Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis

Alexander Crits-Christoph, Spencer Diamond, Cristina N. Butterfield, Brian C. Thomas, & Jillian F. Banfield

Abstract

In soil ecosystems, microorganisms produce diverse secondary metabolites such as antibiotics, antifungals and siderophores that mediate communication, competition and interactions with other organisms and the environment[1,2]. Most known antibiotics are derived from a few culturable microbial taxa[3], and the biosynthetic potential of the vast majority of bacteria in soil has rarely been investigated[4]. Here we reconstruct hundreds of near-complete genomes from grassland soil metagenomes and identify microorganisms from previously understudied phyla that encode diverse polyketide and nonribosomal peptide biosynthetic gene clusters that are divergent from well-studied clusters. These biosynthetic loci are encoded by newly identified members of the Acidobacteria, Verrucomicobia and Gemmatimonadetes, and the candidate phylum Rokubacteria. Bacteria from these groups are highly abundant in soils[5,6,7], but have not previously been genomically linked to secondary metabolite production with confidence. In particular, large numbers of biosynthetic genes were characterized in newly identified members of the Acidobacteria, which is the most abundant bacterial phylum across soil biomes[5]. We identify two acidobacterial genomes from divergent lineages, each of which encodes an unusually large repertoire of biosynthetic genes with up to fifteen large polyketide and nonribosomal peptide biosynthetic loci per genome. To track gene expression of genes encoding polyketide synthases and nonribosomal peptide synthetases in the soil ecosystem that we studied, we sampled 120 time points in a microcosm manipulation experiment and, using metatranscriptomics, found that gene clusters were differentially co-expressed in response to environmental perturbations. Transcriptional co-expression networks for specific organisms associated biosynthetic genes with two-component systems, transcriptional activation, putative antimicrobial resistance and iron regulation, linking metabolite biosynthesis to processes of environmental sensing and ecological competition. We conclude that the biosynthetic potential of abundant and phylogenetically diverse soil microorganisms has previously been underestimated. These organisms may represent a source of natural products that can address needs for new antibiotics and other pharmaceutical compounds.

Main

We reconstructed draft genomes for hundreds of microorganisms from the soil ecosystem of a northern Californian grassland using genome-resolved metagenomic methods, and targeted genomes from four dominant soil

phyla for analysis of their biosynthetic potential (Extended Data Fig. 1). Specifically, we analysed newly reconstructed genomes from 149 Acidobacteria, 135 Verrucomicrobia, 43 Rokubacteria and 49 Gemmatimonadetes species (Supplementary Table 1and Supplementary Methods). We targeted these groups because bacteria from all four phyla are highly abundant at our field sampling site[8] (Fig. 1a) and in globally sampled soils[5]. Specifically, meta-analysis of many 16S rRNA gene sequence studies showed that Acidobacteria and Verrucomicrobia are the first and second most abundant bacterial phyla in soil, respectively[5], and Gemmatimonadetes are also known to be common in soils[9]. There are few reference genomes available for soil-associated bacteria from all four phyla, and their potential for secondary metabolism remains understudied. To our knowledge, the current study represents the largest genomic sampling of soil-associated bacteria from these groups to date and the most detailed analysis of their secondary metabolism.

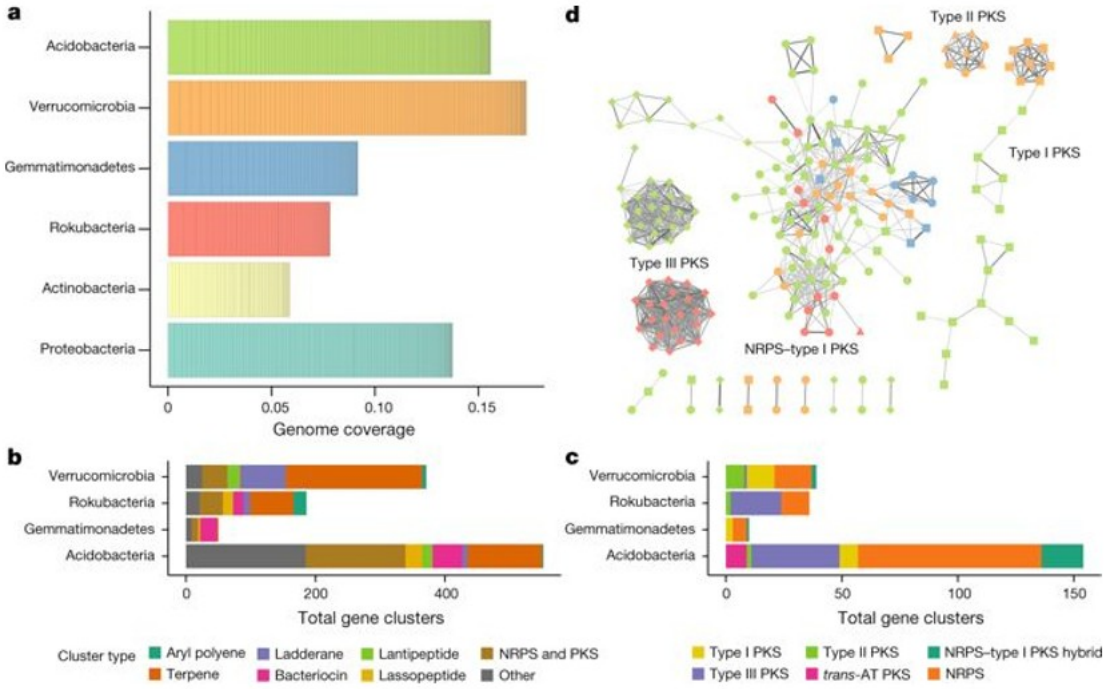

Fig. 1: Diversity of extracted soil genomes and their biosynthetic gene clusters.

a, Mean relative abundances of reconstructed genomes across 60 soil samples as determined by sequencing coverage of the genomes. Genomes from four understudied soil phyla are juxtaposed with recovered genomes from the Actinobacteria and Proteobacteria for comparison. b, Biosynthetic gene clusters found on contigs greater than 10 kb, from each phylum studied, coloured by putative product types as assigned by antiSMASH. c, NRPS and PKS gene clusters found on contigs >10 kb, from each phylum studied. d, Network of biosynthetic gene clusters, in which edges connect clusters that share genes. The line thickness and darkness increase with increasing percentage of genes shared between clusters. *trans*-AT, *trans*-acyltransferase.

Within the genomes, we identified 1,159 biosynthetic gene clusters on contigs at least 10 kb in length (Fig. 1b and Supplementary Table 2) and an additional 440 biosynthetic gene clusters on smaller contigs

(Supplementary Table 3) using antiSMASH 3.0[10], an in silico pipeline that was originally verified against 473 verified biosynthetic gene clusters with a 97.7% reported accuracy[11]. The gene clusters that we identified are inferred to synthesize nonribosomal peptides (NRPs), polyketides, terpenes, bacteriocins, lassopeptides, lantipeptides and metabolites of uncertain function. Most known bacterial natural products—including many of the clinical antibiotics that we use today—have been obtained from microbial isolates[3] of the Actinobacteria, Proteobacteria and *Bacillus*, which represent microorganisms that often comprise a minority in soil microbial communities[4,5]. Previous global analyses based on the few publicly available genomes for Acidobacteria, Verrucomicrobia and Gemmatimonadetes[12,13,14]identified only a handful of biosynthetic clusters, and to our knowledge only the Acidobacteria have previously been suggested to be linked to secondary metabolite production[7,15]. We greatly expand the number of known biosynthetic gene pathways from these soil microorganisms and at the same time confidently link them to their genomic contexts.

Most previous searches for biosynthetic systems from uncultivated microorganisms have randomly cloned environmental DNA into a host organism to screen for function (functional metagenomics)[16]. Other studies[2,17] have used degenerate PCR primers to explore the genetic diversity of novel biosynthetic clusters without the need for cloning, but primers can fail to amplify genetically divergent sequences. Because we reconstructed near-complete genomes de novo, we could identify entire novel biosynthetic gene clusters as well as describe their genomic, phylogenetic and ecological contexts within individual genomes and the environment. We computationally tested the ability of sets of previously used degenerate primers[2,17] to detect genes containing polyketide ketoacyl synthase and NRP amino acid adenylation domains in the clusters reported here, and found that only 5 out of 240 clusters would be likely to amplify properly when using degenerate primers (Supplementary Table 6).

Gene clusters containing nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) were of particular interest, as the products of these enzymes include many antibiotics, antifungals, siderophores and immunosuppressants[14]. These NRPS and PKS biosynthetic pathways use modular enzymatic domains to build molecules with complex chemical structures. We identified 240 NRPS, PKS (types I, II and III, which differ in the organization of their enzymatic domains) and hybrid (NRPS-PKS) gene clusters on contigs from all four phyla of interest (Fig. 1c and Supplementary Table 4) and 86 probably incomplete clusters on smaller genome fragments. Although they are enormously diverse in gene content, these biosynthetic pathways are identifiable owing to their colocalized logical organization of conserved enzymatic domains. Although the majority of these clusters occurred in a wide diversity of Acidobacteria, we also identified 11 NRPS clusters in genomes of the Rokubacteria, a recently

described phylum that was not previously known to produce natural products. The co-linear ‘assembly-line’ regulation of many NRPS and type I PKS systems make predictions of the core scaffold of the molecular product synthesized possible[11,18]. In 136 cases, there were a sufficient number of functional domains with known substrate specificity to predict the core chemical structures of the products using antiSMASH (Supplementary Table 4).

To compare the degrees to which predicted biosynthetic clusters shared genes, we built a relational network of clusters on the basis of shared gene content. This approach revealed substantial genetic variety, with large groups of diverse and sparsely connected NRPS and PKS systems in Verrucomicrobia, Acidobacteria and Rokubacteria and many unique NRPS-based clusters with few close representatives (Fig. 1d). A conserved type III PKS locus that was nearly ubiquitous in the Rokubacteria formed a dense network cluster, as did a conserved type III PKS locus found in a wide clade of the Acidobacteria. The high conservation of these type III PKS loci across taxonomic groups could indicate a broad distribution of a novel group of specialized metabolites.

We compared the 240 NRPS and PKS gene clusters to the reference set described in the ‘Minimum Information about a Biosynthetic Gene’ (MIBiG) repository[19] (Supplementary Table 5). No protein in any cluster shared with reference proteins more than 79.7% amino acid identity across ≥50% of the full protein lengths. Fifty-nine per cent of predicted proteins had no ≥50%-length homologue in MIBiG, and those that did shared an average of only about 39% amino acid identity to the best hit of any MIBiG protein. Using the same thresholds for gene homologues, we found that 220 clusters did not share more than 50% of the genes of any previously described cluster. Although the relationship between gene similarity of biosynthetic genes and structural similarities of their final products can be difficult to discern, previous analyses have shown that structural divergence correlates strongly with genetic divergence, even within families of gene clusters[20].

It is often the case that antibiotic producers will also encode antibiotic resistance genes to avoid self-toxicity, and that these genes will often co-localize with the antibiotic biosynthetic cluster in the genome[21]. Therefore, the presence of antimicrobial resistance genes within a gene cluster could indicate that the cluster is involved in antibiotic production. We mined all NRPS and PKS biosynthetic loci with a set[22] of curated hidden Markov models for antibiotic resistance proteins (in part derived from the Resfams[23] database) (Supplementary Methods). One hundred and fifty-three proteins from 84 different NRPS and PKS clusters most closely matched hidden Markov models for transporters known to be involved in antimicrobial resistance, out of a total of 621 transporter genes within clusters. Annotations that could most confidently be linked to antibiotic resistance included one d-alanine–d-alanine ligase in a Rokubacteria NRPS

cluster, four d-alanine–d-alanine ligases in acidobacterial NRPS clusters, and two modified penicillin-binding protein sequences in Verrucomicrobia NRPS clusters (Supplementary Table 7).

Two near-complete genomes of divergent Acidobacteria were found to encode unusually large repertoires of NRP and PKS gene clusters. We refer to these two organisms as '*Candidatus*Eelbacter' (genome Eelbacter_gp4_AA13) and '*Candidatus*Angelobacter' (genome Angelobacter_gp1_AA117), tentatively placed within the Blastocatellia and the Acidobacteriales, respectively. In the 7-Mb genome of *Candidatus* Eelbacter we identified 17 biosynthetic loci containing 74 NRPS and PKS open reading frames that were 404 kb in total length. In the 6.5-Mb genome of *Candidatus* Angelobacter there were 16 loci containing 54 NRP/PKS open reading frames that were 325 kb in total length. The biosynthetic genes from each species had only distant homology to those from the other. We confirmed the biosynthetic clusters for both genomes by re-analysing with 'Prediction Informatics for Secondary Metabolomes' (PRISM)[24] (Extended Data Figs. 2, 3). In total, each of these organisms contains over 900 kb of genes that are putatively involved in biosynthesis of secondary metabolites (about 12–14% of their recovered genomes). A phylogenetic analysis, using ribosomal protein sequences, of acidobacterial genomes from this study and reference databases revealed that both *Candidatus* Angelobacter and *Candidatus* Eelbacter acquired their unusual arrays of biosynthetic operons independently in evolutionary time (Fig. 2a).

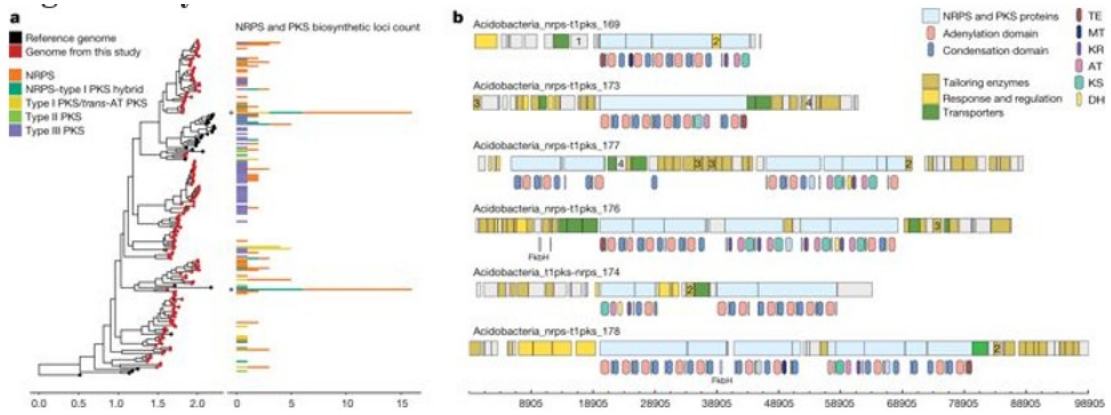

Fig. 2: Biosynthetic NRPS and PKS loci from the Acidobacteria.

a, Concatenated ribosomal protein phylogenetic tree of all acidobacterial genomes from this study (red) and existing reference genomes (black). Scale bar on the tree represents substitutions per site. Adjacent is a chart that reflects the count of NRPS and PKS biosynthetic gene clusters observed in each genome. The phylogenetic placements of *Candidatus* Eelbacter (*) and *Candidatus*Angelobacter (+) are marked. b, Six large PKS–NRPS hybrid biosynthesis gene clusters are encoded in the *Candidatus* Eelbacter genome. Predicted genes and biosynthetic protein domains are coloured by general function, and the genomic positions of polyketide and nonribosomal peptide synthetic domains are shown below each genome track. The following gene annotations are identified by number: 1, penicillin amidase; 2, oxygenase; 3, radical SAM proteins; and 4, betalactamase. AT, acyltransferase; DH, dehydrogenase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; TE, thioesterase.

The *Candidatus* Angelobacter genomes included multiple lantibiotic biosynthesis proteins, a bacteriocin biosynthesis cluster, multigene operons with components for both a type VI and a type II secretion system, and several large RHS-repeat containing proteins, which have been hypothesized to have evolved to mediate microbial competition by facilitating transfer of protein toxins between species[25]. The *Candidatus*Eelbacter genome contained six clusters that were complex type I NRPS-PKS hybrid systems over 45 kb in length (Fig. 2b). Three replicate genomes of *Candidatus* Eelbacter were obtained from independent soil samples and shared the same set of biosynthetic clusters. Both species also possessed CRISPR–Cas loci (31 spacers and repeats in *Candidatus*Angelobacter and 438 across the *Candidatus* Eelbacter genome). The ecological and evolutionary forces that can select for the production of an unusually high number of metabolites in a species are varied, and previously characterized examples are microorganisms with complex cooperative lifestyles[26,27] or an association with a eukaryotic host[28]. The discovery of these two microorganisms establishes that bacterial specialization in secondary metabolite biosynthesis is not limited to known clades in the Actinomycetales, Proteobacteria, Cyanobacteria, Bacilli and the recently discovered Entotheonella[28]. When considered together, the genomic features of these Acidobacteria hint towards an unusually competitive lifestyle mediated by chemical and toxin production.

We tested whether the microorganisms genomically described in this study are active and express biosynthetic NRPS or PKS gene clusters by analysing metatranscriptomics data from 120 soil microcosm samples from two soil depths and two sampling locations from the same field site that were subject to amendment with glucose, methanol or water over 24 h (Supplementary Methods). These experiments were designed to probe the strong biological responses that occur in soils following water addition and nutrient release after a long dry period[29]. Because distinct NRPS or PKS clusters can produce products with very different bioactivities, we tracked expression of each gene cluster as a functional biosynthetic unit by pseudo-aligning exact matches of paired reads to full genomes obtained directly from the environment studied using Kallisto[30]. Overall, we detected expression for 198 NRPS and/or PKS genes across those NRPS and PKS clusters with any level of gene expression (133 out of 180 clusters) (Supplementary Table 8). Expression of NRPS and PKS clusters was detected in all four phyla that we studied, and 84 active clusters were detected in Acidobacteria (Extended Data Fig. 4). We detected the expression of genes within 10 biosynthetic clusters—including 11 genes with NRPS and/or PKS domains within these clusters—of *Candidatus* Eelbacter (Extended Data Fig. 5) and 14 clusters of *Candidatus* Angelobacter—including 25 genes with NRPS and/or PKS domains. We tested for co-expression of genes in all biosynthetic clusters and found that gene clusters were co-expressed more often than were randomized

permutations of genes across each genome (Wilcoxon rank-sum test, $P < 0.001$).

Across all organisms in our dataset, we identified ten NRPS and/or PKS gene clusters from seven genomes with levels of expression that were time-dependent across the 24-h time course of the amendment experiments (permutational multivariate analysis of variance (PERMANOVA); $P < 0.05$, false discovery rate (FDR) = 5%) (Fig. 3a and Extended Data Fig. 6). We confirmed differential expression over time for individual genes within these clusters using a model that accounts for variation in both sequencing library sizes and organism abundances across samples[31] (DESeq2[32]; $P < 0.05$; FDR = 5%) (Supplementary Table 9). Notably, the expression of genes from several gene clusters in *Candidatus* Angelobacter showed a statistically significant increase 12–24 h after substrate addition (Fig. 3a), and we found that the expression of several biosynthetic genes of *Candidatus* Angelobacter was temporally distinct from the expression of core ribosomal genes (Fig. 3b). These results indicate that *Candidatus* Angelobacter populations respond to water and substrate addition, and independently regulate expression of secondary metabolite genes many hours after a period of increased core metabolic gene expression.

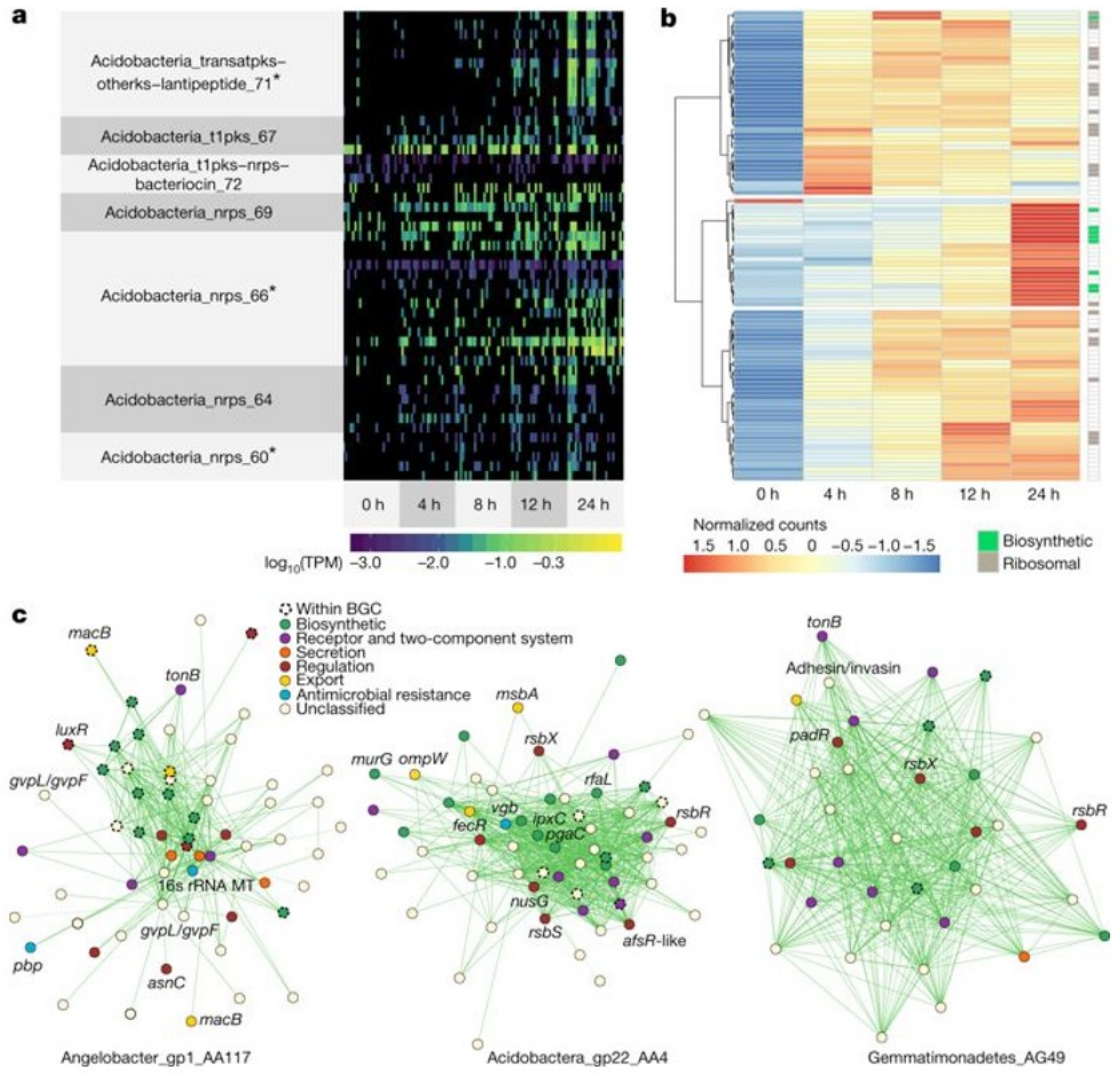

Fig. 3: Metatranscriptomics of biosynthetic genes.

a, Levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the *Candidatus* Angelobacter genome, across 120 microcosm soil samples grouped by extraction times (reported in hours). Expression levels are reported in $\log_{10}$-transformed transcripts per million (TPM). Gene clusters that were significantly differentially expressed across time points (PERMANOVA); $^{*}P < 0.05$, FDR = 5% are marked by an asterisk. b, Hierarchical clustering of expression levels for differentially expressed ($n = 120$; DESeq2; $P < 0.05$; FDR = 5%) genes from the *Candidatus* Angelobacter genome across samples grouped by experimental time point. Differentially expressed genes from biosynthetic clusters and differentially expressed core ribosomal proteins are marked. Values are reported in counts transformed using the rlog transformation from DESeq2 and were normalized by row. c, The transcriptional co-expression network modules ($n = 120$ microcosm time-point samples) significantly enriched in NRPS and PKS biosynthetic genes from three genomes ($P < 0.05$; hypergeometric distribution). Nodes represent gene transcripts and edges between them represent high topological overlap values between the transcripts. Genes outlined are genes found within biosynthetic gene clusters (BGC), and are coloured by assigned function using the Kyoto Encyclopedia of Genes and Genomes and Pfam databases. 16s rRNA MT, gene encoding for a 16S rRNA methyltransferase.

To predict the broader biological and ecological roles of these biosynthetic NRPS and PKS genes, we conducted separate co-expression analyses of all

genes for each of the seven species identified with temporally dependent biosynthetic gene expression, using the WGCNA package[33] (Supplementary Methods), across the 120 microcosm time-point samples. Co-expressed genes often share biological functions and regulation[34]. Modules of co-expressed genes significantly enriched in secondary metabolite genes were identified in four out of seven genomes ($P < 0.05$; hypergeometric distribution) (Fig. 3c, Extended Data Fig. 7 and Supplementary Table 10). These four modules were small (fewer than 69 genes) and very transcriptionally distinct. We found that all four secondary metabolism networks were dominated by genes involved in two-component systems, efflux and transcriptional regulators, and were almost completely devoid of genes for the core processes of transcription, translation and energy metabolism.

For *Candidatus* Angelobacter, genes from five biosynthetic clusters were co-expressed together in a module with a variety of genes involved in environmental sensing and response, including homologues of the gene that encodes for the iron siderophore uptake receptor TonB. Homologues of the gene that encodes for the macrolide export transporter MacB were also found to be co-expressed with the biosynthetic genes, as were two putative antimicrobial resistance genes—those encoding for penicillin-binding protein and for a 16S rRNA methyltransferase. Additional co-expressed genes included an operon for a type VI secretion system and an operon annotated as encoding for gas vesicle proteins. Notably, the Angelobacter population expressed biosynthetic genes from multiple clusters simultaneously, suggesting a concerted response that is linked to ecological competition.

Acidobactera_gp22_AA4 was found to co-express its NRPS gene cluster (Acidobacteria_nrps_112) with response-regulatory genes and a set of genes involved in cell surface structure remodelling, as well as an operon of genes involved in regulating stress response (*rsbX*, *rsbR* and *rsbS*). A homologue of virginiamycin B lyase (*vgb*), which is an inactivator of type B streptogramin antibiotics, was also co-expressed in this module. The same operon of genes involved in the regulation of stress response was found to be co-expressed in the transcriptional network containing a biosynthetic cluster (cluster Gemmatimonadetes_nrps_183) in Gemmatimonadetes_AG49, along with a *tonB* homologue.

In summary, we uncovered extensive evidence for secondary metabolite synthesis in a large collection of bacterial genomes from four phyla of soil bacteria that have not previously been genomically linked to this capacity. Although we cannot confidently predict more than the basic chemical scaffolds of the products derived from the biosynthetic genes reported here, or their biological activities, a large percentage of known polyketide and nonribosomal metabolites isolated from microbial sources have antimicrobial activity[35]. Transcriptional associations between specific NRPS and PKS gene clusters, regulators of iron metabolism and putative

antimicrobial resistance mechanisms suggest that these gene clusters may be involved in competition for iron resources and antibiotic production. The findings underline the utility of genome-resolved metagenomic investigations of soil ecosystems and open the way for laboratory characterization of genes for novel bioactive metabolites with potential ecological and pharmaceutical importance.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessments.

Soil sampling and DNA extraction

Soil samples were collected from the Angelo Coast Range Reserve meadow (39° 44′ 21.4″ N 123° 37′ 51.0″ W) on four dates in 2014 that bracketed the first winter rain of the season. Samples were collected from three depths, 10–20 cm, 20–30 cm and 30–40 cm at six independent sampling sites that were first metagenomically characterized as part of a previous study[8]. Sampling was conducted in biological triplicate, with three of the sites being unamended biological control plots and three being amended with extended spring rainfall from a sprinkler system as described in a previous publication[8]. Sampling was accomplished using a soil coring device that was fitted with sterilized polycarbonate sheaths. Sheaths were removed after each collection event. After collection, samples were flash-frozen in a mixture of dry ice and ethanol, and placed on dry ice for transport. A total of 60 soil cores were sampled across all depth and treatment conditions.

For each depth, DNA was extracted using MoBio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil as previously described[8]. Mean DNA concentration in the extracted samples, quantified by using qubit fluorometric assay, was 388 ng/μl.

Sequencing, genomic assembly and binning

Metagenomic libraries for all 60 samples were prepared and sequenced at the Joint Genome Institute using an Illumina HiSeq 2500 platform to generate 250-bp paired-end reads. Samples were multiplexed for sequencing. Raw sequence data were processed with BBmap[36] to remove Illumina adaptor and phiX sequences, and reads were quality-score trimmed using Sickle with default parameters[37]. Read sets were subsequently analysed for per-base GC content using FastQC[38], and it was determined that GC content increased substantially after 200 bp in some sample read sets. Thus all reads longer than 200 bp were hard-trimmed to 200 bp using BBmap. In total, 6.22×10^9reads were sequenced across all samples, which yielded 1.24 Tb of total sequence information with an average read count of 1.04×10^8 reads per sample.

The 60 samples were individually assembled de novo on a 24-core Intel Xenon Linux cluster node with 256 Gb of RAM using IDBA-UD[39] with the following initial parameters: –pre_correction,–mink 30,–maxk 200,–step 10. In the 13 cases in which assemblies did not complete owing to memory requirements, minimum *k*-mer size was increased to 40 bp. The resulting assemblies averaged 1.15 Gb of assembled sequence with an N50 of 1,609 bp. Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using Bowtie2[40]; 36.4% of reads mapped back to assembled sequence on average. It should also be noted that contigs >100 kb in length were acquired from all 60 assemblies, with a maximum contig size across assemblies of 2.7 Mb.

All resulting assemblies were subsequently clustered into genome bins individually using a hybrid binning approach. Initially, reads from all assemblies were separately cross-mapped to all scaffolds >2 kb in size from a single assembly using Bowtie2 to generate a coverage profile for the scaffolds of that assembly across all samples. Scaffold differential coverage profiles were used to inform five separate automated binning software packages: ABAWCA, ABAWACA2[41], MaxBin2[42], CONCOCT[43] and MetaBAT[44], which were run on all samples individually. The resulting output genome bins for all packages run on a single sample were combined, assessed for completeness using an inventory of 51 universal single-copy genes (SCGs), and dereplicated by selecting the most complete bin of an overlapping set using DAStool[45]. Following automated binning, all genomic bins were manually inspected and curated using our in-house bin visualization and analysis system, ggKbase[46] (http://ggkbase.berkeley.edu). Finally, after manual curation in ggKbase, reads from a given sample were mapped back to the bins derived from that sample to identify and correct assembly and scaffolding errors, as previously described[47]. In total, 10,463 individual genome bins were identified across all samples. Of these bins, 3,334 were then estimated at a completeness of ≥ 70% using CheckM[48]. Taxonomic assignment of bins was performed by looking at the closest known hits and phylogenetic placement of ribosomal marker proteins. Bins were then dereplicated by clustering their ribosomal S3 proteins at 99% amino acid identity and choosing the bin in each cluster with the highest completeness and lowest contamination, which resulted in a final set of 377 nonredundant bins in the bacterial phyla of interest.

Genomic analysis of genomes and biosynthetic gene clusters

Curated genomes were individually processed using antiSMASH 3.0[10] with default parameters. The results are summarized in Supplementary Table 2 for gene clusters on contigs greater than 10 kb, Supplementary Table 1 for gene clusters on contigs smaller than 10 kb and Supplementary Table 4 for all PKS and NRPS clusters on contigs greater than 10 kb. Ribosomal protein phylogenetic trees were built using a concatenated set of 16 ribosomal proteins[49] for all Acidobacteria genomes in this dataset, as well as those that could be obtained from GenBank or the Integrated Microbial Genomes

platform. An *Escherichia coli* genome was used as an outgroup for the tree. These protein sequences were aligned with MUSCLE[50] and then a maximum likelihood phylogeny was built using FastTree2[51] with default parameters.

To test whether existing primer-based methods have the ability to amplify these biosynthetic gene sequences, sets of forward and reverse degenerate primers used by previous analyses of biosynthetic genetic diversity[2,17] for ketosynthase genes and adenylation domain genes were searched for pattern matches against all NRPS and PKS clusters in both reverse and forward reading frames. The inosine nucleotides were substituted with the ambiguous code B, because these nucleotides can base pair with adenine, cytosine and uracil. Only five of our gene clusters had correctly oriented matches to both a forward and reverse primer within 2 kb of each other (Supplementary Table 6).

The network of gene clusters based on shared gene content was built by performing an all-versus-all BLASTP search of predicted biosynthetic protein sequences. Shared proteins were defined as protein alignments with at least 50% of the query sequence covered and amino acid per cent identity >50%. Two clusters (nodes) were connected if either one shared at least 10% of its proteins with the other. The width and colour intensity of the network edges was scaled with the length of the shared protein alignments, normalized to the length in base pairs of the two clusters being compared. Biosynthetic gene clusters were compared to clusters previously reported in the MiBIG repository[19] using BLASTP and the same definition of shared proteins, and the closest hits to MiBIG clusters containing at least five genes were reported. To identify antibiotic resistance genes in clusters, we searched protein products of all biosynthetic gene clusters with a set of hidden Markov models derived from a previous publication[22], using HMMER with the gathering threshold cutoffs specified in this previous study. We then manually curated hits and eliminated matches to ambiguous functions (acetyltransferases, general methyltransferases and amidases) and focused on reporting proteins with functions that are unlikely to be involved in generic biosynthetic pathways. The *Candidatus* Angelobacter and Eelbacter genomes were both subsequently analysed using the PRISM3 webserver[24].

Soil microcosm experiments and RNA extraction

At the Angelo Coast Range Reserve meadow, five holes were bored within a 1-m^2 area to obtain 10-cm-long cores of soil, from depths 10–20 cm and 30–40 cm (permission under APP # 27790). Samples were collected on 21 September 2015. At each depth, five cores were mixed in a large Whirl-Pak bag, then distributed into five capped core liners and stored in individual Whirl-Pak bags at 4 °C. The unsieved soils were mixed a second time in the laboratory to obtain six equally proportioned samples, and the weights were measured. To settle the soil, the core liners were struck with a rubber mallet 50 times each, and then stored at 4 °C. The night before wet-up

experiments, the cores were placed in a cooler alongside the substrate that was to be added, so that the soil and substrate equilibrated to the same temperature and the soil would be kept in the dark. Immediately before adding the substrate, 10 g soil was collected for DNA extraction and 2 g soil with 4 ml LifeGuard RNA Soil Preservation Solution (MoBio) was collected for RNA purification. Both were immediately frozen in liquid N_2 and stored in a freezer at −80 °C. Samples at different time points were collected for nucleic acid extraction in the same manner. Ten millimolar glucose, methanol or water substrate was added to the open-soil core liners and soil in a cooler by pipette 2.5–4 ml at a time over 1 min, and the lid was closed. Substrates were added in amounts that increased the soil moisture to the level of a sample collected from the meadow after 29 cm of rainfall on 5 November 2015 (the moisture level of the field sample was determined by weight loss on drying). RNA was isolated from 2 g soil with RNA PowerSoil Total RNA Isolation kits, following kit protocols. cDNA libraries were prepared and were sequenced to generate 5.9×10^9 150-bp paired-end reads.

Transcriptomics

To test for the expression of clusters of biosynthetic genes within a soil environment, we analysed metatranscriptomics data from experimental soil microcosms. Soil samples from depths of 20 cm and 40 cm from two sampling locations were subject to amendment with glucose, methanol or water, and RNA was extracted from samples at 0, 4, 8, 12 and 24 h after treatment. From the 120 sequenced samples, we generated 5.9×10^9 150-bp paired-end reads. Transcript abundances for all Prodigal-predicted gene sequences from all genomes reconstructed from the project site were quantified using Kallisto[30] exact pseudoalignments of paired reads. Kallisto was run using default parameters. Transcripts that were either found to be expressed in at least 10% of samples or to have at least 100 counts were reported and included in downstream analyses. Differential gene expression analysis was performed using PERMANOVA and DESeq2[32] (see ‘Statistical analysis’).

We mapped RNA reads from one replicate for each sample at the $t = 0$ and $t = 24$ h time points to 16S sequences assembled from our genomic data from the two plots from which the microcosm soil was obtained. A subset of 4,000 RNA reads was compared to the SILVA 16S database using BLAST to determine the percentage of RNA reads that were 16S rRNAs. Of 16S rRNA reads in the RNA data, 47% ± 19% were determined to be at least 98% identical to 16S sequences assembled in the genomic data (Supplementary Table 12), which indicates that the community that we assembled in the genomic dataset is a substantial fraction of the active community in the metatranscriptomic data.

We performed weighted gene co-expression network analyses using the WGCNA package[33] separately and individually on genes from seven

genomes that were identified as having differentially expressed biosynthetic gene clusters over time, reasoning that these genomes will have the strongest signal of secondary metabolite co-expression. Transcripts per million for each gene were log-transformed. A soft network threshold was generated by choosing the lowest value that returned an R^2 fit to a scale-free network greater than 0.8. A signed adjacency matrix was built using Pearson correlations, and a topographical overlap matrix was generated from the adjacency matrix. Module detection was run using the cuttreeDynamic() function with the 'hybrid' method, a minimum cluster size of 15, deepSplit set to TRUE and a cutHeight of 0.95.

Statistical analysis

To test whether cluster genes were significantly more co-expressed than random genes across a genome, we calculated all Spearman correlations between genes within clusters (mean $\rho = 0.063$; $n = 5{,}940$ comparisons), and compared this distribution of correlations to a distribution of all Spearman correlations between 100 randomly chosen genes from each genome (mean $\rho = 0.041$; $n = 503{,}699$ comparisons) using an independent two-group Wilcoxon rank-sum test ($P < 0.001$). We also compared both distributions to a distribution of randomly selected genes from the entire dataset compared (mean $\rho = 0.026$ $n = 4947228$ comparisons) and found random genes to have the lowest levels of co-expression ($P < 0.001$).

To identify differentially expressed clusters of genes between time points, we used the adonis function from the vegan package[52]. Transcript abundances in transcripts per million were $\log_2$-transformed, and adonis tests were run on all clusters with any expression data for at least five proteins. *P*values were corrected for multiple tests using the Benjamini and Hochberg[53] method with a controlled family wise error rate of 5%.

To detect differential expression of individual genes within differentially expressed biosynthetic clusters between time points, we modelled Kallisto counts in the context of all metadata variables (plot, depth, treatment and time) using a negative binomial model implemented in DESeq2[32]. Kallisto count data from each genome were analysed independently so that the DEseq size factors for cross-sample count normalization would reflect the total transcriptomic activity of that genome in each sample. This approach is robust to biases in total transcriptomic activity per organism between samples, with the intention to identify differences in gene expression independent of changes in taxonomic composition, similar to previously reported methods[30]. After size factor normalization, counts were fit to a negative binomial model of the form: count ~ depth + plot + treatment + time. To specifically test whether any genes exhibit differential expression associated with changes in time while accounting for the effects of depth, plot and treatment, we fit count data to a reduced model of the form: count ~ depth + plot + treatment. We then compared fits between the full and reduced model using the likelihood ratio test implemented in DESeq2.

The significant genes (with an FDR-corrected $P < 0.05$) identified by comparing the full and reduced model were grouped, and direct comparisons were made between counts at 0 h and all other time points, to find those time points that exhibited a significant change in expression relative to the 0 h time point. This method confirmed differential expression of several individual genes within each differentially expressed biosynthetic cluster.

When examining modules of co-expression genes, the hypergeometric test was used to determine whether a module was significantly enriched in biosynthetic genes, using the phyper function in R.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability

Custom code used for the analyses (transcriptomics expression, DESeq2 differential expression and WGCNA co-expression analyses) that support this work is available in R Notebook format at http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

Data availability

All genomic data associated with this project has been deposited in BioProject under accession PRJNA449266. DNA sequencing reads for this project have been deposited in the Sequence Read Archive database under PRJNA449266. Genomes analysed as part of this project have been submitted to the Whole Genome Shotgun (WGS) database. Genomes are also available through ggKBase at the following URL: http://ggkbase.berkeley.edu/angelo2014/organisms. Raw data for Fig. 2a and AntiSMASH annotated GenBank files for biosynthetic gene clusters reported on in this Letter are available at: http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

References

1. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Brook Peterson, S. Bacterial competition: surviving and thriving in the microbial jungle. Nat. Rev. Microbiol. 8, 15–25 (2010).

2. Charlop-Powers, Z., Owen, J. G., Reddy, B. V., Ternei, M. A. & Brady, S. F. Chemical–biogeographic survey of secondary metabolism in soil. Proc. Natl Acad. Sci. USA111, 3757–3762 (2014).

3. Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. Biochim. Biophys. Acta 1830, 3670–3695 (2013).

4. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394 (2003).

5. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579–590 (2017).

6. Bergmann, G. T. et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol. Biochem. 43, 1450–1455 (2011).

7. Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A. & Kuramae, E. E. The ecology of Acidobacteria: moving beyond genes and genomes. Front. Microbiol. 7, 744 (2016).

8. Butterfield, C. N. et al. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ 4, e2687 (2016).

9. DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M. & Radosevich, M. Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. Appl. Environ. Microbiol. 77, 6295–6300 (2011).

10. Weber, T. et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 43, W237–W243 (2015).

11. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39, W339–W346 (2011).

12. Hadjithomas, M. et al. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio 6, e00932-e15 (2015).

13. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158, 412–421 (2014).

14. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc. Natl Acad. Sci. USA 111, 9259–9264 (2014).

15. Parsley, L. C. et al. Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. FEMS Microbiol. Ecol. 78, 176–187 (2011).

16. Rondon, M. R. et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541–2547 (2000).

17. Charlop-Powers, Z. et al. Global biogeographic sampling of bacterial secondary metabolism. eLife 4, e05048 (2015).

18. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. 106, 3468–3496 (2006).

19. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol. 11, 625–631 (2015).

20. Medema, M. H., et al. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput. Biol. 10, e1004016 (2014).

21. Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat. Biotechnol. 31, 922–927 (2013).

22. Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. Nat. Chem. Biol. 12, 233–239 (2016).

23. Gibson, M. K., Forsberg, K. J. & Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 9, 207–216 (2015).

24. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 45, W49–W54 (2017).

25. Koskiniemi, S. et al. Rhs proteins from diverse bacteria mediate intercellular competition. Proc. Natl Acad. Sci. USA 110, 7032–7037 (2013).

26. Claessen, D., de Jong, W., Dijkhuizen, L. & Wösten, H. A. Regulation of Streptomyces development: reach for the sky. Trends Microbiol. 14, 313–319 (2006).

27. Zhang, Y., Ducret, A., Shaevitz, J. & Mignot, T. From individual cell motility to collective behaviors: insights from a prokaryote, Myxococcus xanthus. FEMS Microbiol. Rev. 36, 149–164 (2012).

28. Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature506, 58–62 (2014).

29. Unger, S. et al. The influence of precipitation pulses on soil respiration–assessing the “Birch effect” by stable carbon isotopes. Soil Biol. Biochem. 42, 1800–1810 (2010).

30. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).

31. Klingenberg, H. & Meinicke, P. How to normalize metatranscriptomic count data for differential expression analysis. PeerJ 5, e3859 (2017).

32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

33. Langfelder, P & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).

34. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255 (2003).

35. Bérdy, J. Bioactive microbial metabolites. J. Antibiot. (Tokyo) 58, 1–26 (2005).

36. Bushnell, B. BBMap short read aligner. http://sourceforge.net/projects/bbmap (University of California, Berkeley, 2016).

37. Joshi, N. A. & Fass, J. N. sickle - a windowed adapative trimming tool for FastQ files (version 1.33) https://github.com/najoshi/sickle (2011).

38. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

39. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (2012).

40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

41. Brown, C.T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature523, 208–211 (2015).

42. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607 (2016).

43. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (2014).

44. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015).

45. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Methodshttps://doi.org/10.1038/s41564-018-0171-1 (2018).

46. Banfield, J. Development of a Knowledgebase to Integrate, Analyze, Distribute, and Visualize Microbial Community Systems Biology Data. Report No. DOE-UCB-4918) (US Department of Energy, 2015).

47. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. 7, 13219 (2016).

48. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).

49. Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).

50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).

51. Price, M. N., Dehal, P. S. and Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS ONE 5, e9490 (2010).

52. Oksanen, J. et al. vegan: Community ecology package https://cran.r-project.org/package=vegan (2007).

53. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).

Acknowledgements

We thank S. Spaulding for assistance with fieldwork, and M. Traxler and W. Zhang for helpful discussions. Sequencing was carried out under a Community Sequencing Project at the Joint Genome Institute. Funding was provided by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy Grant DOE-SC10010566, the Paul G. Allen Family Foundation and the Innovative Genomics Institute of the University of California, Berkeley.

Author information

Affiliations

1. Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA

- Alexander Crits-Christoph

2. The Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, USA

- Alexander Crits-Christoph
- & Jillian F. Banfield

3. Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA

- Spencer Diamond
- , Cristina N. Butterfield
- , Brian C. Thomas
- & Jillian F. Banfield

4. Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA

- Jillian F. Banfield

5. Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

- Jillian F. Banfield

Contributions

A.C.-C. performed genomic and transcriptomic analysis; S.D. performed metagenome assembly and curation; C.N.B. performed microcosm experiments and RNA extractions; A.C.-C., S.D. and J.F.B. wrote the manuscript; B.C.T. supported the metagenomics bioinformatics work; and J.F.B. supervised the project.

Competing interests

The authors declare no competing interests.

Corresponding author

Correspondence to Jillian F. Banfield.

Extended data figures and tables

Extended Data Fig. 1 Experimental plan and project overview.

Schematic showing major components of microcosm time-point sampling and metagenomic analyses.

Extended Data Fig. 2 NRPS and PKS biosynthetic loci of the *Candidatus* Eelbacter genome.

Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Eelbacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.

Extended Data Fig. 3 NRPS and PKS biosynthetic loci of the *Candidatus* Angelobacter genome.

Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Angelobacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.

Extended Data Fig. 4 Metatranscriptomics of NRPS and PKS proteins.

The graph shows levels of transcriptional expression of genes containing NRPS and PKS protein domains across genomes from the four phyla of interest. Values are reported in $\log_{10}$-transformed transcripts per million and are summed across the 120 soil microcosm samples.

Extended Data Fig. 5 Metatranscriptomics of the *Candidatus*Eelbacter genome.

The levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the *Candidatus* Eelbacter genome across 120 soil microcosm time-point samples grouped by extraction times (reported in hours) are shown. Expression levels are reported in $\log_{10}$-transformed transcripts per million.

Extended Data Fig. 6 Differentially expressed biosynthetic gene clusters over time.

The levels of expression of biosynthetic gene clusters from all organisms studied (excluding *Candidatus* Angelobacter data shown in Fig. 3a) that were found to be significantly differentially expressed between time points (PERMANOVA; $n = 120$; $P < 0.05$, FDR = 5%) across 120 soil microcosm time-point samples are shown. Expression levels are reported in $\log_{10}$ transcripts per million.

Extended Data Fig. 7 Biosynthetic co-expression transcriptional module from Verrucomicrobia_AV7.

A transcriptional network of co-expressed Verrucomicrobia_AV7 genes from a module found to be significantly enriched in genes from the biosynthetic gene clusters Verrucomicrobia_nrps_156 and Verrucomicrobia_nrps_157 ($P <$ 0.05; hypergeometric distribution) is shown. Genes from the biosynthetic locus are outlined with a dashed line.

DOI

https://doi.org/10.1038/s41586-018-0207-y