Chemical-biogeographic survey of secondary metabolism in soil

Zachary Charlop-Powers, Jeremy G. Owen, Boojala Vijay B. Reddy, Melinda A. Ternei, and Sean F. Brady[1]

Laboratory of Genetically Encoded Small Molecules, Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved December 24, 2013 (received for review September 26, 2013)

In this study, we compare biosynthetic gene richness and diversity of 96 soil microbiomes from diverse environments found throughout the southwestern and northeastern regions of the United States. The 454-pyroseqencing of nonribosomal peptide adenylation (AD) and polyketide ketosynthase (KS) domain fragments amplified from these microbiomes provide a means to evaluate the variation of secondary metabolite biosynthetic diversity in different soil environments. Through soil composition and AD- and KS-amplicon richness analysis, we identify soil types with elevated biosynthetic potential. In general, arid soils show the richest observed biosynthetic diversity, whereas brackish sediments and pine forest soils show the least. By mapping individual environmental amplicon sequences to sequences derived from functionally characterized biosynthetic gene clusters, we identified conserved soil type–specific secondary metabolome enrichment patterns despite significant sample-to-sample sequence variation. These data are used to create chemical biogeographic distribution maps for biomedically valuable families of natural products in the environment that should prove useful for directing the discovery of bioactive natural products in the future.

metagenomics | antibiotics | bioprospecting | eDNA

Molecular phylogenetic analyses suggest that soils can contain thousands of unique bacterial species per gram (1, 2), yet only a small fraction of these bacteria has been cultured and studied for their ability to produce bioactive small molecules. Furthermore, this cultured minority of soil bacteria are collectively believed to contain a large number of silent biosynthetic gene clusters that have never been examined for their ability to produce bioactive secondary metabolites (3, 4). Based on these observations and the historical success of bacterial natural products in demonstrating clinically and industrially important bioactivities (5, 6), environmental bacteria are likely to be a rich reservoir of as yet uncharacterized biologically active small molecules (7). The extension of molecular phylogenetic-type analyses to biosynthesis gene diversity in the environment should provide a better understanding of the richness of this hidden biosynthetic reservoir and also help to guide the discovery of additional novel bioactive natural products in the future (8, 9).

Nonribosomal peptides (NRPs) and polyketides (PKs) are two of the largest families of bioactive microbial metabolites, accounting for most of the antibiotic, antifungal, anticancer, and immunosuppressant compounds that have been characterized from cultured bacteria to date (5). Although NRPS and PKS biosynthesis is responsible for producing all biomedically relevant natural products, it leads to many of the metabolites used in the clinic including penicillin, vancomycin, rapamycin, erythromycin, rifamycin, and many others. NRP and PK biosynthesis shares a core biosynthetic logic (10, 11). In both cases, molecules are synthesized by large modular megasynth(et)ase enzymes in an assembly line fashion in which individual modules are responsible for the incorporation of either one acyl-CoA or amino acid building block into the growing metabolite. A minimal NRP module consists of an adenylation (AD) domain for selecting incoming amino acids, a condensation (C) domain for condensing an incoming building block with the peptidyl intermediate from the previous module, and a peptidepeptidyl carrier protein (PCP) domain for carrying the growing polypeptide. Similarly, a minimal PK module consists of an acyltransferase (AT) domain for selecting incoming acyl-CoAs, a ketosynthase (KS) domain for condensing the new building block with the acyl intermediate from the previous module, and an acyl carrier protein (ACP) domain for carrying the growing polyketide (Fig. 1*A*). The repetitive use of conserved domains in these common biosynthetic systems provides an entry point for a large-scale molecular phylogenetic-type analysis of secondary metabolism in the environment using domain-specific, degenerate primer-based PCR and next-generation sequencing (7, 9, 12).

In this study, we surveyed NRP and PK richness and diversity in 96 distinct soil microbiomes through 454 pyrosequencing of AD and KS domain fragments amplified from DNA extracted directly from these soils [environmental DNA (eDNA)]. AD and KS amplicon diversity was then used to compare and contrast the biosynthetic potential of geographically distinct soils with a variety of soil characteristics. By coupling comparative soil composition analyses with AD and KS amplicon diversity measurements, we were able to identify soil types (i.e., soils with similar physiochemical characteristics) with increased biosynthetic potential. AD and KS amplicon data were also used to determine the geographic distribution of gene clusters predicted to encode for metabolites that are evolutionarily related to families of natural products with known bioactivities. In doing so, we observed soil type–specific secondary metabolite gene cluster enrichment patterns that suggest the presence of functionally similar metasecondary metabolomes in similar soil types. Soil type also correlates with species diversity trends observed in these microbiomes. Although the functional consequences of these soil type gene cluster enrichment patterns is not yet clear, it suggests that secondary metabolomes may play a conserved role in the ecology

Significance

A comparative analysis of conserved domain fragments amplified from nonribosomal peptide and polyketide-type gene clusters present in diverse soil microbiomes revealed a link between soil type, species composition, and biosynthetic richness, as well as an unexpected conservation of the metasecondary metabolome present within microbiomes from similar soil types. Although the functional consequences of these enrichment patterns is not yet clear, it suggests that similar soil types contain functionally related collections of secondary metabolites that likely play conserved roles in the ecology of these soils. This work provides a model for the large-scale extension of molecular phylogenetic-type analyses directly to the study, characterization, and comparison of secondary metabolite biosynthetic gene clusters that remain hidden in diverse soil microbiomes.

Author contributions: Z.C.-P., J.G.O., and S.F.B. designed research; Z.C.-P., J.G.O., B.V.B.R., and M.A.T. performed research; Z.C.-P. and B.V.B.R. analyzed data; and Z.C.-P. and S.F.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 3655.

[1] To whom correspondence should be addressed. E-mail: sbrady@rockefeller.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318021111/-/DCSupplemental.

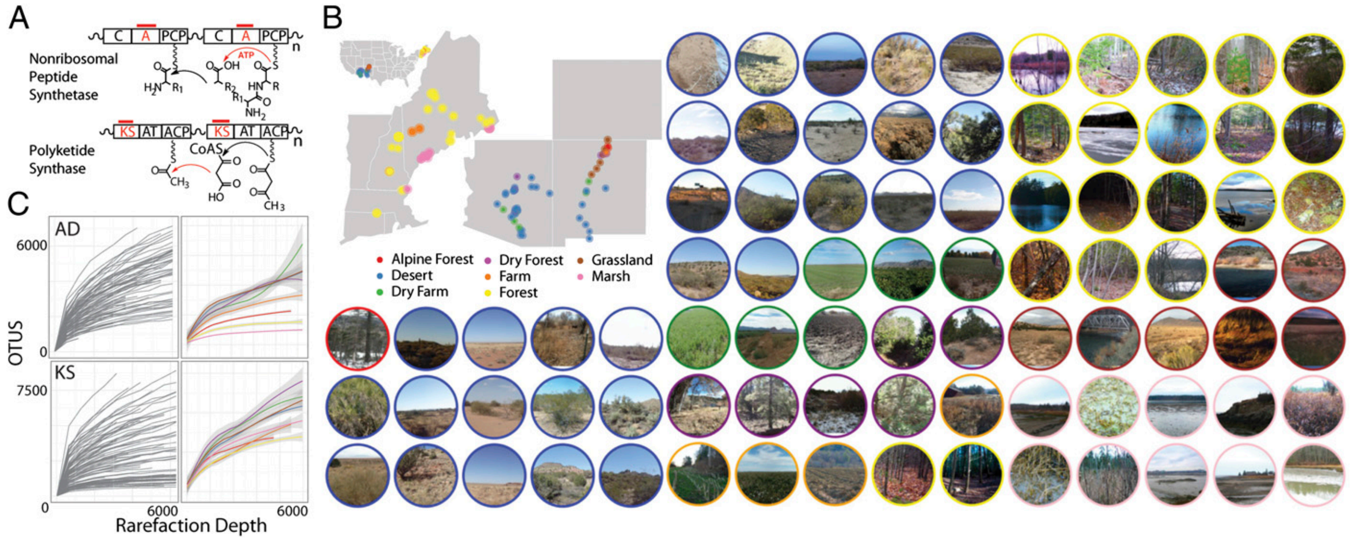

Fig. 1. Profiling of soil AD and KS domain abundances. (*A*) Nonribosomal peptides and polyketides are synthesized by gene clusters sharing a common assembly line–based logic that provides targets for degenerate primer based PCR analysis of biosynthetic diversity. (*B*) For this study, DNA was isolated directly from soils collected in the southwest and northeast regions of the United States. (*C*) Chao1 rarefaction curves generated from pyrosequenced AD and KS domains amplified from these samples can be used to estimate biosynthetic diversity (*Left*) and display groupwise average richness in different soil types (*Right*).

of geographically distant soils that share similar soil type characteristics. The chemical-biogeographic data afforded by biosynthesis gene-driven molecular phylogenetic-type analyses should help to better illuminate the hidden metasecondary metabolomes within diverse soil microbiomes, as well as to direct the future discovery of additional novel bioactive natural products.

Results and Discussion

AD and KS Operational Taxonomic Units and Rarefaction Analyses from Diverse Soil Types. Topsoil was collected from 96 sample locations throughout the southwestern and northeastern United States (Fig. 1*B*). The states of Arizona and New Mexico, from which the southwestern soils were collected, were chosen as collection sites because they are considered to be two of the most biologically diverse regions in the continental United States (13). New England was selected as the second collection site because it is geographically distant and ecologically distinct from the southwestern collection sites. Specific collection sites from within these areas were chosen to optimize variations in altitude, rainfall, dominant flora, soil phenotype, and land use (Dataset S1). Each soil sample was assigned one of eight general descriptors based on our visual assessment of the immediate area surrounding the site from which topsoil was harvested [e.g., desert, arid forest, arid farm, alpine (high-altitude) forest, pine forest, farmland, grassland or salt-water marsh; Fig. 1*B*].

DNA extracted directly from each soil sample (eDNA) was used as template in PCR reactions with AD and KS domain-specific degenerate primers. In preliminary small-scale studies, these degenerate primers were shown to be capable of amplifying diverse collections of AD and KS domains from eDNA (14, 15). The resulting PCR amplicons were 454 pyrosequenced, and cleaned reads were clustered at a genetic distance of 5% to compensate for potential sequencing and PCR errors (Table S1, see *Materials and Methods* for read processing details). Each unique 95% identity cluster was considered to be an AD or KS operational taxonomic unit (OTU) representing a unique natural product biosynthesis sequence tag. Rarefaction curves were generated that display the average Chao1 diversity metric for repeatedly subsampled AD and KS OTU data (Fig. 1*C* and Fig. S1). A potential drawback of using the degenerate primers is the introduction of primer-dependent bias leading to amplicon pools that are skewed relative to the underlying sequence diversity. For the comparative analyses describe here, primer bias should not affect the general conclusions as intersample comparisons are equally biased and global diversity estimates will err on the conservative side. For both KS and AD domains, Chao1 diversity estimates range from less than 1,000 to greater than 7,000 OTUs per soil microbiome. Soils with the highest KS and AD domain richness estimates arise from a subset of southwestern arid environments, whereas brackish water and New England forest environments show the lowest richness estimates (Fig. 2*B*). Surprisingly, no significant differences in biosynthetic richness estimates were seen between cultivated and uncultivated soils harvested from neighboring locations.

Soil Principal Component Analysis. In an attempt to better understand the relationship between soil type and biosynthesis domain richness, we sought to classify each soil microenvironment through a quantitative analysis of basic soil characteristics (e.g., pH, moisture, soil granularity, organic matter, mineral composition; Dataset S1). The aggregated data from this analysis was submitted to a principal component analysis (PCA; Fig. 2*B*), which clustered soil samples into two larger groups (groups A and B) and one smaller group (group C). The two larger groups (A and B) separate along the first principal component axis, largely according to geographic location, with group A representing southwestern soils and group B representing a subset of northeastern soils drawn mostly from forest soils. The third group (group C) is a subset of northeastern samples that were all collected from brackish coastal environments (Fig. 2*B*). Despite major differences in flora and visual soil appearance (Fig. 1*B*), group A soils share a number of key characteristics that result in the observed PCA grouping (Figs. 1*B* and 2*B* and Fig. S2). They are low in moisture and organic content and have elevated calcium, copper, and phosphorus levels. Meanwhile, group B soils are more acidic ($pH < 5$) and organic rich and show elevated concentrations of an orthogonal set of minerals (e.g., iron and aluminum). Group C soils are distinguished by comparatively high concentrations of sodium, sulfur, boron, and vanadium.

Soil PCA-Based Richness Analysis. The number of cleaned 454 sequencing reads varied from sample to sample (Table S1). To permit the most robust direct comparisons between observed domain diversity and PCA soil groupings, we subsampled the

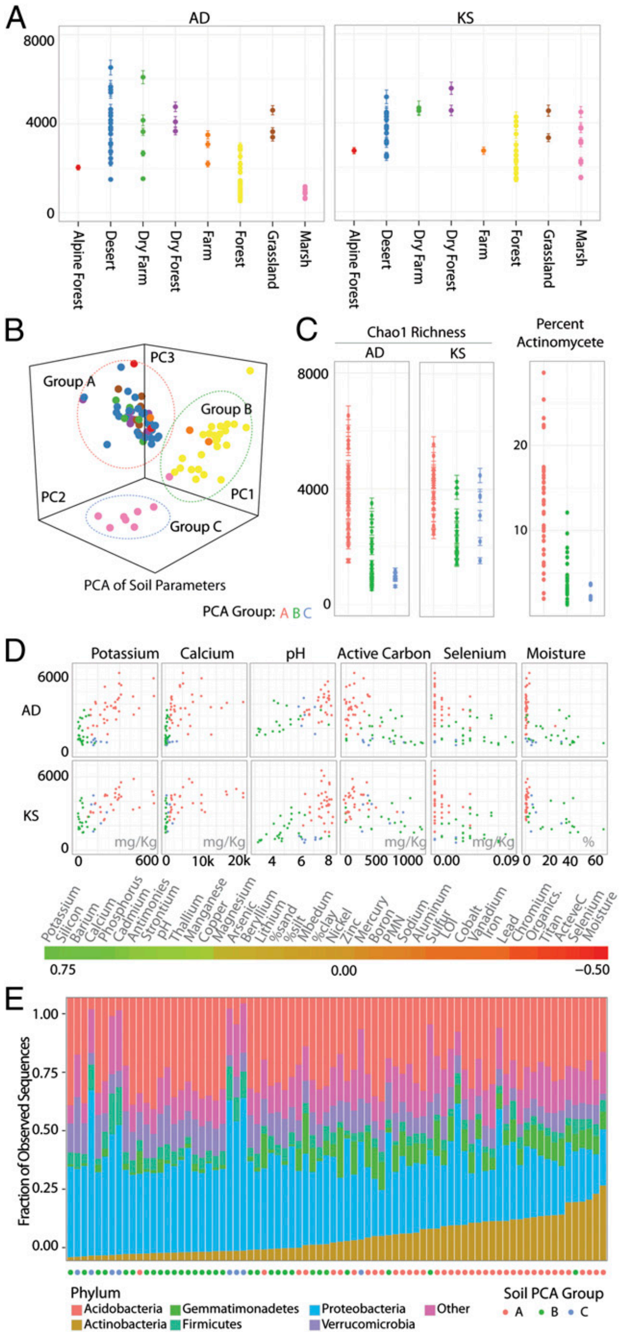

Fig. 2. Comparative richness analysis. (*A*) Chao1 richness estimates of evenly rarefied samples plotted by soil type. (*B*) PCA grouping of soils based on quantitative soil property data. (*C*) Regrouping of Chao1 AD and KS results by PCA groupings (*Left*) and percent of unique 16S OTUs that were annotated as *Actinomycete* (*Right*). (*D*) Pearson correlations of Chao1 richness with soil data (*Lower*; full data available in Table S2) and the linear correlation of a subset of those factors with observed richness (*Upper*). (*E*) 16S-based phylum composition data for each soil sample.

sequencing data obtained from each soil to the same depth. KS and AD domain Chao1 diversity estimates were generated for the subset of 65 soil samples that yielded sufficient sequencing data to permit comparisons of 3,500 cleaned reads for each of these domains. Although domain diversity estimates differ within each PCA soil group, on average group A soils (arid environments) show the highest AD richness estimates and group C soils (brackish environments) show the lowest richness estimates (Fig. 2*C*). Group A soils also tend to show higher KS domain richness estimates than either groups B or C, which are essentially indistinguishable from each other in predicted KS richness.

Differences in the composition of bacterial species found in arid soils compared with that found in forest and brackish soils is likely to partially account for the differences in AD and KS domain richness observed in our analysis. Sequencing data from cultured bacterial genomes suggest that NRP and KS biosynthesis is disproportionally found in a subset of bacterial phyla (e.g., Actinomycetes, γ-Proteobacteria, Cyanobacteria) (16). A population bias toward these species would result in greater observed AD and KS amplicon richness. Many of the low diversity samples in our analysis are derived from the acidic pine forests of Maine (Fig. 2*D*). Low soil pH has not only been correlated with low species diversity but has also been shown to skew soil species composition toward the Acidobacterial phylum and therefore away from more NRP/PK-rich phyla (17). Soil type species composition differences are further highlighted by the fact that NRP and KS rich *Actinomycetes* have been observed to comprise up to 40% of the observed bacterial population in desert soils but as little as 4% of the total bacterial population in forest environments (18).

To assess whether AD and KS richness correlates with *Actinomycetes* species diversity in our samples, 16S gene–based taxonomic information was obtained for 80 of our eDNA samples. These data were used to calculate the relative abundance of major bacterial phyla in each sample. Group A samples do in fact show a higher relative abundance of *Actinomycetes* (Fig. 2*C*) compared with either group B or C samples (Fig. 2*E* and Fig. S3), implicating differences in *Actinomycetes* richness as a driving force behind observed difference in AD- and KS- richness.

Correlations Between Soil Parameters and Biosynthetic Richness. Pearson coefficients were calculated between the KS and AD domain richness of each soil sample and each of the measured soil parameters to look for factors that might show simple linear correlations with observed domain richness (Fig. 2*D*, *Lower*). A number of the soil parameters show linear correlations with observed KS and AD domain richness as measured in this Pearson analysis, including positive linear correlations between biosynthetic domain sequence richness and potassium, calcium, and pH (Fig. 2*D*). The strongest negative linear correlations are seen between domain richness and moisture content, active carbon content, and selenium (Fig. 2*D*).

Global Comparative AD and KS Population Analysis. The similarity between the biosynthetic sequence profile of individual soils was determined using the Jaccard distance (19, 20). The Jaccard distance is determined by pooling the OTUs from two samples and representing their relatedness as the ratio of shared to total OTUs subtracted from 1 (20). A Jaccard distance of 1 therefore represents completely nonoverlapping OTU populations, whereas a distance of 0 indicates that samples have identical OTU populations. Two independent control sequencing experiments that were expected to yield similar KS and AD domain populations were included in our Jaccard analysis. In one set of sequencing controls (resequencing controls: C_1–C_5), the same AD and KS PCR amplicons were sequenced twice. In the second set of sequencing controls (proximity controls: P_{AD}, P_{KS}), eight forest soil samples collected ~10 m from one another were processed and sequenced independently.

When the Jaccard comparisons of AD and KS amplicon pools from different soils are plotted as a network diagram with a Jaccard distance threshold of 0.99 (at least 1% shared OTUs), we see clustering patterns that broadly correspond to the predicted PCA soil groupings (Fig. 3*A*). Largely, KS and AD domain populations from group A (arid) soils cluster together, group B (forest) soils cluster together, and group C (brackish) soils cluster together. When the Jaccard threshold is lowered to 0.85 (at least 15% shared OTUs), the few remaining clusters are dominated by our control studies [e.g., resequencing (c) and

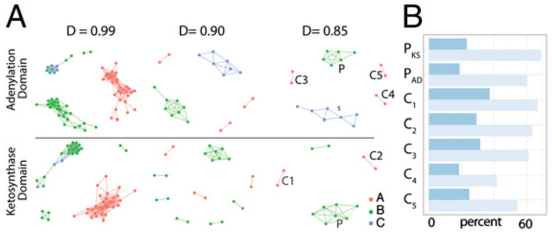

Fig. 3. Beta diversity. (*A*) Network diagram of samples (nodes) that are linked if they are within the specified Jaccard distance. (*B*) Shared OTUs of control samples [C1–C5, duplicate samples; P_{AD}, P_{KS}, proximity controls (pairwise average)] plotted as a percentage of shared OTUs between samples (dark blue) and also as a percentage of shared reads (light blue).

proximity controls (p); Fig. 3*A*]. The only noncontrol relationship cluster that remains at this threshold is composed of AD sequences derived from brackish, high-salt environments (s), suggesting these samples are not only low in observed biosynthetic diversity (Fig. 2*B*) but are also more similar to one another in biosynthesis gene content than soils from within either group A or group B.

Although all soils in our study appear to be quite distinct from one another in specific KS and AD sequence content (with the exception of the proximity controls), soils from within the same soil type PCA group show higher relatedness to each other than they do to soils from other PCA groups. Even among the most closely related control samples, as many as 90% of the OTUs differed. However, at the depth we sequenced, we cannot say for certain that unique OTUs in one sample would not be found in another sample as rare OTUs with additional sequencing. Our soil proximity control study suggests that we have sequenced deeply enough to identify soils with highly similar populations of AD and KS sequences.

To better understand the sequence overlap in related control samples, the percent of sequences common to each control was determined as either a fraction of common OTUs shared by the samples or as the fraction of total reads these common OTUs account for (Fig. 3*B*). Although control samples may only have 20–30% of OTUs in common, these OTUs account for as much as 70% of the reads in these samples (i.e., ∼70% identical sequencing data), thus indicating datasets with significant sequence overlap were generated in both control experiments. The control sequencing experiments therefore suggest that, although we are likely only sampling a subset of the biosynthetic sequence diversity present in any specific soil microbiome, at this sequencing depth (∼70–80% of observed OTUs are not shared in analysis of identical samples), we have obtained sufficient sequence coverage to identify closely related samples.

Predicting the Distribution of Gene Clusters That Are Evolutionarily Closely Related to Known Clusters. A common theme in natural products chemistry is the existence of families of molecules whose related structures derive from evolutionarily related gene clusters. We recently showed that the individual AD and KS natural product sequence tags exhibiting the highest identity to AD and KS domain fragments found in functionally characterized (known) gene clusters can serve as markers for the presence in the environment of novel gene clusters that are functionally related to these known clusters (9). Although functional relationships predicted from sequence tags alone will not be perfect, the systematic examination of such relationships found in amplicon data sets derived from diverse soils should, for the first time, permit at least the low-resolution metafunctional comparison of secondary metabolomes across distinct microbiomes. With this in mind, we probed all amplicon datasets using the bionformatics Environmental Surveyor of Natural Product Diversity (eSNaPD) tool (9). eSNaPD uses a BLAST-based algorithm to identify eDNA-derived natural product sequences that are related to functionally characterized natural product biosynthetic gene clusters (9). A hit from the eSNaPD algorithm is indicative of there being an AD or KS domain sequence in a microbiome that is most closely related to the corresponding domain sequence from a known gene cluster and is taken to be an indication that this particular microbiome has a high likelihood of containing a gene cluster that encodes a metabolite that is both structurally and functionally related to that encoded by the known gene cluster (i.e., a member of the same natural product family).

From the eSNaPD data, it is possible to generate chemical-biogeographic maps representing the predicted occurrence and frequency of natural product gene cluster families in the environment, which should be useful for guiding future natural product discovery efforts. Fig. 3*B* shows the distribution of amplicons that were assigned to three biomedically important natural product families: glycopeptide antibiotics, lipopeptide antibiotics, and rapamycin-like immune regulators. At a lower limit expectation value (e-value) threshold of e^{-45}, ∼10% of soil amplicons mapped to a KS or AD domain from a gene cluster in the current eSNaPD database (Dataset S2). Each soil was found to vary in both the composition and abundance of eSNaPD hits. Taken together, amplicon data from all soils hit to 226 of the gene clusters in eSNaPD (Fig. 4*A* and Fig. S4). As might be expected from the rarefaction analysis (Fig. 2*C*), on average, amplicon pools from group A soils map to a larger number of known gene cluster families than amplicons from either group B or C soils (Fig. 4*A*). Among the soils we examined, arid soils are therefore predicted to be the most productive sources of gene clusters capable of encoding novel members of known biomedically relevant families of natural products.

Global Correlations Between eSNaPD Predictions and Soil Type. Although each soil we examined produced a unique eSNaPD output (Fig. 4*A*, *Lower*), a close examination of these data suggested the presence of several soil type–specific patterns. We therefore looked for possible correlations between the enrichment of specific gene cluster families within each microbiome and the physiochemical properties of the corresponding soils. A pairwise Pearson correlation analysis was performed between each quantitative soil parameter and the eSNaPD hit counts to individual known gene clusters. A heat map displaying this Pearson correlation analysis shows unexpectedly clear enrichment trends between the eSNaPD-predicted occurrence of specific natural product gene clusters and subsets of soil parameters (Fig. 5*A*). These correlations are especially apparent when soil parameters are arranged to mimic the principle components that most influence the PCA soil groupings (Fig. 5*A*). Chemical-biogeographic maps of the eSNaPD hit frequency for sequence tags that recognize individual known gene clusters across all soils examined further illustrate these trends (Fig. 5*B*).

As can be seen in the individual eSNaPD hit frequency maps (Fig. 5*B*), gene cluster families are not predicted to be exclusively found in a single soil type; however, they appear to be in soil type–specific gene cluster enrichment patterns. Although the specific natural product structures encoded by each soil microbiome within a PCA group are undoubtedly different (as exemplified by the large differences in specific domain sequences we observed), our analysis suggests that microbiomes from similar soil types may be enriched for the biosynthesis of functionally related collections of secondary metabolites (as exemplified by similar eSNaPD enrichment patterns). To the best of our knowledge, this study shows a previously unidentified potential microbiome-wide correlation between the metasecondary metabolome (i.e., the collective group of metabolites encoded by a microbiome) of a soil microbiome and the physiochemical characteristics of the soil environment in which the microbiome resides. The metasecondary metabolome of an environmental sample is encoded by the collection of organisms within a microbiome, and therefore, it is not surprising to find that differences in soil species composition also appear to correlate with the physical

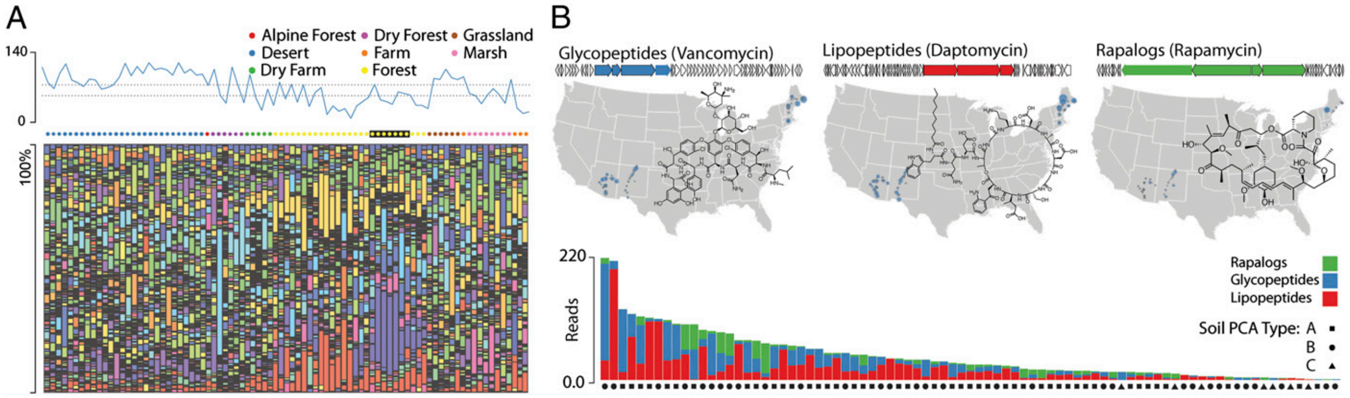

Fig. 4. eSNaPD analysis. (*A*) All hits from the eSNaPD homology search algorithm are aggregated by sample and plotted as the number of eSNaPD domains identified in that sample (*Upper*) or as a normalized bar chart. The stacked bar chart contains 226 bars representing the 226 well-studied natural product gene clusters to which metagenomic AD and KS amplicon sequences are mapped using the eSNaPD analysis (full legend and data are provided in Fig. S4 and Dataset S2). Proximity controls from adjacent areas are highlighted in black. (*B*) eSNaPD can be used to identify and target samples containing gene clusters functionally related to know clusters. eSNaPD-derived chemical-biogeographic maps and raw frequency data for glycopeptide antibiotics (i.e., the antibiotics of last resort, vancomycin, teicoplain, etc.), the lipopeptide antibiotics (i.e., the most recent class of antibiotics to be approved for clinical use, daptomycin, friulimicin, etc.), and the rapamycin-like immune regulators (rapamycin, FK228, etc.).

parameters of soils. Clustering of soil samples by 16S composition (Fig. S5) results in groupings similar to those generated using soil composition data. As natural products are known to play a role in many basic biological processes including inter- and intraspecies signaling, defense and nutrient uptake, predicted soil-specific molecule enrichment patterns likely parallel conserved global ecological aspects of different soil microbiomes.

Conclusion. The use of amplicon sequencing to profile soil microbiomes is an attractive approach for assessing the biosynthetic potential of diverse microbiomes to a depth that is still inaccessible via shotgun sequencing. Previously, investigation of natural product biodiversity using amplicon sequencing had only been applied to a small number of marine and terrestrial environments (21–25). Our study of nearly 100 unique soil samples builds on these earlier small-scale studies to provide a more detailed picture of the distribution of secondary metabolite gene clusters throughout soil microbiomes. In our analysis, we observed a correlation between soil type and both the biosynthetic richness and the predicted secondary metabolomes encoded by soil microbiomes. In general, arid soils show the richest observed biosynthetic diversity, whereas brackish sediments and pine forest soils show the least biosynthetic richness. Mapping individual sequence reads to related sequences from known biosynthetic gene clusters demonstrates soil type–specific secondary metabolome enrichment patterns despite significant sample-to-sample sequence variation. Although the functional consequences of the observed enrichment patterns is not yet clear, it suggests that similar soil types contain functionally related collections of secondary metabolites that likely play conserved roles in the ecology of these soils.

Materials and Methods

Soil Collection and Characterization. Topsoil was collected from 96 sites in the United States: 33 from Arizona, 21 from New Mexico, and 42 from New England (Fig. 1*B*). All samples were collected during the summer and were analyzed for moisture, pH, granularity, and the mineral and organic content

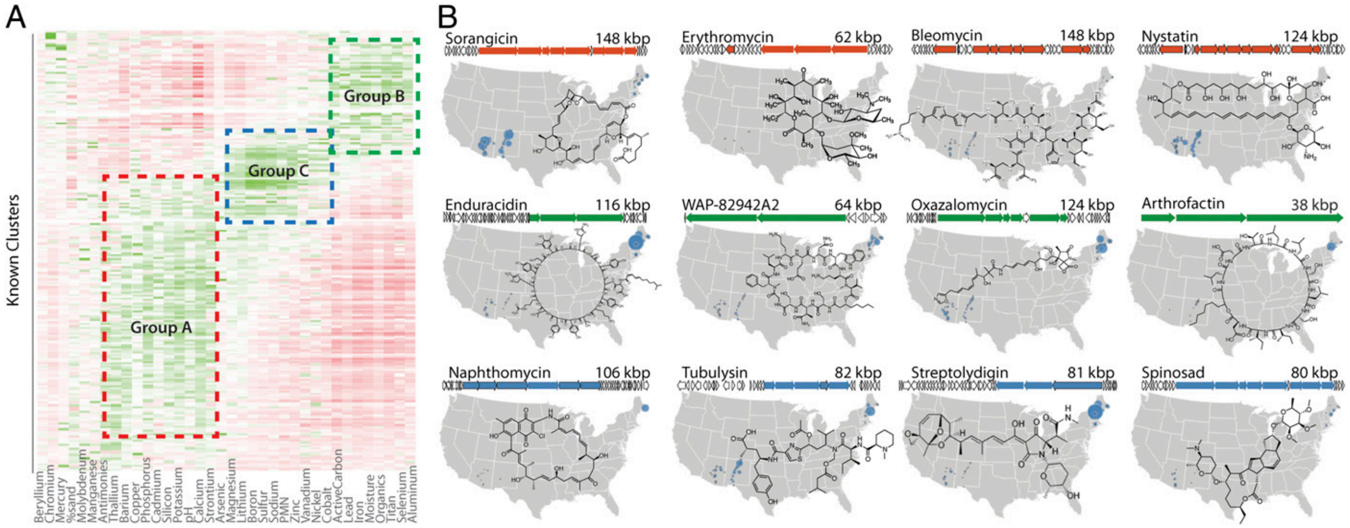

Fig. 5. Regional variations in molecule abundance. (*A*) Pearson coefficient calculations for each soil parameter to eSNaPD abundance count were calculated and plotted as a heat map. eSNaPD molecules are on the vertical axis. Row and columns were arranged to highlight the soil parameters that best define the PCA soil froups A (red), B (green), and C (blue). (*B*) Chemical-biogeographic eSNaPD hit distributions maps for four group A (red), group B (green), and group C (blue) enriched gene clusters.

of the soils by the Soil Health Laboratory at Cornell University (Dataset S1). Numerical data were submitted to PCA within R after first scaling each column by subtracting the column mean and normalizing to a variance of one. The princomp command was used to calculate the principal components, and the first two or three principal components for each sample were used as coordinates to create 2D (Fig. S1) and 3D (Fig. 2*B*) plots using the ggplot2 and rgl packages (26).

Processing 454 Data: AD and KS. Raw reads were assigned to samples using the unique primer barcodes and filtered by size (350–500 bp) and quality (50-bp rolling window PHRED cutoff of 20) using the Qiime pipeline (version 1.6) (27). Insertion/deletion errors due to pyrosequencing were addressed using 200 rounds of denoising with Denoiser (28). Chimeric sequences were removed using the de novo chimera detection tool of USEARCH with the default 1.9 skew value (29). Quality filtering, denoising, and chimera detection reduce the original dataset of ∼2.5 million reads to a final cleaned dataset of ∼1.25 million reads: 629,929 for AD and 558,503 for KS.

Clustering, Rarefaction, and Diversity Analysis: AD and KS. USEARCH (29) was used to cluster cleaned AD and KS datasets using 95% percent identity threshold. Global α diversity curves (Fig. 1*C*) were generated by repeatedly (10×) subsampling each dataset at evenly spaced intervals up to 5,000 reads using the α diversity utility within Qiime (27). The average Chao1 diversity estimates for each sample were plotted as both individual samples and aggregated by soil type (Fig. 1*C* and Fig. S2). The variability in sequencing depth across samples in both AD and KS domain sequencing prohibits straight richness comparison among samples. To allow for intersample comparisons, a subsampling depth of 3,500 reads was chosen for both AD and KS sequences that allowed us to include 65 soil samples. These rarefied OTU tables were imported into the Phyloseq program (19), which was used to calculate and compare species richness (Fig. 2) and to calculate and plot OTU overlap between samples (Fig. 3) using the Jaccard distance metric [1 − ($OTU_{A\&B}$)/(OTU_A + OTU_B)] (20). Intersamples distance of controls were calculated using the OTU tables by expressing shared OTUs as either a fraction of total OTUs or by summing the reads found in shared OTUs and expressing them as fraction of the total per-sample reads (3,500 reads; Fig. 3*B*). For control samples with greater than one sample (P_{AD}, P_{KS}), the average for all pairwise distances was calculated.

Assignment of AD and KS Domains to Known Gene Clusters. AD and KS amplicon reads were assessed for their relationships to known biosynthetic gene clusters using the eSNaPD algorithm at an e-value of 10^{-45} (9). The eSNaPD algorithm is a two-step BLAST-based process that queries a database of known domains and then uses a negative selection step to weed out low-quality reads. It is a semiempirical program that has been used to successfully assign and recover gene cluster homologs of known natural products using only sequence from a single domain amplicon. NRPS/PKS clusters typically have multiple KS or AD domains; hits against any of the domains in a cluster were aggregated together. The eSNaPD OTU table was calculated, and the number of eSNaPD hits per sample was calculated (Fig. 4*A*, *Upper*). The full dataset was normalized on a per-sample basis and displayed as a stacked bar chart where each bar is the fractional representation of individual eSNaPD hits (Fig. 4*A*).

Soil Richness and Soil Molecule Abundance Correlations. To assess possible correlations between individual soil parameters and AD and KS diversity, the Chao1 richness estimates for AD and KS for each subsampled soil were summed to create a single per-sample richness metric. The Pearson correlation of this richness metric with each physical-chemical soil property was calculated across all samples, and the correlation is displayed (Fig. 2*D*, *Lower*) with select parameters plotted (Fig. 2*D*). Pearson correlations between raw eSNaPD hit counts and physico-chemical data were similarly calculated. The resulting coefficient matrix is plotted as a heat map [NeatMap (30)] where rows/columns are positioned adjacent to similar rows/columns. Chemical-biogeographic maps (Figs. 4*B* and 5*B*) of a subset of molecules were generated by plotting circles scaled to match the eSNaPD abundance counts.

ACKNOWLEDGMENTS. This work was supported by National Institutes of Health Grant GM077516. S.F.B. is a Howard Hughes Medical Institute Early Career Scientist.

1. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394.
2. Torsvik V, Goksøyr J, Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56(3):782–787.
3. Bentley SD, et al. (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 417(6885):141–147.
4. Ikeda H, et al. (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* 21(5):526–531.
5. Cragg GM, Newman DJ (2013) Natural products: A continuing source of novel drug leads. *Biochim Biophys Acta* 1830(6):3670–3695.
6. Newman DJ, Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 75(3):311–335.
7. Reddy BV, et al. (2012) Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* 78(10):3744–3752.
8. Banik JJ, Brady SF (2010) Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr Opin Microbiol* 13(5):603–609.
9. Owen JG, et al. (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci USA* 110(29):11797–11802.
10. Wong FT, Khosla C (2012) Combinatorial biosynthesis of polyketides—a perspective. *Curr Opin Chem Biol* 16(1-2):117–123.
11. Finking R, Marahiel MA (2004) Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol* 58:453–488.
12. Ziemert N, et al. (2012) The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* 7(3):e34064.
13. Stein BA (2002) *States of the Union: Ranking America's Biodiveristy* (NatureServe, Arlington, VA).
14. Ayuso-Sacido A, Genilloud O (2005) New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: Detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microb Ecol* 49(1):10–24.
15. Schirmer A, et al. (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge Discodermia dissoluta. *Appl Environ Microbiol* 71(8):4840–4849.
16. Donadio S, Monciardini P, Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics. *Nat Prod Rep* 24(5):1073–1109.
17. Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75(15):5111–5120.
18. Fierer N, et al. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA* 109(52):21390–21395.
19. McMurdie PJ, Holmes S (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8(4):e61217.
20. Oksanen JB, et al. (2013) *Vegan: Community Ecology Package.* Available at http://CRAN.R-project.org/package=vegan. Accessed August 1, 2013.
21. Hochmuth T, Piel J (2009) Polyketide synthases of bacterial symbionts in sponges—Evolution-based applications in natural products research. *Phytochemistry* 70(15-16):1841–1849.
22. Gontang EA, Gaudêncio SP, Fenical W, Jensen PR (2010) Sequence-based analysis of secondary-metabolite biosynthesis in marine actinobacteria. *Appl Environ Microbiol* 76(8):2487–2499.
23. Wawrik B, Kerkhof L, Zylstra GJ, Kukor JJ (2005) Identification of unique type II polyketide synthase genes in soil. *Appl Environ Microbiol* 71(5):2232–2238.
24. Wawrik B, et al. (2007) Biogeography of actinomycete communities and type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Appl Environ Microbiol* 73(9):2982–2989.
25. Woodhouse JN, Fan L, Brown MV, Thomas T, Neilan BA (2013) Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges. *ISME J* 7(9):1842–1851.
26. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York).
27. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.
28. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* 7(9):668–669.
29. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
30. Rajaram S, Oono Y (2010) NeatMap—Non-clustering heat map alternatives in R. *BMC Bioinformatics* 11:45.