	0.03
	Organic vs. conventional	0.03
	Genome size	0.03
	Drought	0.022
	Land use	0.022
	Host association	0.021
	Coding density	0.017
	rrn copy number	0.016
	NPK fertilization	0.013

Color legend

Community-weighted trait	□
Environment-wide association	▨

B Variation in Soil Health Score According to RELAIMPO

	Community-weighted trait / environment-wide association	Rel. Imp.
Category	Genome size	43
	Management	19
	CRISPR	19
	Coding density	7.9
	rrn copy number	5.6
	Disturbance	2.1
	Plant association	2.1
	Biosynthetic gene cluster	0.6
Sub-category	Tillage	24
	Genome size	23
	Organic matter quality	12
	CRISPR	11
	Drought	6.9
	Land use	5.5
	Coding density	4.8
	NPK fertilization	4.8
	rrn copy number	3.2
	Host association	3.1
	Organic vs. conventional	1.9
	Biosynthetic gene cluster	0.5

The analysis was repeated for environment-wide associations grouped at different hierarchal levels designated as category' or 'sub-category' as described in Table S3.

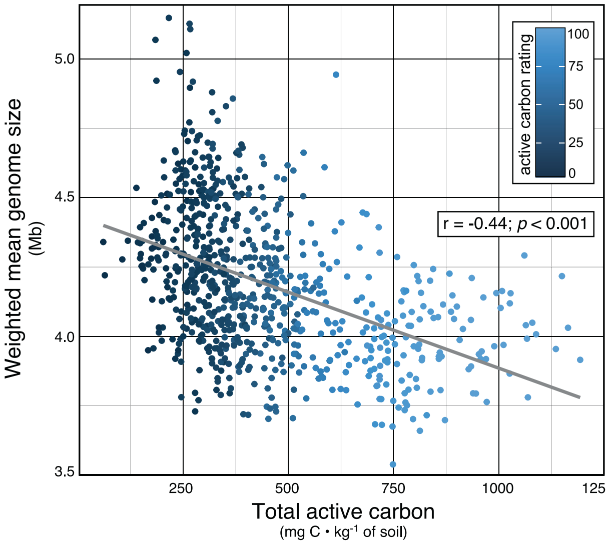

Fig. 5 Community-weighted genome size explains significant variation in active carbon content. Points were colored on the basis of active carbon rating in soil health assessment. Differences in the proportion of unclassified taxa assigned traits was not correlated with active carbon rating ($r = 0.06$, $p = 0.1$; see Fig. S6).

effect of management practices on these representative taxa, which included crop rotation, land-use, and fertilization (Fig. 7C).

DISCUSSION

Our study sought to generate ecological insights into the bacterial bioindicators of biological, chemical, and physical soil properties using microbiome-based analyses of genomic traits and EWAS. Our findings provide further support for the conclusion that changes in microbiome composition are associated with variation in properties relevant to soil health management [19, 21, 32]. A diverse set of 348 bacterial genera were identified as bioindicators of one or more soil health ratings and the majority (62%) of these taxa belonged to, as yet, unclassified genera—a twofold over-representation compared to the whole community. This finding underscores the need for alternative, classification-independent strategies, such as genomic trait-based inference and EWAS, that allow us to gain ecological insights into uncultivated microbes from the representation of their sequences in amplicon, SAG, and MAG public databases.

Genomic traits that correlate with soil disturbance and soil health

The relationship between community-weighted genome size and total health score was among the strongest correlations observed (Fig. 3) and was the most important predictor of variation in total health score (Table 1B). On average, communities with a greater proportion of bacteria with larger genomes occurred in soils with lower overall health rating, lower biological ratings, a history of tillage, and reduced water availability (Fig. S4). This result could indicate that soils of low health select for larger genomes and/or because they select against bacteria with smaller genomes.

Bacteria with larger genomes are hypothesized to have an advantage in habitats characterized by high environmental variability, where their expanded regulatory and metabolic capabilities allow for rapid physiological adaptation to environmental change [70, 71]. This hypothesis might suggest that large genomes are favored in soils with lower soil health because these systems exhibit more environmental instability than healthy soils, with respect to physical disturbance and moisture and nutrient availability. For example, microbes having greater metabolic flexibility are favored in tidal systems that exhibit substantial variation in moisture and nutrient availability over time [72]. Conversely, bacteria with smaller genomes often depend on interspecies interactions and community goods, and these dependencies might render them more sensitive to disturbance. For example, *Udeaobacter* possess a remarkably

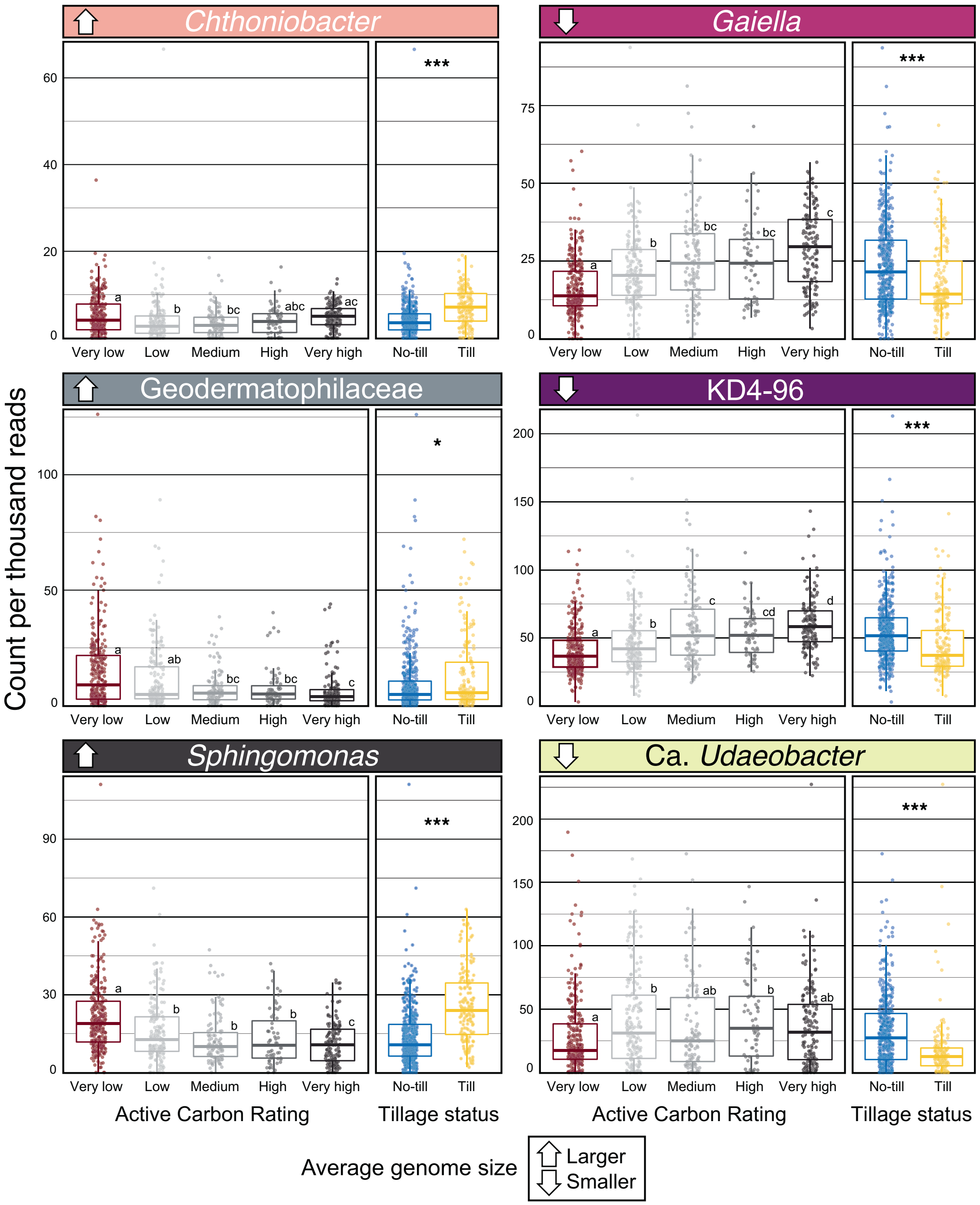

Fig. 6 Variation in the relative abundance of six key bacterial taxa indicates that the effects of active carbon rating and tillage differ with respect to genome size. A set of six taxa were selected to represent extremes of genome size from three of the largest (left size, indicated by upward arrow), to three of the smallest (right size, indicated by down arrow). Collectively, these six taxa comprised 14% of all reads and their relative abundance has a strong impact on relationships indicated in Figs. 4 and 5. Active carbon ratings were divided into categories that range from very low (0–20) to very high (80–100). Pairwise statistically significant differences ($p < 0.05$), according to Tukey HSD, are denoted by lettering.

reduced genome [73] and they were among the strongest bioindicators of higher health ratings and exhibited a strong negative response to tillage (Figs. 6; 7; S4). *Udeaobacter* also exhibits high levels of auxotrophy and antibiotic tolerance [73, 74], which are characteristics associated with the life-history strategies exhibited by dependent organisms [31, 75].

Our analyses also illustrate the challenges of mapping the ecological characteristics of taxa on the basis of inferred genomic traits and environmental associations. For example, *Chthoniobacter* strains tended to have larger genomes (ranging from 3.6–7.8 Mb) and, consistent with other observations, were more prevalent in tilled soils (Fig. 6). However, we also found that *Chthoniobacter* were indicative of high biological ratings, which runs counter to the trends observed in other key bioindicators. The reason for this apparent contradiction remains unclear. It is possible that members of this genus occupy an as yet to be

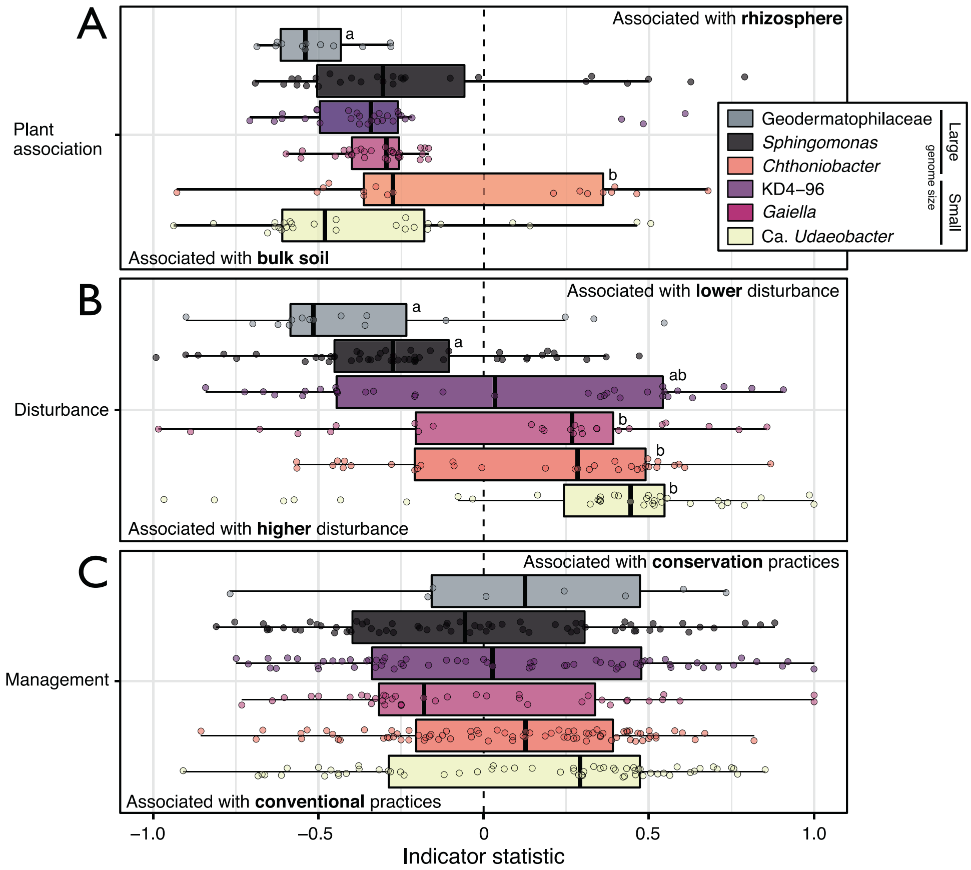

Fig. 7 Environment-wide associations of the taxa indicated in Fig. 6. Each plot shows indicator OTUs for (**A**) plant association (i.e., bulk vs. rhizosphere soil), (**B**) soil disturbance and (**C**) management practices. In (**B**) and (**C**), indicator values were assigned as positive or negative based on whether a factor was a reference or treatment (e.g., no-till vs. till) with details of designations provided in Table S3. Only OTUs shared among the AgroEcoDB and soil health data were included. Pairwise statistically significant differences ($p < 0.05$), according to Tukey HSD, are denoted by lettering.

determined ecological niche that is both favored by tillage and also present in soils of high soil health. However, it is also possible that species within the genus exhibit sufficient ecological differentiation so as to be favored in different types of soils. Such niche differentiation at high phylogenetic resolution would pose a challenge to analyses that build inference from the similarity of taxonomic markers.

Bioindicators of soil health ratings

Most bacterial taxa indicative of physical and chemical health ratings increased in relative abundance at lower soil health ratings (Fig. 1). Such relationships might be expected for stress-tolerant taxa or those that thrive under low nutrient conditions. For example, bioindicators classified to the family *Nitrososphaeraceae* were strongly indicative of soils with poor soil health ratings. *Nitrososphaeraceae* are ammonia-oxidizing archaea that have high substrate affinity and thrive under nutrient limiting conditions [76]. These taxa are commonly enriched in conventionally-managed agricultural soils fertilized with ammonia [19, 77], and they were linked to fertilizer use in our EWAS analyses (Fig. S5). It is important to note that an increase in relative abundance does not necessarily indicate an increase in absolute abundance. An increase in relative abundance does, however, indicate greater relative fitness under a given condition (i.e., potential for contributing genetic material to future generations relative to co-occurring community members), but this increase in fitness could be due to adaptations that promote prolonged survival rather than adaptations that favor reproductive growth.

Most bacterial taxa indicative of biological health ratings increased in relative abundance at higher soil health ratings. Positive correlations between bacterial relative abundance and biological ratings might indicate the enrichment of organoheterotrophs in soils with higher OM quantity and quality. Alternatively, these correlations could result from the sensitivity of these taxa to environmental disturbance, such as tillage, since demographic trends are driven by both growth and death of cells. Soils that have high health ratings tended to have high respiration and high DNA yield, suggestive of higher microbial biomass which is consistent with the hypothesis that members of the organoheterotroph community are enriched in soils with higher health ratings.

Community-weighted *rrn* copy number was significantly higher in tilled fields, which is consistent with previous findings [37]. Community-weighted *rrn* copy number was the only trait that had a significant relationship with hardness ratings and the only trait not correlated with total health score or biological health ratings (Fig. 3). Notably, hardness ratings were negatively correlated with *rrn* copy number, indicating communities in more compacted soils tended to encode a higher number of *rrn* operons. This matched our expectation that higher *rrn* copy number would be associated with more degraded soils, though the nature of the relationship remains unclear. These results confirm the influence of physical disturbance on soil bacteria, also reported by Rieke et al. [20], and their potential to serve as bioindicators of soil properties relevant to the functioning of soils.

Exploring relationships between coding density, CRISPR array abundance, and soil health

Low coding density is a signature of obligate epibionts, endobionts, and parasites, arising from relaxed selection pressure and an accumulation of pseudogenes [78, 79]. Thus, we predicted that lower community-weighted coding density might indicate higher trophic dependency in soils supporting more microbial biomass, which correlate with higher health ratings. However, contrary to expectations, overall community-weighted coding density exhibited the opposite trend (i.e., positively correlated) with total health score. Hence, it is not clear that coding density has a straightforward relationship with soil health status.

We also explored the relationship between community-weighted CRISPR array frequency and soil health with the expectation that phage pressure would select for genomes having more arrays. CRISPR array frequency was the only genomic trait to exhibit strong inverse relationships with water capacity and OM ratings, and a positive correlation with sand content. These observations run counter to expectations that community-weighted CRISPR array abundance would be greatest in OM-rich soils, which retain moisture and would thus support higher average phage abundance [80, 81]. We hypothesize that the high community-weighted frequency of CRISPR loci in sandy soils is driven by the effects of soil texture on diffusive transport and dry-wet dynamics, which promote boom and bust predator-prey dynamics in response to episodic soil wetting events, as observed in soil biocrust communities [82]. That is, we predict that phage pressure might be best predicted from community dynamics and not community composition. However, this relationship requires further study, especially since CRISPR array frequency does not indicate the total length or number of protospacers within a given genome, which may better correlate with phage exposure [83].

CONCLUSIONS

Genomic traits and EWAS represent relatively new strategies for exploring the ecological traits of microorganisms and both provided insight into relationships between the soil microbiome and properties relevant to soil health assessment. We show that community-weighted genome size was the best predictor of the total health rating, and this trait was also linked to tillage, active carbon, and other biological ratings. We observed a large number of bacterial taxa whose abundance was linked to tillage history, where tillage favored microbiomes with high community-weighted genome size and *rrn* copy number. Genome size is highly conserved across broad phylogenetic distances [84], lending support to our conclusions despite the fact community-weighted traits were inferred from reference genomes and at low phylogenetic resolution for many taxa. Future research is needed to confirm these observations using shotgun metagenomic approaches. Furthermore, efforts should be aimed at determining whether the bacterial bioindicators of soil health merely report on existing soil conditions or whether they underlie processes that regulate soil health. In particular, future research should focus on the relationship between genome size, carbon cycling, and soil health, since differences in genome size have been linked to differences in carbon use efficiency [85]. This relationship suggests that low health soils may select for bacteria that promote C loss, possibly causing a negative feedback that works against C accrual and restoration of degraded soils. Our study illustrates an approach for assessing the ecological attributes of bacteria linked to soil health, including unclassified and poorly characterized taxa. This kind of information is needed if we are to understand how microbiome composition is associated with agronomic management decisions that promote soil health.

DATA AVAILABILITY

Raw sequencing data was archived at the National Centre for Biotechnology Information (BioProject: PRJEB35975). All analyses can be reproduced with scripts included in the Supplementary Data package. The **Supplementary Data** package is available through the Open Science Foundation (https://doi.org/10.17605/OSF.IO/UJGQF).

REFERENCES

1. Doran JW. Soil health and global sustainability: translating science into practice. Agric Ecosyst Environ. 2002;88:119–27.
2. Wander MM, Cihacek LJ, Coyne M, Drijber RA, Grossman JM, Gutknecht JLM, et al. Developments in Agricultural Soil Quality and Health: Reflections by the Research Committee on Soil Organic Matter Management. Front Environ Sci. 2019;7:1–9.
3. Stewart RD, Jian J, Gyawali AJ, Thomason WE, Badgley BD, Reiter MS, et al. What we talk about when we talk about soil health. Agric Environ Lett. 2018;3:5–9.
4. Rinot O, Levy GJ, Steinberger Y, Svoray T, Eshel G. Soil health assessment: A critical review of current methodologies and a proposed new approach. Sci Total Environ. 2019;648:1484–91.
5. Hurisso TT, Culman SW, Zhao K. Repeatability and spatiotemporal variability of emerging soil health indicators relative to routine soil nutrient tests. Soil Sci Soc Am J. 2018;82:939–48.
6. Lilburne L, Sparling G, Schipper L. Soil quality monitoring in New Zealand: Development of an interpretative framework. Agric Ecosyst Environ. 2004;104:535–44.
7. Moebius-Clune BN, Moebius-Clune DJ, Gugino BK, Idowu OJ, Schindelbeck RR, Ristow AJ, et al. Comprehensive assessment of soil health - the Cornell framework manual, 3rd ed. Ithaca, NY:Cornell University; 2017.
8. Fierer N, Wood SA, Bueno de Mesquita CP. How microbes can, and cannot, be used to assess soil health. Soil Biol Biochem. 2021;153:108111.
9. Amsili JP, van Es HM, Schindelbeck RR. Cropping system and soil texture shape soil health outcomes and scoring functions. Soil Secur. 2021;4:100012.
10. Wade J, Culman SW, Gasch CK, Lazcano C, Maltais-Landry G, Margenot AJ, et al. Rigorous, empirical, and quantitative: a proposed pipeline for soil health assessments. Soil Biol Biochem. 2022;170:108710.
11. Simonin M, Voss KA, Hassett BA, Rocca JD, Wang SY, Bier RL, et al. In search of microbial indicator taxa: shifts in stream bacterial communities along an urbanization gradient. Environ Microbiol. 2019;21:3653–68.
12. Bissett A, Brown MV, Siciliano SD, Thrall PH. Microbial community responses to anthropogenically induced environmental change: Towards a systems approach. Ecol Lett. 2013;16:128–39.
13. Wilhelm RC, Cardenas E, Maas KR, Leung H, McNeil L, Berch S, et al. Biogeography and organic matter removal shape long-term effects of timber harvesting on forest soil microbial communities. ISME J. 2017;11:2552–68.
14. Gibbons SM, Scholz M, Hutchison AL, Dinner AR, Gilbert JA, Colemana ML, et al. Disturbance regimes predictably alter diversity in an ecologically complex bacterial system. MBio. 2016;7:1–10.
15. Trivedi P, Delgado-Baquerizo M, Anderson IC, Singh BK. Response of soil properties and microbial communities to agriculture: Implications for primary productivity and soil health indicators. Front Plant Sci. 2016;7:1–13.
16. Jiao S, Xu Y, Zhang J, Hao X. Core microbiota in agricultural soils and their potential associations with nutrient cycling. mSystems. 2019;4:1–16.
17. Chang HX, Haudenshield JS, Bowen CR, Allen R, Iii W, Parnell JJ, et al. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. Front Microbiol. 2017;8:519.
18. Trivedi P, Delgado-Baquerizo M, Jeffries TC, Trivedi C, Anderson IC, Lai K, et al. Soil aggregation and associated microbial communities modify the impact of agricultural management on carbon content. Environ Microbiol. 2017;19:3070–86.
19. Armbruster M, Goodall T, Hirsch PR, Ostle N, Puissant J, Fagan KC, et al. Bacterial and archaeal taxa are reliable indicators of soil restoration across distributed calcareous grasslands. Eur J Soil Sci. 2021;72:2430–44.
20. Rieke EL, Cappellazzi SB, Cope M, Liptzin D, Mac Bean G, Greub KLH, et al. Linking soil microbial community structure to potential carbon mineralization: A continental scale assessment of reduced tillage. Soil Biol Biochem. 2022;168:108618.
21. Wilhelm RC, Van Es HM, Buckley DH. Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol Biochem. 2022;164:108472.
22. Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2: An improved and customizable approach for metagenome inference 2. bioRxiv. 2020. https://doi.org/10.1101/672295.
23. Gravuer K, Eskelinen A. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. Front Microbiol. 2017;8:1–16.
24. Chen Y, Maier RM, Barberán A, Neilson JW, Kushwaha P, Maier RM, et al. Life-history strategies of soil microbial communities in an arid ecosystem. ISME J. 2021;15:649–57.
25. Fierer N. Embracing the unknown: Disentangling the complexities of the soil microbiome. Nat Rev Microbiol. 2017;15:579–90.
26. Malik AA, Martiny JBHH, Brodie EL, Martiny AC, Treseder KK, Allison SD, et al. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. ISME J. 2020;14:1–9.
27. Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat Microbiol. 2016;1:1–7.

28. Nunan N, Schmidt H, Raynaud X, Schmidt H. The ecology of heterogeneity: Soil bacterial communities and C dynamics. Philos Trans R Soc B Biol Sci. 2020;375:20190249.
29. Grime JP. Evidence for the existence of three primary strategies in plants and its relevance for ecological and evolutionary theory. Am Nat. 1977;111:1169–94.
30. Barnett S, Youngblut ND, Koechli CN, Buckley DH. Multisubstrate DNA stable isotope probing reveals guild structure of bacteria that mediate soil carbon cycling. PNAS. 2021;118:e2115292118.
31. Wilhelm RC, Pepe-Ranney C, Weisenhorn P, Lipton M, Buckley DH. Competitive exclusion and metabolic dependency among microorganisms structure the cellulose economy of an agricultural soil. MBio. 2021;12:1–19.
32. Schmidt R, Gravuer K, Bossange AV, Mitchell J, Scow K. Long-term use of cover crops and no-till shift soil microbial community life strategies in agricultural soil. PLoS ONE. 2018;13:1–19.
33. Neal AL, Hughes D, Clark IM, Jansson JK, Hirsch PR. Microbiome Aggregated Traits and Assembly Are More Sensitive to Soil Management than Diversity. mSystems 2021;6:e0105620.
34. Lupatini M, Korthals GW, de Hollander M, Janssens TKS, Kuramae EE. Soil microbiome is more heterogeneous in organic than in conventional farming system. Front Microbiol. 2017;7:1–13.
35. Koechli C, Campbell AN, Pepe-ranney C, Buckley DH. Assessing fungal contributions to cellulose degradation in soil by using high- throughput stable isotope probing. Soil Biol Biochem. 2019;130:150–8.
36. Furtak K, Grządziel J, Gałązka A, Niedźwiecki J. Prevalence of unclassified bacteria in the soil bacterial community from floodplain meadows (fluvisols) under simulated flood conditions revealed by a metataxonomic approachss. Catena. 2020;188:104448.
37. Schmidt R, Mitchell J, Scow K. Cover cropping and no-till increase diversity and symbiotroph: saprotroph ratios of soil fungal communities. Soil Biol Biochem. 2019;129:99–109.
38. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.
39. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43.
40. Levy R, Borenstein E. Reverse Ecology: From systems to environments and back. Adv Exp Med Biol. 2012;751:329–45.
41. Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, et al. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecol. 2016;20:241–8.
42. Hamilton JP, Neeno-Eckwall EC, Adhikari BN, Perna NT, Tisserat N, Leach JE, et al. The Comprehensive Phytopathogen Genomics Resource: a web-based resource for data-mining plant pathogen genomes. Database. 2011;2011:bar053.
43. Detheridge AP, Brand G, Fychan R, Crotty FV, Sanderson R, Griffith GW, et al. The legacy effect of cover crops on soil fungal populations in a cereal rotation. Agric Ecosyst Environ. 2016;228:49–61.
44. McKenna TP, Crews TE, Kemp L, Sikes BA. Community structure of soil fungi in a novel perennial crop monoculture, annual agriculture, and native prairie reconstruction. PLoS ONE. 2020;15:1–15.
45. Rocca JD, Simonin M, Blaszczak JR, Ernakovich JG, Gibbons SM, Midani FS, et al. The Microbiome Stress Project: Toward a global meta-analysis of environmental stressors and their effects on microbial communities. Front Microbiol. 2019;9:3272.
46. Ramirez KS, Knight CG, De Hollander M, Brearley FQ, Constantinides B, Cotton A, et al. Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat Microbiol. 2018;3:189–96.
47. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457–63.
48. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci Rep. 2016;6:1–9.
49. Jurburg SD, Konzack M, Eisenhauer N, Heintz-Buschart A. The archives are half-empty: a field-wide assessment of the availability of microbial community sequencing data. Commun Biol. 2020;3:474.
50. Emerson JB, Everhart SE, Eversole K, Frost KE, Herr JR, Huerta AI, et al. Community-driven metadata standards for agricultural microbiome research. Phytobiomes J. 2020; 4:115-121.
51. Anderson TH, Martens R. DNA determinations during growth of soil microbial biomasses. Soil Biol Biochem. 2013;57:487–95.
52. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the Miseq Illumina sequencing platform. Appl Environ Microbiol. 2013;79:5112–20.
53. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
54. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 2013;41:590–6.
55. Harrell F, Dupont C. Hmisc: Harrell miscellaneous. *R Package* 2015.
56. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc. 1995;57:289–300.
57. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10:1669–81.
58. Mills RH, Dulai PS, Vázquez-Baeza Y, Sauceda C, Daniel N, Gerner RR, et al. Multi-omics analyses of the ulcerative colitis gut microbiome link *Bacteroides vulgatus* proteases with disease severity. Nat Microbiol. 2022;7:262–76.
59. De Cáceres M, Legendre P, De Caceres M, Legendre P. Associations between species and groups of sites: indices and statistical inference. Ecology. 2009;90:3566–74.
60. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: A data management and analysis system for metagenomes. Nucleic Acids Res. 2008;36:534–8.
61. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015;43:593–8.
62. R Core Team. R: a language and environment for statistical computing. R Foundation. 2020.
63. Wickham H. Reshaping data with the reshape package. J Stat Soft. 2007;21:1–20.
64. Wickham H. The split-apply-combine strategy for data analysis. J Stat Soft. 2009;40:1–29.
65. Wickham H. Elegant graphics for data analysis. Media. 2009;35:211.
66. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 2013;8:e61217.
67. Grömping U. Relative importance for linear regression in R: the package relaimpo. J Stat Softw. 2006;17:1–27.
68. Bastian M, Heymann S. Gephi: an open source software for exploring and manipulating networks. Proc Int AAAI Conf Web Soc Media. 2009:361–2.
69. Hu Y. Efficient, high-quality force-directed graph drawing. Math J. 2006;10:37–71.
70. Ranea JAG, Grant A, Thornton JM, Orengo CA. Microeconomic principles explain an optimal genome size in bacteria. Trends Genet. 2005;21:21–5.
71. Nielsen DA, Fierer N, Geoghegan JL, Gillings MR, Gumerov V, Madin JS, et al. Aerobic bacteria and archaea tend to have larger and more versatile genomes. Oikos. 2021;130:501–11.
72. Chen Y, Leung PM, Wood JL, Bay SK, Kessler AJ, Shelley G, et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. ISME J. 2021;15:2986–3004.
73. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus *Udaeobacter copiosus*'. Nat Microbiol. 2016;2:16198.
74. Willms IM, Rudolph AY, Göschel I, Bolz SH, Schneider D, Penone C, et al. Globally Abundant "Candidatus *Udaeobacter*" Benefits from Release of Antibiotics in Soil and Potentially Performs Trace Gas Scavenging. mSphere. 2020;5:1–17.
75. Kaboré OD, Godreuil S, Drancourt M. *Planctomycetes* as host-associated bacteria: a perspective that holds promise for their future isolations, by mimicking their native environmental niches in clinical microbiology laboratories. Front Cell Infect Microbiol. 2020;10:1–19.
76. Martens-Habbena W, Berube PM, Urakawa H, De La Torre JR, Stahl DA. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. Nature. 2009;461:976–9.
77. Zhalnina K, De Quadros PD, Gano KA, Davis-Richardson A, Fagen JR, Brown CT, et al. Ca. *Nitrososphaera* and *Bradyrhizobium* are inversely correlated and related to agricultural practices in long-term field experiments. Front Microbiol. 2013;4:1–13.
78. Land M, Hauser L, Jun S, Nookaew I, Leuze MR, Ahn T, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genom. 2015;15:141–61.
79. Gil R, Latorre A, Postal A. Factors behind junk DNA in bacteria. Genes (Basel). 2012;3:634–50.
80. Williamson KE, Radosevich M, Wommack KE. Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol. 2005;71:3119–25.
81. Williamson KE, Corzo KA, Drissi CL, Buckingham JM, Thompson CP, Helton RR. Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biol Fertil Soils. 2013;49:857–69.
82. Van Goethem MW, Swenson TL, Trubl G, Roux S, Northen TR. Characteristics of wetting-induced bacteriophage blooms in biological soil crust. MBio. 2019;10:e02287-19.
83. Westra ER, Van Houte S, Gandon S, Whitaker R, Van Houte S, Gandon S, et al. The ecology and evolution of microbial CRISPR-Cas adaptive immune systems. Philos Trans R Soc B Biol Sci. 2019;374:20190101.
84. Martinez-Gutierrez CA, Aylward FO. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. PLoS Genet. 2022;18:1–17.

85. Saifuddin M, Bhatnagar JM, Finzi AC, Segrè D, Finzi AC. Microbial carbon use efficiency predicted from genome-scale metabolic models. Nat Commun. 2019;10:1–10.

ACKNOWLEDGEMENTS

We thank Christopher DeRito at Cornell University for his assistance in preparing amplicon sequencing libraries. This work was supported by the USDA National Institute of Food and Agriculture, Hatch project 1010520; by the U.S. Department of Energy, Office of Biological & Environmental Research Genomic Science Program under award number DE-SC0016364; and by the USDA Natural Resource Conservation Service Conservation Innovation Grant, agreement number: 69-3A75-16-039. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the National Institute of Food and Agriculture (NIFA), the United States Department of Agriculture (USDA), or the United States Department of Energy.

AUTHOR CONTRIBUTIONS

RCW performed all data analysis, research, and writing. JPA, KSMK, and HMV managed sample collection and soil health testing. DHB guided all research efforts, including analyses and writing.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-022-00209-1.

Correspondence and requests for materials should be addressed to Roland C. Wilhelm.

Reprints and permission information is available at http://www.nature.com/reprints

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