Title
Ecology and evolution of specialized metabolism in uncultivated bacteria

Author
Crits-Christoph, Alexander

Publication Date
2021

Peer reviewed|Thesis/dissertation

Ecology and evolution of specialized metabolism in uncultivated bacteria

By

Alexander J Crits-Christoph

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor Jillian F. Banfield, Chair
Professor Matthew Traxler
Professor Ronald Amundson

Spring 2021

Abstract

Ecology and evolution of specialized metabolism in uncultivated bacteria

By

Alexander J Crits-Christoph

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian Banfield, Chair

A wide range of specialized (or "secondary") metabolites are produced by bacteria in natural ecosystems with varying functions, including antibiotics, siderophores, signalling molecules, and antifungals. These specialized metabolites are produced using operonic sets of genes that work in concert, known as biosynthetic gene clusters. In this work, genome-resolved metagenomic approaches were applied to understand the distribution and ecology of biosynthetic genes and the bacteria that possess them. Bacterial genomes were assembled and binned from deeply sequenced metagenomes from a California grassland meadow soil and a permanently wet vernal pool soil, and clusters of biosynthetic genes were identified in each genome. Genomes were reconstructed for novel species that belong to rarely cultivated but ubiquitous soil phyla, including the Acidobacteria, Verrucomicrobia, and the candidate phylum Rokubacteria. Bacteria from the grassland meadow soil were shown to have unexpectedly large numbers of biosynthetic gene clusters. In particular, two novel lineages of Acidobacteria were identified that possessed an unusual genomic capacity for specialized metabolite biosynthesis - up to 15% of their genomes were predicted to be dedicated to the production of nonribosomal peptides and polyketides. Sampling a second study site of soils from a vernal pool, Another three species were obtained from one of these uncultivated lineages — the candidate genus *Angelobacter* — which also possessed a large genomic repertoire of diverse biosynthetic genes. By mining public soil metagenomes, additional high quality draft genomes from this candidate genus were also analyzed, confirming that species in this genus are widespread across soil environments. It was therefore established that *Angelobacter spp.* with a substantial capacity for specialized metabolite biosynthesis are widespread in soils with a range of moisture contents and vegetation types.

Transcriptional activity of nonribosomal peptide synthetase and polyketide synthase genes of abundant organisms from the grassland soil was tracked over time using 120 metatranscriptomic samples from soil microcosms. For several bacterial species within the samples, unsupervised clustering of genes by co-expression across samples identified modules of biosynthetic genes that were tightly co-expressed with genes involved in transcriptional regulation, environmental sensing, and secretion. For some vernal pool samples where *Angelobacter* were the most abundant microbial community members, metatranscriptomics demonstrated clear transcriptional activity *in situ*.

Transcription of many *Angelobacter* biosynthetic genes was detected, extending findings from the grassland soil microcosms.

Genetic variation in soil bacteria and their biosynthetic genes was investigated using population genomics methods that leverage genetic variation within sequencing reads that map to genomes from metagenomes. Metagenomic methods to track genetic variation within populations in a spatial context were applied to study the most abundant bacterial species across the grassland meadow study site. Genetic variation specifically within biosynthetic genes was elevated, indicating that there can be substantial allelic diversity in the biosynthetic genes of an abundant species in a local soil ecosystem. For about half of the bacterial populations studied, strong genetic population structure associated with spatial scale was observed. Genomes and gene variants were more genetically similar if they were from the same meadow plot. Simultaneously, while genetic gradients were observed across the meadow, within sample genetic diversity was also found to be high. Genomic signatures of recombination and gene-specific selection were also identified, indicating that ongoing selection and recombination may shape genetic divergence of populations on local spatial scales in soils.

While biosynthetic gene clusters can be outlined and annotated with confidence in microbial genomes, prediction of the function of the metabolites produced for novel gene clusters is often an unsolved problem. Colocalized transporter genes associated with biosynthetic gene clusters may help predict metabolite function, due to their intimate association with the metabolite(s) they are transporting. This hypothesis was tested and benchmarked on a dataset of characterized biosynthetic gene clusters. In particular, a strong specificity of transporter genes for siderophore export and re-uptake was quantified as a signal of siderophore production. Using this specific genomic signal, putative siderophore BGCs were annotated across bacterial genomes recovered from soil, as well as from better characterized microbes from the adult and premature infant microbiomes. Surprisingly few genomes from soil bacteria contained transporter genes associated with siderophore biosynthesis. While 23% of microbial genomes from premature infant microbiomes possess at least one siderophore-like biosynthetic gene cluster, only 3% of those from adult gut microbiomes do.

In sum, this thesis presented a metagenomic perspective on specialized metabolisms, contributed to discovery of novel species, examined evolutionary processes, and improved genomic functional predictions. The strength of this approach lies in its ability to investigate microbes in *in situ* community contexts and detect ecological trends among the uncultivated microbial majority.

Table of Contents

Introduction

The vast majority of microbes living on our planet are still unknown. The unknowns of microbial ecology span many scientific perspectives: there are novel microbial species, genes, biochemical functions, and ecological interactions remaining to be characterized or understood in natural ecosystems. One of the biggest obstacles for characterization of novel microbes is difficulty in cultivation - it is estimated that we only currently have been able to culture representations of less than 3% of microbial species detected in natural environments (Steen et al. 2019), and often entire lineages appear recalcitrant to existing or commonly employed cultivation methods (Lloyd et al. 2018).

Since its invention in 2004, genome-resolved metagenomics has become a guiding candle in the dark of the microbial world (Tyson et al; 2004; Sangwan et al. 2016). Genome-resolved metagenomics is the assembly, curation, and quantification of draft-quality or complete microbial genomes directly from the environment. Where-as both 16s rRNA sequencing and reference-based metagenomics methods technologies track the abundances of characterized microbes and genes in the environment, genome-resolved metagenomics provides insight into whole genomes of uncultivated microbes that are abundant in natural environments. Using microbial genomes extracted from metagenomes, we can predict the genomic capacities and functional capabilities of novel microbes. While genomic potential does not perfectly predict phenotypic function, it is also possible to combine metatranscriptomics with genome-resolved metagenomics to identify genes that are actively being expressed in natural ecosystems.

Soils are the most complex microbial ecosystem, as measured by any number of metrics: species diversity, genetic diversity, and biogeochemical complexity (Fierer 2017). Due to their genetic complexity, until recently soils were considered difficult to study using genome-resolved metagenomics techniques (Howe et al. 2014). However, improvements in sequencing technology and informatics approaches have made it possible to assemble high-quality genomes out of soil metagenomes. For the first time, this allows us to interrogate the genomic capacity of the uncultivated majority of soil microorganisms for functional traits of interest, and to better understand their lifestyles and/or ecological roles. This is particularly important in soil ecosystems, where it is estimated that <3% of all microbial species have a cultivated representative, and intra-species genetic diversity is also high. With these genome-resolved tools, we can begin to build pictures of individual species in soil ecosystems that have never been reported before. We can also begin to answer questions about any functional roles with a known genomic basis, and also questions of population genetics, evolution, and ecological interactions for the microorganisms in soils.

Soils have also long been an environmental source of microbes that produce specialized metabolites, including antibiotics, antifungals, siderophores and other molecules with therapeutic or industrial value. For example, the soil bacterium *Amycolatopsis orientalis* is a natural producer of the antibiotic Vancomycin (Xu et al. 2014), and *Streptomyces hygroscopicus* is the source of the immunosuppressant Rapamycin (Aparicio et al. 1996). A large number of important natural products have thus been discovered through a process of isolating bacteria from the natural environment, screening their exudates for biological activities of interest, and then identifying the active metabolite (Genilloud 2017). Interbacterial interactions involving cultivated producers of specialized metabolites have been studied in laboratory settings (Traxler and Kolter 2015), and to the development of the 'competition sensing' hypothesis that predicts that microbes turn on genes for production of specialized metabolites in response to signals of local competition from others microbes (Cornforth and Foster 2013). However, how do these principles and findings extend to the uncultivated bacterial majority in soils?

This open question forms the guiding framework for this work, and is approached through a genome-resolved metagenomics lens. Firstly, we examine the ecology of specialized metabolisms, and the phylogenetic distribution of specialized metabolite production. Which uncultivated microbes in soil ecosystems make specialized metabolites? Are they more abundant than cultivated lineages of specialized metabolite producers? Metatranscriptomics of soil communities is used also to answer questions about the contexts in which genes for specialized metabolite production are expressed after rain events in soil. Secondly, we examine the evolution of specialized metabolisms, and the genetic diversity of soil bacterial populations that possess genes for specialized metabolite biosynthesis is explored. Are strains of the same species genetically similar across a local soil ecosystem, or highly variable? And are the genes for specialized metabolite production more clonal, or less clonal, than the average gene in a population? Finally, we take a closer look at the kinds of transporter proteins associated with specialized metabolisms, and ask whether these genes can help predict the ecological functions of unknown specialized metabolites. Taken together, these results provide a genome-centric perspective on specialized metabolisms of uncultivated microbes with the aim of drawing parallels to and extending upon what is known from the cultivated microbial world.

Acknowledgements

I would first like to thank my advisor Jill Banfield for supporting me with time, energy, knowledge, resources, and guidance on this work and all of my other experiences while in her laboratory. I think it is no secret that I can be an unusual student to mentor and Jill has supported my ideas and initiatives nonetheless. I would also like to thank the entire Banfield lab who have provided so much mentorship, guidance, and friendship over the years. Matthew Olm, Patrick West, Keith Bouma-Gregson, Chris Brown, and Alex Thomas have contributed so much to this work with collaborations, discussions, and friendship. I would especially like to thank Spencer Diamond, who has perfectly walked the fine line of being both an effective mentor and a truly dear friend. Outside of the lab, I would like to thank our collaborators, especially Matt Traxler, Rita Pessotti, and Mira Liu. They have always approached our shared projects with passion and enthusiasm, and I have learned much from them.

I thank my family - my parents, who have always supported my desire to go to graduate school, and my siblings, Avery and Blaire. And of course my partner, Jill Hakim, who has been there with me for the entire journey, and been my best friend all the way through.

There is nobody I need to thank more than the entire PMB community. PMB is such a wonderful group of people from whom I have learned so, so much. I would like to thank the professors, postdocs, and graduate students. While I can't list everyone, thank you so much for sharing your science with me. I cannot imagine a better community to do graduate work with.

1. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis

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

Published in *Nature*, 2018.

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 (Hibbing et al. 2010; Charlop-Powers et al. 2014). Most known antibiotics are derived from a few culturable microbial taxa (Cragg and Newman 2013), and the biosynthetic potential of the vast majority of bacteria in soil has rarely been investigated (Rappé and Giovannoni 2003). 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 (Fierer 2017; Bergmann et al. 2011; Kielak et al. 2016), 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 (Fierer 2017). 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.

1.1 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 (Butterfield et al. 2016). 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 (Butterfield et al. 2016). 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. 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 (Bushnell 2016) to remove Illumina adaptor and phiX sequences, and reads were quality-score trimmed using Sickle (Joshi and Fass 2011) with default parameters. Read sets were subsequently analysed for per-base GC content using FastQC (Andrews 2010), 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, 6.22×10^9 reads 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 (Peng et al. 2012) 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 (Langmead and Salzberg 2012); 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 (Brown et al. 2015), MaxBin2 (Wu et al. 2016), CONCOCT (Alneberg et al. 2014) and MetaBAT (Kang et al. 2015), 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 (Sieber et al. 2018). Following automated binning, all genomic bins were manually inspected and curated using our in-house bin visualization and analysis system, ggKbase (Banfield 2015) (http://ggkbase.berkeley.edu.libproxy.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 (Anatharaman et al. 2016) . 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 (Parks et al. 2015). 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 (Weber et al. 2015) with default parameters. Ribosomal protein phylogenetic trees were built using a concatenated set of 16 ribosomal proteins (Hug et al. 2016) 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 (Edgar 2004) and then a maximum likelihood phylogeny was built using FastTree2 (Price et al. 2010) 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 (Charlop-Powers et al. 2014; Charlop-Powers et al. 2015) 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.

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 (Medema et al. 2015) 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 (Johnston et al. 2016), 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 (Skinnider et al. 2017).

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 (Bray et al. 2016) 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 (Love et al. 2014) (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, 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 (Langfelder et al. 2008) 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 (Oksanen et al. 2007). 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 method (Benjamini and Hochberg 1995) 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. 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 (Klingenberg and Meinicke 2017). 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.

1.2 Results

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 (**Fig. 1.4**). Specifically, we analysed newly reconstructed genomes from 149 Acidobacteria, 135 Verrucomicrobia, 43 Rokubacteria and 49 Gemmatimonadetes species (see *Methods*). We targeted these groups because bacteria from all four phyla are highly abundant at our field sampling site (Butterfield et al. 2016) (**Fig. 1.1a**) and in globally sampled soils (Fierer 2017). 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 (Fierer 2017) , and Gemmatimonadetes are also known to be common in soils (DeBruyn et al. 2011). 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.

Within the genomes, we identified 1,159 biosynthetic gene clusters on contigs at least 10 kb in length (**Fig. 1.1b**) and an additional 440 biosynthetic gene clusters on smaller contigs using antiSMASH 3.0 (Weber et al. 2015), an in silico pipeline that was originally verified against 473 verified biosynthetic gene clusters with a 97.7% reported accuracy (Medema et al. 2011). 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 of the Actinobacteria, Proteobacteria and *Bacillus* (Cragg and Newman 2013), which represent microorganisms that often comprise a minority in soil microbial communities (Rappé and Giovannoni 2003; Fierer 2017). Previous global analyses based on the few publicly available genomes for Acidobacteria, Verrucomicrobia and Gemmatimonadetes (Hadjithomas et al. 2015; Cimermancic et al. 2014; Wang et al. 2014) 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 (Kielak et al. 2016; Parsley et al. 2011). 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) (Rondon et al. 2000). Other studies (Charlop-Powers et al. 2014; Charlop-Powers et al. 2015) 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 (Charlop-Powers et al. 2014; Charlop-Powers et al. 2015) 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.

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 (Wang et al. 2014). 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. 1.1c**) 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 (Medema et al. 2011; Fischbach and Walsh, 2006). 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.

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. 1.1d**). 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 (Medema et al. 2015). 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 (Medema et al. 2014).

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 (Thaker et al. 2013). 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 (Johnston et al. 2016) of curated hidden Markov models for antibiotic resistance proteins (in part derived from the Resfams (Gibson et al. 2015) database) (see *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.

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) (Skinnider et al. 2017) (**Fig 1.5; Fig 1.6**). 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. 1.2a**).

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 (Koskiniemi et al. 2013). The *Candidatus* Eelbacter genome contained six clusters that were complex type I NRPS-PKS hybrid systems over 45 kb in length (**Fig. 1.2b**). 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 (Claessen et al. 2006; Zhang et al. 2012) or an association with a eukaryotic host (Wilson et al. 2014). 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 (Wilson et al. 2014). 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 (see *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 (Unger et al. 2010). 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 (Bray et al. 2016). 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). Expression of NRPS and PKS clusters was detected in all four phyla that we studied, and 84 active clusters were detected in Acidobacteria. 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 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. 1.3a; Fig 1.7**). 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 (Klingenberg and Meinicke

2017) (DESeq2 (Love et al. 2014); $P < 0.05$; FDR = 5%). Notably, the expression of genes from several gene clusters in *Candidatus* Angelobacter showed a statistically significant increase 12–24 h after substrate addition (**Fig. 1.3a**), and we found that the expression of several biosynthetic genes of *Candidatus* Angelobacter was temporally distinct from the expression of core ribosomal genes (**Fig. 1.3b**). 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.

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 (Langfelder et la. 2008) (see *Methods*), across the 120 microcosm time-point samples. Co-expressed genes often share biological functions and regulation (Stuart et al. 2003). 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. 1.3c**). 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 a TonB-dependent iron siderophore uptake receptor. 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-dependent receptor 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 (Berdy et al. 2005). 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.

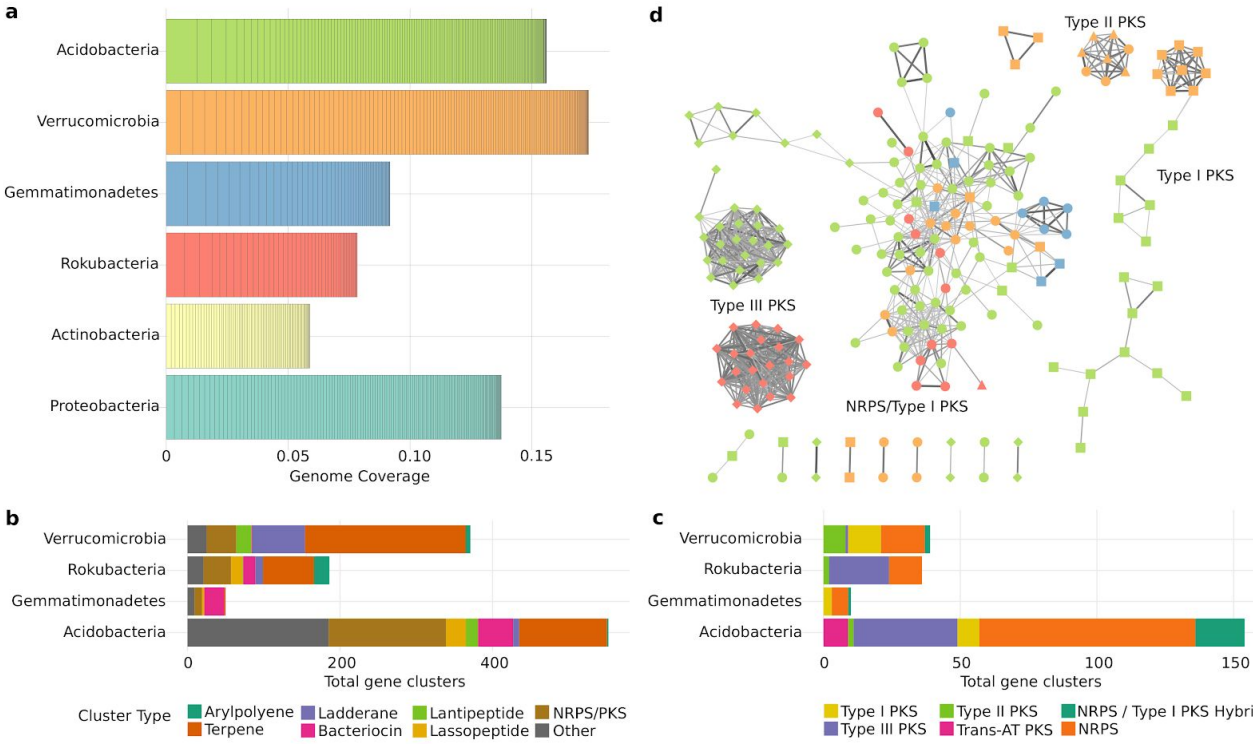

Fig. 1.1 | Diversity of extracted soil genomes and their biosynthetic gene clusters. **(A)** Mean relative abundances of reconstructed genomes across 60 soil samples as determined by sequencing coverage of the genomes. Genomes from four understudied soil phyla are juxtaposed with recovered genomes from 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.

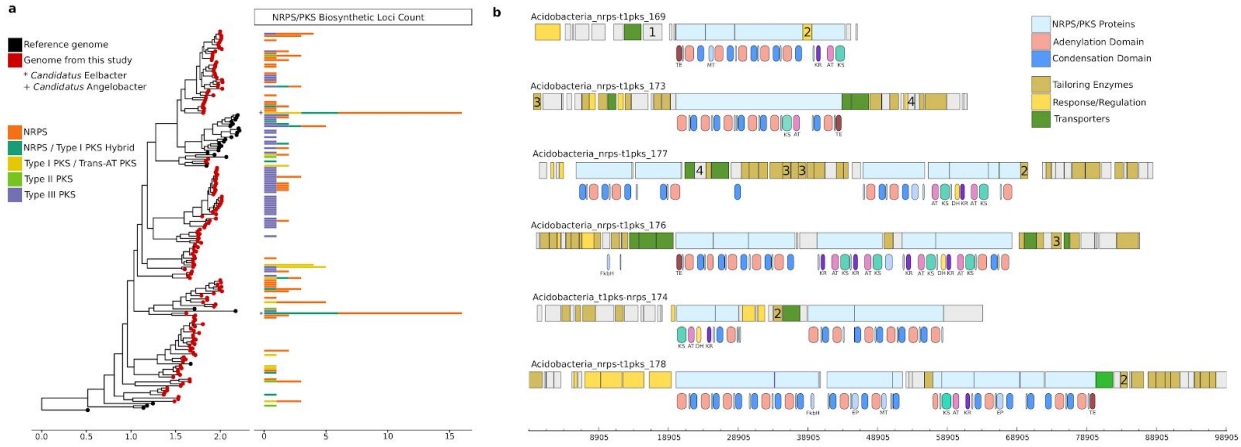

Fig. 1.2 | Biosynthetic NRPS and PKS loci from the Acidobacteria. **(A)** Concatenated ribosomal protein phylogenetic tree of all acidobacterial genomes from this study (red) and existing reference genomes (black). Scale bar on the tree represents substitutions per site. Adjacent is a chart that reflects 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.

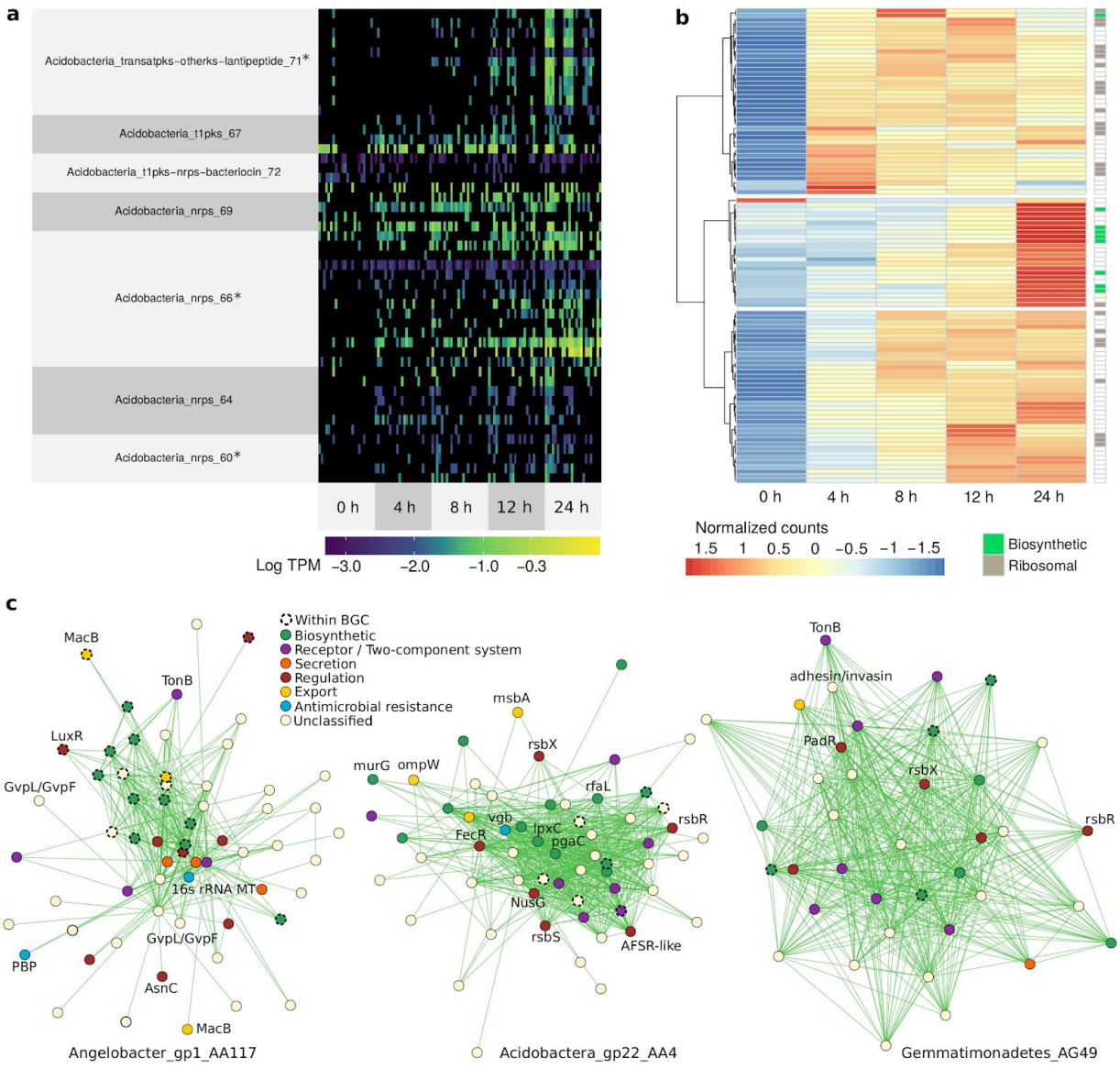

Fig. 1.3 | Metatranscriptomics of biosynthetic genes. (A) Levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the Candidatus Angelobacter genome, across 120 microcosm soil samples grouped by extraction times (reported in hours). Expression levels are reported in log 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.

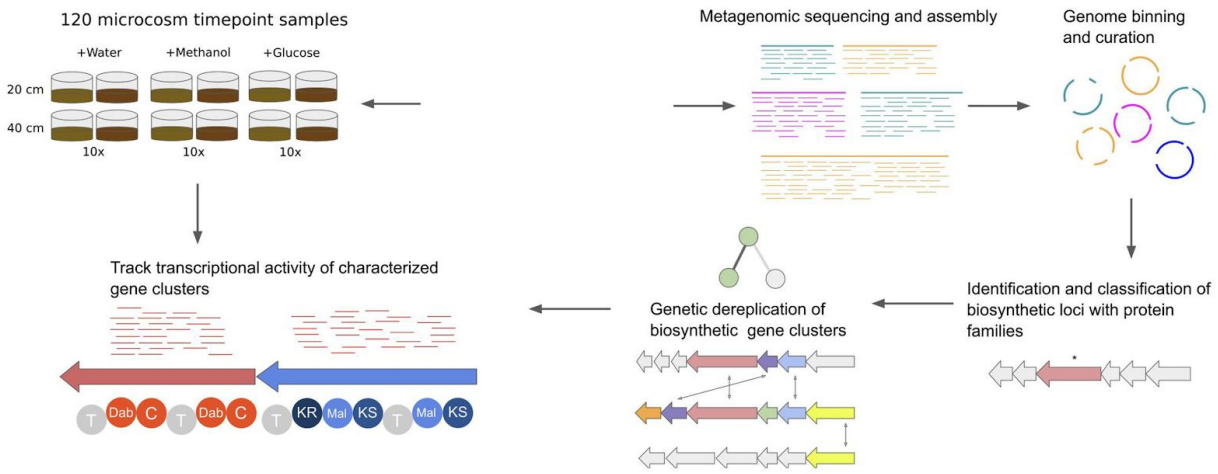

Fig. 1.4 | Experimental plan and project overview. Schematic showing major components of microcosm time-point sampling and metagenomic analyses.

Fig. 1.5 | 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.

Fig. 1.6 | 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.

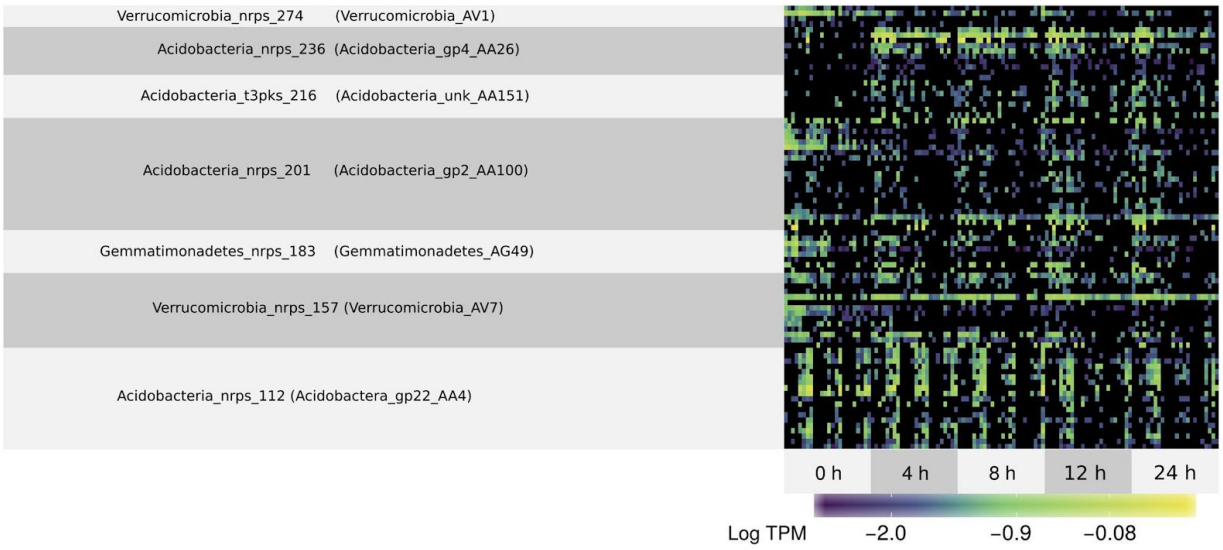

Fig. 1.7 | Differentially expressed biosynthetic gene clusters over time. The levels of expression of biosynthetic gene clusters from all organisms studied (excluding Candidatus Angelobacter data shown in Fig. 3a) that were found to be significantly differentially expressed between time points (PERMANOVA; 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.

2. A widely distributed genus of soil Acidobacteria genomically enriched with biosynthetic gene clusters

Alexander Crits-Christoph, Spencer Diamond, Basem Al-Shayeb, Luis Valentin-Alvarado, and Jillian F. Banfield

Unpublished, 2020

Bacteria of the phylum Acidobacteria are the most abundant microorganisms across soil ecosystems, yet they are represented by few genome sequences, leaving gaps in our understanding of their metabolic diversity. Recently, a surprising repertoire of specialized metabolite gene clusters was reported in genomes reconstructed from uncultivated soil and marine Acidobacteria, but the degree to which this is widespread is still unknown. To investigate this, we augmented a dataset of publicly available Acidobacteria genomes with 68 genomes recovered from the saturated soils of a vernal (spring) pool ecosystem in Northern California. We recovered high quality genomes for three novel species from *Candidatus Angelobacter* (a Group 1 Acidobacteria genus) which were previously shown to be genomically enriched in genes for specialized metabolite biosynthesis. One of these was the most abundant bacterial species in some samples. We detected numerous and highly diverse biosynthetic gene clusters (BGCs) in these genomes, and also in publicly available genomes for other related *Angelobacter*. Metabolic analysis indicates that *Angelobacter* have an aerobic lifestyle and are capable of complex carbon degradation, with some carbohydrate active enzymes rarely found in other Acidobacteria. Using metatranscriptomics, we identified in situ expression of both primary metabolic and specialized metabolic traits for two species from this genus. In conclusion, we expand genomic sampling of the uncultivated *Angelobacter* genus, and show that it represents a common and sometimes highly active member of some soil communities, with a high degree of capacity for synthesis of diverse specialized metabolites.

2.1 Introduction

It is estimated that an overwhelming majority of soil bacterial species have thus far been recalcitrant to cultivation (Steen et al. 2019), and these uncultivated bacteria are not evenly distributed across the tree of life (Lloyd et al. 2018). While many phyla of bacteria found in soils have few cultivated representatives, there are few as ubiquitous and diverse as the Acidobacteria (Kielak et al. 2016). It

is from metagenomic and 16S rRNA surveys that we have learned that Acidobacteria are collectively the most abundant phylum in soils (Fierer 2017), and there is significant taxonomic diversity in soils (Lee et al. 2008; Kielak et al. 2009), with over 26 accepted subdivisions (Barns et al. 2007). However, they are woefully undersampled in cultivation efforts, with only 72 sequenced isolate genomes from the entire phylum deposited into the RefSeq database as of the start of 2021. Some reported isolates have also not been genomically sequenced or deposited into public strain collections, complicating the study of even previously cultivated members of the phylum. Isolate-based studies of soil Acidobacteria indicate that they are heterotrophic, aerobic and capable of complex carbon degradation, and are thought to be mostly oligotrophic (Kielak et al. 2016). However, it is unclear to what degree these findings extrapolate to the entire phylum.

More recently, genome-resolved metagenomics, or the process of assembling and curating genomes directly from metagenomes, has been applied to soil bacterial communities and resulted in the assembly of hundreds of novel soil Acidobacteria genomes (Diamond et al. 2019; Woodcroft et al. 2018; Xue et al. 2020). Genome annotation and functional prediction from these genomes have improved our understanding of the phylum's metabolic potential. Metabolic characteristics predicted from genomes of uncultivated Acidobacteria include the ability to degrade many complex carbohydrate compounds via carbohydrate active enzymes, the capacity for nitric oxide reduction, and the ability to oxidize methanol (Diamond et al. 2019). It also was previously shown that many Acidobacteria genomes from metagenomic data from a single soil ecosystem encode numerous gene clusters for the biosynthesis of specialized metabolites (Crits-Christoph et al. 2018). In particular, two particular lineages of Acidobacteria in subgroup 1 and 4, designated Candidatus *Angelobacter* and Candidatus *Eelbacter*, respectively, were found to possess unusually large repertoires of biosynthetic genes. Specifically, the genomes each encoded for 300-400 Kb of nonribosomal peptides synthetases (NRPSs) and polyketides synthases (PKSs). A third lineage of uncultivated Acidobacteria was reported which also possessed similarly large NRPS and PKS gene clusters, and was sequenced from ocean biofilm samples (Zhang et al. 2019). Additional genomes from this third lineage have since been noted for their unusually large nonribosomal peptide genes as well (Nayfach et al. 2020). However, from the genomes reported thus far, it remains unclear how these three lineages are related, or whether they are widespread in soil environments. Here, we extend these results using publicly available Acidobacteria genomes from a variety of soil types, and augmented the dataset by sampling and new metagenomic analysis of saturated soil, which remains largely understudied to date. Of the previously reported Acidobacteria lineages enriched in biosynthetic gene clusters, we found Candidatus *Angelobacter* to be the most broadly distributed, and focused our analysis on genomes from this group. Our results substantially increase the number of reported genomes from the Candidatus *Angelobacter* genus and suggest that a significant investment in secondary metabolism is a common feature of soil bacteria from this lineage.

2.2 Results

Sampling and metagenomic sequencing

We collected 29 soil samples from a seasonal vernal pool in Lake County, California, USA in October 2018 and October 2019. The elevation of the site is approximately 600 m and the pool is surrounded by Douglas Fir and Oak (**Fig 2.1b**). The soils at the site are fine grained, organic-rich mud and clay-rich at depth, surrounded by gravelly loam Inceptisols formed in material weathered from rhyolitic tuff. Sampling occurred when the pool was at its driest before the first major autumn rainfall, along a transect in the pool bed that would be covered by water for a majority of the year. Total nitrogen and total carbon measured at the site averaged 1% and 13% respectively, both decreased with increasing soil depth. All samples were saturated with water at the time of collection. Samples were collected from 25, 35, 45, 60, and 80 cm depths, and either stored on dry ice for DNA extraction or flash frozen in ethanol cooled with dry ice for RNA extraction.

Sample metagenomes were deeply sequenced on the NovaSeq, with an average of 9 Gb sequenced per sample for 2018 and 20 Gb per sample for samples collected in 2019. Metagenomic assembly resulted in XX Gb of assembled sequence in contigs > 1 Kb. Using these assemblies, we generated 230 dereplicated bacterial and archaeal genomes of at least medium-quality (>90% complete, <10% contaminated).

Community composition and assembled genomes

In order to assess the community composition, phylogenetic inference of the L6 ribosomal protein was used as a taxonomic marker. The bacterial communities at the site were found to be much less complex than the soil communities observed by a previous metagenomic effort in a arid grassland meadow, with an average of 50 species-level L6 ribosomal sequences assembled per sample. Dominant taxa included species in the phyla Acidobacteria, Proteobacteria, Chloroflexi, with a variety of Archaea dominating the community at deeper depths of 60-80 cm (**Fig 2.1a**). Among bacteria at high taxonomic ranks, community composition was fairly consistent across depths. Curiously, members of the Candidate phyla radiation consistently composed ~10% of the community, when they have been observed to be rarer in less saturated soil environments.

Of particular interest was the high abundance of the phylum Acidobacteria within the vernal pool soil microbial community, in some samples reaching 25% relative abundance. We assembled 68 near-complete species-dereplicated Acidobacteria genomes from the site. We placed these genomes in a phylogenetic tree of all 370 known Acidobacteria species in the NCBI Assembly database and the Genome Taxonomy Database (GTDB) using a concatenated set of ribosomal proteins (**Fig 2.2**). Genomes were obtained from five different Acidobacteria classes (group 1, 3, 5, 7, and 11), indicating a wide diversity of species abundant in this soil.

Phylogenetic analysis of single copy marker genes identified three novel genomes related to an uncultivated group 1 Acidobacteria, *Candidatus Angelobacter* (GTDB genera g__Gp1-AA17; NCBI taxon 'Acidobacteria bacterium AA117'). Besides the previously published genome from another site in Northern California, there is only one other metagenome assembled genome from this clade, recovered from a thawing permafrost peatland located in arctic Sweden. The three new genomes obtained from the vernal pool study site were all near-complete with low estimated contamination and ranged in size from 6-7 Mb, with an average GC content of 55%. One of these genomes, *SRVP-Angelobacter-2*, was 6.44 Mb total across 39 contigs, which is the most contiguous assembly of a *Candidatus Angelobacter* species, and significantly more contiguous than the previously published genome. *Angelobacter* were at reasonably high abundances in all samples, and the *SRVP-Angelobacter-3* species was the most abundant organism samples from 20 cm depth.

To further expand our characterization of the *Angelobacter* genus, we searched the IMG database of assembled metagenome bins for additional genomes from the genus by phylogenetic placement of all Acidobacterial genomes in the dataset. We found additional high quality genomes for five more species in the *Candidatus Angelobacter*, all from soils. Three were from a large metagenomic study of corn and switchgrass rhizosphere in Michigan (Howe et al. 2014), one from a metagenomic study of soils amended with Pyrogenic organic matter in New York (Whitman 2016), and one genome was obtained from a mini-metagenomic selection approach from Massachusetts forest soils (Alteio et al. 2020)

Many species of *Candidatus Angelobacter* genus possess diverse biosynthetic gene clusters

The previous genome reported from *Candidatus Angelobacter* was notable in its substantial genomic capacity for production of specialized metabolites, particularly via large biosynthetic gene clusters of nonribosomal peptide synthetases and polyketide synthases. To understand how consistent this trait is across this genus, we ran antiSMASH 5.0 on all of the recovered genomes to identify Biosynthetic Gene Clusters (BGCs) and predicted both polyketide keto-synthase (KS) and NRPS condensation (CD) protein domains across antiSMASH BGCs. When visualizing the number of these biosynthetic domains per genome, it is clear that the *Angelobacter* genus stands out amongst the acidobacterial phylum (**Fig 2.2**). We also identify the phylogenetic placement of two other independent acidobacterial clades with large numbers of biosynthetic enzymatic domains: Candidatus Eelbacter is a group 4 Aacidobacterial genome previously reported. There is also a lineage represented by genomes previously reported from ocean biofilm and soil metagenomes that cluster phylogenetically with each other and are identified by similarity to the UBA5704 genome. Within the *Angelobacter* clade, we also identified a minority of genomes with few biosynthetic gene clusters, indicating that genomic capacity may vary widely within this lineage. This is also the case for Actinomycetes, which are renowned for specialized metabolite production. It is also possible that the BGCs were not assembled or binned properly in some genomes.

The total number of BGCs in newly recovered *Angelobacter* genomes rivals, and in many cases outnumbers, the previously reported *Angelobacter* genome's biosynthetic gene content (**Fig 2.3a**). The genome recovered with the most BGCs, Angelobacter-SRVP1, was also one of the most contiguous genomes reported (with only 39 scaffolds), indicating that this high BGC count was also not due to contig fragmentation. The majority of all *Angelobacter* BGCs were NRPS or NRPS-PKS hybrids with unusual complexity and length. Many of these NRPS genes in the clusters were unusually large, with the largest ORF in the genus ranging up to 24 Kbp in length.

To clarify whether *Angelobacter* species tend to share similar BGCs, we ran the program BiG-SCAPE on the recovered *Angelobacter* BGC collection to identify gene cluster families, or groups of related BGCs. We found that the majority of BGCs in the collection were singletons, indicating substantial genetic diversity and comparatively few BGCs shared between species (**Fig 2.3b; Fig 2.3c**). Of BGC families that were shared by species, the majority were only shared by two species; only five clusters were shared amongst more than three *Angelobacter* species. The gene cluster families that were commonly shared include Terpene, Type I PKS, NRPS, and a ribosomally synthesized peptide gene cluster family. While most of the shared gene cluster families had fewer genes than average, intriguingly, two *Angelobacter* species from different study sites shared a highly similar large NRPS cluster with similar, but non-identical adenylation domain structure (**Fig 2.3b**).

Candidatus Angelobacter species are heterotrophic aerobes with genes for carbon degradation

Genomic inferences about primary metabolisms can help inform understanding of a microorganisms' lifestyle and trophic niche, while also possibly guiding cultivation efforts. We characterized key metabolic genes of the three Candidatus Angelobacter species from the vernal pool soil, as well as the other 65 acidobacterial genomes obtained from the site. We found that the primarily metabolisms of the three Angelobacter species were not particularly unusual amongst the acidobacterial genomes obtained from the vernal pool. We predict that they are heterotrophic, without an identified pathway for carbon fixation, and possess formate dehydrogenases. We identified operonic cytochrome c genes (CoxABCDE) likely involved in aerobic respiration, and also an operon with a nitrate reductase operon with a narG homolog that may indicate facultative anaerobic nitrate reduction. One of the Angelobacter genomes (SRVP-Angelobacter-2) from the vernal pool soil contained a Type IC CRISPR-Cas array and another (SRVP-Angelobacter-3) contained a Type IIID CRISPR-Cas system.

Candidatus Angelobacter species are transcriptionally active in situ

To track transcriptional activity of Angelobacter *in situ*, we flash-froze soil samples in the field to preserve for metatranscriptomic RNA extraction, extracting and sequencing 20 Gbp of RNA for 10 samples taken from the vernal pool study site in 2019. Mapping both DNA and RNA reads back to genomes obtained from the site, we were able to track both relative abundance (DNA) and relative

transcriptional activity (RNA) for microbes of interest. Across organisms, we found that microbial genomes that recruited more DNA reads tended to also recruit more RNA reads. Two *Candidatus Angelobacter* species (SRVP-2 and SRVP-3) were found to be more abundant and more transcriptionally active than the mean or median bacterial/archaeal species genome obtained from the site (**Fig 2.4a**). We next compared total expression for these two *Candidatus Angelobacter* species by soil depth of the sample, and found that the species were most active at 5 cm depths, followed by 30 cm depths, and were least active in 80 cm deep soils (**Fig 2.4b**).

We next identified and annotated the 50 most abundant transcripts from either *Candidatus Angelobacter* species in the dataset at both shallow (5-20 cm) and deeper depths (30-80 cm) (**Fig 2.4c**). Intriguingly, we observed a strongly unique transcriptomic profile in one sample taken from a 5 cm depth, in which several stress-related genes were highly expressed that remained largely unexpressed across the other samples. The highly expressed set of genes were clustered by their transcriptional abundances across samples, and we did not observe uniform patterns clustering by sample depth. However, three samples collected at a depth of 30 cm did cluster together, with high expression for a set of genes that seemed to be involved in growth and general metabolism: ribosomal proteins, formate dehydrogenase, and a chromosomal segregation protein. These data are consistent with Candidatus Angelobacter transcriptional activity in situ, with activity highest at shallower soil depths, and possibly strong changes in transcriptional regulation between samples, similar to the kind previously reported in a previously metatranscriptomic experiment performed on soils with the originally reported strain.

2.3 Methods

Field sampling

We collected soil samples from a seasonal vernal pool in Lake County, California, in October 2018 and October 2019. Samples were frozen in the field using dry ice, and kept at -80 C until extraction. The Qiagen PowerSoil Max DNA extraction kit was used to extract DNA from 10 g of soil, and the Qiagen AllPrep DNA/RNA extraction kit was used to extract RNA from 2 g of soil. Samples were sequenced by the QB3 sequencing facility at the University of California, Berkeley on a NovaSeq 6000. Read lengths for the 2018 DNA samples and the RNA samples were 2x150 bp and 2x250 bp for the 2019 DNA samples. A sequencing depth of 10 Gb was targeted for each of the 2018 samples, and 20 Gbp for each of the 2019 samples.

Metagenomic assembly and annotation

Metagenomes were quality trimmed using Sickle (Joshi and Fass 2011) and assembled using the IDBA_UD assembler (Peng et al. 2012) . Contigs greater than 2.5 Kb were retained and sequencing

reads from all samples were cross-mapped against each resulting assembly using Bowtie2 (Langmead and Salzberg 2012). The resulting differential coverage profiles were filtered at a 95% read identity cutoff, and then used for genome binning with MetaBAT2 (Kang et al. 2015). Resulting genome bins were assessed for completeness and contamination using CheckM (Parks et al. 2015), and were manually curated using taxonomic profiling with ggKBase. Taxonomy was assigned to bins using phylogenetic placement of single copy marker genes with GTDB-Tk (Parks et al. 2018). Community relative abundance profiles were determined for each sample using the GraftM metagenomic classifier and the ribosomal protein L6 marker gene.

Biosynthetic gene clusters were annotated in genomes using antiSMASH 5.0 (Blin et al. 2019). The number of Condensation domains and Ketoacyl synthase domains was determined by using HMMER3 (Eddy 1998) and querying all antiSMASH predicted biosynthetic proteins using the PF00109 and PF00668 Pfam HMMs. BiG-SCAPE (Navarro-Muñoz et al. 2020) was used to generate gene cluster families of BGCs. KEGG functional annotations for genes across the entire genomes were obtained using METABOLIC.

Metatranscriptomic data was mapped to all genome bins using Bowtie2, and filtered to only paired end reads with >95% identity using Pysam. Read counts were then log-normalized in R.

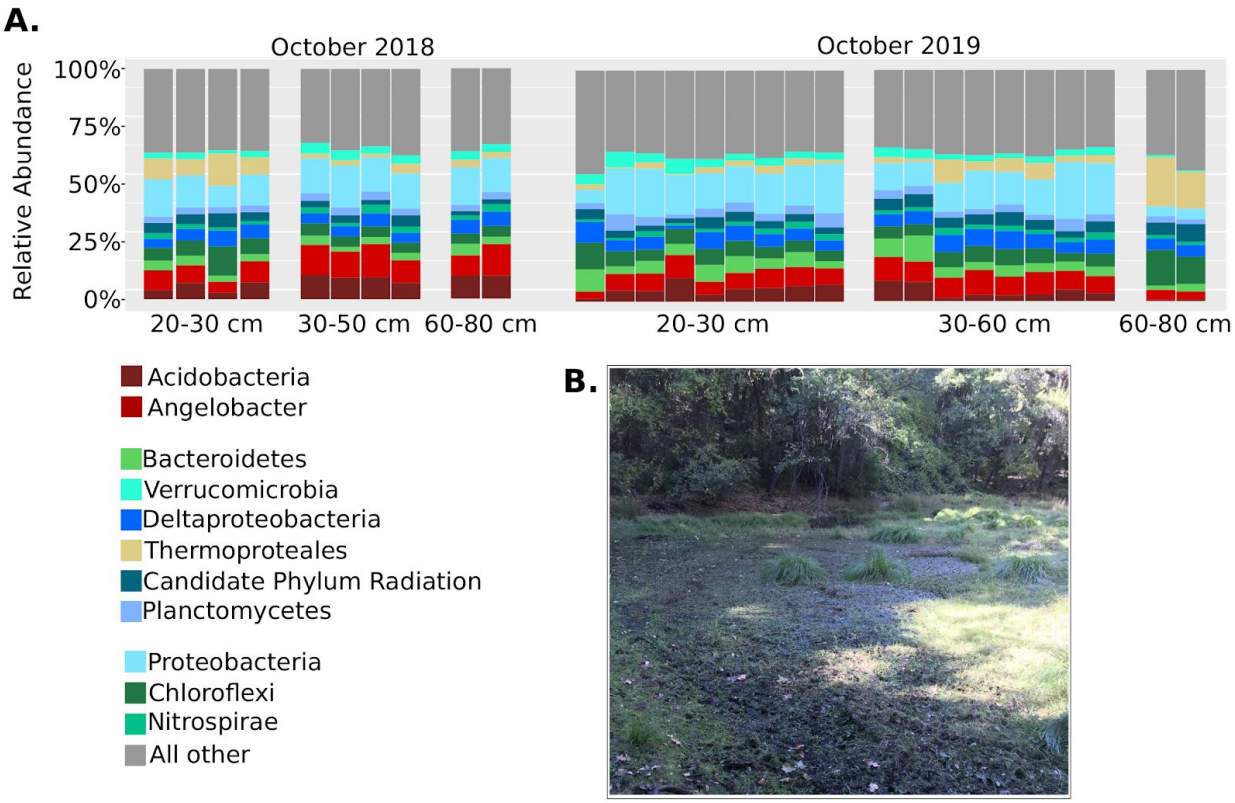

Fig. 2.1 | Bacterial/Archaeal community compositions of soils from a vernal pool. (**A**) Ribosomal protein (L6) abundances and taxonomic classifications across all metagenomic samples obtained in this study. The abundances of *Candidatus Angelobacter* (*Gp1-AA117)* are shown separately from all other hits in phylum Acidobacteria. (**B**) Photograph of the vernal pool that was metagenomically sampled in this study, in Lake County, California, USA.

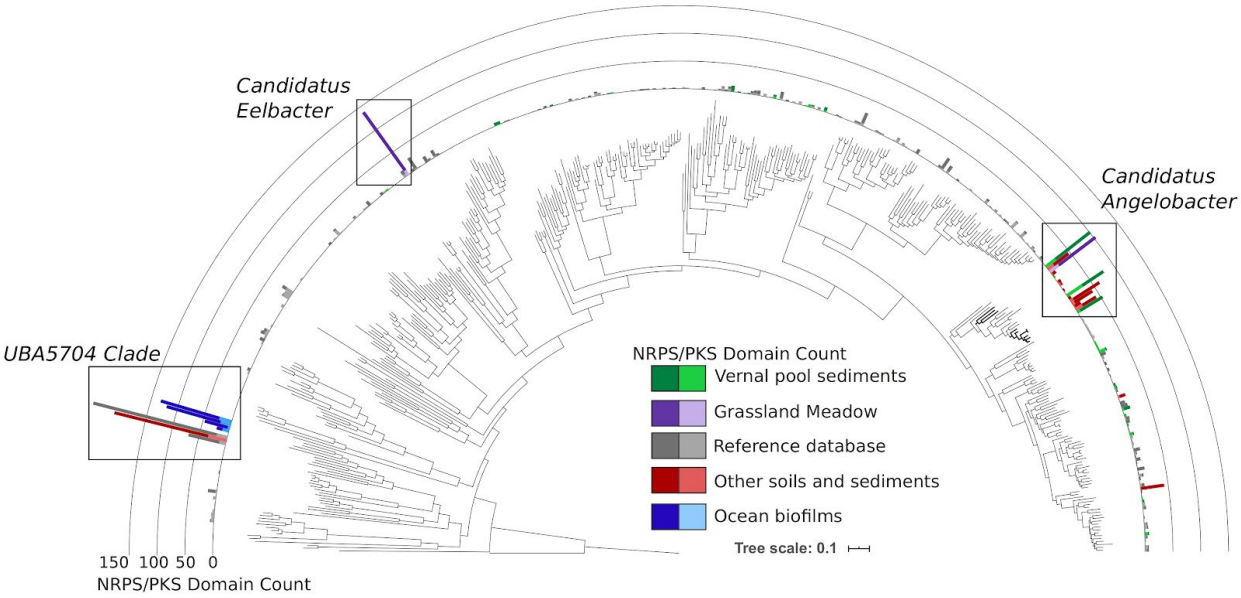

Fig. 2.2 | Phylogeny of the Acidobacteria annotated with biosynthetic domain counts per genome. Concatenated ribosomal protein phylogeny of all Acidobacteria genomes in NCBI GenBank, additional genomes obtained from IMG, and genomes obtained from this study ("Vernal pool sediments"). Plotted is the number of biosynthetic NRPS CD /PKS KS domains per genome, and genomes are colored by their ecosystem of origin.

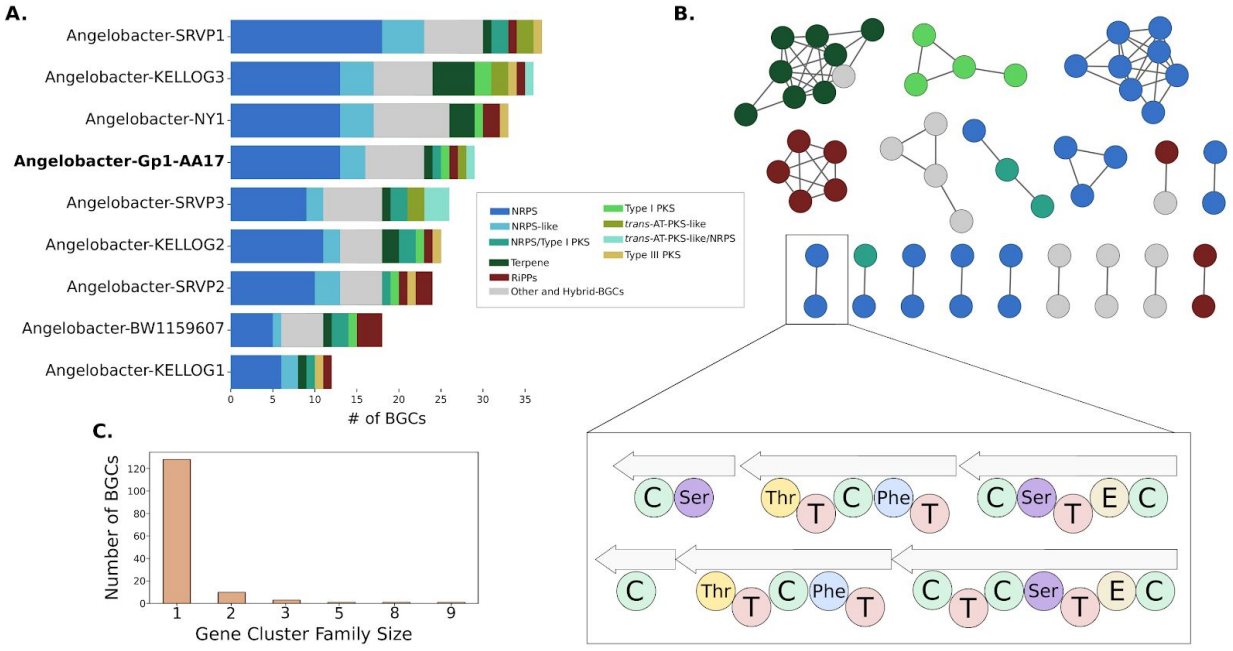

Fig. 2.3 | Biosynthetic gene clusters from genomes in Candidatus Angelobacter. **(A)** The number and class of BGCs in each genome in Candidatus Angelobacter. The previously published reference genome for this family is highlighted in contrast to new genomes obtained in this study. **(B)** A BiG-SCAPE gene cluster family network of BGCs from Candidatus Angelobacter. Each node is a BGC, connected to other similar BGCs by genomic similarity. Two core NRPS genomes from different species are shown in the inset. Nomenclature: C, Condensation Domain; Ser, Adenylation Domain (Serine); The, Adenylation Domain (Threonine); T, Peptidyl Carrier Protein; Phe, Adenylation Domain (Phenylalanine). E, Epimerase. **(C)** The number of Angelobacter BGCs in a gene cluster family of a certain size. The vast majority of BGCs are novel singletons.

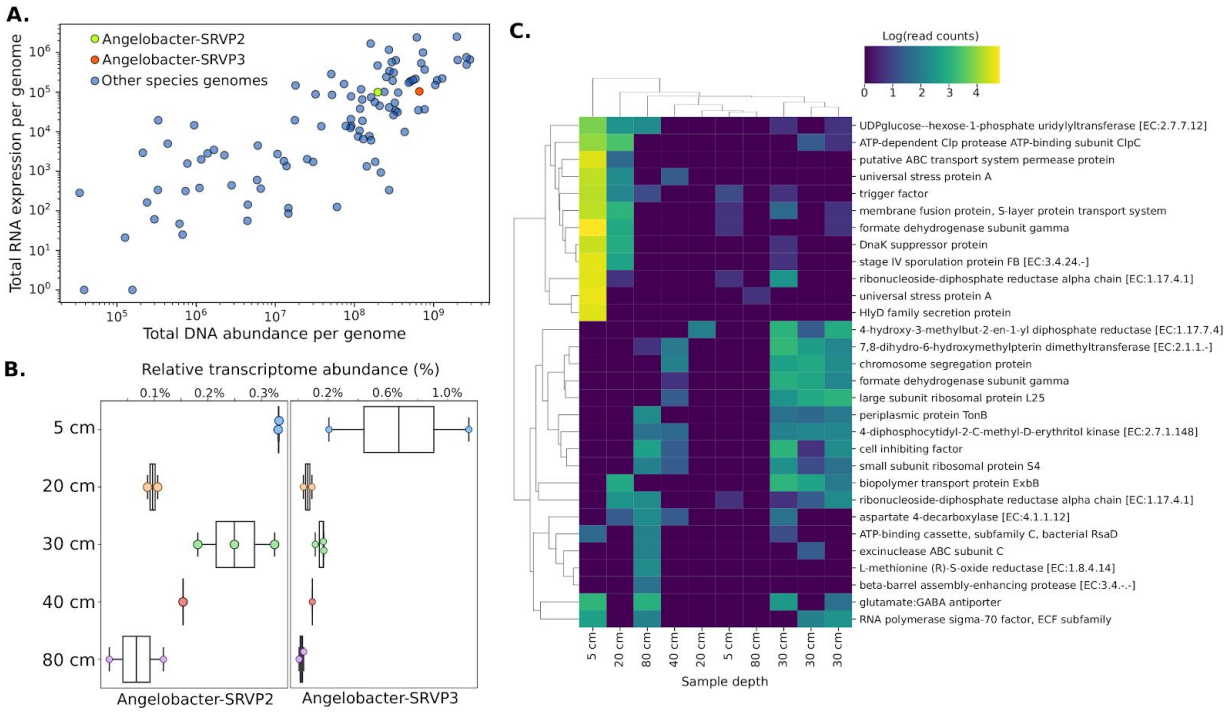

Fig. 2.4 | Metatranscriptomic activity of *Candidatus* Angelobacter species in situ. **(A)** Total vernal pool sediments metatranscriptomic reads mapping to two *Candidatus* Angelobacter species, compared to total metatranscriptomic activity for all other microbes with genomes obtained from the site. **(B)** Total transcriptional activity for two *Candidatus* Angelobacter species, compared by sediment depth of sampling. **(C)** Counts of highly expressed Kegg Orthologs in each KEGG category with at least 10 highly expressed KOs. **(D)** Counts of KOs in each KEGG module with at least 3 KOs with above-average expression.

3. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow

Alexander Crits-Christoph, Matthew R. Olm, Spencer Diamond, Keith Bouma-Gregson & Jillian F. Banfield

Published in *The ISME Journal*, 2020

Soil microbial diversity is often studied from the perspective of community composition, but less is known about genetic heterogeneity within species. The relative impacts of clonal interference, gene-specific selection, and recombination in many abundant but rarely cultivated soil microbes remain unknown. Here we track genome-wide population genetic variation for 19 highly abundant bacterial species sampled from across a grassland meadow. Genomic inferences about population structure are made using the millions of sequencing reads that are assembled de novo into consensus genomes from metagenomes, as each read pair describes a short genomic sequence from a cell in each population. Genomic nucleotide identity of assembled genomes was significantly associated with local geography for over half of the populations studied, and for a majority of populations within-sample nucleotide diversity could often be as high as meadow-wide nucleotide diversity. Genes involved in metabolite biosynthesis and extracellular transport were characterized by elevated nucleotide diversity in multiple species. Microbial populations displayed varying degrees of homologous recombination and recombinant variants were often detected at 7–36% of loci genome-wide. Within multiple populations we identified genes with unusually high spatial differentiation of alleles, fewer recombinant events, elevated ratios of nonsynonymous to synonymous variants, and lower nucleotide diversity, suggesting recent selective sweeps for gene variants. Taken together, these results indicate that recombination and gene-specific selection commonly shape genetic variation in several understudied soil bacterial lineages.

3.1 Introduction

Soil microbial communities play key biogeochemical roles in terrestrial ecosystems (Fierer 2017). Within a single hectare of temperate grassland soil, there can be over 1000 kg of microbial biomass and corresponding large microbial population sizes (Fierer et al. 2009). Recent progress has been made in cataloging the diversity of 16S rRNA genes in soils (Thompson et al. 2017), which is useful

for understanding microbial community composition, but this technique is incapable of discerning most genetic variation within populations (Chevrette et al. 2019). In addition, many of the most common soil microorganisms, such as the highly abundant *Acidobacteria*, *Verrucomicrobia*, and *Gemmatimonadetes* phyla, are underrepresented in or nearly absent from culture collections and genomic databases, even at the level of class or phylum (Fierer 2017; Lloyd et al. 2018; Bergmann et al. 2011). For these reasons, the processes of recombination and selection in many of the most globally prolific soil microbial phyla remain unstudied. However, genome-resolved metagenomics, in which shotgun sequencing of metagenomic DNA is assembled and binned into draft genomes, has recently resulted in whole genome characterization of these rarely cultivated but widespread soil bacteria (Hultman et al. 2015; Butterfield et al. 2016; White et al. 2016; Ji et al. 2017; Woodcroft et al. 2018). Sequence variation within genome-resolved metagenomic datasets can therefore be used to track changes in allele frequencies, and to infer the operation of evolutionary forces of genetic drift, natural selection, and homologous recombination in these natural populations (Garud et al. 2019; Whitaker and Banfield 2006).

Homologous recombination can vary dramatically in its importance relative to the other processes for different microbial species. Analyses of reference genomes have shown that homologous recombination frequently occurs in bacteria populations, both globally and locally (Gonzalez-Torres et al. 2019; Sakoparnig et al. 2019; Lin and Kussell 2019). For example, certain populations of hotspring *Cyanobacteria* approach panmixia, where recombination is so frequent that individual cells are unlinked random mixtures of alleles (Rosen et al. 2015). In other species-like oceanic *Vibrio*, recombination is high but large blocks of alleles important for ecological niche differentiation are co-inherited and may remain linked due to selection (Cui et al. 2015; Shapiro et al. 2012). In soils, *Streptomyces flavogriseus* isolates were also found to approach a freely recombining panmixia (Doroghazi and Buckley 2010). In contrast, *Myxoccocus xanthus* isolates recovered from a series of soil samples were distinct but highly clonal, implying recombination between strains was low (Wielgoss et al. 2016). However, despite their frequent cultivation, cultivation-independent studies show that those taxa are comparatively rare in soils compared with undercultivated members of the *Acidobacteria*, *Gemmatimonadetes*, and *Verrucomicrobia* (Fierer 2017). The degree of recombination in these rarely cultivated but abundant soil bacterial lineages has not been investigated, but recombination may be widespread, as high cell densities could promote the sharing of genetic material via transformation (Thomas and Nielsen 2005), conjugation (Rocha et al. 2005), or the uptake of extracellular vesicles (Tran and Boedicker 2019).

When recombination rates are low or selection is extremely strong, several clonal strains compete until one or more beneficial alleles is highly selected for, resulting in a single clonal genotype increasing in abundance or even sweeping to fixation. However, when recombination unlinks gene variants within a population, beneficial alleles can sweep through a population independent of genomic context in a selective sweep (Smith and Haigh 2007) with gene/locus-specific effects. One

genome-resolved metagenomic study observed a single clonal sweep over a 9 year period for one *Chlorobium* population in a freshwater lake (Bendall et al. 2016), while most of the other bacterial populations studied possessed genomic loci with unusually few SNPs (single nucleotide polymorphisms), an observation interpreted as evidence for gene-specific selective sweeps. However, positive selection acting on a genomic locus in a recombining population can also leave additional locus-specific signals, including higher linkage disequilibrium (a strong association between alleles) and different allele frequencies between populations (Krause and Whitaker 2015; Shapiro et al. 2009). In soil microbial populations, comparative frequencies of gene-specific sweeps versus genome-wide clonal strain competition and replacement are largely unknown, as are the spatial scales at which these evolutionary processes can occur.

Previously, we conducted a large scale genome-resolved metagenomics study of soils from a grassland meadow in the Angelo Coast Range Reserve in northern California that established a dataset of 896 phylogenetically diverse microbial genomes dereplicated by species, and reported on community composition (Diamond et al. 2019). The soil at the site is a sandy loam mixture of ~45% clay-, ~45% silt-, and 10% sand-sized particles, with pH values in the range of 4.6–4.9 (Butterfield et al. 2016). The mineralogy of the soils at the site was previously found to be predominantly vermiculite, with plagioclase and alkali feldspars and minor apatite (Berhe et al. 2012). The soils were classified as Ultic Haploxeralfs of the Holohan-Hollowtree-Casabonne complex, and at a depth of 30 cm were found to have a bulk density of ~2.0 g/cm^3, a cation exchange capacity between 17 and 19 meq per 100 g soil, C:N ratios from 10–12, and total C concentrations of 10 mg g^{-1} up to 18 mg g^{-1} (Berhe et al. 2012). The grassland is dominated by annual Mediterranean grasses and forbs (Butterfield et al. 2016; Subtle et al. 2007). The meadow has been part of a rainfall amendment climate change study ongoing for over 17 years and has also been studied in the context of plant diversity and productivity (Sullivan et al. 2016), invertebrate herbivores and predators (Suttle et al. 2007), fungal communities (Hawkes et al. 2011), soil organic matter (Behre et al. 2012), metabolomics, and metaproteomics (Butterfield et al. 2016). By analyzing the population genomics of 19 highly abundant bacterial species across this meadow, we found high nucleotide diversity within samples, intrapopulation genetic structure often shifting over local spatial scales, varying degrees of homologous recombination for different species, and gene-specific population differentiation partially driven by selection.

3.2 Materials & Methods

Sampling, genome sequencing, and metagenomic assembly

The sampling scheme, local soil characteristics, and study design that were utilized in this analysis have been previously described (Diamond et al. 2019). Previously, 60 soil samples were collected at

depths of 10–20, 20–30, and 30–40 cm near the respective centers of six 10 m diameter plots, spaced 5 m apart. Sampling plots were spatially arranged in "blocks" of two plots across the meadow, and one of the two plots in each block received extended spring rainfall (Suttle et al. 2007). Samples were collected over a period of 2 months before and following autumn rainfall, resulting in 10 samples per plot (**Fig. 3.1a**). Briefly, DNA was extracted from 10 g of soil for each sample using the PowerMax Soil DNA isolation kit (MoBio Laboratories) from individual soil cores. Metagenomic libraries were prepared and sequenced with 2 × 250 bp paired read sequencing on the Illumina HiSeq2500 platform at the Joint Genome Institute. Reads were quality filtered to a maximimum 200 bp in length using BBduk (Bushnell 2014). Metagenomes were assembled using IDBA_UD (Peng et al. 2012) and individual genomes were binned using differential coverage binning and a suite of metagenomic binners as previously described (Diamond et al. 2019).

Genome dereplication, filtering, and comparison

The 10,538 genome bins previously described obtained from the study site were dereplicated using dRep (Olm et al. 2017) with the secondary clustering threshold -sa 0.97, and were filtered with CheckM (Parks et al. 2015) to generate a dereplicated species-level (97% ANI) genome set to be used for read mapping with >70% completeness and <10% contamination. Representative genomes for each species cluster were chosen based on the highest CheckM completeness and lowest contamination using the scoring algorithm described by (Olm et al. 2017). For this analysis, we then used the 19 species-level genome clusters with at least 12 replicate genomes that were estimated to be at least >80% complete with <10% contamination, independently assembled and binned out of different samples. Each of the 19 representative genomes was assembled from >100,000 short reads, and millions of reads could be assigned to each population from all 60 samples. Because microbial populations within the 10 g soil samples used for DNA extraction are comprised of orders of magnitude more cells than were sequenced, most read pairs are likely from unique cells (or DNA molecules) in the population. Each genome assembled from each sample was sequenced at around 10× coverage, but meadow-wide coverages for each population ranged from 224× to 908×.

Open reading frames were predicted using Prodigal (Hyatt et al. 2010) and annotated using USEARCH against UniProt (UniProt Consortium 2018), Uniref90 (Suzek et al. 2015), and KEGG (Kanehisa and Goto 2000), and HMM-based annotation of proteins using PFAM (Finn et al. 2013), and antiSMASH 4.0 (Blin et al. 2017) for biosynthetic gene prediction. PERMANOVA tests for association of ANI matrices with environmental data were run using the adonis2 function with the parameter 'by' set to 'margin' in the vegan package in R. Multidimensional scaling (MDS) plots were made with the mds function from the smacof package in R, with the ndim parameter set to 4. Hypergeometric tests for statistical enrichment of protein families was performed using HMMER annotated PFAM features in R.

Possible contamination was further removed from representative genomes using the assembled replicate genomes for each species population. The pan-genomic analysis pipeline Roary (Page et al. 2015) was run with default settings on the set of genomes for each species to identify protein clusters across genomes. Contigs with at least 50% of their protein clusters found in less than 25% of each genome set were then discarded as potential contamination (generally fewer than 20 contigs, often small, were removed per genome). Therefore, the final set of contigs used in this analysis contained only contigs that reliably assembled and binned independently for a species in multiple samples. Organism DNA relative abundances were calculated using the total number of reads that mapped to each genome in each sample and dividing by the total number of reads per sample.

Read mapping, SNP calling, and nucleotide diversity

All metagenomic reads were mapped using Bowtie 2 (Langmead and Salzberg 2012) with default parameters except for the insert size parameter -X 1000 to an indexed database of all of the 664 dereplicated genomes obtained from the environment, a database which also contains representative genomes from each of our 19 species for study. Reads that mapped uniquely to the representative species of interest were then used for analysis. Read filtering of the resulting BAM files was performed using a custom script, filter_reads.py, available at the link under code availability. Mapping files were then filtered for reads that meet the following criteria: (1) both reads in a pair map within 1500 bp of each other to the same scaffold (as a maximum possible end to end insert size), (2) the combined read pair maps with a percent identity of at least 96% to the reference, at least one of the read pairs has a mapq score >1, indicating that this is a uniquely best mapping for this read pair in the index. We further compared nucleotide diversity of genes calculated with a 96% ANI_{r2} cutoff to those at a 98% ANI_{r2} cutoff and found a strong correlation. SNPs were then called at frequencies >5% in each population using reads from all samples, using a simple null model that assumes a false discovery rate $<\sim 10^{-6}$.

For all population diversity metrics, we used reproducible custom python scripts (available under Code Availability) that calculated metrics, each explained below, from all filtered cross-sample read mappings. For each representative genome in our set of 19, we analyzed its data in samples that passed a cutoff of at least 50% of the genome being covered with at least 5× coverage. Five hundred eighty-six out of 1140 sample genome comparisons (19 genomes × 60 samples) passed this minimum requirement. Base pairs of reads with Phred scores less than 30 were not used in SNP or linkage analyses. Nucleotide diversity was calculated as the expected frequency of a difference between two sequencing reads at a position, equation $\pi = 1 - (A^2 + C^2 + G^2 + T^2)$, where A, C, G, T and the observed proportions of each respective nucleotide. This is equivalent to the definition from Nei and Li 1979 calculated separately on each genomic position with at least 5× coverage within each sample, and then averaged across genes. Sample read mappings were pooled by replicates, by plot of origin, by block and origin, and finally by all samples in the meadow, and nucleotide diversity was recalculated

on each pooled set of samples. For downstream analyses, nucleotide diversity for all samples (meadow-wide) and nucleotide diversity within each of the three sampling blocks (block-wide) was analyzed. To quantify the impact of changing sequencing coverage on nucleotide diversity, we recalculated nucleotide diversity for each genome, subsampling coverage at each genomic position. We found that the bias in nucleotide diversity due to low sequencing coverage was minimal above 50×, and our block-wide and meadow-wide coverages are often well above this coverage. We see a larger bias in nucleotide diversity when subsampling to only 5x coverage, but this bias is small compared with the biological variation observed between samples, and many of our individual soil samples are over 10×.

SNPs were called on the combined meadow-wide population set of filtered reads. We constructed a simple null model based on an error rate of 0.01% (the Phred 30 cutoff) and simulated simple sampling with replacement to construct estimated rates of erroneous SNP counts at genomic positions of varying coverages. Positions with an alternative allele that occurred with counts that had a false positive rate of ~10^{-6} with the given coverage of that site in the null model and a minimum allele frequency (MAF) of 5% were called as SNPs. Because meadow-wide coverages were at least 224× (and ranged up to 908×) the MAF cutoff alone would likely be a stringent cutoff for Phred 30 error rates. SNPs were assigned as synonymous or nonsynonymous using a custom BioPython script and the gene calls annotated by Prodigal.

Linkage disequilibrium, FST, and tests for selection

Linkage was calculated within mapped reads for all pairs of segregating sites that were spanned by at least 30 read pairs with high quality base pair data. R^2 and D' linkages were calculated using formulas described by (VanLiere and Rosenberg 2008). The relative rate of recombination to mutation (gamma/mu) for each population was calculated using the mcorr package (cite) on synonymous third position codon sites across filtered mapped reads from all samples. We analyzed 10 out of 19 genomes, as these had normally distributed residuals for the model fit and the bootstrapping mean was within 2× of the final estimate for gamma/mu. F_{ST}, (a measure of differences in allele frequencies between two populations), was calculated on sites segregating across both blocks being compared (for all three block comparisons) using the Hudson method (Hudson et al. 1992) as recommended by (Bhatia et al. 2012), as implemented in the scikit-allel package (Miles and Harding 2017). A site had to have a coverage of at least 20× in each block in order to calculate F_{ST}, and genes which had coverages in a block outside of the range of two standard deviations were excluded from the analysis. A ratio of averages was then used to determine mean F_{ST} for each gene. A two-sample Wilcoxon test was used to determine if average linkage of highly differentiated loci differed from the genomic average for each species, and two-sample *t*-tests in R were used to determine if average nucleotide diversity of highly differentiated loci differed from the genomic

average. Both sets of tests were corrected for multiple hypotheses using the Benjamini–Hochberg (Benjamini and Hochberg 1995) method.

3.3 Results

Genomic similarity within bacterial populations is spatially organized across a meadow

Starting with an undereplicated but quality filtered dataset of 3215 draft quality genomes assembled from samples collected from across the meadow (Diamond et al. 2019), we calculated all pairwise genome-wide average nucleotide identities (ANI) and alignment coverages (roughly analogous to shared gene content; **Fig. 3.1b**). We observed a sharp decrease in pairwise ANI for all of the genomes from the meadow around 96.5–97%, similar to the threshold for bacterial species delineation reported recently (Jan et al. 2018). There was a more gradual decline in shared gene content. We used the 97% ANI cutoff to cluster genomes into groups of species-like populations and found that some species-like groups contained dozens of near-complete draft genomes, each assembled from a different sample independently. To focus on the populations that were most abundant in the metagenomic data, we selected species clusters with at least 12 genomes estimated to be >80% complete with <10% contamination for population genetics analysis, which resulted in a final set of 467 genomes from 19 widespread species populations (312 genomes are estimated to be >90% complete). The bacterial species in this set included many commonly reported highly abundant soil bacteria from phyla including *Chloroflexi*, *Acidobacteria*, *Verrucomicrobia*, and *Candidatus* Rokubacteria, which are known to be abundant globally in soils, but remain understudied. Most of the species in this set were likely novel at the taxonomic rank of class, and one likely represents a novel candidate phylum tentatively designated *Candidatus* ANGP1 (Diamond et al. 2019). Based on measurement of the relative DNA abundance of each population across the entire meadow, these bacteria are some of the most abundant species in the soil, although no individual species contributed >1% of the DNA in a sample (**Fig. 3.1c**).

For each of the 19 meadow-wide populations, we tested to see if the meadow plot of origin predicted genetic similarity of the assembled genomes (PERMANOVA; FDR ≤5 %; adjusted $p \leq 0.05$). Further, we tested if genomes obtained from the same soil depth were more similar than those collected from different depths. We found that the genetic variation of genomes from 12 of the 19 populations were significantly associated with sampling plot, and that genetic variation within 5 of the 19 populations were significantly associated with sampling depth (**Fig. 3.2a**). MDS of the nucleotide identity matrices of genomes from each population shows clear associations with both plot of origin and depth (**Fig. 3.2b**). Because the genome assembly from each sample reflects the most abundant sequence variant in each population, this implies that major allele frequencies varied across the meadow for a majority of the populations. While local spatial heterogeneity has

been shown to highly explain microbial community composition in soils (O'Brien et al. 2016), here we demonstrate that there are also spatial patterns within the genetic variation of some individual species.

Population nucleotide diversity is high meadow-wide and within soil samples

To assess the genetic variability within each population across the meadow, we calculated the per-site nucleotide diversity of the sequencing reads at each locus for that population. Metagenomic studies have sometimes used either the average similarity of reads to a reference or the total number of SNPs/Mbp as metrics of genetic diversity (Bendall et al. 2016; Anderson et al. 2017). We chose to measure nucleotide diversity, because (1) it is less sensitive to large changes in coverage (2), it can be calculated both for a single site and averaged over genes or windows, and (3) it considers not only the number of SNPs but also their frequencies in the population. We found a wide range of per-gene nucleotide diversity values for the 19 different populations (**Fig. 3.3a**). Because nucleotide diversity is less sensitive to changes in coverage, we could use it to track how nucleotide diversity changes between the plots spread across the meadow. We calculated nucleotide diversity in pooled mapped reads for each population sampled from the same location and soil depth, within plot, within block (pairs of plots), and across the entire meadow for each species (**Fig. 3.3a**). Although nucleotide diversity tends to be higher at the meadow scale compared with the sample scale, the nucleotide diversity within some samples was comparable to that across the entire meadow for many populations, indicating that in some cases high nucleotide diversity persists within soils even at the centimeter scale.

In almost all populations, the ribosomal genes consistently had lower nucleotide diversity than the average genes in the genome, consistent with these genes being strongly conserved and under higher purifying selection (Jordan et al. 2002) (**Fig. 3.3a**). Biosynthetic genes involved in the production of small molecules, annotated with *antiSMASH* (Blin et al. 2017), were found to have significantly higher nucleotide diversity than the genomic average in *Chloroflexi* and one *Acidobacteria* species, while having significantly lower nucleotide diversity in one *Gemmatimonadetes* species (Welch two-sample *t* test; $q < 0.05$)(**Fig. 3.3a**). Examining all genetically diverse genes with nucleotide diversities greater than 2.5 standard deviations above the mean within each population, we find that protein families for biosynthesis of small molecules and extracellular secretion were significantly over-enriched compared with the genomic averages (hypergeometric test; $p < 0.05$) (**Fig. 3.3b**). Short chain dehydrogenase enzymes and outer membrane beta-barrel domains were significantly enriched among highly diverse genes across several taxa, whereas multiple transposon families were diversifying within *Acidobacteria* genomes. Across the *Acidobacteria*, *Gammaproteobacteria*, and *Gemmatimonadetes* species studied, secretion system proteins were also diversifying. These genes involved in biosynthesis and secretion may likely have experienced local selective pressures for diversification across soil microbial species.

Homologous recombination is common, but populations exist far from panmictic equilibria

Measuring the impact of homologous recombination on the observed genetic diversity in a population can be accomplished with metagenomic data by measuring linkage disequilibrium of SNPs spanned by paired reads (Lin and Kussell 2019; Rosen 2015). When recombination occurs within a population, the chance for a recombination event to occur between two sites on the genome increases with the distance between them, resulting in a characteristic signal known as linkage decay. Given ~200 bp reads and intra-read pair distances with a median of 383 bp and a 95th percentile of 500 bp, we could reliably assess genomic linkage of SNPs from 772 bp (median) to 846 bp (95th percentile) bp apart in each population. Consistent with the expectation that natural bacterial populations can undergo extensive homologous recombination, we observed the r^2 metric of linkage disequilibrium decay as the genomic distance between two polymorphisms increased (**Fig. 3.4a**). Using the *mcorr* package, we could estimate the neutral rate of recombination relative to mutation on synonymous third position codon sites for 10 out of 19 populations. While the estimated confidence intervals for these relative rates were large and varied between species, they were well within the ranges reported in the literature (Lin and Kussell 2019) for many known highly recombinogenic species, but generally were below the rate reported for a *S. flavogriseus* population considered to be approaching panmixia (Doroghazi and Buckley 2010).

In the less genetically diverse populations, there was a noticeable higher r^2 of synonymous variants linked to other synonymous variants than for nonsynonymous variants linked to other nonsynonymous variants (**Fig. 3.4c**). This has been previously observed in hotspring *Cyanobacteria*, and was explained as a decrease in coupling linkage for slightly deleterious nonsynonymous variants, where recombinants inheriting the doubly deleterious haplotype (of a pair of variants) are selected against (Rosen et al. 2018). In more genetically diverse populations, this ratio shifts toward nonsynonymous polymorphisms (r^2_N) having higher r^2 than synonymous (r^2_S) (**Fig. 3.4c**). One recent study reported a similar positive r_N-r_S ratio as a signature of positive balancing selection in *Neisseria gonorrhoeae* (Arnold et al. 2019); for six of the 19 populations analyzed here, the genomic average r^2_N/r^2_S ratio was greater than 1 (**Fig. 3.4c**), also indicating a greater degree of coupling linkage for nonsynonymous variants in these populations. The increase in the r^2_N/r^2_S ratio with nucleotide diversity (linear regression; $R^2 = 0.29$; $p = 0.009$) suggests an increase in the ratio of beneficial to slightly deleterious nonsynonymous SNPs as diversity increases.

Although r^2 is often used as a signal to identify the presence or absence of recombination, r^2 values <1 can occur with or without recombination. For example, three of four possible haplotypes (pairs of variants) for two biallelic sites could occur due to lineage divergence prior to mutation occurring at one of the sites. D', an alternative metric of linkage equilibrium, is only <1 if all possible combinations of a pair of biallelic sites are observed, which can only occur in the presence of recombination or recurrent mutation. Generally, we found that average D' for a population linearly

correlated with mean r^2 (**Fig. 3.4d**). Inspection of the distribution of pairs of SNPs separated by <~1 kb revealed that 3 of 4 possible haplotypes is most common, but there was detection of all four possible biallelic haplotype combinations ($D' < 1$), at 7% to 36% of all site pairs in each population (**Fig. 3.6**). The observed frequencies of the least common of the four SNP combinations also were higher than expected based on sequencing error, although much less frequent than expected from linkage equilibrium, as evidenced by mean D' values above 0.8. Nonetheless, the extensive appearance of $D' < 1$ at a significant fraction of loci, along with a signal of linkage decay with genomic distance across all populations, provides firm evidence for ongoing processes of within population homologous recombination, albeit to different degrees between organisms.

Given evidence for recent homologous recombination, we searched the genomes for genes that could confer natural competence, such as homologs of the *ComEC* gene with all three functional domains (Pimental and Zhang 2018), and identified loci with additional genes involved in DNA uptake and recombination (Cassier-Chauvat et al. 2016). We found that the presence of a ComEC homolog with multiple adjacent operonic recombination-related genes was strongly associated with the lowest values of D' (**Fig. 3.4d**). Thus, it is likely that natural competence is a common mechanism that facilitates homologous recombination for abundant soil bacteria.

Gene-specific selective sweeps contribute to divergence of alleles across the meadow

In neutrally evolving local populations, nucleotide diversity is expected to increase monotonically with population size (Nei and Li 1979). Across populations, as we did not see a relationship between nucleotide diversity and relative abundance (**Fig. 3.7**; linear regression; $p = 0.88$), purely neutral growth and processes cannot explain the observed differences in nucleotide diversity between these populations. Similarly, we also did not observe a significant relationship between diversity and abundance within species for 14 of the 19 populations (**Fig 3.7**). Except for populations with the lowest nucleotide diversity, ratios of nonsynonymous to synonymous polymorphisms for each population are consistently low. This trend, as also observed in lake metagenomics (Shapiro 2016) and whole genome comparisons (Rocha et al. 2006), further indicates that purifying selection has eliminated slightly deleterious mutations in the populations that have accrued more nucleotide diversity (**Fig. 3.8**; linear regression; $R^2 = 0.25$; $p = 0.018$). In all species except the least diverse population (an *Acidobacterium*), nonsynonymous variants also had consistently higher values of D' than synonymous variants. Further, as genome-wide D' decreased (more recombination), the degree to which nonsynonymous variants were more linked than synonymous (D'_N/D'_S) increased (**Fig. 3.6**). As the number of observed recombination events increased, nonsynonymous linkage increased in comparison to synonymous linkage. This effect is consistent with stronger selection on nonsynonymous variants with an increase in diversity: both purifying selection and positive selection would increase D' for nonsynonymous SNPs.

Soil ecosystems are exceptionally heterogeneous, and environmental factors can change over millimeter distances, potentially due to changes in aboveground plant productivity, soil geochemistry, plant litter composition, and soil particulate structure (Vos et al. 2013). While it is difficult to tease apart the effects of changing abiotic parameters over spatial scales, it is possible to examine how allele frequencies change over the scales within our study design. We calculated the pairwise fixation index F_{ST} for each gene between allele frequencies from the three meadow blocks for every species-group (**Fig. 3.1b**). For most populations, mean gene F_{ST} values were low (<5%), consistent with dispersal of most alleles between blocks. For a minority of populations, mean gene F_{ST} values were consistently >10%, indicating that there was significant geographic organization of genetic structure at most loci across the genome. Therefore, while the total variation in genome-wide major consensus alleles is often well explained by meadow geography, most individual alleles have a high chance of being found at fairly similar frequencies across the meadow.

When specific loci are characterized by significantly higher F_{ST} than the background average for the genome, it is characteristic of population-specific (in this case, spatially defined) selective pressures acting on that locus (Holsinger and Weir 2009). To identify genomic regions of unexpectedly high F_{ST}, we scanned over a moving 5 gene window and tested if that region had a mean F_{ST} greater than 2.5 standard deviations above the genomic mean. We removed genes with either coverage 2 standard deviations above or below the mean coverage in either block from this analysis. To define the length of the genomic region with elevated F_{ST} values, we extended successful windows until the mean F_{ST} fell below this cutoff. This test at first identified 48 loci of elevated F_{ST} within some microbial genomes, despite those genomes having low average F_{ST}. To further test for evidence of recent selection at these loci we looked for a statistically significant average increase in linkage and a significant change in nucleotide diversity compared with the genomic average in one or both of the blocks (**Fig. 3.5a**). We noticed that both signals of purifying selection (characterized by low N:S ratios) and a reduction in genomic coverage (potentially indicating gene loss in some portion of the population) often correlated with low nucleotide diversity in genomic regions, and based on this, caution against identifying gene-specific selective sweeps in metagenomic data based solely on a reduction in nucleotide diversity or SNP frequency. While we found many loci with unusual F_{ST} or strong changes in nucleotide diversity, our stringent criteria narrowed that set down to 8 high F_{ST} loci with significantly increased rates of linkage compared with the genomic background (**Fig. 3.5c**). These loci also had significant changes in nucleotide diversity within blocks when compared with their genomic averages (**Fig. 3.5b, c**). All of these loci showed decreases in nucleotide diversity, consistent with selective sweep events in one or multiple meadow blocks. Genes at these loci also had higher N:S ratios than genomic averages, possibly consistent with either recent selection acting on beneficial nonsynonymous mutations or a local accumulation of slightly deleterious nonsynonymous genetic hitchhikers.

Some loci with evidence of recent differential selection across the meadow contained transporter genes (**Fig. 3.5c**), which could indicate selective pressures for uptake of different compounds between sites. A *Verrucomicrobia* population also showed evidence of a selective sweep occurring at a putative hopene biosynthesis operon, often involved in regulating membrane stability. Within a *Deltaproteobacteria* population, a highly differentiated locus encoded for numerous genes involved in two-component systems and histidine kinases, potentially related to environmental sensing and response. Taken together, these multiple genomic signals suggest that gene-specific selection partially drives differences in population genetic structure across meadow soils.

3.4 Discussion

Developing a cohesive picture of the genetic structures of soil bacterial populations is crucial for understanding the evolution and distribution of genes that can play critical ecosystem and niche-specific functions. Doing so has been difficult because the most abundant soil bacteria are difficult to cultivate, and when cultivation is possible, it is uncertain if microbial isolates are truly random samples of a population, considering the intense selective pressure put on cells during the isolation process. Here, we show that recent advances in sequencing technologies and software tools enable the monitoring of heterogeneity within genomically defined bacterial populations needed to address these questions in soil. Compared with a previously published study on freshwater lake microbial populations (Bendall et al. 2016), we observed far more polymorphisms per species (4721 to 43,225 SNPs/Mb), despite having a MAF cutoff of 5%. Most populations also had higher rates of SNPs/Mb than a similar analysis of metagenome assembled genomes found for microbial populations in deep-sea hydrothermal vents (Anderson et al. 2017). Nucleotide diversity was heterogeneous within individual samples but was often still high, implying that within each 10 g sample of soil many alternative alleles are frequently encountered.

The results of this study suggest that recombination and gene-specific selection are important modes of evolution across the most abundant soil microbes, and are capable of structuring populations at a scale of meters in soils. Even within a single meadow, our data demonstrate that there are likely thousands of combinatorial genetic mixtures for each species, with recombination resulting in no easily measurable numbers of irreducible strain-like lineages. We were able to calculate the relative rate of recombination to mutation for half of our species studied, and comparing these values to those reported for other species in the literature places them among other known recombinogenic species. Thus, the dynamics of dominant soil bacterial populations may be partially described using 'quasi-sexual' models (Garud et al. 2019; Rosen et al 2015). Although we only observed the importance of these dynamics in structuring populations at the 1–10 m scale, it is still uncertain how evolutionary dynamics in soil bacterial populations develop on other spatial and temporal scales. We conclude that future work on soil microbial ecology would

benefit by considering the role of substantial unlinked allelic diversity within species in shaping local gene content and allele frequencies.

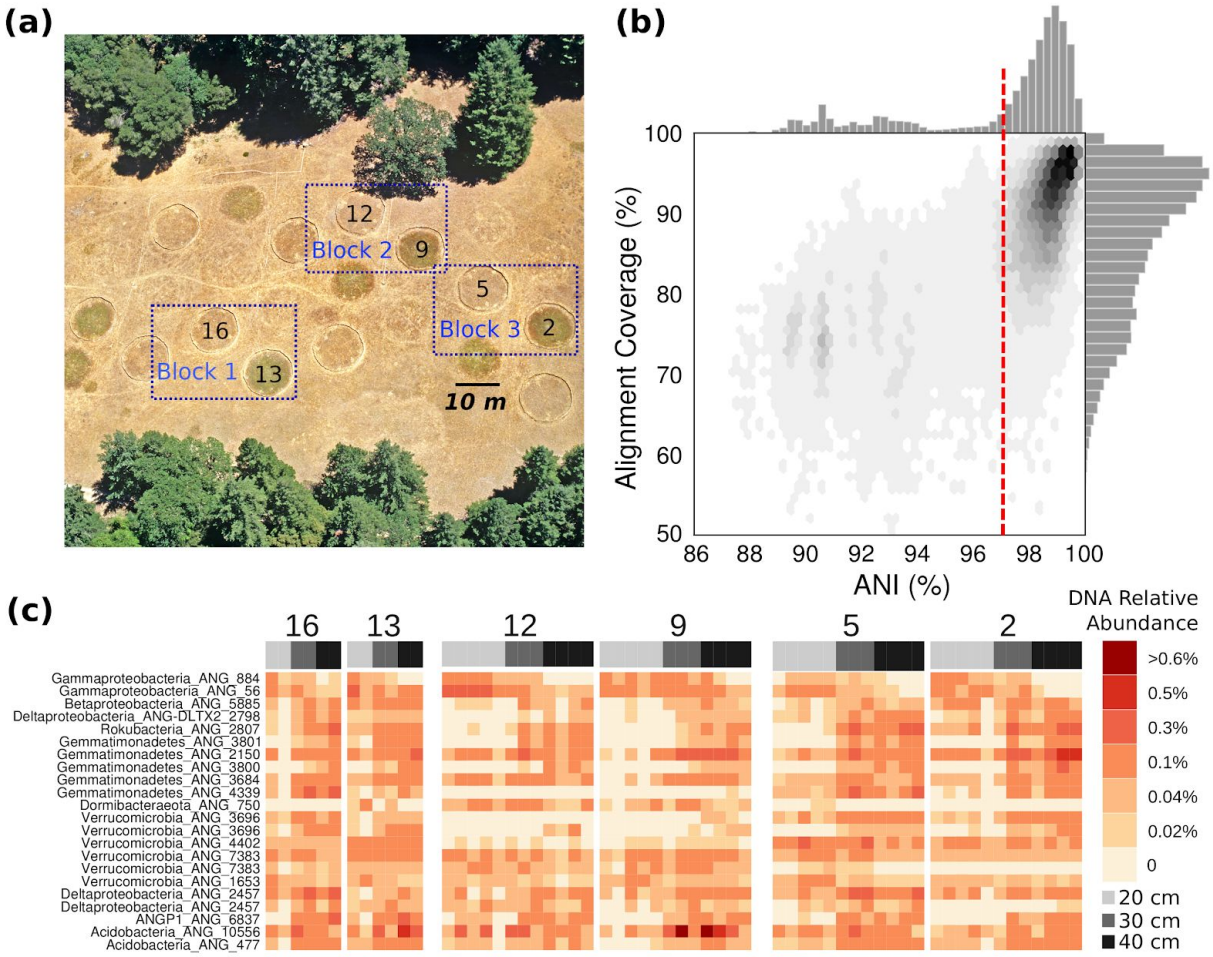

Fig. 3.1 | Meadow overview and population DNA relative abundances. (A) A bird's eye view of the grassland meadow located at the Angelo Coast Range Reserve in Mendocino County, California (39° 44′ 17.7″ N, 123° 37′ 48.4″ W). **(B)** Histograms of average nucleotide identity (ANI) and alignment coverage (an approximation of % shared gene content) values for all-vs-all comparisons between genomes assembled from the meadow soils. **(C)** Relative DNA abundances of the 19 bacterial populations analyzed in this study in each sample, organized by experimental plot where sample was collected, ordered by sample depth from left to right.

Fig. 3.2 | Spatial variation in genetic differences within species. (A) The percentage of variation in genetic similarity (ANI) of consensus genomes explained by plot of origin (red) and sampling depth (blue). **(B)** Multidimensional scaling ordinations of genetic dissimilarities between genomes within each species, for all 12 populations for which sampling plot explained a significant fraction of the variation in genetic dissimilarity (PERMANOVA; FDR = 5%; $p < 0.05$). The first two axes are plotted, there is a single point for each genome independently assembled for a population, and genomes are colored by sample plot of origin and the point shape indicates the sample depth.

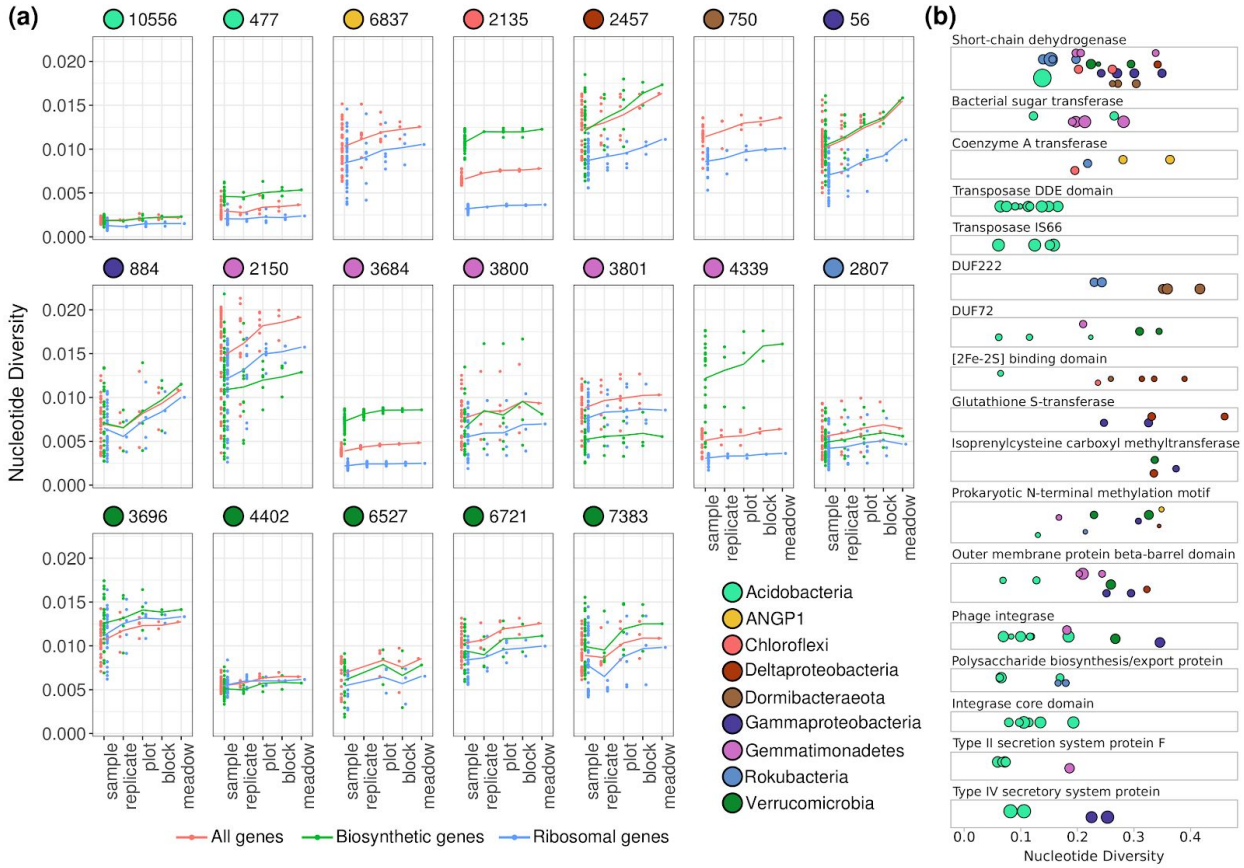

Fig. 3.3 | Nucleotide diversity of 19 highly abundant bacterial populations. (A) Distributions of average per-gene nucleotide diversity for each population, measured across increasing scales of sampling. Separated by all genes (red), ribosomal genes (blue), and biosynthetic genes (green) for species that have them (all except species *Candidatus* ANGP1 6837 and *Dormibacteraeota* ANG 750). Lines connect the means of each distribution of points across scales. **(B)** Nucleotide diversities of genes from protein families that were found to be enriched among genetically diverse genes compared with the average genomic frequencies. Each gene is a point arranged by protein family from the PFAM database, and point size scales with the gene's length.

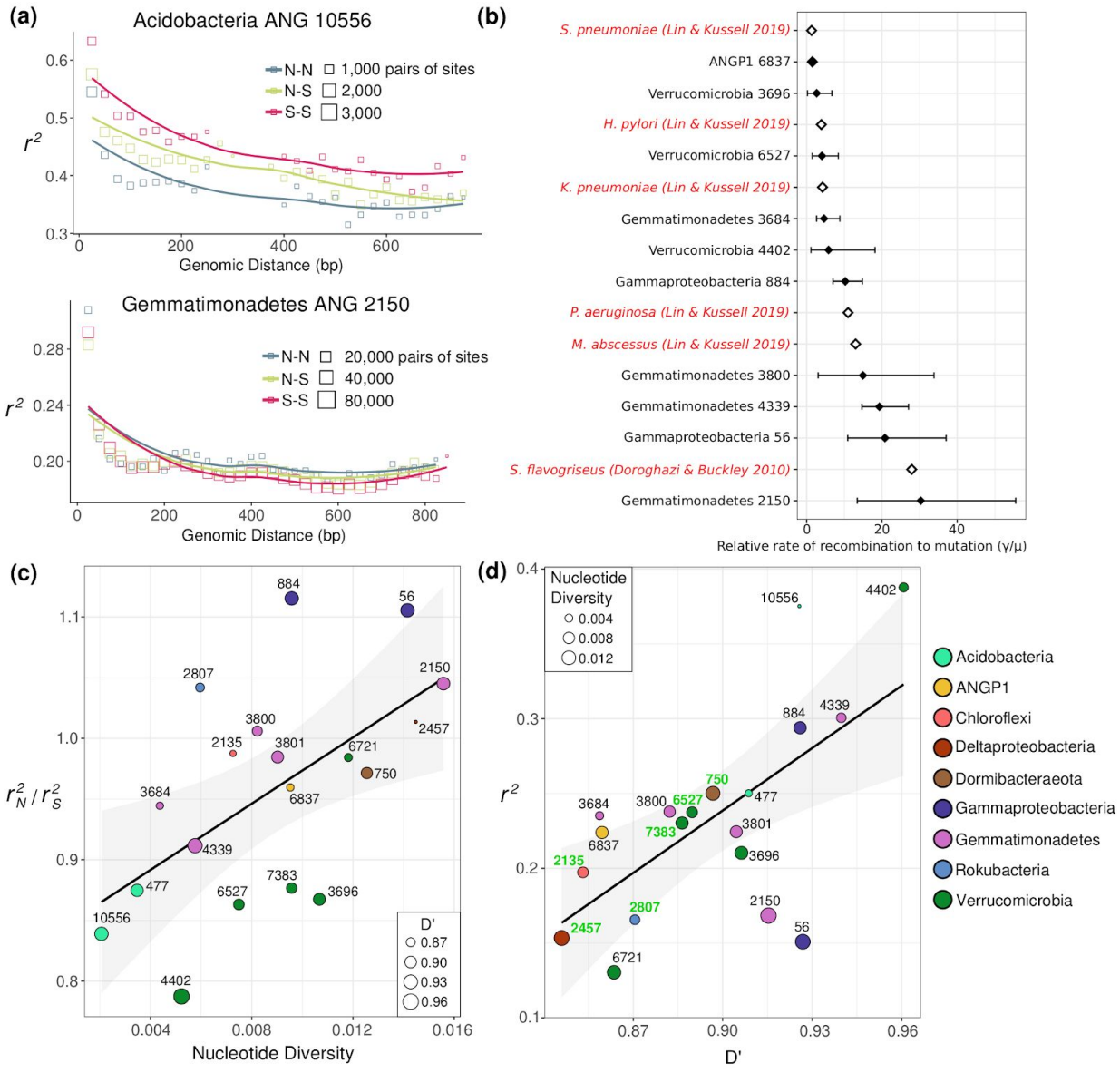

Fig. 3.4 | Varying rates of linkage disequilibrium within populations. (A) Linkage decay of r^2 for pairs of loci within the population with the lowest nucleotide diversity (top) and the highest nucleotide diversity (bottom). Each square is an average of pairs of biallelic sites at that distance, with the area of the square point proportional to the number of pairs of biallelic sites that went into the mean. Haplotypes (site pairs) are binned by the predicted function of the mutations of each of the paired SNPs (nonsynonymous: N, synonymous: S). **(B)** Relative rates of recombination to mutation calculated across the entire meadow for ten populations on synonymous third position codon sites, compared with previous values (red) reported by Lin and Kussell and the value reported for Streptomyces flavogensis by Doroghazi and Buckley. Error

bars represent the 95% confidence interval across 1000 bootstraps. **(C)** The relationship between nucleotide diversity and r^2_N/r^2_S. The mean nucleotide diversity and the mean ratio of the linkage of nonsynonymous-nonsynonymous vs synonymous-synoymous pairs of mutations across species is shown. The size of each point represents the mean D' value for that species. A linear regression is shown (linear regression; $R^2 = 0.29$; $p = 0.009$). **(D)** The relationship between mean r^2 and mean D' across the 19 bacterial populations studied. Genomes with evidence for multiple operonic competence related genes are labeled in green. A linear regression model is shown (*F*-statistic: 11.9, Adjusted *R*-squared: 0.38, $p = 0.003$).

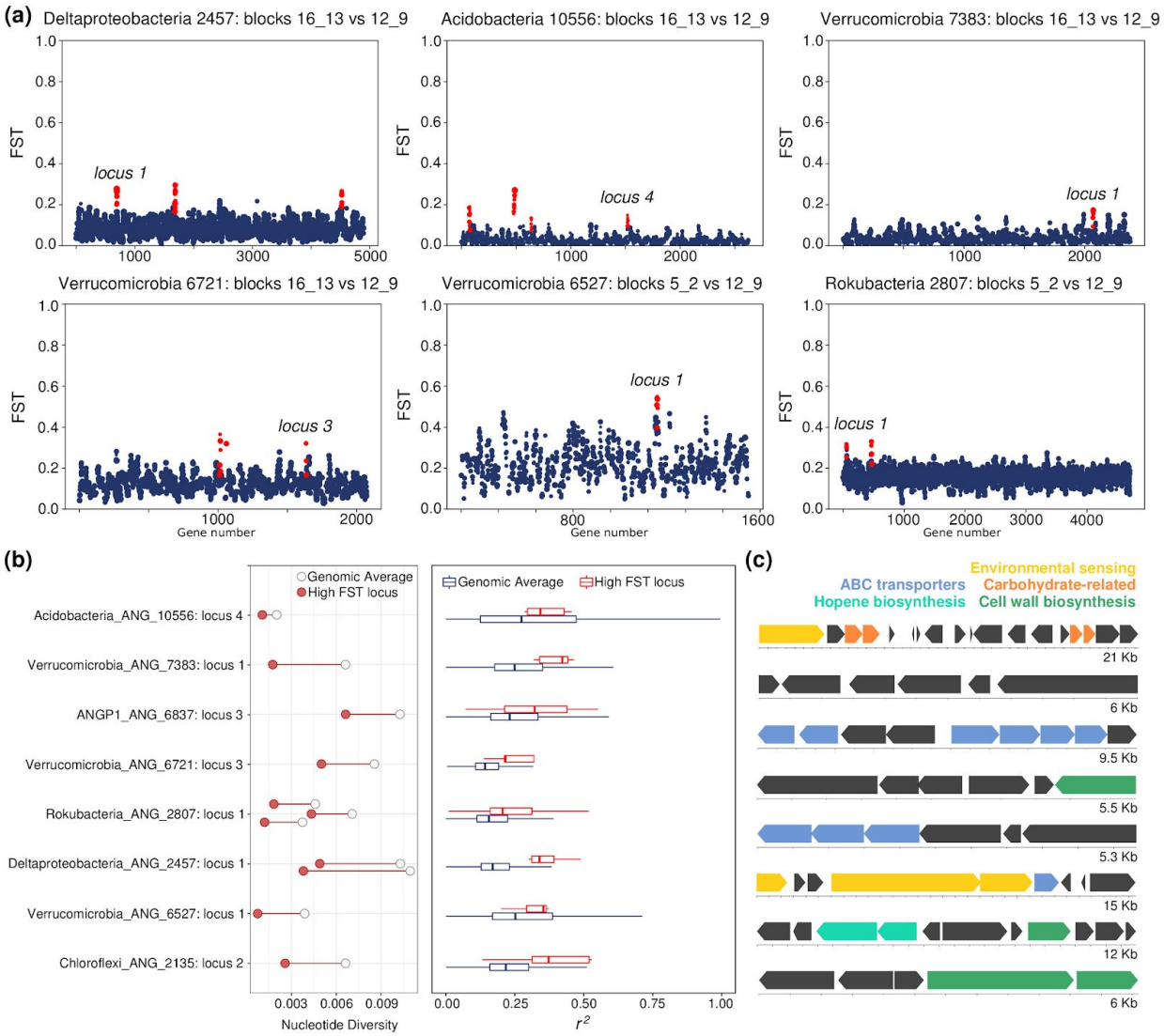

Fig. 3.5 | Highly differentiated genomic loci between sites within a meadow. (A) Values of F_{ST} for genes across the genomes of six bacterial populations. Each point is a gene, and the size of the point is determined by the number of SNPs within that gene. Plotted is the mean F_{ST} for that gene. Loci with significantly higher F_{ST} than the background are highlighted in red, and those that passed further filtering are labeled by their genome-specific locus numbers used in part (b). **(B)** Left: Nucleotide diversity at highly differentiated loci (red circle) compared with the average (empty circle) for each population. Right: The extent of linkage disequilibrium at highly differentiated loci (red) compared with the genomic average (black) for each population. **(C)** Gene diagrams and annotations of highly differentiated loci (genome and loci identities given in b).

Each block indicates an open reading frame, and blocks are colored by a subset of predicted functions.

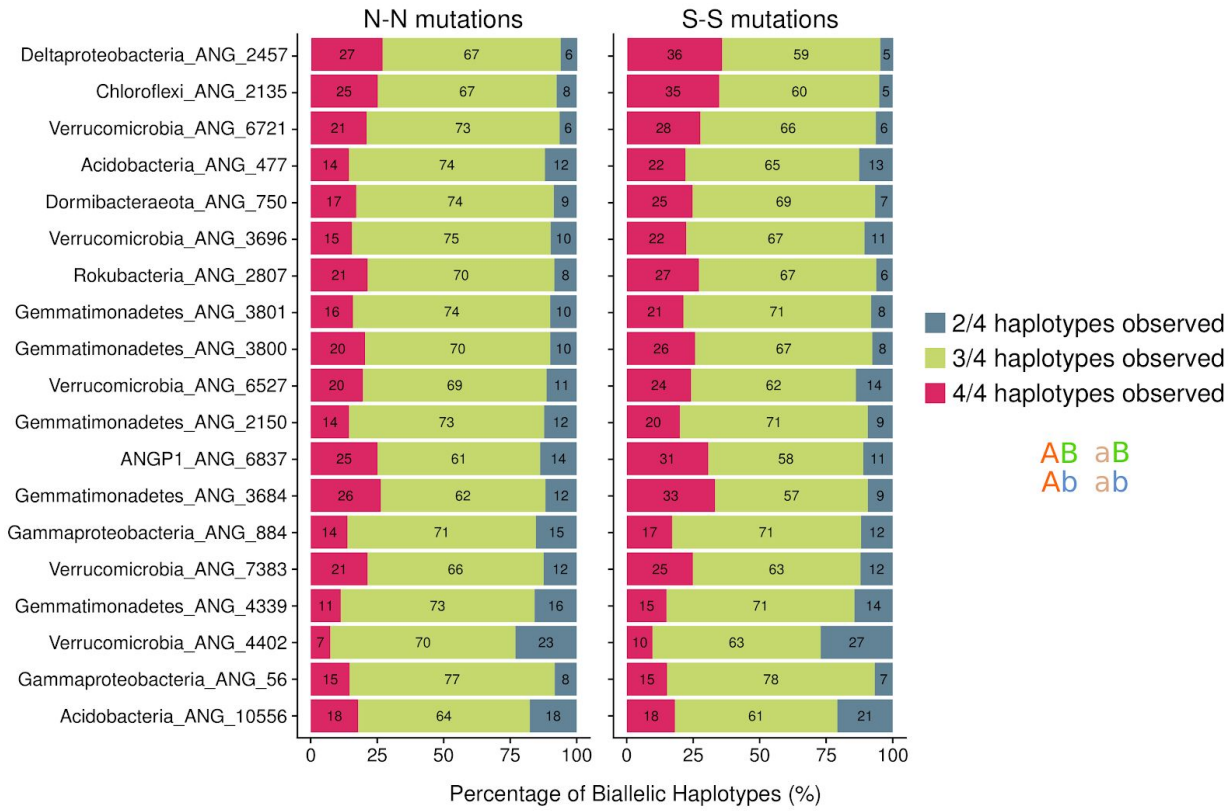

Fig. 3.6 | Biallelic haplotype counts for each species. The percentage of haplotype counts for each pair of biallelic sites within ~1 Kb of each other is shown for both nonsynonymous-nonsynonymous and synonymous-synoymous pairs of segregating sites.

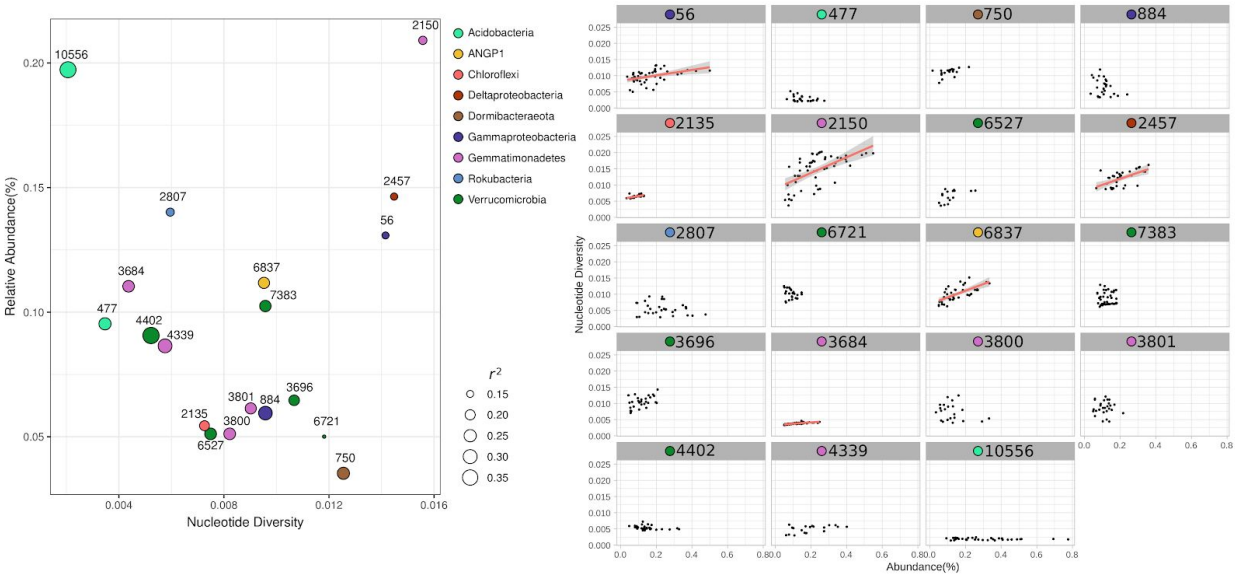

Fig. 3.7 | Relationships between nucleotide diversity and relative abundance. On the left is the relationship between the mean nucleotide diversity and relative abundances across species for the entire meadow. On the right are the relationships for each species across all of the individual samples from the meadow; linear regressions are drawn for all significant correlations ($p<0.05$, bonferroni corrected).

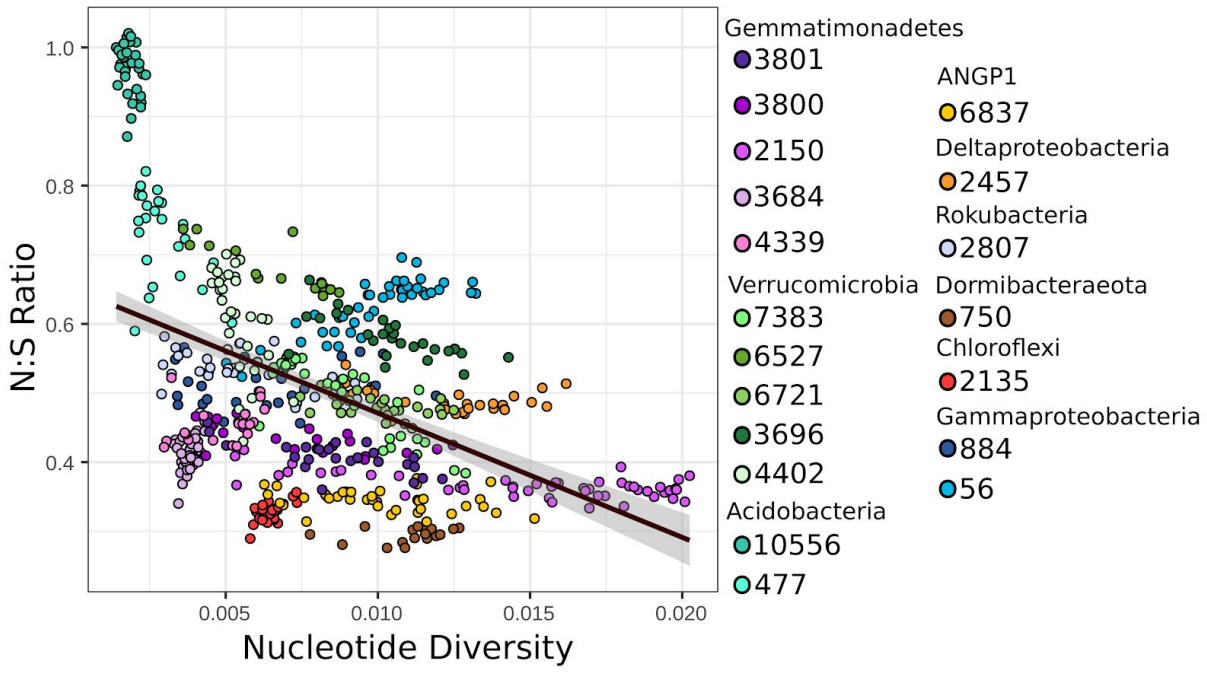

Fig. 3.8 | Relationship between nucleotide diversity and the N:S SNP ratio across species. Each point represents the nucleotide diversity observed for each genome within each sample. A linear regression is shown (R2=0.21; p<2.2*10-16).

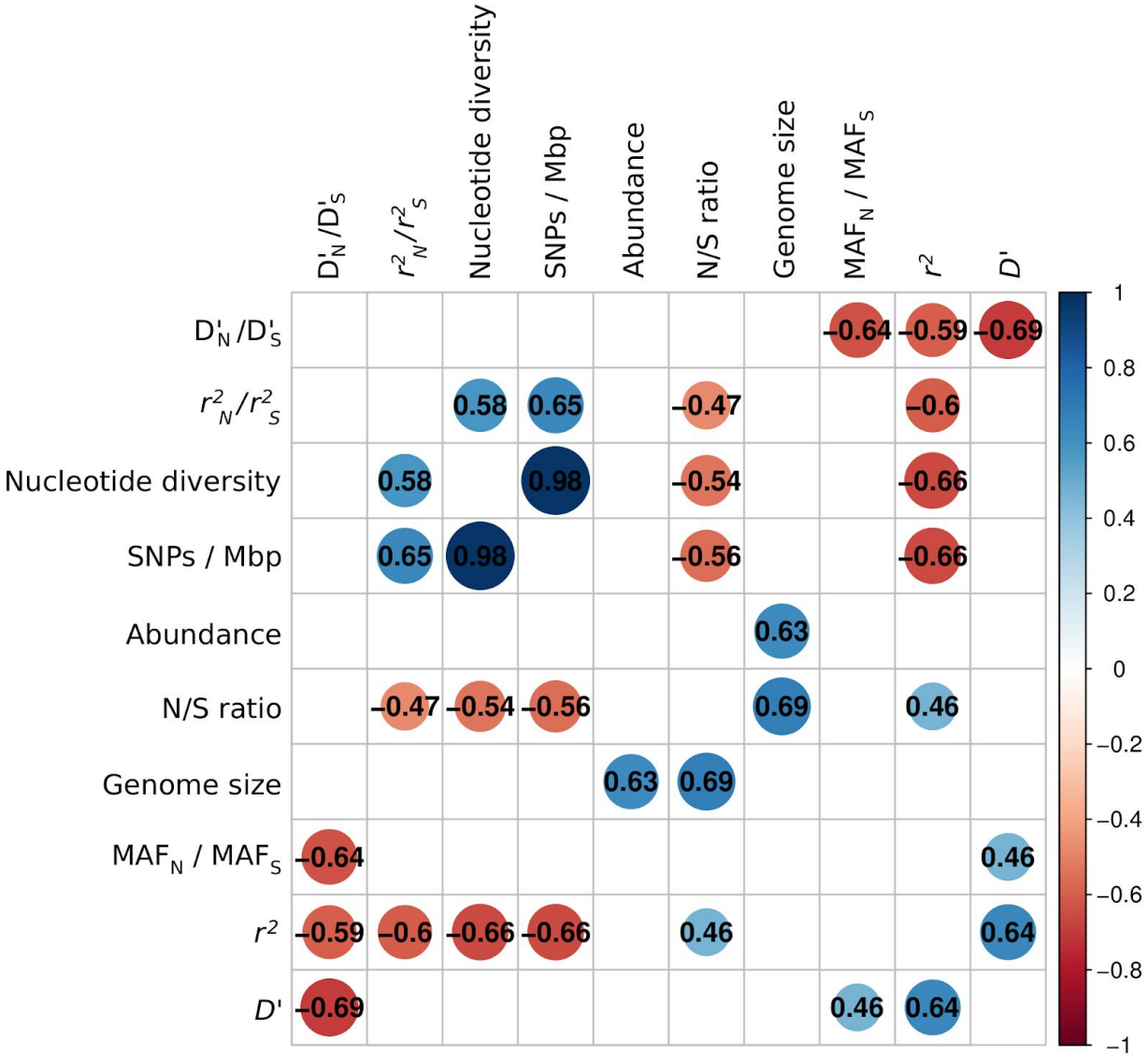

Fig. 3.9 | Cross-population correlations between means of all population genetics summary statistics in this study. In each box, the pearson correlation coefficient between two metrics is shown. Only correlation coefficients with a p-value <0.01 are shown.

4. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity

Alexander Crits-Christoph, Nicholas Bhattacharya, Matthew R. Olm, Yun S Song, and Jillian F Banfield

Published in *Genome Research*, 2021

Biosynthetic gene clusters (BGCs) are operonic sets of microbial genes that synthesize specialized metabolites with diverse functions, including siderophores and antibiotics, which often require export to the extracellular environment. For this reason, genes for transport across cellular membranes are essential for the production of specialized metabolites, and are often genomically colocalized with BGCs. Here we conducted a comprehensive computational analysis of transporters associated with characterized BGCs. In addition to known exporters, in BGCs we found many importer-specific transmembrane domains that co-occur with substrate binding proteins possibly for uptake of siderophores or metabolic precursors. Machine learning models using transporter gene frequencies were predictive of known siderophore activity, molecular weights, and a measure of lipophilicity (log *P*) for corresponding BGC-synthesized metabolites. Transporter genes associated with BGCs were often equally or more predictive of metabolite features than biosynthetic genes. Given the importance of siderophores as pathogenicity factors, we used transporters specific for siderophore BGCs to identify both known and uncharacterized siderophore-like BGCs in genomes from metagenomes from the infant and adult gut microbiome. We find that 23% of microbial genomes from the infant gut have siderophore-like BGCs, but only 3% of those assembled from adult gut microbiomes do. While siderophore-like BGCs from the infant gut are predominantly associated with *Enterobactericaee* and *Staphylococcus*, siderophore-like BGCs can be identified from taxa in the adult gut microbiome that have rarely been recognized for siderophore production. Taken together, these results show that consideration of BGC-associated transporter genes can inform predictions of specialized metabolite structure and function.

4.1 Introduction

Microbes produce specialized metabolites with diverse functions, including siderophores, ionophores, antibiotics, antifungals, and signalling molecules (Osbourn 2010). Specialized

metabolites therefore often underlie both cooperative and competitive interactions between microbes, and microbial interactions with the physiochemical environment (Davies 2013; Sharon et al. 2014; Tyc et al. 2017). The vast majority of specialized metabolites in bacteria are produced by biosynthetic gene clusters (BGCs), which are sets of genomically colocalized genes.

Colocalization of genes into BGCs is thought to occur because of selection for co-inheritance and co-regulation (Fischbach, Walsh, and Clardy 2008). While thousands of microbial natural products have been characterized, genomic BGC predictions made using programs such as antiSMASH (Blin, Shaw, et al. 2019) and ClusterFinder (Cimermancic et al. 2014) suggest that characterized molecules represent just a small fraction of all existing microbial natural products (Medema and Fischbach 2015)(Kim et al. 2017). Many of these unknown metabolites may be highly novel due to enzymatic and combinatorial diversity of genes in BGCs (Jenke-Kodama, Börner, and Dittmann 2006; Chevrette et al. 2019).

Because of the sheer number of sequenced but otherwise uncharacterized BGCs and the time and costs required for chemical characterization, there is a pressing need for predictions of BGC metabolite structures or functions to enable prioritization of targets for laboratory study (Tran et al. 2019). Prediction of metabolite structure or function for a novel BGC from gene content alone is challenging. For many biosynthetic nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), there is a 'co-linear' assembly-line regulation in which the order of genes relates to the order of enzymatic modifications on the metabolite during synthesis (Fischbach and Walsh 2006). Using this co-linearity rule can help predict some degree of structural detail in NRPSs and PKSs, as is done by by antiSMASH and PRISM (Skinnider et al. 2017), but there are many known exceptions to this rule (Wenzel and Müller 2005), and the accuracies of these software predictions have not been formally assessed using a large training dataset.

Prediction of BGC metabolite function generally relies on contextual genes associated with BGCs. The observation that genes conferring resistance to the produced metabolite are also colocalized with the BGC motivates investigation of putative resistance genes (self-resistance gene mining (Yan, Liu, and Tang 2020; Mungan et al. 2020)) for functional prediction. For siderophore activity prediction, AntiSMASH assigns a functional 'siderophore' label for BGCs that contain the IucA / IucC gene family, but this gene is only specific for siderophores with biosynthetic pathways similar to aerobactin (Hider and Kong 2010). More recently, (Hannigan et al. 2019) trained neural networks to both identify BGCs in genomes and classify BGCs by known metabolite functions. These networks used Protein Families from the Pfam database (El- Gebali et al. 2019) (Pfams) found in each BGC as features. They predicted activity labels of antibacterial, antifungal, cytotoxic, and inhibitor, with precisions of 36%, 47%, 61%, and 69% on each class respectively.

Many specialized metabolites perform their ecological roles extracellularly, and thus require transport across cellular membranes. Transporter genes often colocalize in BGCs and have been

shown to be compound specific and necessary for export of the product in many cases (Severi and Thomas 2019; Martín, Casqueiro, and Liras 2005; Méndez and Salas 2001). Therefore, transporters may also inform predictions of BGC metabolite structure and function. The distribution of transporters associated with biosynthetic gene clusters has so far been assessed only in characterized BGCs with experimental validation, a small fraction of the total number BGCs sequenced. At least 40 BGC-associated exporters have been characterized, mostly in the Actinomycetes, with varying degrees of experimental validation (Severi and Thomas 2019).

Transporters associated with BGCs are commonly either ATP-dependent active transporters or ion-gradient dependent transporters (Martín, Casqueiro, and Liras 2005). ATP- dependent transporters include the ATP-binding cassette (ABC) superfamily of both importers and exporters (Rees, Johnson, and Lewinson 2009), and the MacB tripartite efflux pump (Greene et al. 2018b). Examples of characterized structures of each transporter class and their substrates are shown in **Fig 4.1a**. In brief, Type I ABC importers are characterized by the BPD_transp_1 transmembrane (TM) protein family, and include MalFGK and MetNI for malate and methionine import in E. coli (Beek et al. 2014). Type II ABC importers are characterized by the FecCD protein family, and examples include BtuCD and HmuUV (Beek et al. 2014), and the FecBCDE system for Iron(III) dicitrate import in Escherichia coli (Staudenmaier et al. 1989). Both types of ABC importers often associate with Substrate Binding Proteins (SBPs), small membrane or periplasmic proteins for substrate uptake (Beek et al. 2014; Berntsson et al. 2010). Periplasmic binding proteins, Type II ABC importers, and TonB-dependent receptors are also known to play key roles in siderophore uptake in multiple bacterial species (Ellermann and Arthur 2017).

Meanwhile, examples of ABC exporters include McjD for Lasso peptide microcin J25 export (Romano et al. 2018) and the Staphylococcus aureus multidrug exporter Sav1866 (Dawson and Locher 2007), composed of the ABC_mem TM protein family, while the O-antigen polysaccharide exporter is composed of the ABC2_membrane TM protein family (Bi et al. 2018). Export of Nystatin, Doxorubicin, and Mccj25 was found to be dependent upon ATP-dependent transporters (Severi and Thomas 2019). The vast majority of ribosomally synthesized and post- translationally modified peptides (RiPPs) and a number of antibiotics from Actinomycetes also rely on characterized ATP-dependent ABC transporters (Gebhard 2012) (Méndez and Salas 2001).

Ion-gradient dependent transporters (also known as secondary active transport systems) do not require ATP and facilitate transport of small molecules in response to chemiosmotic gradients (Quistgaard et al. 2016). Those found in BGCs are often examples of the major facilitator superfamily (MFS) and occasionally, the resistance nodulation division (RND) family or the multidrug and toxic compound extrusion (MATE) family. Examples of characterized secondary active transporters for antibiotics include an RND transporter for pyoluteorin, and MFS transporters specific for mitomycin C, virginiamycin S, and landomycin (Severi and Thomas 2019).

Thousands of BGCs with chemically characterized metabolites open up the possibility for a broad genomic and computational analysis of phylogenetically and functionally diverse BGCs. Here, we used a curated version of the Minimum Information about a Biosynthetic Gene cluster database (MIBiG 2.0) of BGCs (Kautsar et al. 2020) and selected transporter-specific protein domain hidden markov models (HMMs) to perform a wide genomic assessment of the distribution of transporters in BGCs. We found clear correlations between transporter domains and corresponding metabolite features, especially siderophore activity, that indicate underlying logical structure to transporter associations and can inform functional and structural prediction of specialized metabolites from genomics alone.

4.2 Materials & Methods

Curation and selection of BGCs and transporter annotations

We parsed the MIBiG 2.0 database of biosynthetic gene clusters metadata and extracted information including host genus, compound count, chemical structures of the metabolite product, known metabolite activities, and the number of open reading frames for each BGC. Using Entrez and NCBI, we assigned the expected Gram status for each BGC based on phylum, coded 0=Gram-negative, 1=Gram-positive, 2=Fungal, 3=other. For the purpose of this manuscript, we only analyzed BGCs from Gram-positive and Gram-negative bacteria. We noticed that the activity labels in MIBiG 2.0 were often incomplete, and manually added a set of antibacterial, siderophore, and antifungal labels derived from the literature. We found that 28 BGCs (1.8%) in MIBiG were unusually large in length, and comparisons to published papers on these BGCs showed that their MIBiG counterparts were overextended in comparison to the validated BGC. For this reason, we eliminated the 28 BGCs over 60 ORFs in length. Using Python and RDKit, we calculated molecular weights and partition coefficients (log P) using the algorithm described in (Wildman and Crippen 1999) for all 1,042 MIBiG BGCs with a single associated compound structure. 238 BGCs have more than 1 associated metabolite structure, and these multi-structure BGCs were not used in our structural association analyses. We annotated biosynthetic genes in MIBiG with HMMER hmmsearch (Eddy 1998) and cath-resolve-hits (Lewis, Sillitoe, and Lees 2019) on a set of the 99 most commonly represented biosynthetic Pfams in antiSMASH BGCs obtained from (Cimermancic et al. 2014), to generate a counts table of biosynthetic protein families for each BGC.

To obtain a comprehensive overview of the distribution of transport-associated protein domains in biosynthetic gene clusters, we generated two separate feature tables: (a) using CATHDB (Sillitoe et al. 2019) HMMs and (b) using Pfam (El-Gebali et al. 2019) HMMs. For the first, we downloaded all proteins in the Transporter Classification Database (TCDB) (Saier et al. 2016) and annotated them with hmmsearch and cath-resolve-hits, using all CATHDB Functional Family HMMs. We then selected all CATHDB HMMs which were represented at least 5 times. We then manually curated

this list down to 180 final CATHDB HMMs that were transport-specific. We then calculated the specificities of each CATHDB HMM for TCDB families; 80 were specific for exactly one TCDB family.

For the second set of features, we took all protein sequences with an annotation including "Transport" in the antiSMASH database v2 (Blin, Pascal Andreu, et al. 2019) and annotated these proteins with the Pfam-A set of HMMs using hmmsearch and the option --cut_ga. We then selected highly represented HMMs and manually curated this list to be transporter-specific and representative of the major transporter classes in TCDB. We then also selected the Pfam Substrate Binding Protein and Periplasmic Binding Protein HMMs that were represented more than 5 times in MIBiG. When comparing both the Pfam and CATHDB set of HMMs, we found substantial overlap, but the CATHDB set is composed of 166 domain features while the Pfam set only contains 18.

Machine Learning to predict metabolite structural and functional characteristics

To identify associations between metabolite functional classes and structural properties with BGC gene content, we used traditional statistical tests and different machine learning models. Our classification tasks were (a) siderophore (n=16 Gram-positive; n=24 Gram-negative) vs other activity (n=142 Gram-positive; n=42 Gram-negative), (b) antibiotic and antifungal (n=131 Gram-positive; n=27 Gram-negative) vs other activity (n=57 Gram-positive; n=37 Gram-negative), (c) metabolite molecular weight > 1000 Daltons (n=149) vs metabolite molecular weight < 1000 Daltons (n=421), and (d) predicted partition coefficient log P <0 (n=220) vs log P >= 0 (n=350). For the functional classification tasks, we noticed a strong class imbalance with Gram-status, so we performed functional classification separately for BGCs from Gram-positive and Gram-negative bacteria. We first tested for univariate differences in proportional representation of each transporter BGC between classes for classification tests using Fisher's exact test in Python ($q<0.05$, Benjamini-Hochberg (Hochberg and Benjamini 1990) correction).

We then assessed the predictive power of the three sets of features for each BGC: transporter Pfam HMMs, transporter CATHDB HMMs, and biosynthetic Pfam HMMs. Features were counts of protein families, which were standardized using the StandardScaler function in the scikit-learn package. Given the nature of our study, we used simple models to ensure reliability and interpretability of our results. We fit two classes of machine learning models (a)

LASSO-penalized logistic regression, which fits a linear model with a sparsity penalty on weights, and (b) shallow decision trees (of depth one or two), which can classify based on splitting at most two features (Franklin 2005). All models were trained using the Python package scikit-learn. Due to data size and class imbalance, we fit models using repeated, stratified k-fold cross-validation ("RepeatedStratifiedKFold" in scikit-learn) with five repeats and five folds. On each cross-validation split of our data, we computed the area under the precision-recall curve (AUPRC) to evaluate

performance for our class-imbalanced tasks. Thus for the final output of this procedure we reported the mean of accuracy, precision, recall, and AUPRC each generated from five repeats of different random 5-fold partitions of the data. We further use these repeated cross-validation splits to see which features are consistently used for classification across repeats.

Annotating metagenomic siderophore-like BGCs from the human microbiome

To assess the distribution of siderophore-like BGCs in the human microbiome, we downloaded two sets of genomes assembled from metagenomes obtained from the human gut microbiome: 2,425 genomes from a neonatal intensive care unit (NICU) premature infant microbiome (Olm et al. 2019) and 24,345 genomes from a diverse set of mostly adult human cohorts (Nayfach et al. 2019). We ran antiSMASH 5.0 on these genomes and then scanned predicted BGCs for at least two of the Pfams that were found to be specific for siderophore BGCs (FecCD, Peripla_BP_2, and Ton_dep_Rec). We dereplicated these BGCs and compared them to known MIBiG siderophore BGCs using the software BiG-SCAPE (Navarro-Muñoz et al. 2020) run with default settings. We then considered BGCs with either set of hits and reported their genomic taxonomic distribution based on the closest BLAST hit representatives of genomic ribosomal proteins to taxonomic genera defined by GTDB (Parks et al. 2018) (minimum percent identity of hits >80%).

Data Access

All Python code used in this paper, along with the data analyzed, antiSMASH BGCs, and additional data tables, is available at GitHub (https://github.com/nickbhat/bgc_tran) and as Supplemental Code and Data.

4.3 Results

Genome mining of transporters associated with biosynthetic gene clusters

Using two compiled sets of transporter-specific HMMs (Pfam and CATHDB), we cataloged all classes of transporters across the MIBiG 2.0 database of characterized and experimentally validated biosynthetic gene clusters. We found that 56% of the bacterial BGCs in MIBiG contained at least one Pfam transporter hit and an additional 6% contained a CATHDB transporter hit without a Pfam domain (**Fig. 4.1c**). These percentages increased among BGCs that produce antibiotics (71%) and siderophores (78%), indicating that BGCs with these activities are more likely to contain at least one transporter. BGCs with transporters contained 2.5 transporter- associated domains across transport-annotated genes on average (**Fig. 4.1d**), which is expected as many ATP-dependent transporter systems have at least 2 domain complexes.

However, some BGCs contain considerably more transporter ORFs and domains, indicating that sometimes multiple transport systems can be associated with one BGC, although the number of protein domains that function as one transport system can often vary. The number of transporters in a BGC had no association with the number of metabolite structures reported for that BGC. The ATP-binding ABC transporter (ABC_tran) domain and the Major Facilitator Superfamily 1 (MFS_1) domain were the two most common transporter domains found in BGCs (**Fig. 4.1b**). A variety of proteins had nucleotide-binding domains along with several different transmembrane domains – ABC_membrane and ABC2_membrane domains were most common but ABC2_membrane_2, -_3, and -_4 domains were also represented.

Examining domains specific for export, the ABC_membrane domain is often characteristic of exporters (e.g., Sav1866 (Velamakanni et al. 2008)), but recently has been reported in the genes for siderophore uptake (YbtPQ) in Yersinia (Wang, Hu, and Zheng 2020), and is therefore not necessarily indicative of export or import alone. The second most common transmembrane domain, ABC2_membrane, has been observed in the O-antigen polysaccharide exporter (Bi et al. 2018). The MacB-FtsX tripartite efflux pump was found in 60 BGCs, while the RND (ACR_tran) efflux pump was less common, and found in only 10 BGCs. Other known efflux systems, such as SMR, MatE, and the MFS families 2 through 5 were comparatively rare across BGCs.

We next calculated co-occurrence correlations between all transporter protein families across BGCs and observed a strong negative correlation between MFS transporters and ATP- dependent transporters relying on the nucleotide binding domain, and a weaker negative correlation of MatE domains from the ATP-dependent NBD (**Fig. 4.2a**). This points towards a dichotomous choice between ATP-dependent and ATP-independent transport associated with a BGC.

Multiple lines of annotation evidence indicated that many of the transporter genes associated with BGCs were likely to be importers. Importers can be involved in the uptake or re- uptake of molecules like siderophores, and may also play roles in importing precursor metabolites for a BGC. The membrane domains specific to Type I importers (BPD_transp_1) and Type II importers (FecCD) were the most often observed ATP-dependent transmembrane domains besides ABC_membrane and ABC2_membrane (**Fig. 4.1b**). CATH Protein Structure Classification database HMMs (CATHDBs) that were specific for importer families in the Transporter Classification Database and found in BGCs included permeases for sugars, oligopeptides, and iron siderophores (**Fig. 4.2c**). 11% of BGCs with a transporter also contained a substrate binding protein. Among the substrate binding proteins we searched for, the most common contained domain was Peripla_BP_2, also found in the E. coli B12 importer complex BtuCDF, and variants of this SBP (cluster A-II) are specific for siderophores and cobalamin (Berntsson et al. 2010). We also observed many substrate binding proteins with specificities predicted to include carbohydrates, oligopeptides, and peptide uptake (Berntsson et al. 2010) (**Fig. 4.2b**). For example, the gene family SBP_bac_1 (SBP cluster D-I) (Berntsson et al. 2010) and is specific for uptake of sugars, was found in the BGCs for the

glycopeptide Mannopeptimycin and the aminoglycoside spectinomycin and may play a role in sugar precursor uptake (**Fig. 4.7**). It was also found in the acarbose and acarviostatin BGCs (**Fig. 4.7**), consistent with their putative roles as carbophors (Guo et al. 2012).

In the co-occurrence data, we observed pairing of different substrate binding proteins with different transmembrane domains. Peripla_BP_2 positively correlated strongly with FecCD and TonB_dep_Rec, genes known to be involved in siderophore uptake. BPD_trans_1 co-occurred with either SBP_bac_5 (SBP cluster C) or SBP_bac_1, while BPD_transp_2 co-occurred with Peripla_BP_4 (SBP cluster B-I). Taken together, these results show a logical organization of importer-specific transporter domains within BGCs that may be involved in either siderophore uptake, precursor uptake, or other roles. Regardless of the substrate specificity of these proteins, care must be taken when assuming that a transporter in a BGC is definitively for export of the matured product.

Prediction of siderophore and antibacterial activity from biosynthetic transporters

Because transporters are required for the ecological functions of biosynthesized specialized metabolites, we used machine learning to test if transporter classes were predictive of BGC-synthesized metabolite structures and functions. We noticed that metabolite activity labels in MIBiG were strongly associated with phylogeny: 83% of antibiotic BGCs were from Gram- positive bacteria while only 40% of siderophore BGCs were from Gram-positive bacteria in the dataset of curated MIBiG BGCs. To reduce the impact of this potential bias, we created separate training and testing datasets for activity prediction for Gram-positive and Gram-negative organisms.

Using our curated set of BGCs with transporters from the MIBiG 2.0 database, we tested for associations between BGC transport genes and metabolite function. We generated two activity classification tasks: (a) distinguishing siderophores (including known ionophores) (n=16 Gram-positive, n=24 Gram-negative), from non-siderophores (n=142 Gram-positive, n=42 Gram-negative), and (b) distinguishing antibiotics and antifungals (n=131 Gram-positive, n=27 Gram-negative) from non-antibiotics (n=57 Gram-positive, n=37 Gram-negative). We observed several statistically significant (Fisher's exact test; $q<0.05$) associations in the distribution of transporter types between both the siderophore and other activity classes. Among Gram-positive bacteria, 60% of siderophore BGCs contained the SBP Peripla_BP_2 and 55% contained the FecCD importer, while no BGCs with other activities had either (**Fig. 4.3a; Fig 4.8**). The situation was similar for Gram-negative siderophore BGCs. The TonB-dependent receptor (completely absent from Gram-positive bacteria) was the strongest signal, found in almost 80% of Gram- negative siderophore BGCs with a transporter and never in BGCs with other activities.

To assess siderophore predictability from BGC gene content, we used decision trees with only two layers applied to different feature sets of protein domain annotations- transport-affiliated Pfams, transport-affiliated CATHDB HMMs, and biosynthetic Pfams. To avoid issues with class imbalance,

we report precision and recall on the siderophore class, as siderophore prediction requires searching for a minority class (siderophores) within a background of mostly non-siderophores. With the transport-only features, we found that just with two gene decisions, it is possible to achieve 100% precision with over 80% recall using either Pfam or CATHDB transporter annotations for Gram-negative siderophores, and 100% precision with over 80% recall using CATHDB transporter annotations for Gram-positive siderophores (**Fig. 4.3b**) On the other hand, when using all biosynthetic annotations within BGCs we found that 2-layer decision trees trained on biosynthetic genes performed substantially worse at predicting siderophore activity than those trained on transporter genes (**Fig. 4.3b**). We further validated our results by training LASSO linearized regression models, which do not model interactions between features. These models obtained a slightly improved area under the precision-recall curve, indicating that very simple transporter patterns are highly predictive of whether a BGC is siderophore producing or not in our dataset (**Fig. 4.3b**). Transporter features predictive of siderophores were consistently selected by LASSO across stratified cross-validation repeats, giving evidence that these patterns are robust. The top predictive biosynthetic features of siderophores were the IucA/IucC protein family (used by antiSMASH to label siderophores and which is known to be involved in aerobactin biosynthesis) and condensation domains (likely to capture non-ribosomal peptide siderophores), but predictive effect sizes were smaller than those for transporters.

Using siderophore-specific transporter genes, we attempted to predict siderophore classes for any remaining gene clusters that have no annotated function in MIBiG2 (that we had not already hand curated). We searched for gene clusters containing the siderophore-predictive genes FecCD and Peripla_BP_2 and found 6 additional BGCs with no annotated activity, three of which were experimentally validated by the literature to be siderophores (Matsuo et al. 2011; Y. Chen et al. 2013; Carran et al. 2001) (**Fig. 4.3c**). The remaining three identified BGCs were false positives. One of them, the BGC for the antibiotic Ficellomycin, only contained these transport genes in flanking regions not known to be involved in biosynthesis (Liu et al. 2017) while the herbimycin A BGC contains the transporters genes in the reverse reading frame from the BGC, separated by an unusual 16 kb intergenic region and the genes appear to be fragmented (Rascher et al. 2005). The other false positive was Lividomycin, an antibiotic that does seem to be a rare non-siderophore BGC with FecCD and Peripla_BP_2 transporters in the MIBiG2.0 database.

To understand the extent of transporter specificity for particular classes of BGCs, we expanded our analysis to 95,293 BGCs in the antiSMASH database, which is a set of predicted BGCs in microbial genomes from the RefSeq database. Frequencies of general transporters and transporter genes that were specific for siderophore biosynthesis in MIBiG (FecCD and the TonB-dependent Receptor) were calculated across all antiSMASH BGCs by bacterial genus (**Fig. 4.4**). The general ABC transporter ATP-binding domain and MFS superfamily transporter genes varied in their frequencies across different classes of BGCs, but there were some consistent patterns. Thiopeptides,

NRPS-independent siderophores, and arylpolyene BGCs tended to have an MFS exporter while ribosomally synthesized products (e.g., Lasso peptides and Lantipeptides) consistently had ATP-dependent transport mechanisms (**Fig. 4.4a**).

Alternatively, the TonB-dependent receptor and the FecCD Type II importer were highly restricted to specific classes of BGCs. They only consistently appeared in NRPS-independent siderophore and NRPS clusters and were nearly absent from entire other classes of BGCs (**Fig. 4.4b**). This is consistent with the known NRPS-independent and NRPS biosynthetic pathways for siderophores. The TonB-dependent receptor was also found associated with Lasso peptides (possibly functioning as a resistance gene (Mathavan et al. 2014)), but FecCD was not. 21% of NRPS clusters contained a FecCD gene, and 28% contained a TonB-dependent receptor, possibly indicating that at least one in five uncharacterized NRPSs may function as siderophores. Thus, even in an uncurated dataset of thousands of BGCs, it appears as though siderophore-specific transporters only rarely occur in biosynthetic gene clusters that are unlikely to have siderophore functions.

Classifying antibiotics and antifungals from either transporters or biosynthetic genes proved more challenging than classifying siderophores. In Gram-negative bacteria, we observed positive associations between the MacB-FtsX tripartite efflux pump with antibacterial or antifungal activity (Fisher's exact test; $q<0.05$). MacB and associated components (FtsX and OEP) were positively associated with antibiotic activity by LASSO logistic regression. This result was stable across both cross-validation folds and repeats of cross-validation. We found that 7 out of 27 Gram-negative antibacterial BGCs contained a MacB, while no BGCs in our classes of other activities contained MacB. Although MacB is involved in export of the siderophore Pyoverdine (for which there is not an accurate BGC in MIBiG) in Pseudomonas (Greene et al. 2018a), in general MacB may be a strong indicator of antibacterial activity for a BGC. Previously, we identified a number of MacB-FtsX exporters in BGCs from novel Acidobacteria (Crits-Christoph et al. 2018), possibly indicating a role in antibacterial activity for these BGCs. LASSO effect sizes for individual biosynthetic genes were substantially lower.

Association of biosynthetic transporter classes with the molecular weights and lipophilicity of their putative substrates

We next hypothesized that transporter classes could be predictive of other molecular features beyond functional activity. There was no strong correlation between the molecular size of the metabolite produced and Gram status of the bacteria encoding each corresponding BGC in the MIBiG dataset. Thus, we tested for differences in transporter classes in BGCs producing metabolites that were (a) less than and (b) greater than 1000 Da in size across all bacteria. There was a significant difference in the frequencies of some transporters between BGCs with different metabolite molecular weights (**Fig. 4.5a**). The strongest difference was in the distribution of MFS transporters - found in 57% of BGCs with products under 1000 Da, but only 14% of those over 1000 Da (Fisher's exact test; $q<0.05$). The 95th percentile of metabolite molecular weights for clusters

with MFS_1 and without ABC_tran was 1082 Da, which may be approaching a biological limit for the molecular weights of substrates for these transporters. Conversely, the ATP-dependent ABC_tran domain was found in 89% of BGCs producing high molecular weight compounds but only 42% of those producing low molecular weight compounds (Fisher's exact test; $q<0.05$). MacB/FtsX and the rarer transmembrane domains were also associated with higher molecular weight compounds.

We tested for an association within the two largest chemical classes in the dataset, PKs and NRPs, and found that also within the PKS and NRPS biosynthetic classes ATP-dependent transporters were associated with larger metabolite molecular weights than the MFS family (**Fig. 4.5c**). We also observed that the ABC2_membrane_4 was associated almost exclusively with large RiPPs, with almost all of the BGCs in which it is found in producing compounds over 1500 Da in size. After training both LASSO logistic regression and 2-layer decision tree models to classify whether produced molecules are greater than 1000 Da, we found that transporter genes were able to distinguish large from small metabolites with moderate precision and recall and an AUPRC of up to 42%, while biosynthetic genes performed similarly (**Fig. 4.5b**). Top transporter features consistently had larger effect sizes than top biosynthetic features, indicating that transporter-based features provided clearer signals. This result was stable across both cross-validation folds and repeats. The biosynthetic protein family most associated with high molecular weight metabolites was Glycos_transf_2, likely due to the addition of sugar groups to metabolites by these enzymes in BGCs.

It has previously been reported that transporters can be specific for compounds with similar hydrophilicity (Rempel et al. 2020). The lipophilicity of a metabolite is often considered critical for its success in clinical development for human therapeutics (Arnott and Planey 2012). With LASSO logistic regression, we predicted partition coefficients (log P), a measure of lipophilicity, for all of the metabolites in MIBiG, and tested how well metabolite partition coefficients could be predicted by gene content. We found that the presence of 5 transporter classes was significantly associated with increased lipophilicity (Fisher's exact test; $q<0.05$). In particular, we observed an association between varying ATP-dependent transmembrane domains and log P, with ABC2_Membrane_3 domain co-occurring with BGC-metabolites with a high log P (median 4.4) (**Fig. 4.5d**) and ABC2_Membrane_4 domain co-occurring with BGC-metabolites with a low log P (median -6.2). While the ABC2_Membrane_4 association is likely due to its exclusive association with large RiPP products, the ABC2_Membrane_3 domain occurred in multiple BGC classes, mostly polyketides, and was still associated with a decrease in log P just within the polyketide class. LASSO logistic regression distinguished log $P > 0$ from log $P < 0$ with 77% AUPRC. On this task biosynthetic genes were distinctly superior over transporters at prediction of log P, obtaining a 83% AUPRC with a LASSO logistic regression trained on biosynthetic genes.

Identifying novel siderophore-like biosynthetic gene clusters in the human microbiome

To demonstrate the predictive utility of BGC-associated transporters, we mined BGCs with siderophore-specific transporters in metagenomic genomes (dereplicated per species) from (a) the gut microbiomes of neonatal infants in the intensive care unit (Olm et al. 2019) and (b) a cross-study collation of genomes assembled from multiple human gut studies (Nayfach et al. 2019). Identified "siderophore-like" BGCs putatively produce siderophores, as they contained the transporter classes that can achieve near 100% siderophore specificity in the MIBiG database- (a) Peripla_BP_2 and FecCD in Gram-positive bacteria, and (b) Peripla_BP_2, FecCD, and TonB_dep_Rec in Gram-negative bacteria. We identified 1442 BGCs with siderophore-like transporter classes (**Fig. 4.6a**) and then grouped them into novel gene cluster families using BiG-SCAPE, resulting in 75 siderophore-like gene cluster families (**Fig. 4.6b**).

Most siderophore-like BGCs were in large gene cluster families with other BGCs which also contained the same set of transporter hits. 23% of microbial genomes from the neonatal infant gut microbiomes had siderophore-like BGCs, but only 3% of those assembled from adult gut microbiomes did. Siderophore-like BGCs were identified across a range of bacterial genera, but the majority were from the Staphylococcus or Enterobacteriales, which are known to be in high abundance in the neonatal gut microbiome. The genera with the most siderophore-like BGCs were Staphylococcus and Klebsiella, common hospital-acquired pathogens of neonates. Five of the 75 identified siderophore-like BGC families included a known representative gene cluster in MIBiG, and four of the known representatives were siderophores - again pointing to the specificity of these transporter classes.

The rest of the siderophore-like gene cluster families that were identified had no closely characterized representative in MIBiG, indicating that there is likely capacity for production of multiple novel siderophores in the human gut microbiome (**Fig. 4.6c**). Almost all siderophore-like BGCs were either NRPS or NRPS-independent siderophore classes, the latter of which is based on the presence of the IuA/IuC gene family, as in aerobactin biosynthesis. Of the NRPS siderophore-like gene clusters, the majority had adenylation domain specificities for serine (ser) and 2,3-dihydroxybenzoate (dhb), indicating similar catechol-containing nonribosomal biosynthetic pathways to siderophores like enterobactin and salmochelin. We observed substantial genetic diversity between gene cluster families containing similar NRPS domains, which may indicate the existence of possible unknown derivatives of these siderophores in the human microbiome. In adult gut microbiome samples, one large novel NRPS siderophore-like gene cluster with unknown adenylation specificity was identified in Coprococcus, members of the Lachnospiraceae, often considered to be important commensals in the human gut (Duvallet et al. 2017; L. Chen et al. 2017). Thus, while a majority of siderophore- like BGCs in the human microbiome contained core enzymes similar to known siderophore biosynthetic pathways, there was substantial genetic diversity that could indicate further unexplored structural variation.

4.4 Discussion

We uncovered several strong associations between transporters within characterized BGCs and features of the corresponding BGC-synthesized metabolites. With regards to prediction of metabolite activity, we quantified the specificity of TonB-dependent receptors, FecCD, and Periplasmic-binding protein 2 for siderophore-producing BGCs. This complements existing literature indicating that genes in these families are specific for siderophore import in both Gram-positive and Gram-negative bacteria (Chu et al. 2010). We also identified a putative association between the MacB tripartite efflux pump and antibacterial/antifungal activity. Based on these findings, a strategy of targeting novel BGCs containing MacB for characterization may be useful for antibiotic prospecting. In addition to activity prediction, we used metabolite structural information in MIBiG in order to predict metabolite molecular weight and lipophilicity from BGC gene content. We discovered a strong relationship between the transporters in characterized BGCs and the molecular weight of their synthesized metabolites. The strong dichotomy between ATP-dependent transporters (utilizing the ABC_tran nucleotide binding domain) and MFS family transporters points towards required ATP-dependence for transporting metabolites larger than 1000 Da. We also identified relationships between two understudied membrane components (ABC2_Membrane_3 and ABC2_Membrane_4) and substrate log P, possibly indicating tradeoff in membrane domains for molecules of different chemical properties. Future phylogeny-based subdivision of these families may improve upon general protein family annotations to increase the predictive power of transporter substrate characteristics.

There are multiple caveats to our work. Molecular activity of specialized metabolites based on functions proven in the laboratory may be very different from the ecological roles that metabolites play in natural settings (van der Meij et al. 2017; Kramer, Özkaya, and Kümmerli 2020; Behnsen and Raffatellu 2016). Further, many BGCs may produce multiple variants of a metabolite (Fischbach and Clardy 2007), only some of which may be reported. There may also e reference-database biases in our gene searches - while they are sensitive, it is possible that phylogenetically divergent microbes use transporter genes that are not hit by our sequence models. Finally, as reported, a significant proportion of BGCs contain no transporter at all or a transporter gene genomically adjacent to a BGC may not be functionally linked. There are both technical and biological reasons why a BGC might not contain a transporter gene. Firstly, the transporter(s) for the metabolite produced may be encoded elsewhere in the genome. Secondly, the BGC's genomic boundaries may be misannotated, and the transporter may be downstream of annotated genes. Thirdly, it is possible that the metabolite being produced performs its primary function intracellularly and does not require a transporter for export. It is also possible that there are unannotated transport systems in BGCs: To further investigate this, we identified unannotated proteins with transmembrane domains in BGCs, and found that 18% of BGCs in MIBiG without a transporter contained one unknown membrane protein. Despite these caveats, it appears as though transporter genes provide

simple and strong signals for inferring both activity and chemical properties of metabolites produced by BGCs.

Siderophores are both considered critical pathogenicity factors for many human-associated microbes (Weakland et al. 2020) and are also known to facilitate interactions with other microbes and the innate immune system in the human gut microbiome (Behnsen and Raffatellu 2016; Holden et al. 2016; Zhu et al. 2020; Lam et al. 2018). Therefore, being able to annotate genes for the production of siderophores across diverse bacterial species may be critical for understanding the distribution of virulence factors, yet it is difficult to do using traditional annotation pipelines alone. We observed a high prevalence of siderophore-like BGCs in bacterial genomes from NICU premature infant guts, suggesting that the premature infant gut could be more prone to invasion by pathogens with siderophore virulence factors. Potentially novel siderophore-like BGCs were most consistently found to be encoded in the genomes of members of the Enterobacteriaceae and Staphylococcus in the premature infant microbiome. Only in the adult microbiome datasets did we identify siderophore-like BGCs in the Lachnospiraceae, that are often considered important commensals, indicating that there may also be commensal siderophore production in adult gut microbiomes. Importantly, we identified siderophore-like BGCs in these taxa that are not homologous to known siderophore clusters, indicating that there is still substantial unknown chemical diversity of siderophores, even within well-studied lineages.

In general, here we demonstrated that consideration of transporter genes can aid holistic functional prediction of BGC products. A transporter guided approach could be especially useful for identification of siderophore targets for medical (Nagoba and Vedpathak 2011) and biotechnological applications(Ahmed and Holmström 2014). Given the large diversity of BGCs and that chemical characterization of their products can be time and resource intensive, better functional prediction of BGCs for targeted study can improve selection of targets for antimicrobial discovery and downstream activity tests.

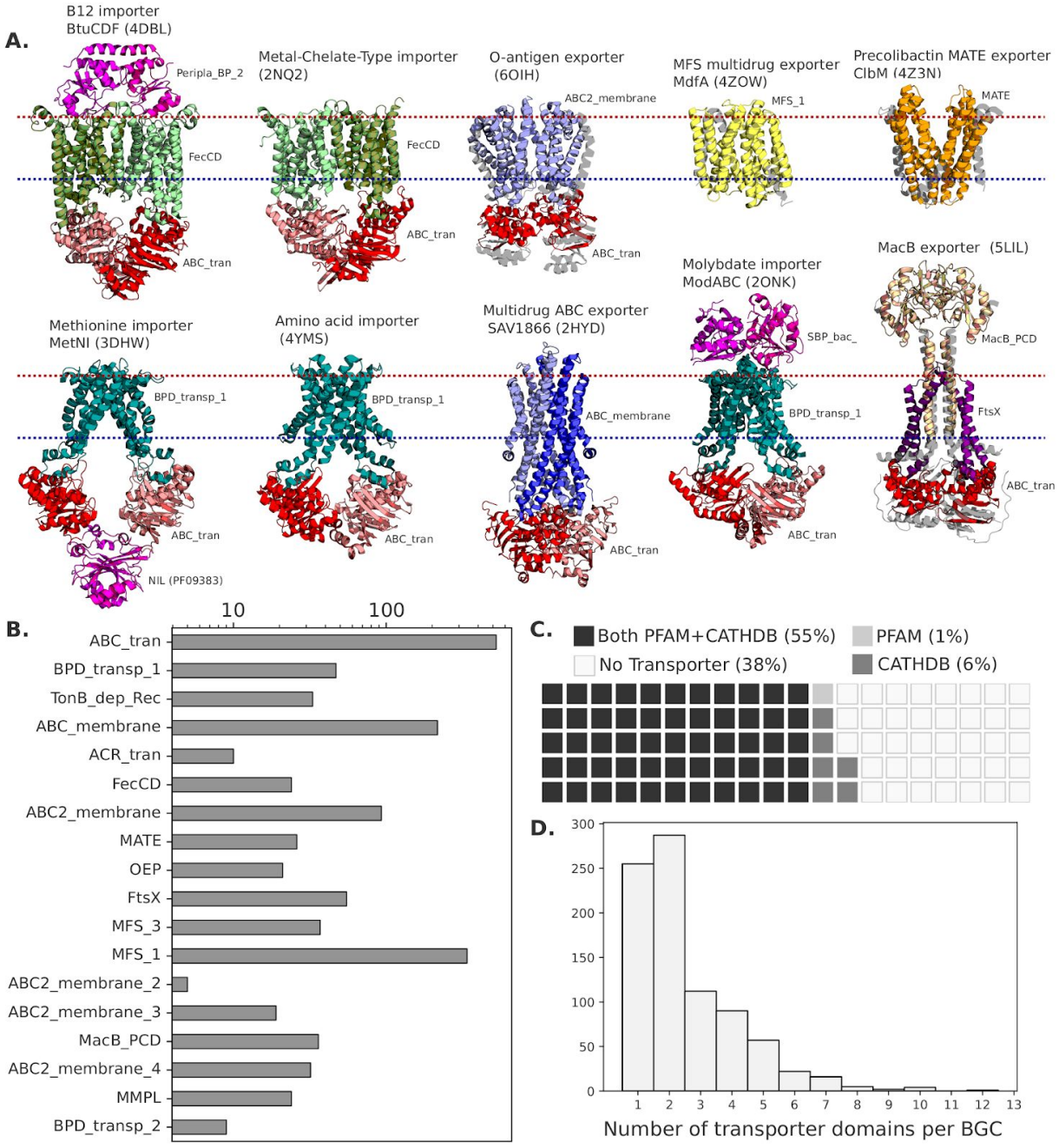

Fig. 4.1 | Distributions of transporter classes in biosynthetic gene clusters. (A) Structures of characterized examples of major transporter classes often found in BGCs, colored and labeled by Pfam domains. The extracellular/periplasmic side of the membrane is shown as a red line, and the intracellular side is in blue. **(B)** The frequencies of common Pfam transporter domains across the bacterial BGCs in the MIBiG database. **(C)** The percentages of bacterial BGCs in MIBiG that do and do not contain transporter domains. Each square represents 1% of BGCs. **(D)** The counts

of transporter domains per each bacterial BGC that contains at least one transporter gene across MIBiG.

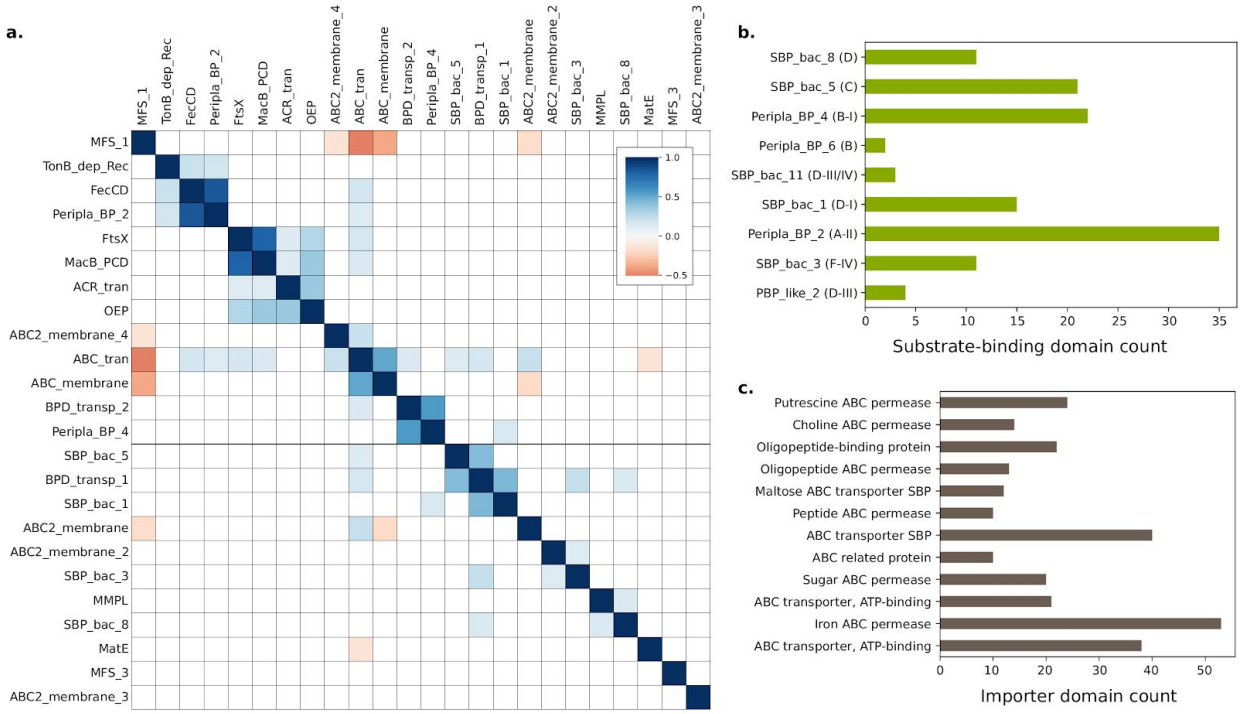

Fig. 4.2 | Presence of importer-specific domains and co-occurrence between transporters across BGCs. (A) Spearman's correlations between commonly occurring Pfam transporter domains across MIBiG BGCs—only correlations with P<0.001 are shown. **(B)** Counts of Pfam transporter substrate binding domain families and corresponding substrate binding protein clusters described by Berntsson et al. (2010). **(C)** Counts of importer-specific CATH domainsacross MIBiG BGCs. The CATH functional family "Iron ABC permease"is essentially synonymous with the FecCD Pfam.

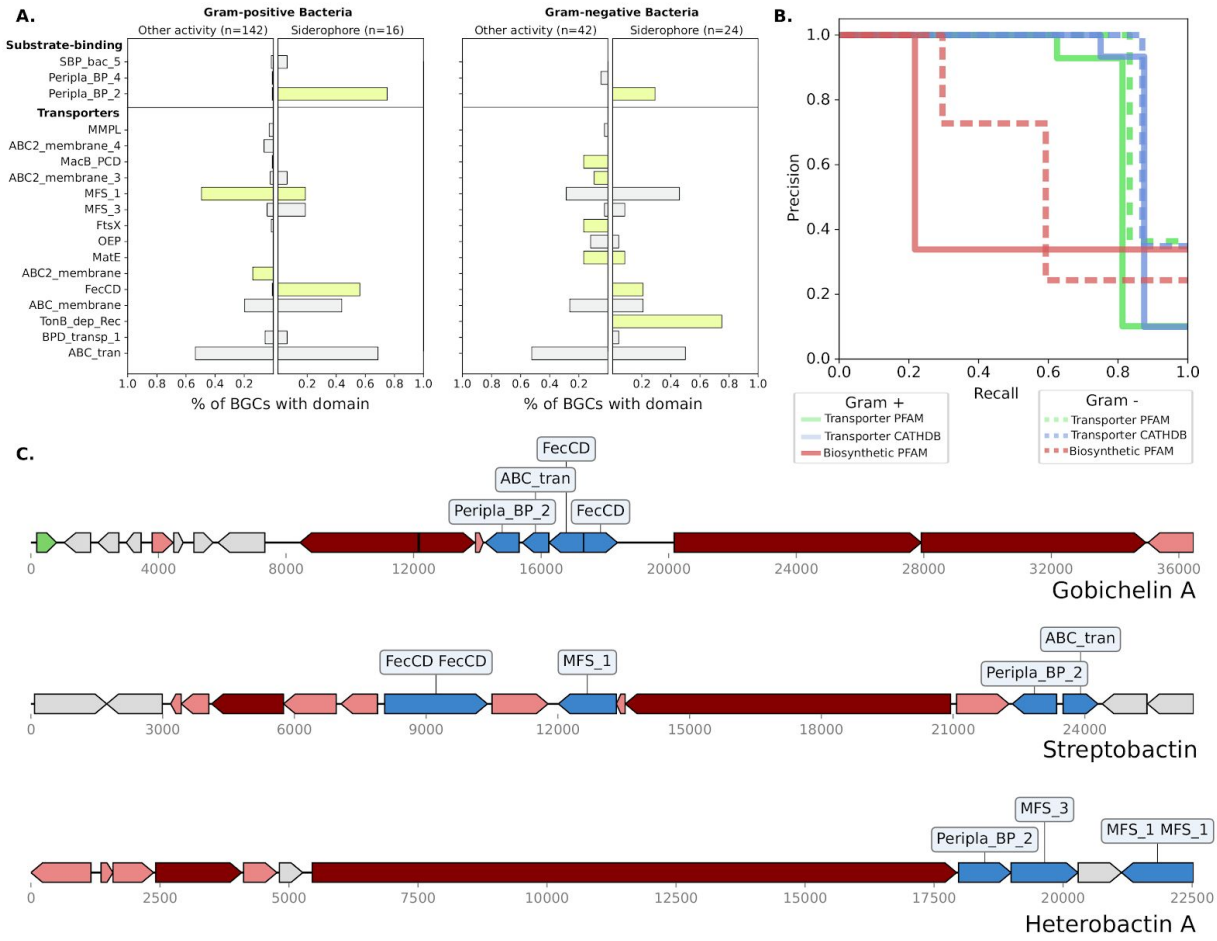

Fig. 4.3 | Transporter domains are predictive of siderophore BGCs. (A) The frequencies of common transporter Pfam domains across siderophore BGCs and BGCs of other known activities in Gram-positive and Gram-negative bacteria. Bars in green were significantly different in frequency between the two classes (Fisher's exact test;Q<0.05) **(B)** Precision-recall curves for two-layer decision trees classifying siderophore BGCs using Pfam transporter, CATH transporter, and Pfam biosynthetic gene features in Gram-negative and Gram-positive bacteria. **(C)** Examples of three siderophore BGCs without activity labels in MIBiG 2.0, which could be identified using transporter frequencies. Transporter genes are blue, core biosynthetic genes (NRPS and PKS) are dark red, accessory biosynthetic genes are light red, and regulatory genes are green.

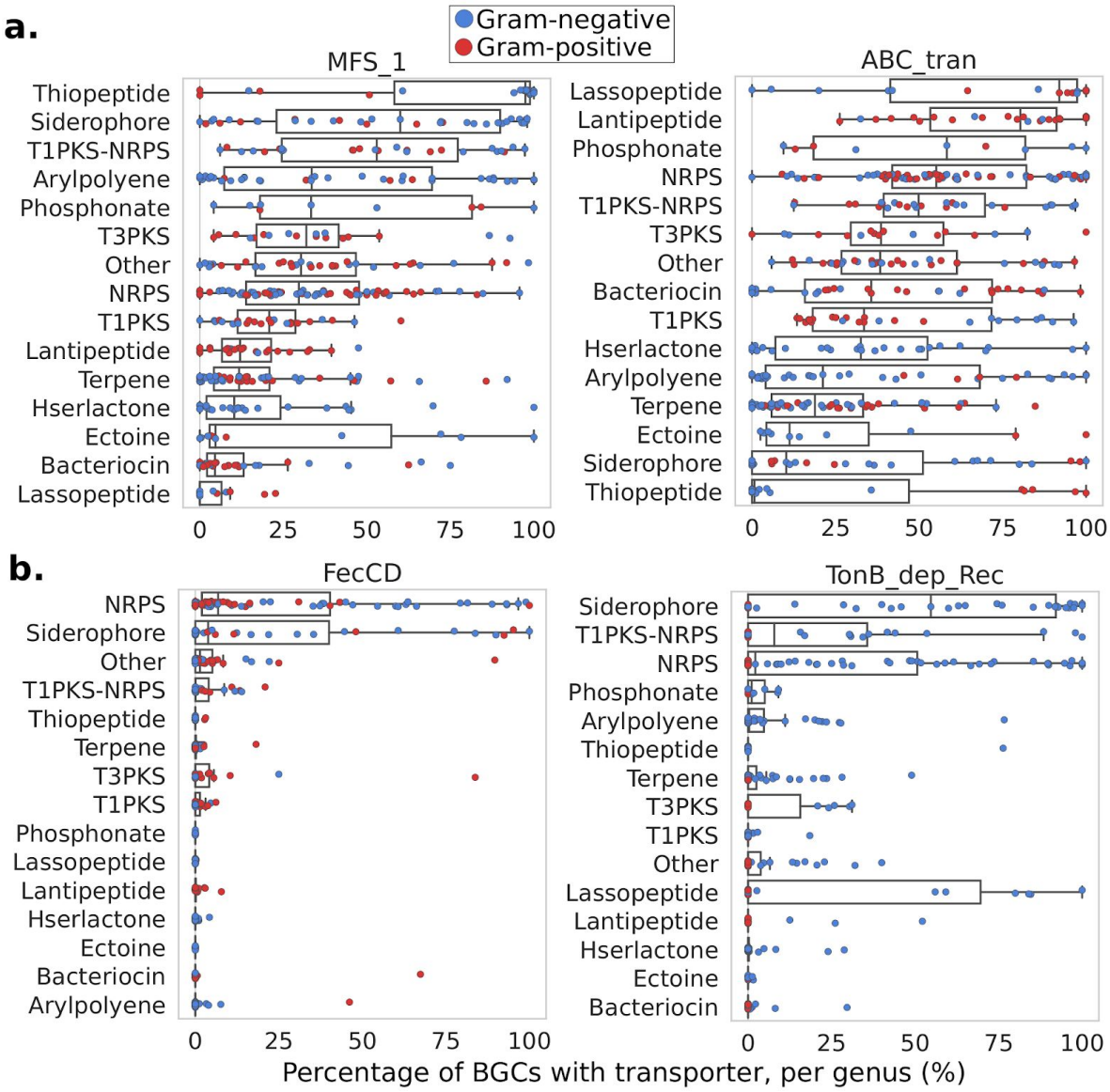

Fig. 4.4 | Presence of general and siderophore-specific transporters by biosynthetic class and bacterial genus across the antiSMASH database. (A) ATP-dependent and ATP-independent (MFS) transporters are commonly associated with a variety of BGCs in the antiSMASH database across a wide range of genera. Each point is the percentage of a BGC class with a transporter within a particular genus. Each genus is colored by its Gram status, and genera with fewer than 20 BGCs of a particular class are excluded. **(B)** Siderophore-specific transporters are associated with few BGC classes in the antiSMASH database. Each point is the percentage of a BGC class with a transporter within a particular genus. Each genus is colored by its Gram status, and genera with fewer than 20 BGCs of a particular class are excluded.

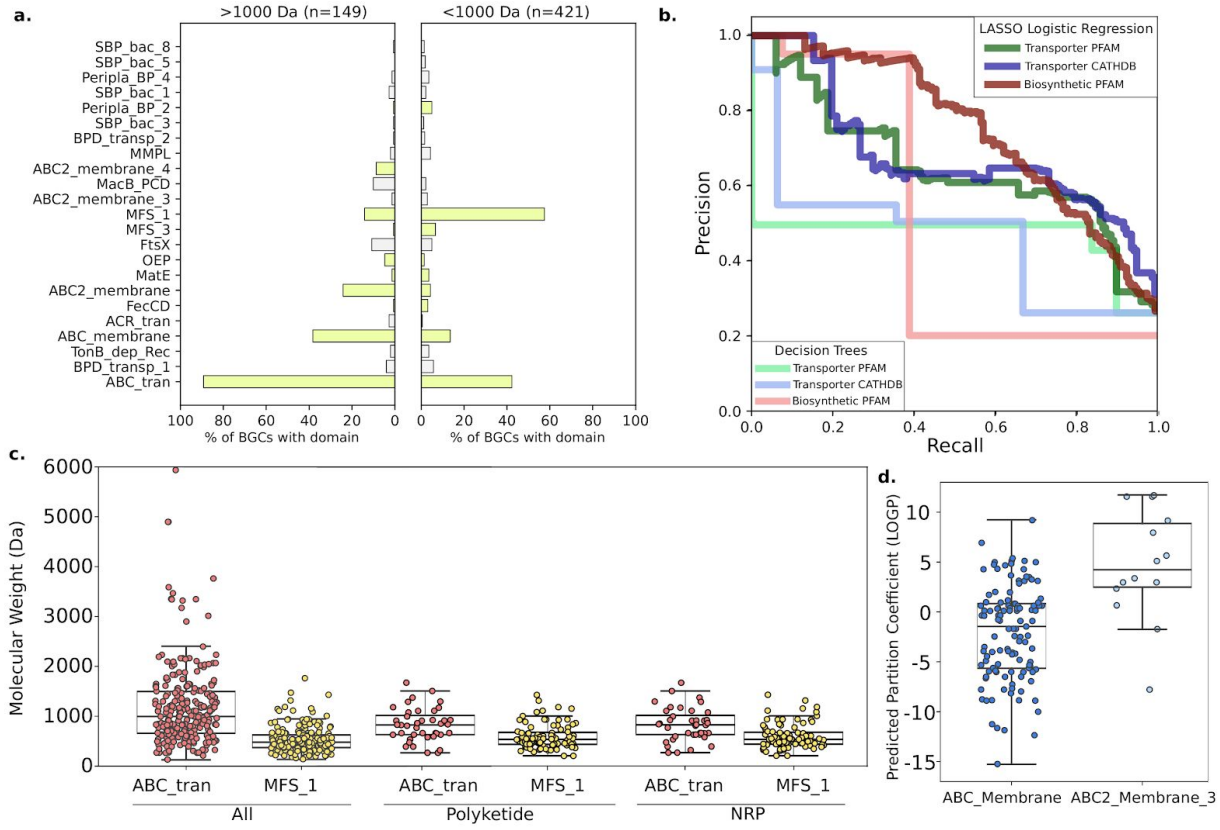

Fig. 4.5 | Transporter domains associated with molecular size and partition coefficient. **(A)** The frequencies of common transporter Pfam domains in BGCs that synthesize metabolites >1000 Da (left) and <1000 Da (right). Bars in green were significantly different in frequency between the two classes(Fisher's exact test;Q<0.05). **(B)** Precision-recall curves for two-layer decision trees and LASSO logistic regression models classifying BGCs producing metabolites >1000 Da using Pfam transporter, CATH transporter, and Pfam biosynthetic gene features. **(C)** The distribution of metabolite molecular weight synthesized by BGCs with at least one NBD-binding ABC transporter domain, at least one MFS domain, and the ABC2_membrane_3 transmembrane domain. **(D)** Predicted partition coefficients (logP) for metabolites synthesized by BGCs that contain at least one variant of two different ABC transporter transmembrane domains.

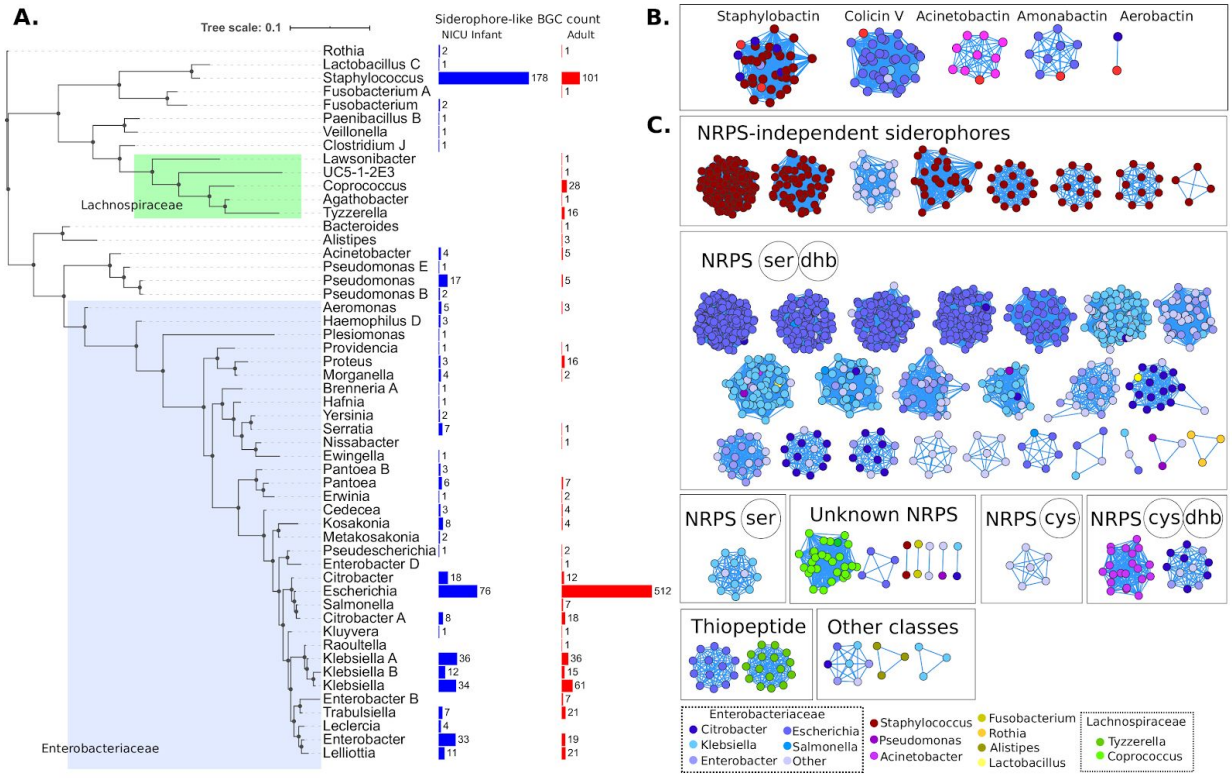

Fig. 4.6 | BGCs with siderophore-like transporters from human gut microbiomes. (A) Concatenated ribosomal protein tree (collapsed to the genus level) for high-quality genomes from the infant and adult gut microbiomes that encode siderophore-like BGCs. On the right are counts of siderophore BGCs from infant gut genomes (blue) and adult gut genomes (red). **(B)** Gene Cluster Families of BGCs containing known siderophores (bright red) and human microbiome–derived BGCs with siderophore-like transporters. BGCs are connected by similarity to other BGCs in the same gene cluster family, calculated usingBiG-SCAPE. **(C)** Families of siderophore-like BGCs without any similarity to existing known BGCs. BGCs in the network are colored by the taxonomy of the genome of origin and are grouped and labeled by the antiSMASH reported biosynthetic class: for NRPS gene clusters, the adenylation domain specificities are reported.

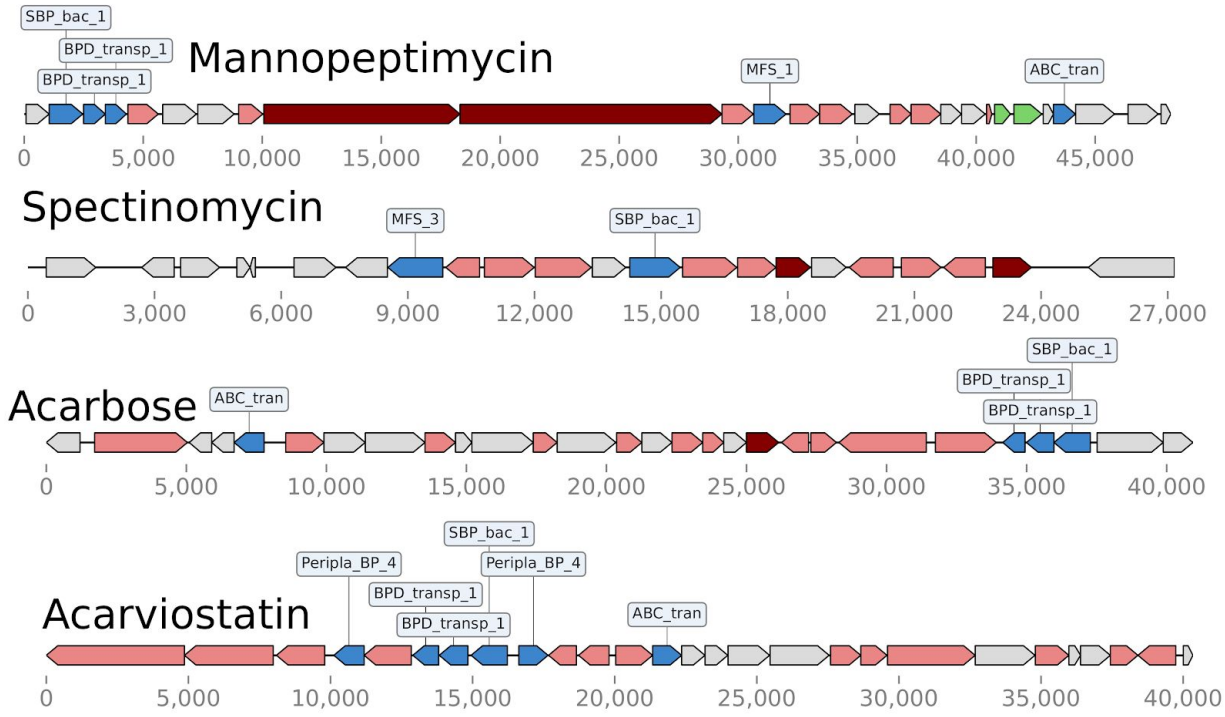

Fig. 4.7 | Examples of BGCs with the SBP_bac_1 substrate binding protein domain.

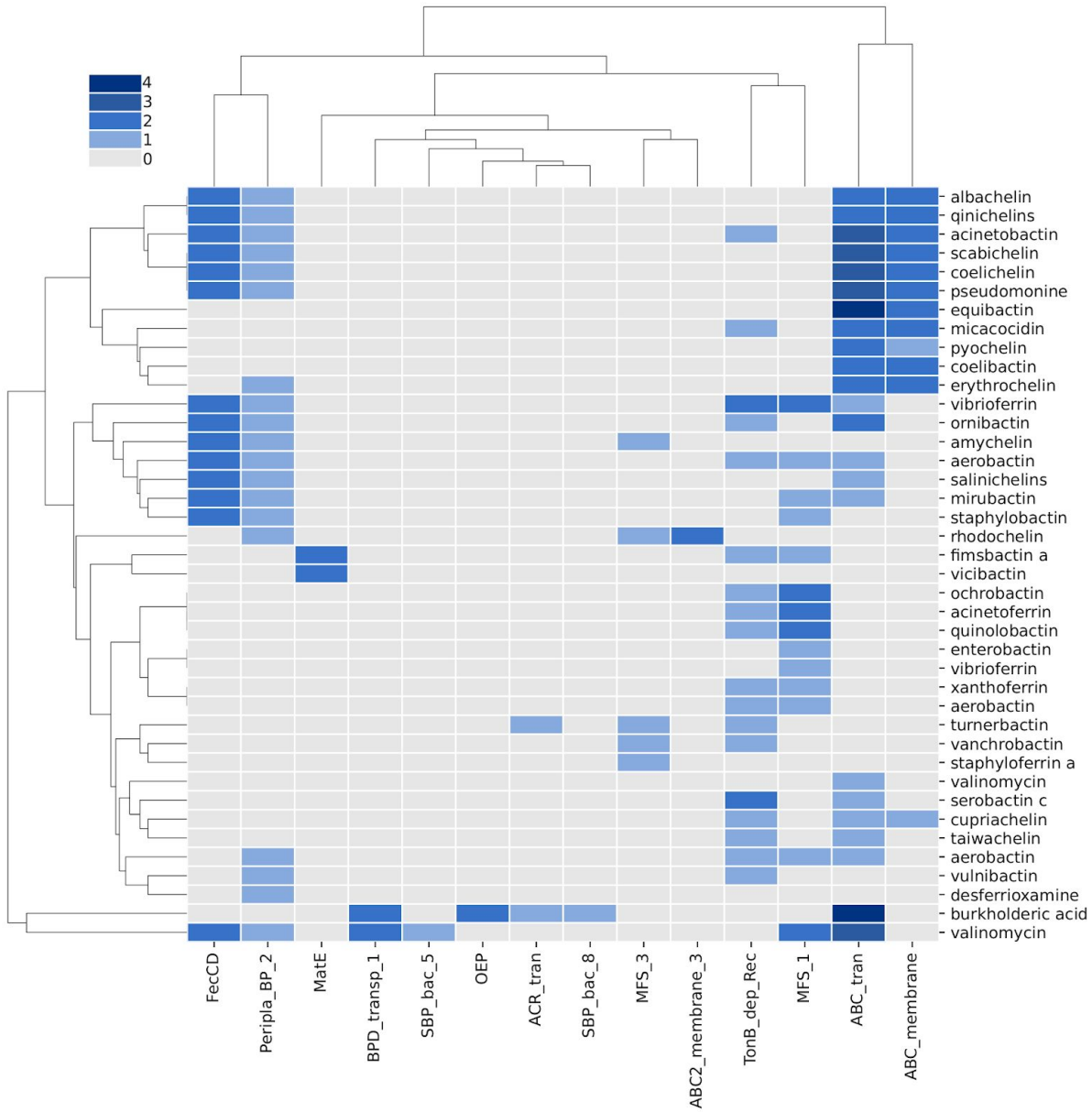

Fig. 4.8 | Heatmap of PFAM transporter domain abundances in annotated siderophore BGCs in the MIBiG database.

Conclusions

A primary conclusion from this work was the identification and genomic characterization of uncultivated lineages of Acidobacteria that possess large genomic repertoires for specialized metabolite biosynthesis. Since the publication of Chapter 1, this primary conclusion has been replicated, with additional reports of Acidobacteria in these lineages from ocean biofilms and soil microbiomes. In Chapter 2, we followed up on this conclusion by discovering several more genomes from one of these lineages - demonstrating that these microorganisms are not uncommon in soils, and can be found in many soil metagenomes. Intriguingly, these specialized metabolite producers also appear to be more abundant in the bulk soils studied than many commonly cultivated lineages - for example, we only recovered marker genes from *Streptomyces* or *Pseudomonas* at low abundances in these studies. While genome-resolved metagenomics does present its own biases as a window in soil microbial community composition - particularly sequencing bias against high GC% genes, poor DNA extraction from spores, and difficult assembly of high strain complexity species - this still may indicate that these yet uncultivated organisms may play an outsized role in chemical ecology in the ecosystems in which we find them. The difficulty herein therefore lies with bridging the gap between these results and cultivation. While some Acidobacteria have been cultivated from soils, it is relatively rare to do so - and this phylum is as diverse metabolically and phylogenetically as the Proteobacteria, and woefully undersampled in comparison. In brief attempts at cultivation we were unsuccessful in obtaining cultures on solid media, even when isolating from samples in which the species were at high abundances (data not shown). Additionally, no cultures of these particular Acidobacteria lineages have been reported elsewhere - indicating that there are still some unknown factors missing from common cultivation efforts waiting to be understood.

Using metatranscriptomics, we were able to characterize the natural contexts in which biosynthetic genes of uncultivated organisms were expressed in soils. We identified transcriptomic activity for the genus Angelobacter in soils from two different sites studied. We also identified initial transcriptomic signals that are consistent with competition sensing. Firstly, we observed a distinct rise in genes associated with growth and core metabolism (Figure 1.3) early in the experiment, with a delayed transcriptomic response for many biosynthetic genes. This could be consistent with biosynthetic gene expression following - and possibly in response to - an initial stage of general community growth after wet-up. The second was a signal of co-expression: across several species of bacteria, we identified modules of genes co-expressed with biosynthetic gene clusters, and found that they were enriched in genes also involved in environmental sensing and response. While this is indirect evidence, it is also one of the first times a transcriptomic approach has been taken to understand the regulation of specialized metabolisms in natural ecosystems. Future work combining both isolated organisms and their expression in natural and synthetic

communities will help build on our understanding of when and how specialized metabolism is regulated in an ecosystem context.

Using a strain-resolved metagenomics approach, we tracked genetic variation in abundant bacterial species, demonstrating that intra-species genetic diversity is often spatially structured for multiple abundant bacterial populations in soils. We also showed that genetic variants are often recombined, and the frequencies of these recombinant variants implies that for the soil bacterial species studied, there is no easily measurable number of clonal strain genotypes at our study site - rather, with high genetic diversity and combinatorial recombination, the number of unique genotypes across the populations is likely quite large. We also identified genomic signatures of gene-specific selection in these populations - indicating that due to recombination unique environmental conditions across a soil landscape may select for specific genes and gene variants, as opposed to selection for entire clonal lineages. There are two primary implications that arise. The first is that it may be numerically infeasible to isolate all genetically unique genotypes for a species (even ignoring extremely rare variants) in a local soil ecosystem. This is unlike microbial colonists of the human gut, where it can be feasible to isolate all of the major genotypes in a population. However, it is reminiscent of eukaryotic populations, where due to the frequency of sexual recombination every generation, every individual in the population is a unique genetic mixture. Soil microbial populations may therefore exist somewhere on this continuum. We also were able to track genetic variation within specific genes - and for some populations, we saw that biosynthetic genes were often more genetically diverse than the average gene within species. This may indicate that biosynthetic genes are under diversifying selective pressures in some natural ecosystems. Future studies may want to isolate cultures from an ecosystem and characterize their biosynthetic gene clusters, and then use strain-resolved metagenomics to track specific variants in these genes across natural settings.

Finally, we were able to gain some insight into prediction of biosynthetic gene cluster function using transporter gene sequences. For a while it has been colloquially known that transporter genes in a biosynthetic gene cluster can be useful for guessing the resulting metabolite's function. However, we were able to quantify the predictive power of these relationships, and demonstrate that especially for siderophores, the relationship is highly specific. We were then able to use those predictions to annotate un-annotated biosynthetic genes in genomes from the human microbiome. However, when applied to more divergent genomes of novel organisms from soil, we found the sets of annotations used were often absent entirely from biosynthetic gene clusters. This may be because many highly divergent soil organisms use transporter genes that are correspondingly more divergent, and difficult to annotate - indicating that there is yet still a lot of work to be done for prediction of biosynthetic gene function in many of the novel organisms we reported on here. This will likely be the biggest challenge for research on specialized metabolism going forward. While we have shed light on the genomics of specialized metabolism in uncultivated

microbes, bridging the gap between genomes and molecular characterization will likely be the next step forward to build on these findings.

References

Ahmed, E., and S. J. M. Holmström. 2014. “Siderophores in Environmental Research: Roles and Applications.” Microbial Biotechnology 7 (3): 196–208.

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

Alteio, L. V., F. Schulz, R. Seshadri, N. Varghese, W. Rodriguez-Reillo, E. Ryan, D. Goudeau et al. "Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil." Msystems 5, no. 2 (2020).

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

Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, et al. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat Commun. 2017;8:1114.

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

Aparicio, Jesús F., István Molnár, Torsten Schwecke, Ariane König, Stephen F. Haydock, Lake Ee Khaw, James Staunton, and Peter F. Leadlay. "Organization of the biosynthetic gene cluster for rapamycin in Streptomyces hygroscopicus: analysis of the enzymatic domains in the modular polyketide synthase." Gene 169, no. 1 (1996): 9-16.

Arnold BJ, Sohail M, Wadsworth C, Corander J, Hanage WP, Sunyaev S, et al. Fine-scale haplotype structure reveals strong signatures of positive selection in a recombining bacterial pathogen. https://www.biorxiv.org/content/10.1101/634147v1. 2019.

Arnott, John A., and Sonia Lobo Planey. 2012. “The Influence of Lipophilicity in Drug Discovery and Design.” Expert Opinion on Drug Discovery 7 (10): 863–75.

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).

Barns, Susan M., Elizabeth C. Cain, Leslie Sommerville, and Cheryl R. Kuske. "Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum." Applied and environmental microbiology 73, no. 9 (2007): 3113-3116.

Beek, Josy Ter, Josy ter Beek, Albert Guskov, and Dirk Jan Slotboom. 2014. “Structural Diversity of ABC Transporters.” The Journal of General Physiology. https://doi.org/10.1085/jgp.201411164.

Behnsen, Judith, and Manuela Raffatellu. 2016. “Siderophores: More than Stealing Iron.” mBio. https://doi.org/10.1128/mBio.01906-16.

Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodol). 1995;57:289–300.
Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem. 2011;43:1450–5.
Berhe AA, Suttle KB, Burton SD, Banfield JF. Contingency in the direction and mechanics of soil organic matter responses to increased rainfall. Plant Soil. 2012;358:371–83.
Berntsson, Ronnie P-A, Ronnie P. -A. Berntsson, Sander H. J. Smits, Lutz Schmitt, Dirk-Jan Slotboom, and Bert Poolman. 2010. "A Structural Classification of Substrate-Binding Proteins." FEBS Letters. https://doi.org/10.1016/j.febslet.2010.04.043.
Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Res. 2013;23:1514–21.
Bi, Yunchen, Evan Mann, Chris Whitfield, and Jochen Zimmer. 2018. "Architecture of a Channel-Forming O-Antigen Polysaccharide ABC Transporter." Nature 553 (7688): 361–65.
Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45:W36–41.
Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. "antiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline." Nucleic Acids Research 47 (W1): W81–87.
Blin, Kai, Victòria Pascal Andreu, Emmanuel L. C. de Los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. "The antiSMASH Database Version 2: A Comprehensive Resource on Secondary Metabolite Biosynthetic Gene Clusters." Nucleic Acids Research 47 (D1): D625–30.
Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
Brown, C.T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211 (2015).
Bushnell B. BBTools software package. http://sourceforge.net/projects/bbmap. 2014.
Bushnell, B. BBMap short read aligner. http://sourceforge.net/projects/bbmap (University of California, Berkeley, 2016).
Butterfield CN, Li Z, Andeer PF, Spaulding S, Thomas BC, Singh A, et al. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ. 2016;4:e2687.
Bérdy, J. Bioactive microbial metabolites. J. Antibiot. (Tokyo) 58, 1–26 (2005).
Carran, C. J., M. Jordan, H. Drechsel, D. G. Schmid, and G. Winkelmann. 2001. "Heterobactins: A New Class of Siderophores from Rhodococcus Erythropolis IGTS8 Containing Both Hydroxamate and Catecholate Donor Groups." Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine 14 (2): 119–25.

Cassier-Chauvat C, Veaudor T, Chauvat F. Comparative genomics of DNA recombination and repair in cyanobacteria: biotechnological implications. Front Microbiol. 2016;7:1809.
Charlop-Powers, Z. et al. Global biogeographic sampling of bacterial secondary metabolism. eLife 4, e05048 (2015).
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. USA 111, 3757–3762 (2014).
Chen, Liang, Justin E. Wilson, Mark J. Koenigsknecht, Wei-Chun Chou, Stephanie A. Montgomery, Agnieszka D. Truax, W. June Brickey, et al. 2017. "NLRP12 Attenuates Colon Inflammation by Maintaining Colonic Microbial Diversity and Promoting Protective Commensal Bacterial Growth." Nature Immunology 18 (5): 541–51.
Chen, Yunqiu, Michelle Unger, Ioanna Ntai, Ryan A. McClure, Jessica C. Albright, Regan J. Thomson, and Neil L. Kelleher. 2013. "Gobichelin A and B: Mixed-Ligandsiderophores Discovered Using Proteomics." MedChemComm. https://doi.org/10.1039/c2md20232h.
Chevrette MG, Carlos-Shanley C, Louie KB, Bowen BP, Northen TR, Currie CR. Taxonomic and metabolic incongruence in the ancient genus Streptomyces. Front Microbiol. 2019;10:2170.
Chevrette, Marc G., Karina Gutiérrez-García, Nelly Selem-Mojica, César Aguilar-Martínez, Alan Yañez-Olvera, Hilda E. Ramos-Aboites, Paul A. Hoskisson, and Francisco Barona-Gómez. 2019. "Evolutionary Dynamics of Natural Product Biosynthesis in Bacteria." Natural Product Reports, December. https://doi.org/10.1039/c9np00048h.
Chu, Byron C., Alicia Garcia-Herrero, Ted H. Johanson, Karla D. Krewulak, Cheryl K. Lau, R. Sean Peacock, Zoya Slavinskaya, and Hans J. Vogel. 2010. "Siderophore Uptake in Bacteria and the Battle for Iron with the Host; a Bird's Eye View." Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine 23 (4): 601–11.
Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. "Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters." Cell 158 (2): 412–21.
Claessen, D., de Jong, W., Dijkhuizen, L. & Wösten, H. A. Regulation of Streptomyces development: reach for the sky. Trends Microbiol. 14, 313–319 (2006).
Cornforth, Daniel M., and Kevin R. Foster. "Competition sensing: the social side of bacterial stress responses." Nature Reviews Microbiology 11, no. 4 (2013): 285-293.
Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. Biochim. Biophys. Acta 1830, 3670–3695 (2013).
Crits-Christoph, Alexander, Spencer Diamond, Cristina N. Butterfield, Brian C. Thomas, and Jillian F. Banfield. 2018. "Novel Soil Bacteria Possess Diverse Genes for Secondary Metabolite Biosynthesis." Nature 558 (7710): 440–44.
Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al. Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 2015;32:1396–410.

Davies, Julian. 2013. “Specialized Microbial Metabolites: Functions and Origins.” The Journal of Antibiotics 66 (7): 361–64.
Dawson, Roger J. P., and Kaspar P. Locher. 2007. “Structure of the Multidrug ABC Transporter Sav1866 fromStaphylococcus Aureusin Complex with AMP-PNP.” FEBS Letters. https://doi.org/10.1016/j.febslet.2007.01.073.
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).
Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, Anantharaman K, et al. Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. Nature Microbiol. 2019.
Doroghazi JR, Buckley DH. Widespread homologous recombination within and between Streptomyces species. ISME J. 2010;4:1136–43.
Duvallet, Claire, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm. 2017. “Meta-Analysis of Gut Microbiome Studies Identifies Disease-Specific and Shared Responses.” Nature Communications 8 (1): 1–10.
Eddy, S. R. 1998. “Profile Hidden Markov Models.” Bioinformatics. https://doi.org/10.1093/bioinformatics/14.9.755.
Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi et al. "The Pfam protein families database in 2019." Nucleic acids research 47, no. D1 (2019): D427-D432.
Ellermann, Melissa, and Janelle C. Arthur. 2017. “Siderophore-Mediated Iron Acquisition and Modulation of Host-Bacterial Interactions.” Free Radical Biology & Medicine 105 (April): 68–78.
Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC. Global patterns in belowground communities. Ecol Lett. 2009;12:1238–49.
Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol. 2017;15:579–90.
Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2013;42:D222–30. Nov
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).
Fischbach, Michael A., and Christopher T. Walsh. 2006. “Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms.” Chemical Reviews 106 (8): 3468–96.

Fischbach, Michael A., and Jon Clardy. 2007. “One Pathway, Many Products.” Nature Chemical Biology. https://doi.org/10.1038/nchembio0707-353.
Fischbach, Michael A., Christopher T. Walsh, and Jon Clardy. 2008. “The Evolution of Gene Collectives: How Natural Selection Drives Chemical Innovation.” Proceedings of the National Academy of Sciences of the United States of America 105 (12): 4601–8.
Franklin, James. 2005. “The Elements of Statistical Learning: Data Mining, Inference and Prediction.” The Mathematical Intelligencer. https://doi.org/10.1007/bf02985802.
Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. 2019;17:e3000102.
Gebhard, Susanne. 2012. “ABC Transporters of Antimicrobial Peptides in Firmicutes Bacteria - Phylogeny, Function and Regulation.” Molecular Microbiology 86 (6): 1295–1317.
Genilloud, Olga. "Actinomycetes: still a source of novel antibiotics." Natural product reports 34, no. 10 (2017): 1203-1232.
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).
González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. Impact of homologous recombination on the evolution of prokaryotic core genomes. MBio. 2019;10:e02494–18.
Greene, Nicholas P., Elise Kaplan, Allister Crow, and Vassilis Koronakis. 2018a. “Antibiotic Resistance Mediated by the MacB ABC Transporter Family: A Structural and Functional Perspective.” Frontiers in Microbiology 9. https://doi.org/10.3389/fmicb.2018.00950.
Guo, X., P. Geng, F. Bai, G. Bai, T. Sun, X. Li, L. Shi, and Q. Zhong. 2012. “Draft Genome Sequence of Streptomyces Coelicoflavus ZG0656 Reveals the Putative Biosynthetic Gene Cluster of Acarviostatin Family α-Amylase Inhibitors.” Letters in Applied Microbiology 55 (2): 162–69.
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).
Hannigan, Geoffrey D., David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. 2019. “A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction.” Nucleic Acids Research 47 (18): e110.
Hawkes CV, Kivlin SN, Rocca JD, Huguet V, Thomsen MA, Suttle KB. Fungal community responses to precipitation. Glob Change Biol. 2011;17:1637–45.
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).
Hider, Robert C., and Xiaole Kong. 2010. “Chemistry and Biology of Siderophores.” Natural Product Reports. https://doi.org/10.1039/b906679a.
Hochberg, Yosef, and Yoav Benjamini. 1990. “More Powerful Procedures for Multiple Significance Testing.” Statistics in Medicine. https://doi.org/10.1002/sim.4780090710.

Holden, Victoria I., Paul Breen, Sébastien Houle, Charles M. Dozois, and Michael A. Bachman. 2016. "Klebsiella Pneumoniae Siderophores Induce Inflammation, Bacterial Dissemination, and HIF-1α Stabilization during Pneumonia." mBio 7 (5). https://doi.org/10.1128/mBio.01397-16.
Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F ST. Nat Rev Genet. 2009;10:639.
Howe, Adina Chuang, Janet K. Jansson, Stephanie A. Malfatti, Susannah G. Tringe, James M. Tiedje, and C. Titus Brown. "Tackling soil diversity with the assembly of large, complex metagenomes." Proceedings of the National Academy of Sciences 111, no. 13 (2014): 4904-4909.
Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132:583–9.
Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).
Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521:208–12.
Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma. 2010;11:119.
Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
Jenke-Kodama, Holger, Thomas Börner, and Elke Dittmann. 2006. "Natural Biocombinatorics in the Polyketide Synthase Genes of the Actinobacterium Streptomyces Avermitilis." PLoS Computational Biology 2 (10): e132.
Jesse Shapiro B, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, et al. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336:48–51.
Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, et al. Atmospheric trace gases support primary production in Antarctic desert surface soil. Nature. 2017;552:400.
Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. Nat. Chem. Biol. 12, 233–239 (2016).
Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12:962–8.
Joshi, N. A. & Fass, J. N. sickle - a windowed adapative trimming tool for FastQ files (version 1.33) https://github.com/najoshi/sickle (2011).
Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(Jan):27–30.
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).
Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J. J. van der Hooft, Jeffrey A. van Santen, et al. 2020. "MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function." Nucleic Acids Research 48 (D1): D454–58.

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).
Kielak, Anna, Agata S. Pijl, Johannes A. Van Veen, and George A. Kowalchuk. "Phylogenetic diversity of Acidobacteria in a former agricultural soil." The ISME journal 3, no. 3 (2009): 378-382.
Kim, Hyun Uk, Kai Blin, Sang Yup Lee, and Tilmann Weber. 2017. "Recent Development of Computational Resources for New Antibiotics Discovery." Current Opinion in Microbiology 39 (October): 113–20.
Klingenberg, H. & Meinicke, P. How to normalize metatranscriptomic count data for differential expression analysis. PeerJ 5, e3859 (2017).
Koskiniemi, S. et al. Rhs proteins from diverse bacteria mediate intercellular competition. Proc. Natl Acad. Sci. USA 110, 7032–7037 (2013).
Kramer, Jos, Özhan Özkaya, and Rolf Kümmerli. 2020. "Bacterial Siderophores in Community and Host Interactions." Nature Reviews Microbiology. https://doi.org/10.1038/s41579-019-0284-4.
Krause DJ, Whitaker RJ. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 2015;64:926–35.
Lam, Margaret M. C., Kelly L. Wyres, Louise M. Judd, Ryan R. Wick, Adam Jenney, Sylvain Brisse, and Kathryn E. Holt. 2018. "Tracking Key Virulence Loci Encoding Aerobactin and Salmochelin Siderophore Synthesis in Klebsiella Pneumoniae." Genome Medicine 10 (1): 77.
Langfelder, P & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357.
Lee, Sang-Hoon, Jong-Ok Ka, and Jae-Chang Cho. "Members of the phylum Acidobacteria are dominant and metabolically active in rhizosphere soil." FEMS microbiology letters 285, no. 2 (2008): 263-269.
Lewis, T. E., I. Sillitoe, and J. G. Lees. 2019. "Cath-Resolve-Hits: A New Tool That Resolves Domain Matches Suspiciously Quickly." Bioinformatics 35 (10): 1766–67.
Lin M, Kussell E. Inferring bacterial recombination rates from large-scale sequencing datasets. Nat Methods. 2019;16:199.
Liu, Yang, Meng Li, Huiyan Mu, Shuting Song, Ying Zhang, Kun Chen, Xihong He, et al. 2017. "Identification and Characterization of the Ficellomycin Biosynthesis Gene Cluster from Streptomyces Ficellus." Applied Microbiology and Biotechnology 101 (20): 7589–7602.
Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems. 2018;3:e00055–18.
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).
Martín, Juan F., Javier Casqueiro, and Paloma Liras. 2005. "Secretion Systems for Secondary Metabolites: How Producer Cells Send out Messages of Intercellular Communication." Current Opinion in Microbiology 8 (3): 282–93.

Mathavan, Indran, Séverine Zirah, Shahid Mehmood, Hassanul G. Choudhury, Christophe Goulard, Yanyan Li, Carol V. Robinson, Sylvie Rebuffat, and Konstantinos Beis. 2014. “Structural Basis for Hijacking Siderophore Receptors by Antimicrobial Lasso Peptides.” Nature Chemical Biology 10 (5): 340–42.
Matsuo, Yoshihide, Kaneo Kanoh, Jae-Hyuk Jang, Kyoko Adachi, Satoru Matsuda, Osamu Miki, Toshiaki Kato, and Yoshikazu Shizuri. 2011. “Streptobactin, a Tricatechol-Type Siderophore from Marine-Derived Streptomyces Sp. YM5-799.” Journal of Natural Products 74 (11): 2371–76.
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).
Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol. 11, 625–631 (2015).
Medema, M. H., et al. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput. Biol. 10, e1004016 (2014).
Medema, Marnix H., and Michael A. Fischbach. 2015. “Computational Approaches to Natural Product Discovery.” Nature Chemical Biology. https://doi.org/10.1038/nchembio.1884.
Meij, Anne van der, Sarah F. Worsley, Matthew I. Hutchings, and Gilles P. van Wezel. 2017. “Chemical Ecology of Antibiotic Production by Actinomycetes.” FEMS Microbiology Reviews 41 (3): 392–416.
Miles A, Harding N.cggh/scikit-allel: v1. 1.8. https://doi.org/10.5281/zenodo.822784. 2017
Mungan, Mehmet Direnç, Mohammad Alanjary, Kai Blin, Tilmann Weber, Marnix H. Medema, and Nadine Ziemert. 2020. “ARTS 2.0: Feature Updates and Expansion of the Antibiotic Resistant Target Seeker for Comparative Genome Mining.” Nucleic Acids Research 48 (W1): W546–52.
Méndez, C., and J. A. Salas. 2001. “The Role of ABC Transporters in Antibiotic-Producing Organisms: Drug Secretion and Resistance Mechanisms.” Research in Microbiology 152 (3-4): 341–50.
Nagoba, Basavraj, and Deepak Vedpathak. 2011. “Medical Applications of Siderophores.” Electronic Journal of General Medicine. https://doi.org/10.29333/ejgm/82743.
Navarro-Muñoz, Jorge C., Nelly Selem-Mojica, Michael W. Mullowney, Satria A. Kautsar, James H. Tryon, Elizabeth I. Parkinson, Emmanuel L. C. De Los Santos, et al. 2020. “A Computational Framework to Explore Large-Scale Biosynthetic Diversity.” Nature Chemical Biology 16 (1): 60–68.
Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu et al. "A genomic catalog of Earth’s microbiomes." Nature biotechnology (2020): 1-11.
Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. “New Insights from Uncultivated Genomes of the Global Human Gut Microbiome.” Nature 568 (7753): 505–10.
Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA. 1979;76:5269–73.
Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’hara RB, et al. Package ‘vegan’. Community Ecol Package, Vers. 2013;2:1–295.
Oksanen, J. et al. vegan: Community ecology package https://cran.r-project.org/package=vegan (2007).

Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
Olm, Matthew R., Nicholas Bhattacharya, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Yun S. Song, Michael J. Morowitz, and Jillian F. Banfield. 2019. "Necrotizing Enterocolitis Is Preceded by Increased Gut Bacterial Replication, Klebsiella, and Fimbriae-Encoding Bacteria." Science Advances. https://doi.org/10.1126/sciadv.aax5727.
Osbourn, Anne. 2010. "Secondary Metabolic Gene Clusters: Evolutionary Toolkits for Chemical Innovation." Trends in Genetics: TIG 26 (10): 449–57.
O'Brien SL, Gibbons SM, Owens SM, Hampton-Marcell J, Johnston ER, Jastrow JD, et al. Spatial scale drives patterns in soil bacterial diversity. Environ Microbiol. 2016;18:2039–51.
Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
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).
Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." Nature Biotechnology 36 (10): 996–1004.
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).
Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
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).
Pimentel ZT, Zhang Y. Evolution of the natural transformation protein, ComEC, in bacteria. Front Microbiol. 2018;9:2980.
Price, M. N., Dehal, P. S. and Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS ONE 5, e9490 (2010).
Quistgaard, Esben M., Christian Löw, Fatma Guettou, and Pär Nordlund. 2016. "Understanding Transport by the Major Facilitator Superfamily (MFS): Structures Pave the Way." Nature Reviews. Molecular Cell Biology 17 (2): 123–32.
R Development Core Team. The R reference manual: base package. Network Theory. 2003. 736 p.
Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394 (2003).

Rascher, Andreas, Zhihao Hu, Greg O. Buchanan, Ralph Reid, and C. Richard Hutchinson. 2005. “Insights into the Biosynthesis of the Benzoquinone Ansamycins Geldanamycin and Herbimycin, Obtained by Gene Sequencing and Disruption.” Applied and Environmental Microbiology 71 (8): 4862–71.
Rees, Douglas C., Eric Johnson, and Oded Lewinson. 2009. “ABC Transporters: The Power to Change.” Nature Reviews. Molecular Cell Biology 10 (3): 218–27.
Rempel, S., C. Gati, M. Nijland, C. Thangaratnarajah, A. Karyolaimos, J. W. de Gier, A. Guskov, and D. J. Slotboom. 2020. “A Mycobacterial ABC Transporter Mediates the Uptake of Hydrophilic Compounds.” Nature 580 (7803): 409–12.
Rocha EPC, Cornet E, Michel B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet. 2005;1:e15.
Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239:226–35.
Romano, Maria, Giuliana Fusco, Hassanul G. Choudhury, Shahid Mehmood, Carol V. Robinson, Séverine Zirah, Julian D. Hegemann, et al. 2018. “Structural Basis for Natural Product Selection and Export by Bacterial ABC Transporters.” ACS Chemical Biology 13 (6): 1598–1609.
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).
Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.
Rosen MJ, Davison M, Fisher DS, Bhaya D. Probing the ecological and evolutionary history of a thermophilic cyanobacterial population via statistical properties of its microdiversity. PLoS ONE. 2018;13(Nov):e0205396.
Saier, Milton H., Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. 2016. “The Transporter Classification Database (TCDB): Recent Advances.” Nucleic Acids Research. https://doi.org/10.1093/nar/gkv1103.
Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species. https://www.biorxiv.org/content/10.1101/601914v1. 2019.
Sangwan, Naseer, Fangfang Xia, and Jack A. Gilbert. "Recovering complete and draft population genomes from metagenome datasets." Microbiome 4, no. 1 (2016): 1-11.
Severi, Emmanuele, and Gavin H. Thomas. 2019. “Antibiotic Export: Transporters Involved in the Final Step of Natural Product Production.” Microbiology 165 (8): 805–18.
Shapiro BJ, David LA, Friedman J, Alm EJ. Looking for Darwin’s footprints in the microbial world. Trends Microbiol. 2009;17:196–204.
Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016;31:116–23.
Sharon, Gil, Neha Garg, Justine Debelius, Rob Knight, Pieter C. Dorrestein, and Sarkis K. Mazmanian. 2014. “Specialized Metabolites from the Microbiome in Health and Disease.” Cell Metabolism 20 (5): 719–30.

Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Methods https://doi-org.libproxy.berkeley.edu/10.1038/s41564-018-0171-1 (2018).
Sillitoe, Ian, Natalie Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, Paul Ashford, Adeyelu Tolulope, et al. 2019. "CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences." Nucleic Acids Research 47 (D1): D280–84.
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).
Skinnider, Michael A., Nishanth J. Merwin, Chad W. Johnston, and Nathan A. Magarvey. 2017. "PRISM 3: Expanded Prediction of Natural Product Chemical Structures from Microbial Genomes." Nucleic Acids Research 45 (W1): W49–54.
Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 2007;89:391–403.
Staudenmaier, H., B. Van Hove, Z. Yaraghi, and V. Braun. 1989. "Nucleotide Sequences of the fecBCDE Genes and Locations of the Proteins Suggest a Periplasmic-Binding-Protein-Dependent Transport Mechanism for iron(III) Dicitrate in Escherichia Coli." Journal of Bacteriology 171 (5): 2626–33.
Steen, Andrew D., Alexander Crits-Christoph, Paul Carini, Kristen M. DeAngelis, Noah Fierer, Karen G. Lloyd, and J. Cameron Thrash. "High proportions of bacteria and archaea across most biomes remain uncultured." The ISME journal 13, no. 12 (2019): 3126-3130.
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).
Sullivan MJP, Thomsen MA, Suttle KB. Grassland responses to increased rainfall depend on the timescale of forcing. Glob Chang Biol. 2016;22:1655–65.
Suttle KB, Thomsen MA, Power ME. Species interactions reverse grassland responses to changing climate. Science. 2007;315:640–2.
Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926–32.
Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat. Biotechnol. 31, 922–927 (2013).
Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3:711.
Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457.
Tran F, Boedicker JQ. Plasmid characteristics modulate the propensity of gene exchange in bacterial vesicles. J Bacteriol. 2019;201:e00430–18.
Tran, Phuong Nguyen, Ming-Ren Yen, Chen-Yu Chiang, Hsiao-Ching Lin, and Pao-Yang Chen. 2019. "Detecting and Prioritizing Biosynthetic Gene Clusters for Bioactive Compounds in Bacteria and Fungi." Applied Microbiology and Biotechnology 103 (8): 3277–87.

Traxler, Matthew F., and Roberto Kolter. "Natural products in soil microbe interactions and evolution." Natural Product Reports 32, no. 7 (2015): 956-970.
Tyc, Olaf, Chunxu Song, Jeroen S. Dickschat, Michiel Vos, and Paolina Garbeva. 2017. "The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria." Trends in Microbiology 25 (4): 280–92.
Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature 428, no. 6978 (2004): 37-43.
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).
UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(Mar):2699.
VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol. 2008;74:130–7.
Velamakanni, Saroj, Yao Yao, Daniel A. P. Gutmann, and Hendrik W. van Veen. 2008. "Multidrug Transport by the ABC Transporter Sav1866 from Staphylococcus Aureus." Biochemistry 47 (35): 9300–9308.
Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev. 2013;37:936–54.
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).
Wang, Zhiming, Wenxin Hu, and Hongjin Zheng. 2020. "Pathogenic Siderophore ABC Importer YbtPQ Adopts a Surprising Fold of Exporter." Science Advances 6 (6): eaay7997.
Weakland, Danelle R., Sara N. Smith, Bailey Bell, Ashootosh Tripathi, and Harry L. T. Mobley. 2020. "The Serratia Marcescens Siderophore, Serratiochelin, Is Necessary for Full Virulence during Bloodstream Infection." Infection and Immunity. https://doi.org/10.1128/iai.00117-20.
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).
Wenzel, Silke C., and Rolf Müller. 2005. "Formation of Novel Secondary Metabolites by Bacterial Multimodular Assembly Lines: Deviations from Textbook Biosynthetic Logic." Current Opinion in Chemical Biology 9 (5): 447–58.
Whitaker RJ, Banfield JF. Population genomics in natural microbial communities. Trends Ecol Evol. 2006;21:508–16.
White RA, Bottos EM, Chowdhury TR, Zucker JD, Brislawn CJ, Nicora CD, et al. Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems. 2016;1:e00045–16.

Whitman, Thea, Charles Pepe-Ranney, Akio Enders, Chantal Koechli, Ashley Campbell, Daniel H. Buckley, and Johannes Lehmann. "Dynamics of microbial community composition and soil organic carbon mineralization in soil following addition of pyrogenic and fresh organic matter." The ISME journal 10, no. 12 (2016): 2918-2930.
Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ, et al. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. ISME J. 2016;10:2468–77.
Wildman, Scott A., and Gordon M. Crippen. 1999. "Prediction of Physicochemical Parameters by Atomic Contributions." Journal of Chemical Information and Computer Sciences. https://doi.org/10.1021/ci990307l.
Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506, 58–62 (2014).
Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, et al. Genome-centric view of carbon processing in thawing permafrost. Nature. 2018;560:49.
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).
Xu, Li, He Huang, Wei Wei, Yi Zhong, Biao Tang, Hua Yuan, Li Zhu et al. "Complete genome sequence and comparative genomic analyses of the vancomycin-producing Amycolatopsis orientalis." BMC genomics 15, no. 1 (2014): 1-18.
Xue, Yaxin, Inge Jonassen, Lise Øvreås, and Neslihan Taş. "Metagenome-assembled genome distribution and key functionality highlight importance of aerobic metabolism in Svalbard permafrost." FEMS microbiology ecology 96, no. 5 (2020): fiaa057.
Yan, Yan, Nicholas Liu, and Yi Tang. 2020. "Recent Developments in Self-Resistance Gene Directed Natural Product Discovery." Natural Product Reports, January. https://doi.org/10.1039/c9np00050j.
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).
Zhu, Wenhan, Maria G. Winter, Luisella Spiga, Elizabeth R. Hughes, Rachael Chanin, Aditi Mulgaonkar, Jenelle Pennington, et al. 2020. "Xenosiderophore Utilization Promotes Bacteroides Thetaiotaomicron Resilience during Colitis." Cell Host & Microbe 27 (3): 376–88.e8.