TaxiBGC: a Taxonomy-guided Approach for the Identification of Experimentally Verified Microbial Biosynthetic Gene Clusters in Shotgun Metagenomic Data

Utpal Bakshi[1,2], Vinod K. Gupta[1,2], Aileen R. Lee[3], John M. Davis III[4], Sriram Chandrasekaran[5,6,7,8], Yong-Su Jin[9,10], Michael F. Freeman[3], and Jaeyun Sung[1,2,4,*]

[1]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
[2]Division of Surgery Research, Department of Surgery, Mayo Clinic, Rochester, MN 55905, USA
[3]Department of Biochemistry, Molecular Biology, and Biophysics and BioTechnology Institute, University of Minnesota-Twin Cities, St. Paul, MN 55108, USA
[4]Division of Rheumatology, Department of Medicine, Mayo Clinic, Rochester, MN 55905, USA
[5]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, 48109, USA
[6]Program in Chemical Biology, University of Michigan, Ann Arbor, MI, 48109, USA
[7]Center for Bioinformatics and Computational Medicine, University of Michigan, Ann Arbor, MI, 48109, USA
[8]Rogel Cancer Center, University of Michigan Medical School, Ann Arbor, MI, 48109, USA
[9]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[10]Department of Food Science and Human Nutrition, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[*]To whom correspondence should be addressed. Tel: +1-507-538-3210; Email: Sung.Jaeyun@mayo.edu

ABSTRACT

Biosynthetic gene clusters (BGCs) in microbial genomes encode for the production of bioactive secondary metabolites (SMs). Given the well-recognized importance of SMs in microbe-microbe and microbe-host interactions, the large-scale identification of BGCs from microbial metagenomes could offer novel functional insights into complex chemical ecology. Despite recent progress, currently available tools for predicting BGCs from shotgun metagenomes have several limitations, including the need for computationally demanding read-assembly and prediction of a narrow breadth of BGC classes. To overcome these limitations, we developed TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters), a computational pipeline for identifying experimentally verified BGCs in shotgun metagenomes by first pinpointing the microbial species likely to produce them. We show that our species-centric approach was able to identify BGCs in simulated metagenomes more accurately than by solely detecting BGC genes. By applying TaxiBGC on 5,423 metagenomes from the Human Microbiome Project and various case-control studies, we identified distinct BGC signatures of major human body sites and candidate stool-borne biomarkers for multiple diseases, including inflammatory bowel disease, colorectal cancer, and psychiatric disorders. In all, TaxiBGC demonstrates a significant advantage over existing techniques for systematically characterizing BGCs and inferring their SMs from microbiome data.

INTRODUCTION

Microbial-derived secondary metabolites (SMs) are a group of low molecular-weight and structurally diverse, bioactive, chemical compounds (1–3). Although not essential for the primary growth, development, and reproduction of microorganisms, these compounds—which include pigments, bacteriocins, and siderophores—are known to play key biological roles in mediating interactions within complex ecosystems (4–6). Interestingly, these SMs have recently been shown to play critical roles in regulating many aspects of human health and disease (7–11). For example, pyrazinones of *Staphylococcus aureus* could induce bacterial virulence (12); colibactin, which is a genotoxin produced by *Escherichia coli*, was found to contribute to colon cancer (13–16); and polysaccharide A from *Bacteroides fragilis* was shown to suppress gut mucosal inflammatory response (17, 18). In addition, *B. fragilis* was able to produce the canonical CD1d ligand α-galactosylceramide, revealing a specific mechanism by which the gut microbiota is capable of modulating host natural killer T cell function (19). Lastly, microbial-derived SMs are a major source of antibiotics, antifungals, anticancer agents, immunosuppressants, and other pharmaceutical drugs (20–24). Major categories of SMs and a few examples of their biomedical applications are summarized in **Supplementary Information**.

For the large-scale mining of SMs from microbial communities, researchers have harnessed recent advances in molecular biology and sequencing techniques to connect SMs to their genetically encoded biosynthetic machinery (25). The genes responsible for the synthesis, export, and regulation of SMs in microbes are often found in biosynthetic gene clusters (BGCs), which are physically clustered groups of genes in a microbial genome that together encode a biosynthetic pathway for the production of a SM and its chemical variants (26, 27). Recent advances in genomic and metagenomic sequencing technologies have enabled the prediction and discovery of novel SMs (9, 28). As such, recent studies have newly identified a vast number of previously unknown BGCs and their corresponding SMs (27, 29–34), including from human microbiomes (23, 35).

Computational detection of BGCs (which allows the inference of SMs) from microbial genomes has evolved in parallel with the exponential rise in publicly available microbial genomes and metagenome sequences. Early approaches of BGC detection from microbial genomes (e.g., antiSMASH (36), CLUSEAN (37), PRISM (38)) used heuristic, rule-based algorithms based upon gene/protein domain compositional similarity to annotated reference BGCs. Other tools that utilize machine-learning models have emerged, demonstrating a greater ability to discover novel BGCs. One such widely used method is ClusterFinder (27), which employs a Hidden Markov Model-based approach for BGC detection. Additionally, DeepBGC (39) implements recurrent neural networks for BGC identification. These rule-based and machine learning-based approaches have been shown to detect a wide variety of BGCs. Moreover, Navarro-Muñoz *et al*. provided an integrated computational workflow, which consists of two software tools (BiG-SCAPE and CORASON), for identifying novel gene cluster families and their phylogenetic relationships from microbial genomes (40). Recently, the primary specialized metabolic gene clusters (MGCs) of anaerobic bacteria were extensively studied using gutSMASH (41); this web server determines the metabolic potential of anaerobic bacteria by predicting both known and putative MGCs using the same Pfam domain detection rules employed by

antiSMASH. Notably, gutSMASH is specifically designed to uncover MGCs from the gut microbiome that have been associated with microbe-microbe and host-microbe interactions.

Approaches to predict BGCs have expanded beyond assembled or complete genomes to enable BGC prediction directly from culture-independent analyses of microbial communities via shotgun metagenomics. One such method is BiosyntheticSPAdes (42), which uses assembly graphs to detect clusters encoding nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) from whole-genome shotgun sequences and low-complexity metagenomes with long reads. Recently, MetaBGC (35) demonstrated the characterization of BGCs directly from metagenome reads. This method first detects BGC reads from shotgun metagenome sequences based upon sequence-scoring models; afterwards, the identified reads are binned for targeted assembly for the reconstruction of novel BGCs.

Despite the variety of approaches to detect BGCs from microbial genomes and metagenomes, currently available tools have several shortcomings: Approaches such as antiSMASH (36), CLUSEAN (37), PRISM (38)) can predict BGCs of known pathway classes from only individual microbial genomes. Machine-learning-based methods can predict unknown BGCs, but these methods can generate much more false-positive predictions than rule-based approaches (43). Moreover, these rule-based and machine-learning-based approaches require assembled or complete genome sequences to detect BGCs, which limits the scalability of their application towards metagenomic sequencing datasets. Among metagenome-based BGC detection methods, BiosyntheticSPAdes is limited by the requirement of read-assembly before BGC prediction, and thus requires large computational time and memory consumption. MetaBGC, which is an assembly-independent method, can be used directly on metagenome reads for BGC prediction; however, as mentioned in its publication (35), the requirement of laborious parameter optimization and unavailability of pre-built models for a BGC of interest makes it difficult to use across a wide range of metagenome datasets. Furthermore, neither metagenome-based BGC detection methods pinpoint the specific microbes that harbor the predicted BGC genes. Clearly, there is a need to improve upon current BGC detection methods to enable the large-scale prediction of BGCs (and their corresponding SMs) from microbiomes.

Realizing the above limitations, we set out to develop a method for accurate and rapid detection of BGCs from real-world, complex microbiomes. Herein, we introduce TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters), an original computational pipeline that identifies experimentally verified BGCs from shotgun metagenomic data and infers their known SM products. TaxiBGC is a read-based, assembly-independent BGC prediction method that can be applied to any given metagenome. Importantly, our novel method enables the prediction of experimentally verified BGCs by first considering the microbial species from which they are derived, thereby allowing us to trace the predicted BGCs' likely taxonomic origin. To evaluate its prediction accuracy, we test our TaxiBGC pipeline on simulated metagenomes constructed from varied combinations of species identities, species numbers, and library sizes. Furthermore, we demonstrate TaxiBGC on publicly available microbiomes from healthy human body niches, as well as on gut microbiomes from several disease-control studies. We expect that our taxonomy-guided approach provides a robust analytical method for the large-scale identification of BGCs in microbial metagenomes, and carves a path towards systematically characterizing the chemical ecology of microbiomes.

MATERIALS AND METHODS

Microbial species from the MetaPhlAn2 marker database and their genomes from GenBank database

For BGC prediction, the TaxiBGC pipeline relies on the TaxiBGC database, which is the background database that links BGCs and SMs with microbial species. Microbial species for the construction of the TaxiBGC database were obtained from the MetaPhlAn2 v2.7.8 (44) marker gene database (mpa_v20_m200), which is comprised of clade-specific marker genes derived from 16,903 microbial species (12,926 bacterial, 300 archaeal, 3,565 viral, and 112 eukaryotic species). Viral species were excluded from our analysis, as their genomes are not widely known to harbor BGCs. Uncharacterized species were also excluded from further analysis. The remaining 13,338 species were searched in GenBank (as of November 2019) (45), which led to the download of a total of 380,184 genomes from 4,187 species of archaea, bacteria, and fungi.

BGCs and their SMs for the construction of the TaxiBGC database

BGCs in all 380,184 microbial genomes were predicted using the 'KnownClusterBlast' function in antiSMASH v4.0 (46) on the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database v2.0 (as of November 2019) (47), which is a data repository for BGCs. Of note, MIBiG encompasses annotations of only experimentally verified BGCs that encode known SM-products. In brief, antiSMASH identifies BGCs in genomes using profile Hidden Markov Models that identify BGC enzymatic domains encoded within the same genomic neighbourhoods (46). In antiSMASH, a similarity criterion (i.e., proportion of genes of the BGCs that have significant KnownClusterBlast hits) of 100% was used to predict only the BGCs whose entire repertoire of MIBiG genes were identified in the genomes. A KnownClusterBlast hit was considered significant if e-value < 1E-05; sequence identity > 30%; and alignment coverage > 25%.

Simulated metagenomes

For the construction of simulated metagenomes, microbial communities of 25, 75, and 125 unique species were used to represent different sizes of community compositions (**Supplementary Table S1**). Each community was composed of randomly selected species from the TaxiBGC database (**Supplementary Table S2**). An even distribution of species' abundances was considered for each microbial community, thereby providing equal weight to all species. All strains of selected species (from GenBank) were considered during the creation of microbial communities. The genome sequences of all species in a given microbial community were combined and then converted into paired-end metagenome reads of 150 bp read-length using WGSIM v1.9 in the SAMtools software package (48). The number of reads generated from each genome was proportional to the size of the respective genome. Seven different sets of randomly selected species (from the TaxiBGC database) for each microbial community size (25, 75, and 125 species) were considered for the creation of simulated metagenomes. For each species set, simulated metagenomes with seven different sequencing library sizes (1.25M, 2.5M, 5M, 10M, 20M, 40M, and 80M reads) were constructed (**Supplementary Table S3**). In all, a total of 147 simulated metagenomes ((3 community sizes) × (7 species

sets) × (7 library sizes)) were created, resulting in unique combinations of BGCs (**Supplementary Tables S2** and **S4**)

BGC prediction through direct gene-detection of BGC genes

As an alternative to the taxonomy-guided BGC identification approach, a direct BGC gene-detection strategy was used to predict BGCs in a given metagenome without taxonomic profiling. In this approach, by considering all BGCs in the MIBiG v2.0 database (1,927 total BGCs), we performed prediction by mapping reads of BGC genes using Bowtie 2 v2.3.5 with the 'sensitive' preset for 'end-to-end' alignment (49) without any taxonomic guidance. Accuracies for BGC prediction were calculated based upon known compositions of BGCs (in the case of simulated metagenomes) in the form of the F_1 score.

BGC prediction accuracy

Accuracies for both BGC prediction approaches (TaxiBGC and direct BGC gene-detection approach) were calculated based upon the known composition of BGCs in each simulated metagenome. Both accuracies were then compared to identify the more robust approach. Determining whether a BGC is present or absent in a metagenome is largely dependent on the proportion of its constituent genes found to be available; as such, BGC prediction accuracy was evaluated over a range of thresholds (selected *a priori*) for the minimum proportion of genes that must be detected as present. To select the best gene-presence threshold for predicting BGCs from the metagenomes, multiple thresholds (20%, 40%, 60%, 80%, and 100%) were tested on the simulated metagenomes. The F_1 score, which is defined as the harmonic mean of precision and recall, was used to evaluate prediction accuracy. The threshold that yielded the highest F_1 scores was selected as the optimal BGC gene-presence threshold.

Metagenomic datasets for TaxiBGC pipeline demonstration

The following metagenomic datasets were downloaded and used for TaxiBGC pipeline demonstration: 2,355 metagenomes from the Human Microbiome Project (HMP), whose samples were collected across five distinct body sites (gut, oral, anterior nares, retroauricular crease, and vagina) from healthy subjects (50, 51) (**Supplementary Table S5**); 2,418 stool metagenomes from seven inflammatory bowel disease (IBD) studies (six studies with data from both Crohn's disease (CD) and ulcerative colitis (UC) patients, and one study with only CD) (52–58); 277 stool metagenomes from three colorectal cancer (CRC) studies (59–61); and 189, 74, and 110 stool metagenomes from one study on rheumatoid arthritis (RA) (62), autism spectrum disorder (ASD) (63), and schizophrenia (64), respectively (**Supplementary Table S6**). In summary, the TaxiBGC pipeline was applied on 5,423 human-derived metagenomes from 14 independent studies.

Prior to their downloading, metagenome samples from the following were excluded from further analysis: i) healthy individuals (i.e., controls) from case-control studies if their reported BMI fell outside the range of normal BMI (18.5–24.9), regardless of whether they had been determined as healthy in the original studies; ii) subjects undergoing dietary interventions; and iii) patients with other comorbidities. Raw sequence files (.fastq) for all metagenomes were downloaded from the NCBI Sequence Read Archive. All runs of a sample were merged together if multiple runs were available for a particular sample. If multiple samples were

collected from the same individual at different time-points, then all of those samples were included in the analysis.

Association between BGCs and cohort-specific features

Body-site-specific variations in the presence of microbiome BGCs were identified using 2,355 metagenome samples from the HMP dataset. Hierarchical clustering of the BGC presence/absence profiles from different body sites was performed using the 'pheatmap' function in R based upon pairwise, euclidean distances. In addition, BGCs associated with different pathologies were identified using 3,068 case-control stool metagenomes. The significance of association was calculated by a two-tailed Fisher's exact test using the 'fisher.test' function in R.

RESULTS

Development of the TaxiBGC pipeline for prediction of biosynthetic gene clusters from metagenomes

TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters) is a computational strategy for profiling the composition of experimentally verified BGCs and their annotated SMs from metagenomic shotgun sequencing data (.fastq files). The TaxiBGC pipeline includes three major steps (**Fig. 1A**): the first step of the TaxiBGC pipeline performs species-level, taxonomic profiling on the metagenome using MetaPhlAn2; the second step performs a first-pass prediction of BGCs through querying these species (identified in the first step) in the TaxiBGC database (**Fig. 1B**)—a pre-defined collection of 744 species with their experimentally verified BGCs and associated SMs (**Table 1** and **Supplementary Table S7**); the last step of the TaxiBGC pipeline confirms (*in silico*) the prediction of BGCs (from the second step) based on the detection of BGC genes in the metagenome (via gene-mapping) using Bowtie 2 with a predetermined minimum gene-presence threshold followed by retrieval of their corresponding SMs from the TaxiBGC database. TaxiBGC is fully open access on GitHub (https://github.com/jaeyunsung/TaxiBGC_2021) for anyone to apply on their own metagenomic dataset.

The gene presence of a BGC is defined as the proportion of its total number of genes that were found to be available in a metagenome via gene-mapping (Bowtie 2). We determined the optimum gene-presence threshold of the TaxiBGC pipeline based upon our prediction results from simulated metagenomes: microbial communities of 25, 75, and 125 species (**Fig. 1C**). We considered seven different species combinations per specified community size (i.e., number of species) to account for variability in community compositions. Simulated metagenomes were constructed with varying library sizes: 1.25M, 2.5M, 5M, 10M, 20M, 40M, and 80M paired-end reads. Afterwards, analogous to the steps described in **Fig. 1A**, we predicted BGCs and their SMs from the simulated metagenomes accordingly: i) species-level taxonomic profiling; ii) initial prediction of experimentally verified BGCs based upon the identified species; and iii) *in silico* confirmation of the BGC predictions where BGC genes were searched in the metagenomes using a range of different gene-presence thresholds. We obtained prediction accuracy in the form of an F_1 score by comparing the predicted BGCs with the known (actual) BGCs of the microbial species comprising the simulated metagenomes. The threshold

yielding the highest average F_1 scores across various species sets and library sizes was selected as the optimal BGC gene-presence threshold for our TaxiBGC pipeline (**Methods**).

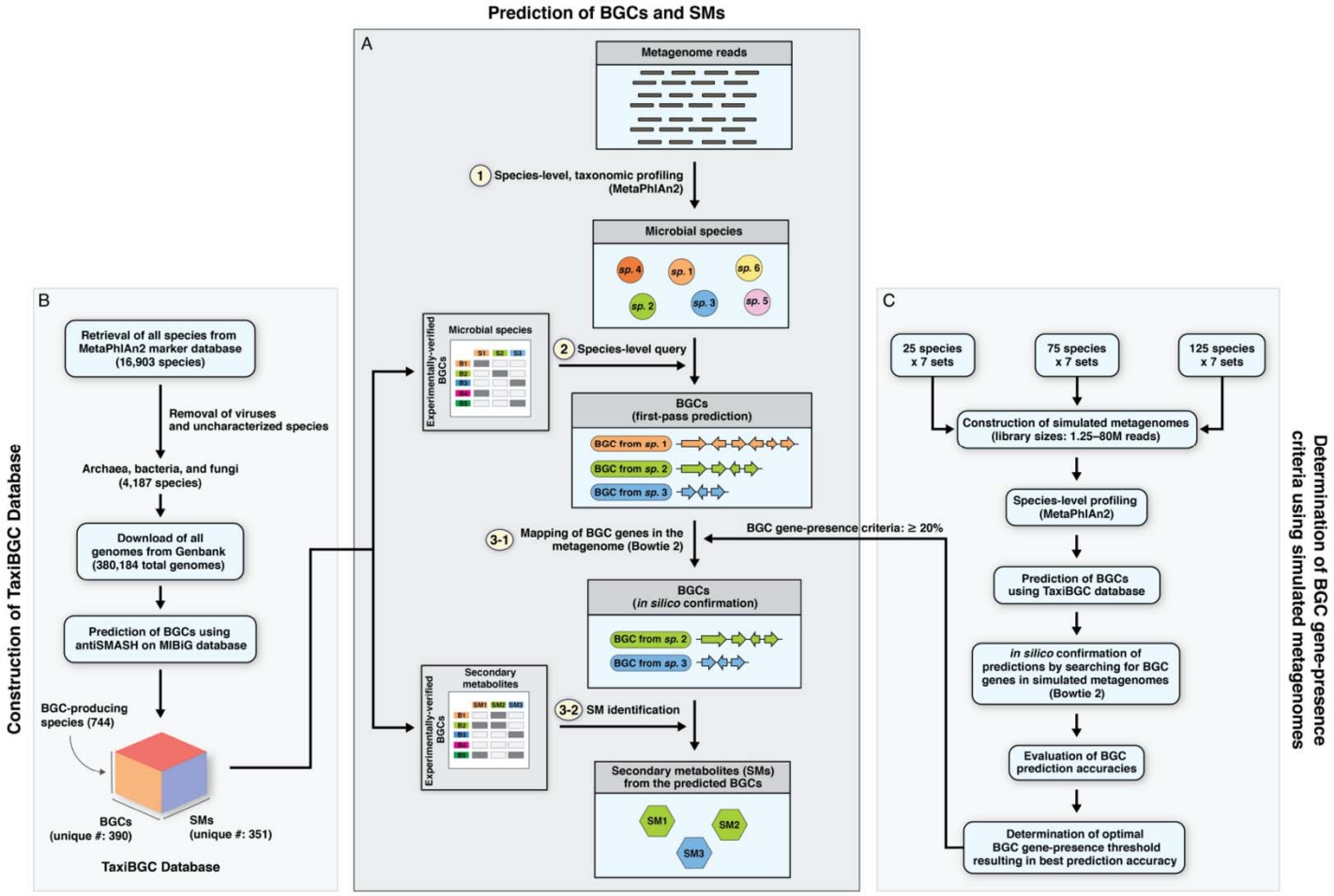

Figure 1. Schematic overview of TaxiBGC for the identification of microbial BGCs in shotgun metagenomes. A) The prediction of BGCs and their corresponding SMs from a metagenome (i.e., microbiome) sample by TaxiBGC can be summarized in three main steps: first, species-level, taxonomic profiling is performed on a metagenome using MetaPhlAn2 with default parameters; second, all detected species are queried in the TaxiBGC database (see **B**) as a first-pass prediction of which experimentally provided BGCs may be present in the metagenome sample; and third, as an *in silico* confirmation of the predicted BGCs, genes of those BGCs are searched in the metagenomes using Bowtie 2 (with the 'sensitive' preset for 'end-to-end' alignment) with a pre-determined minimum gene-presence threshold (see **C**). For the BGCs whose gene presence are confirmed, their corresponding SMs are retrieved from the TaxiBGC database. **B)** The TaxiBGC database provides annotated information on microbial species' experimentally verified BGCs and their SMs (see **Methods**). Genomes of all archaeal, bacterial, and fungali species in the MetaPhlAn2 marker database (while excluding viruses and uncharacterized species) were obtained from GenBank (November 2019). BGCs were predicted in all genomes by using antiSMASH (v4.0) on experimentally verified BGCs from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository. BGCs and information regarding their SM products were retrieved from MIBiG. The final TaxiBGC database encompasses 744 BGC-encoding microbial species, 390 experimentally verified BGCs, and 351 SMs. **C)** Determining the optimal BGC gene-presence threshold using simulated metagenomes of microbial communities. Simulated metagenomes were created for microbial communities of 25, 75, and 125 species (**Methods**). To account for variability in community compositions, seven different species combinations

per specified community size (i.e., number of species) were considered. Simulated metagenomes were created with varying library sizes: 1.25M, 2.5M, 5M, 10M, 20M, 40M, and 80M paired-end reads. Afterwards, as described in part by **A**, BGCs and their SMs were predicted from the simulated metagenomes by TaxiBGC in three main steps: i) species-level taxonomic profiling; ii) prediction of initial prediction of experimentally verified BGCs based upon the identified species; and iii) for *in silico* confirmation of the BGC predictions, where genes of the BGCs were searched in the metagenomes (via gene-mapping using Bowtie 2 with the 'sensitive' preset for 'end-to-end' alignment) using a range of different gene-presence thresholds (**Methods**). Here, gene presence of a BGC is defined as the proportion of its total number of genes that were mapped (i.e., found to be available, in principle) in a metagenome. Prediction accuracy in the form of an F_1 score can be obtained by considering the predicted BGCs and the actual BGCs comprising the simulated metagenomes of known composition. A '20%' gene-presence threshold, which resulted in the highest collection of accuracies (average F_1 scores) across various species sets and sequencing library sizes, was selected as the optimal threshold to be used in **A** (**Methods**).

Table 1. Phylogenetic summary of the TaxiBGC database.

Kingdom	Phylum	Class	# of species	# of strains[a]	# of BGC producing strains[b]	# of unique BGCs	# of unique SMs[c]
Bacteria	Actinobacteria	Actinobacteria	262	9,724	5,980	131	118
	Bacteroidetes	Cytophagia	3	4	4	1	1
		Sphingobacteriia	1	6	3	2	2
		Flavobacteriia	1	11	2	1	1
	Chlamydiae	Chlamydia	1	69	1	1	1
	Cyanobacteria	Cyanobacteria	24	127	81	22	20
	Firmicutes	Clostridia	3	286	46	3	2
		Bacilli	112	26,003	13,346	83	81
		Negativicutes	1	10	1	1	1
	Planctomycetes	Planctomycetia	1	1	1	1	1
	Proteobacteria	Alphaproteobacteria	93	796	422	18	14
		Betaproteobacteria	35	3,012	2,588	26	24
		Gammaproteobacteria	179	51,187	10,733	86	74
		Deltaproteobacteria	10	38	37	20	19
Archaea	Euryarchaeota	Methanomicrobia	1	15	1	1	1
Eukaryota	Ascomycota	Eurotiomycetes	12	124	115	29	29

		Sordariomycetes	4	28	11	4	4
		Dothideomycetes	1	3	3	2	2
		Total unique	744	91,444	33,375	390	351

[a] Strain information from Genbank (November 2019). [b] Identified after running antiSMASH (at 100% similarity criterion) on strain genome sequences. [c] According to MIBiG database (November 2019).

The TaxiBGC database provides annotated information on the experimentally verified BGCs and their corresponding SMs found in microbial species. This database, which is relied on by the TaxiBGC pipeline during BGC prediction, is pre-defined and not assembled anew for each metagenome sample. For the construction of this database (**Methods**; **Fig. 1B**), we obtained genomes of all species in the MetaPhlAn2 marker database (while excluding viruses and uncharacterized species) from GenBank (as of November 2019). More specifically, we downloaded a total of 380,184 genomes from 4,187 species of archaea, bacteria, and fungi. We predicted BGCs in these genomes by using antiSMASH v4.0 (46) on experimentally verified BGCs from the MIBiG) repository (47), which provides information regarding BGCs and their SM products. In sum, the final TaxiBGC database encompasses 744 BGC-encoding microbial species, 390 experimentally verified BGCs, and 351 SMs (**Table 1**, **Supplementary Tables S7–S9**). Of note, a unique SM can be encoded by more than one BGC.

Depending on their chemical structure and biosynthetic origin, the 390 experimentally verified BGCs in the TaxiBGC database can be categorized into following classes (26): alkaloids, nonribosomal peptides (NRPs), ribosomally synthesized and post-translationally modified peptides (RiPPs), polyketides, saccharides, terpenes, hybrid classes, and an undefined class (‘others’) (**Fig. 2**). Among these 390 BGCs, NRPs were the most prevalent (102 of 390, 26.2%), followed by RiPPs (80 of 390, 20.5%), polyketides (69 of 390, 17.7%), others (42 of 390, 10.8%), terpenes (31 of 390, 7.9%), alkaloids (4 of 390, 1.0%), and saccharides (1 of 390, 0.3%). In addition, a major portion of the BGC classes in the TaxiBGC database are hybrids of two BGC classes (61 of 390, 15.6%) (**Supplementary Table S9**).

BGCs for producing terpenes are the most prevalent among the 744 species (258 of 744 species, 34.7%), which include those of Actinobacteria, Cyanobacteria, and Proteobacteria phyla. NRPs were found in 230 of the 744 species (30.9%) from mostly Actinobacteria, Ascomycota, Cyanobacteria, Firmicutes, and Proteobacteria phyla. In addition, species with polyketides (172 of 744 species, 23.1%) were found mostly in Actinobacteria, Ascomycota, Cyanobacteria, and Proteobacteria phyla. RiPPs were in 160 of the 744 species (21.5%) spanning Actinobacteria, Cyanobacteria, Firmicutes, and Proteobacteria phyla. Alkaloids (6 of 744 species, 0.8%) were found in Actinobacteria, Ascomycota, and Proteobacteria phyla. A saccharide was found only in a single species (1 of 744, 0.1%) of Proteobacteria. Furthermore, hybrid BGCs, which are defined as secondary metabolites having hybrid structures of two (or more) classes, are in 117 of 744 species (15.7%) from the Actinobacteria, Ascomycota, Cyanobacteria, Firmicutes, and Proteobacteria phyla. Lastly, the undefined ‘others’ class (340 of 744 species, 45.7%) was found in mostly Actinobacteria, Cyanobacteria, Firmicutes, and Proteobacteria. BGC classes in our database and the taxa from which they are derived are described in **Supplementary Tables S8** and **S9**.

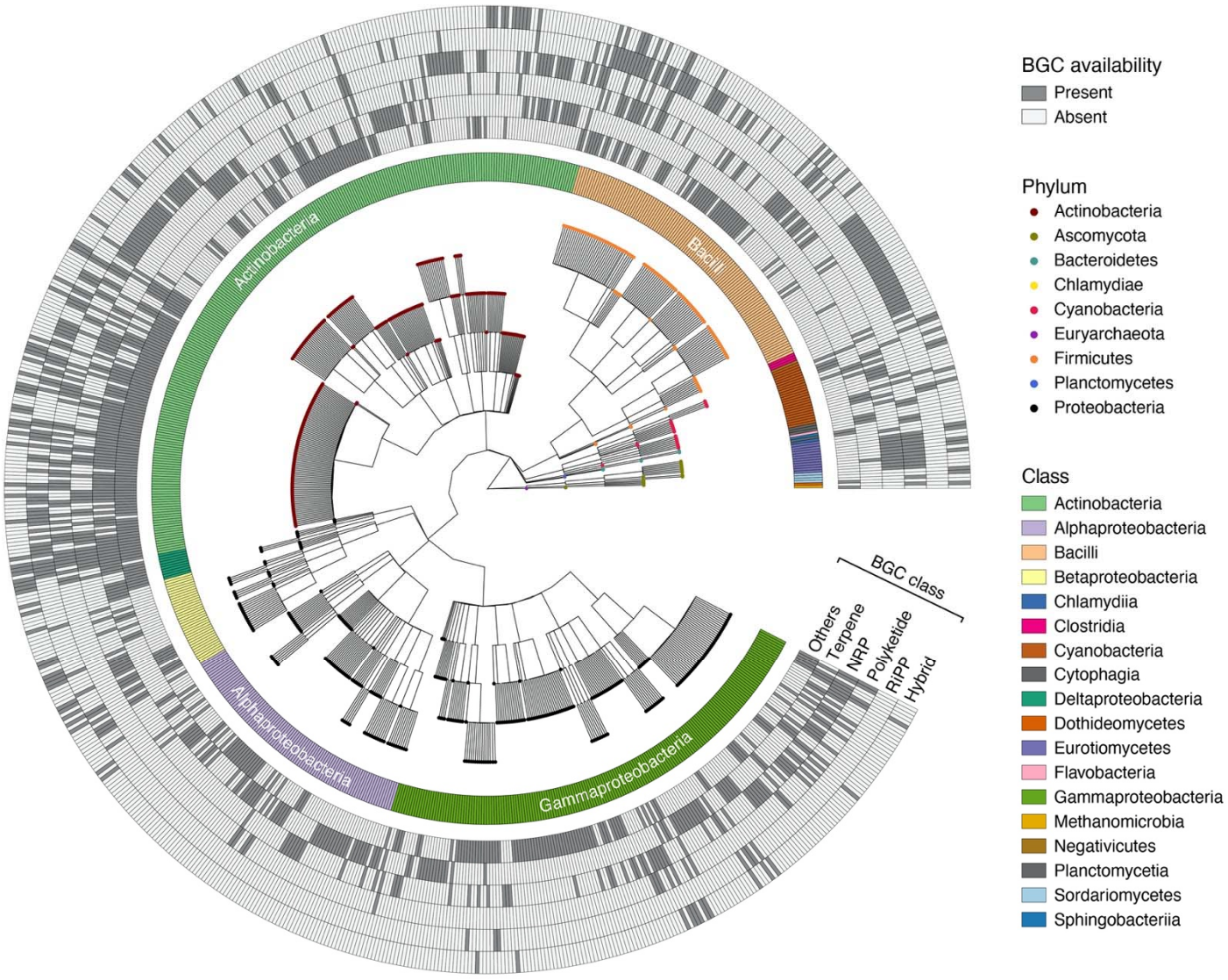

Figure 2. Distribution of taxonomic ranks and BGC classes in the TaxiBGC database. A phylogenetic tree (center) shows the evolutionary relationships (distances) among the 744 microbial species comprising the TaxiBGC database. These species are spread across nine phyla (outermost colored nodes of the tree) and eighteen classes (innermost colored ring). Encoded in the genomes of these species are a total of 390 unique experimentally verified BGCs, which can all be categorized into nonribosomal peptides (NRPs), ribosomally synthesized and post-translationally modified peptides (RiPPs), polyketides, terpenes, alkaloids, saccharides, hybrid classes, and undefined 'others'. Alkaloids and saccharides are not shown due to simplicity, as each were found in only six and one of the 744 total species, respectively. Gray and white in the outer six rings reflect the presence and absence, respectively, of a particular BGC class (or a hybrid class) in the genome of the corresponding

species. Full details regarding BGCs, their classes, their SM products, and their taxonomic origins, as provided by the TaxiBGC database, are available in **Supplementary Tables S7–S9**.

The number of genes comprising the BGCs in the TaxiBGC database widely varies, ranging from 2 to 135 genes (**Supplementary Table S10**). Ideally, when a BGC is present in a metagenome, all the genes of that BGC should be detected in the metagenome; however, in reality, accurate and comprehensive computational detection of the genes depends upon several technical issues (65, 66), including coverage (67) and sequencing depth in the metagenome (68, 69). Therefore, we evaluated a range of gene-presence criteria (20%, 40%, 60%, 80%, and 100%) in TaxiBGC to decide on the optimum threshold (**Fig. 3**). Using simulated metagenomes composed of seven sets of 25, 75, and 125 microbial species ranging across seven library sizes (1.25M, 2.5M, 5M, 10M, 20M, 40M, and 80M reads), we predicted BGCs in each metagenome and then compared them with the actual BGC composition. Our predictions in simulated metagenomes showed that, on average, F_1 scores increased with increasing library sizes up to 20M reads and then remained consistent for higher library sizes (**Fig. 3**). Notably, prediction accuracies gradually increased with smaller BGC gene-presence thresholds: average F_1 scores of 0.27–0.72 for 100% BGC gene-presence; 0.47–0.75 for 80% BGC gene-presence; 0.61–0.78 for 60% BGC gene-presence; 0.66–0.81 for 40% BGC gene-presence; and 0.70–0.82 for 20% BGC gene-presence. Furthermore, we detected no further increase in F_1 scores with BGC gene-presence thresholds below 20% (**Supplementary Fig. S1**). Taken together, we selected '20%' as the optimal BGC gene-presence threshold, as it clearly resulted in the most robust and accurate predictions.

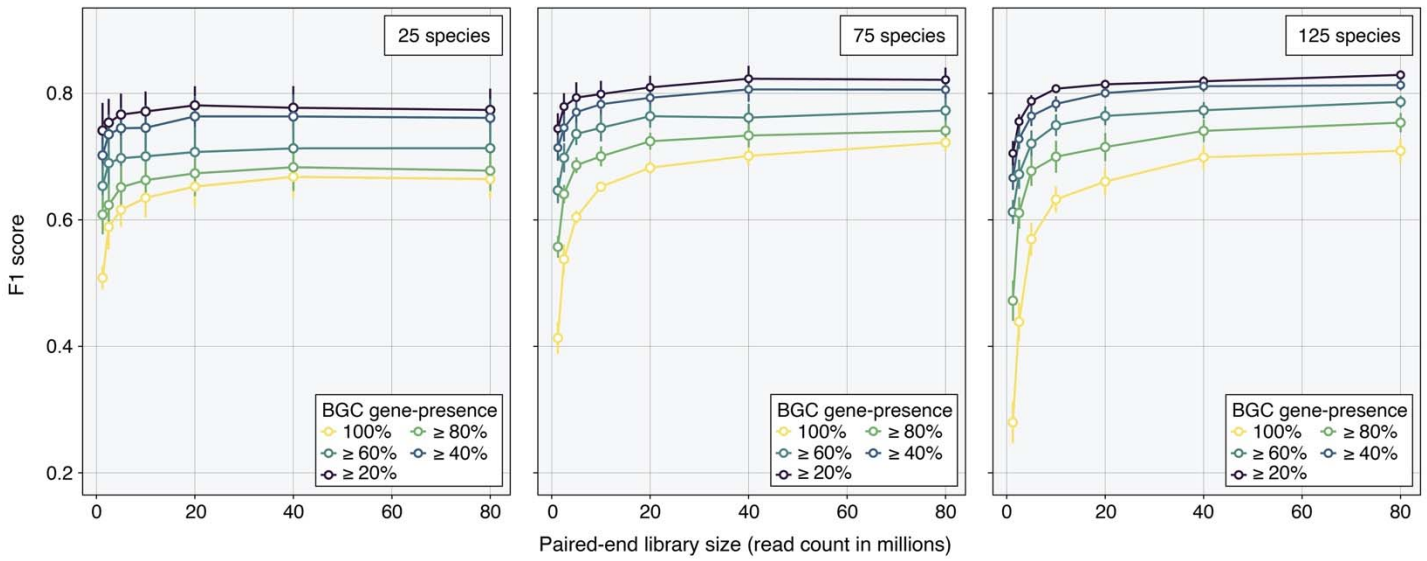

Figure 3. Best accuracy in predicting BGC presence in simulated metagenomes is consistently achieved at a minimum BGC gene-presence threshold of 20%. BGCs were predicted from simulated metagenomes composed by seven sets of **A)** 25, **B)** 75, and **C)** 125 microbial species ranging across seven sequencing library sizes (1.25M, 2.5M, 5M, 10M, 20M, 40M, and 80M reads) (**Methods**). The predicted BGCs of each simulated metagenome were then compared with the actual BGC composition, and F_1 scores were used to assess prediction accuracy. In general, prediction accuracy markedly increased with increasing library size up to 20M reads and remained nearly consistent for higher library sizes. In all three community sizes, a minimum BGC gene-presence threshold of 20% was found to provide the best range of prediction accuracies (average F_1

scores: 0.70–0.82). F_1 scores resulting from different thresholds are shown in different colors. Mean and standard errors of the F_1 scores are shown in circles and error bars, respectively.

Taxonomy-guided BGC prediction provides higher prediction accuracy than direct BGC gene-detection from metagenome reads

As described above, the TaxiBGC pipeline predicts BGCs from metagenomes by first identifying the microbial species that encode BGCs. As an alternative to this taxonomy-guided approach, we can also predict BGCs from metagenomes through direct detection of BGC genes without considering their taxonomic origins. In this approach, we first used a range of BGC gene-presence thresholds for directly mapping reads of BGC genes (**Methods**). By and large, higher gene-presence thresholds led to less false-positive predictions, and thereby improved prediction accuracies (F_1 score). As such, we found that the BGC gene-presence threshold of 100% led to the best F_1 scores in all three community sizes of 25, 75, and 125 microbial species (**Supplementary Fig. S2**). This was due, in part, to the lowest number of false positive predictions. With this BGC gene-presence threshold, we evaluated the predictive performance of the direct BGC gene-detection approach on the same aforementioned simulated metagenomes.

The taxonomy-guided approach (TaxiBGC) clearly demonstrated higher predictive performances on the simulated metagenomes. More specifically, the average F_1 scores of TaxiBGC varied between 0.74–0.78, 0.74–0.82, and 0.70–0.82 for community sizes of 25, 75, and 125 species, respectively (**Fig. 4**). On the other hand, in the direct BGC gene-detection approach, average F_1 scores varied between 0.43–0.52, 0.36–0.56, and 0.26–0.57 for community sizes of 25, 75, and 125 species, respectively (**Fig. 4**). Of note, we observed higher positive predictive values (PPV) in the taxonomy-guided approach compared to the direct BGC gene-detection approach (**Supplementary Fig. S3**). In all, these data suggest that the taxonomy-guided strategy is a more sensitive and precise method for predicting species-specific, experimentally verified BGCs from metagenome reads.

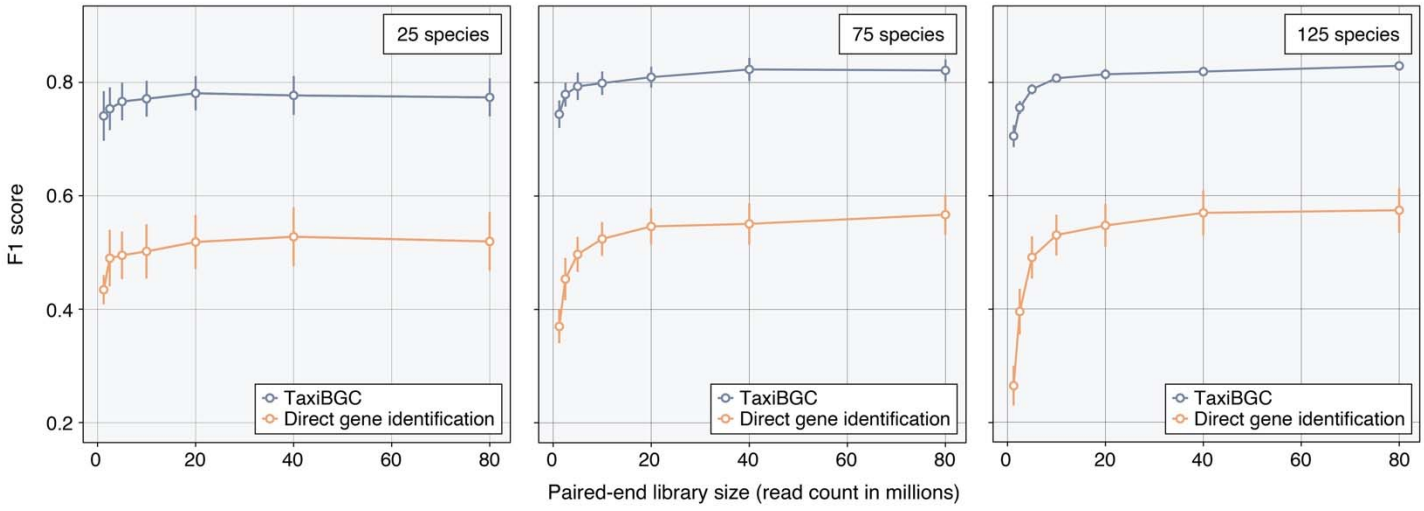

Figure 4. TaxiBGC outperforms a direct BGC gene-detection method in all simulated metagenome combinations. The optimal BGC gene-presence thresholds for TaxiBGC and a direct BGC gene-detection approach (**Methods**) was set to 20% and 100%, respectively; these cut-offs led to the best prediction accuracies for their respective methods. In simulated metagenomes of seven different combinations of **A)** 25, **B)** 75, and **C)** 125 microbial species members and library sizes, TaxiBGC provided higher prediction accuracies, with average F_1 scores varying between 0.74–0.78, 0.74–0.82, and 0.70–0.82 for community sizes of 25, 75, and 125 species, respectively. On the other hand, in the direct BGC gene-detection approach, average F_1 scores varied between 0.43–0.52, 0.36–0.56, and 0.26–0.57 for community sizes of 25, 75, and 125 species, respectively.

Human body sites harbor distinct BGC signatures in their microbiomes

We first demonstrated our TaxiBGC pipeline on 2,355 metagenomic samples from the Human Microbiome Project (HMP-1-I and HMP-1-II). These samples originated from five major human body sites (anterior nares (253 samples), gut (553 samples), oral (1,259 samples), retroauricular crease (left/right; 56 samples), and vagina (234 samples)) collected from 265 healthy human individuals (**Supplementary Table S5**). We investigated microbiome samples of each body-site to determine the frequency of predicted BGCs (**Methods**). Briefly, a BGC that was predicted in 2% or more of the samples ($\log_2$(proportion of samples) ≥ 1) in at least one body-site was considered for further analysis; this resulted in a total of 55 unique BGCs, which were used to form body-site-specific BGC profiles. Clustering of these profiles revealed similar patterns between body sites (**Fig. 5**). For example, possibly due to the aerobic/anaerobic nature of their respective environments, BGC profiles of the gut, oral cavity, and vagina clustered together (gut more closely with the oral cavity, as expected); whereas BGC profiles of anterior nares closely clustered with those of the retroauricular crease.

Figure 5. Distribution of BGCs across metagenomes derived from five major human body sites. Microbiome samples of each body-site (provided by the Human Microbiome Project (HMP)) were investigated to determine how frequently a BGC was predicted. Color intensity in the heatmap reflects proportions of samples in a body-site wherein the corresponding BGC was predicted to be present. Only the BGCs detected in 2% or more of the samples ($\log_2$(proportion of samples) ≥ 1) in at least one body-site were considered. Clustering (Euclidean distance) based upon similarity of BGC distributions shows gut and oral cavity as the two closest body sites. The next two body sites to cluster together were the anterior nares and retroauricular crease. BGCs mostly present across human body sites are non-ribosomal peptides (NRPs) and ribosomally synthesized and post-translationally modified peptides (RiPPs).

In **Figure 5**, we show that 44 of the 55 unique BGCs identified from the 2,355 HMP metagenomes were distributed into three major BGC classes: twenty-three RiPPs, fourteen NRPs, and seven terpenes. Among the remaining eleven BGCs, eight were of an undefined class ('other') and one each were of NRP/polyketide, polyketide, and saccharide. We found eleven core BGCs present in all body sites (**Supplementary Table S11**): six RiPPs (cytolysin ClyLl/cytolysin ClyLs, enterocin W alpha/enterocin W beta, microcin J25, microcin J26, mutacin II, and nisin U), two NRPs (aureusimine and pseudomonine), a hybrid NRP/polyketide class (N-myristoyl-D-asparagine), a saccharide (lipopolysaccharide), and a terpene (carotenoid). Interestingly, twenty BGCs were identified exclusively in the microbiomes of particular body sites (**Supplementary Table S12**): three RiPPs (BhtA1/BhtA2, mutacin III, and Smb A/SmbB) in the oral cavity; three NRPs (acinetoferrin, fimsbactin A, and massetolide A), one RiPP (paeninodin), four terpenes (all code for carotenoids) and three BGCs of an undefined class (one for bacilysin and two for ectoines) in retroauricular crease; two NRPs (andrimid and tilivalline) and one terpene (carotenoid) in anterior nares; and one NRP (equibactin) and two RiPPs (salivaricin A and streptin) in the vagina. Furthermore, two BGCs, one for an NRP (lugdunin) and another for a RiPP (nukacin ISK-1), were found only in the two outer body areas (anterior nares and retroauricular crease).

The SM products of these BGCs were reported to have a beneficial influence on host health. For example, microcins, two of which (microcin J25 and microcin J26) were found in all body sites, can play a significant role in eliminating enteric pathogens (e.g., *Salmonella enterica*) (70). More specifically, in a murine model, microcins produced by commensal *Escherichia coli* strains bind to siderophore receptors of gut pathogenic strains, and thereby restrict their iron uptake; Mutacin III, a bacteriocin produced mainly by oral streptococcal species, is known to inhibit the growth of a wide spectrum of gram-positive bacteria (71). Mutacin III produced from oral streptococci can have a protective role against oral pathogens, such as group A streptococci (72, 73); furthermore, we found a BGC for lugdunin to be present exclusively in sites of the skin (anterior nares and retroauricular crease). Lugdunin has been reported to inhibit *Staphylococcus aureus* colonization by increasing the expression of human-derived antimicrobial peptides (AMPs) in keratinocytes (74, 75). Lastly, although not found exclusively in the vaginal microbiome, lactocillin was nevertheless observed in a high proportion (48.3%) of vaginal microbiome samples. By combining genome-mining strategies with empirical validation, Donia *et al*. discovered this SM in human-associated bacteria and demonstrated its potent antibacterial activity against a range of gram-positive pathogens (23).

During the identification of taxonomic sources for the body-site-specific BGCs, we observed various noteworthy associations between species and BGCs in different human body sites. For example, three BGCs encoding RiPPs (BhtA1/BhtA2, mutacin III, and Smb A/SmbB) were exclusively found in the oral cavity. These three compounds are bacteriocins isolated from different members of Streptococci, and have been shown to contribute towards bacterial colonization in the oral cavity (73, 76, 77). The source species of all three BGCs was found to be *Streptococcus mutans,* which is a well-known member of the human oral microbiome (78). Two of the three BGCs exclusively present in the anterior nares (one andrimid and one terpene) were produced by two species of the *Pantoea* genus (*P. agglomerans* and *P. ananatis*); five of the eleven BGCs found exclusively in retroauricular crease were from three species of *Acinetobacter* genus (*A. lwoffii, A. baumannii*, and *A. indicus*) and one species each of *Bacillus* and *Sphingobium* genera (*B. megaterium* and *S. yanoikuyae*, respectively); the remaining six BGCs produced from two species of *Pseudomonas* (*P. stutzeri* and *P. syringae*)

and one species each of *Massilia* and *Serratia* genus (*M. time* and *S. marcescens*), respectively. Interestingly, we detected two exclusively present BGCs (lugdunin and nukacin ISK-1) in two skin body sites (anterior nares and retroauricular crease), produced from three species of genus *Staphylococcus* (*S. lugdunensis*, *S. pasteuri*, *S. warneri*) present in TaxiBGC database. In vaginal microbiome samples, among three experimentally verified BGC, the source species of equibactin is *Anaerococcus lactolyticus*. One strain of the *Anaerococcus* genus was first isolated from vaginal discharges and ovarian abscesses (79). Two other vagina specific BGCs, salivaricin A and streptin, were found to be produced from *Streptococcus agalactiae*.

Meta-analysis of gut microbiome studies reveals BGCs associated with human disease

We can also apply TaxiBGC in the context of translational research for the discovery of potential disease biomarkers. To this point, we investigated for differences in the prevalence of BGCs in the gut microbiome between patients with disease and control individuals. For this, we performed a meta-analysis by applying our TaxiBGC pipeline to 3,068 stool metagenome samples from the following thirteen independent case-control studies: seven inflammatory bowel disease (IBD) studies (six of both Crohn's disease (CD) and ulcerative colitis (UC) (52–57), and one study with only CD (58); three colorectal cancer (CRC) studies (59–61); and one study each for rheumatoid arthritis (RA) (62), autism spectrum disorder (ASD) (63), and schizophrenia (64). Detailed information regarding these studies is available in **Supplementary Table S6**. As we elaborate further below, we identified a total of fifteen unique BGCs at significantly different frequencies when comparing case and control ($P < 0.05$, Fisher's exact test; **Methods**; and **Fig. 6**).

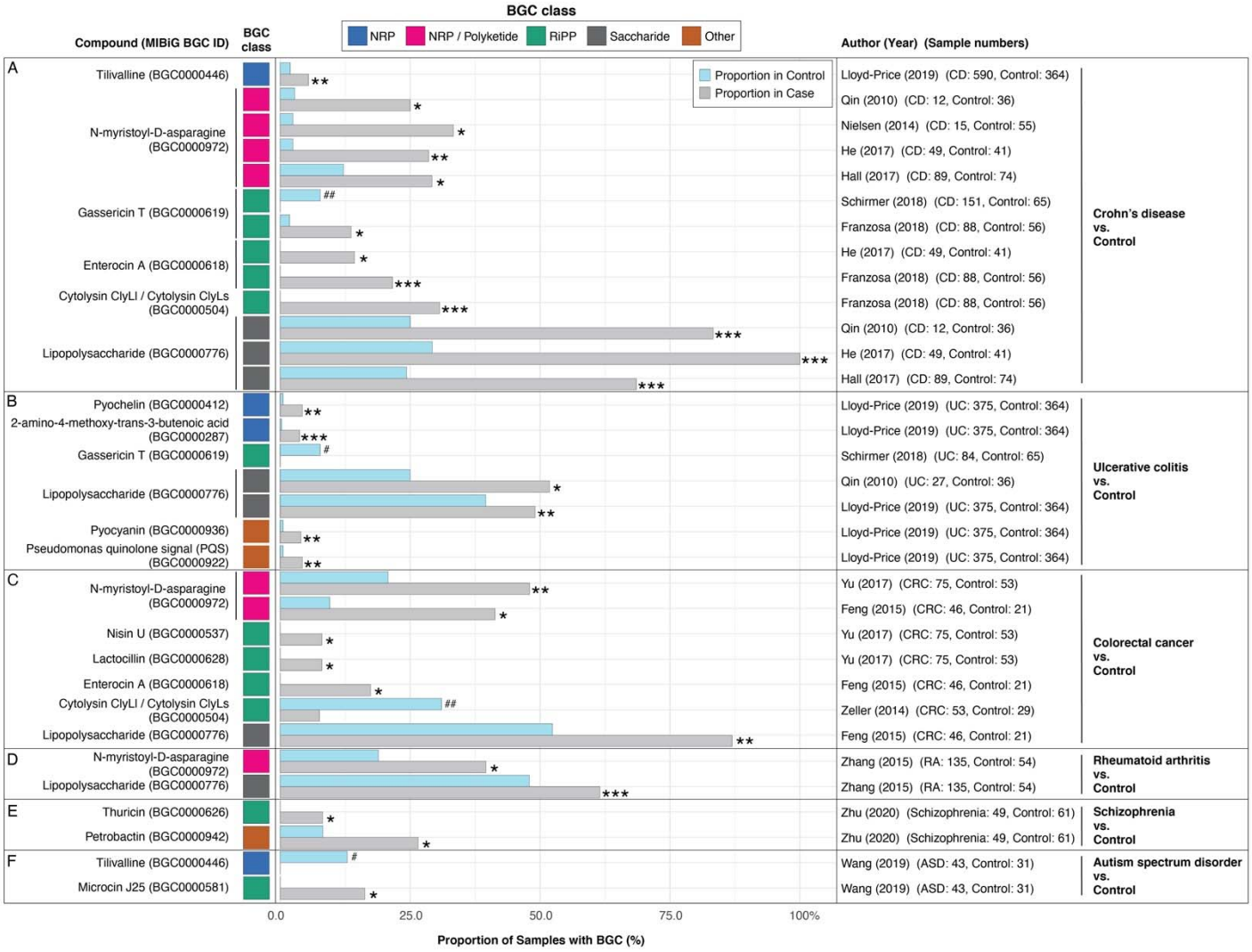

Figure 6. Analysis of gut microbiomes from published studies reveals BGCs associated with several human pathologies. TaxiBGC was used to predict the presence of BGCs in microbiome samples from multiple case-control studies. **A)** Crohn's disease (CD). BGCs encoding for six different SMs (tilivalline, N-myristoyl-D-asparagine, gassericin T, enterocin A, cytolysin ClyLl/cytolysin ClyLs, and lipopolysaccharide) were predicted to be associated with either case or control. In four ('Hall 2017', 'He 2017', 'Nielsen 2014', and 'Qin 2010') of seven CD studies, a BGC encoding for N-myristoyl-D-asparagine was predicted significantly more often in CD compared to control. A saccharide BGC that encodes for lipopolysaccharide was found in significantly higher proportions in CD in three ('Hall 2017', 'He 2017', and 'Qin 2010') studies. **B)** Ulcerative colitis (UC). BGCs encoding for six different SMs (pyochelin, 2-amino-4-methoxy-trans-3-butenoic acid, gassericin T, lipopolysaccharide, pyocyanin, and Pseudomonas quinolone signal (PQS)) were predicted to be associated with either case or control. Five of these six were identified as having significantly higher proportions in UC than in controls. Four BGCs were predicted exclusively in one study ('Lloyd-Price 2019'). Remarkably, as also seen for CD, one BGC encoding for gassericin T was found in higher proportions in the control samples in the 'Schirmer 2018' study. **C)** Colorectal cancer (CRC). Six BGCs were found to be associated with either case or control. More specifically, five BGCs encoding N-myristoyl-D-asparagine, enterocin A, lactocillin, lipopolysaccharide, and nisin U were predicted more frequently in CRC. Of note, three (N-myristoyl-D-asparagine, enterocin A, and lipopolysaccharide) of these five were also associated with CD, suggesting gut microbiome-derived BGC signatures shared across multiple diseases. Alternatively, one BGC (cytolysin ClyLl/cytolysin ClyLs) was found significantly

more often in control samples. Only the N-myristoyl-D-asparagine-producing BGC was found to be associated with CRC in multiple studies (in 'Feng 2015' and 'Yu 2017'). **D)** Rheumatoid arthritis (RA). In 'Zhang 2015', two BGCs encoding for N-myristoyl-D-asparagine and lipopolysaccharide were found in significantly higher proportions in case than in control. **E)** Schizophrenia. In a study concerning mental health patients with schizophrenia ('Zhu 2020'), two BGCs that were undetected in previously mentioned disorders were found: a BGC encoding for thuricin and another encoding for petrobactin were both found in a significantly higher proportion in case than in control samples. **F)** Autism spectrum disorder (ASD). In a study with ASD patients and control subjects ('Wang 2019'), one microcin J25-encoding BGC was found in a significantly higher proportion in case samples, while one tilivalline-encoding BGC was found to have a significantly higher proportion in control. Case-control studies are designated with the lead author's (last) name and publication year. Number of samples for case and control are shown for each study. Further details on all case-control studies are provided in **Supplementary Table S6**. Levels of significance for the two-tailed Fisher's exact test are indicated by the following symbols: higher abundance in case: *, $0.01 \leq P < 0.05$; ** for $0.001 \leq P < 0.01$; and *** for $P < 0.001$; higher abundance in control: #, $0.01 \leq P < 0.05$; ## for $0.001 \leq P < 0.01$.

One BGC encoding for the pre-colibactin metabolite N-myristoyl-D-asparagine (hybrid NRP/polyketide) was predicted significantly more often in CD in four ('Hall 2017', 'He 2017', 'Nielsen 2014', and 'Qin 2010') of the seven CD studies. We obtained one saccharide BGC that encodes for lipopolysaccharide in significantly higher proportions in CD than in controls in three ('Hall 2017', 'He 2017', and 'Qin 2010') CD studies. Other BGCs found to be more prevalent in CD were cytolysin ClyLl/cytolysin ClyLs (RiPP, 'Franzosa 2018'), enterocin A (RiPP, 'Franzosa 2018'), and tilivalline (NRP, 'Lloyd-Price 2019'). In contrast, the sole BGC found in higher proportions in control samples was for the one encoding for gassericin T (RiPP, 'Schirmer 2018').

BGCs that were found in significantly higher proportions of UC samples (compared to those of controls) include 2-amino-4-methoxy-trans-3-butenoic acid (NRP, 'Lloyd-Price 2019'), lipopolysaccharide (saccharide, 'Lloyd-Price 2019' and 'Qin 2010'), Pseudomonas quinolone signal (PQS) (other, 'Lloyd-Price 2019'), pyochelin (NRP, 'Lloyd-Price 2019'), and pyocyanin (other, 'Lloyd-Price 2019'). Interestingly, as was observed in CD, a BGC for gassericin T was found in significantly higher proportions in controls than in UC patients (RiPP, 'Schirmer 2018').

In two ('Feng 2015' and 'Yu 2017') of the three CRC studies, a BGC encoding for N-myristoyl-D-asparagine (hybrid NRP/polyketide) was predicted significantly more often in the gut microbiome of patients with CRC than in controls. Five other BGCs found to be associated with CRC were those for cytolysin ClyLl/cytolysin ClyLs (RiPP, 'Zeller 2014'), enterocin A (RiPP, 'Feng 2015'), lactocillin (RiPP, 'Yu 2017'), lipopolysaccharide (saccharide, 'Feng 2015'), and nisin U (RiPP, 'Yu 2017'). In the RA study ('Zhang 2015'), BGCs for N-myristoyl-D-asparagine (hybrid NRP/polyketide) and lipopolysaccharide (saccharide) were found in higher proportions in patients with RA than in controls. Finally, in gut microbiomes of major psychiatric disorders, we found two BGCs (petrobactin (other) and thuricin (RiPP)) more often in schizophrenic patients; and in patients with ASD ('Wang 2019'), a microcin J25-encoding BGC (RiPP) was found to be more prevalent in ASD, whereas one tilivalline-encoding BGC (NRP) was found in a significantly higher proportion in controls.

Notably, the SM product of two BGCs identified in our case-control gut microbiome meta-analysis have previously established links with their respective diseases: i) One BGC that was identified in significantly higher proportions in CD, CRC, and RA stool metagenomes (compared to controls) was N-myristoyl-D-

asparagine, a structural part of pre-colibactin that gets cleaved by the peptidase ClbP to form the active colibactin (80). In regard to colibactin, this small-molecule is synthesized by an NRPS/PKS hybrid BGC, and is commonly produced by *E. coli* and other members of the Enterobacteriaceae family (14). Colibactin induces double-strand DNA breaks, eukaryotic cell cycle arrest, and chromosome aberrations (14). Previous studies reported that colibactin-producing *E. coli* can have a deteriorative effect on human intestinal mucosa in CRC and IBD patients, with profound tissue damage, increased antigen delivery from the lumen into the gut submucosa, and chronic inflammation (81–83); ii) the frequency for a BGC encoding lipopolysaccharide (LPS) was found to be higher in all three auto-immune or inflammatory diseases (CD, RA, and UC) and CRC. Possibly in line with these findings, LPS has previously been shown to be elevated in serum and plasma of CRC and IBD patients (84–86). Moreover, in the context of RA pathogenesis, LPS has shown to physically interact with collagen type II in the extracellular matrix of *in vitro* human cell models, thus developing procollagen-endotoxin complexes that trigger joint cartilage inflammation and degeneration (87).

Our BGC predictions provide two candidate SMs that may warrant further investigation into their possible beneficial or protective effects in specific diseases. More specifically, a BGC for gassericin T (a RiPP) was found in significantly higher proportions in controls than in IBD (both CD and UC) patients. Although we were not able to find studies that associate gassericin T with either CD or UC, this SM produced by *Lactobacillus gasseri* (a microbe commonly present in the human gut) has been shown to have strong bactericidal activity against gram-positive bacteria that cause food poisoning (e.g., *Listeria monocytogenes*, *Staphylococcus aureus*) (88). In addition, we more frequently observed a BGC for cytolysin (ClyLl/ ClyLs) in controls than in CRC. Cytolysins are a large family of enterococcal toxins and bacteriocins that are known to be lethal to a broad range of prokaryotic and eukaryotic microorganisms (89). However, to the best of our knowledge, their role in CRC is currently unknown.

Taken together, we uncover a widespread distribution of small-molecule-encoding BGCs across a range of pathologies. Accordingly, our findings warrant the use of TaxiBGC as a computational tool for the unbiased identification of novel BGC- or SM-based ‘molecular signatures’ of disease (90) or small molecules with possible clinical roles. A notable example of such is gliotoxin, which is an immunosuppressive cytotoxin produced by fungi and has been suggested for the detection of aspergillosis (91).

DISCUSSION

Accurately predicting the biosynthetic potential of BGCs from uncultivated microbiome sequencing data is currently one of the biggest challenges in the field of microbial natural product discovery. Thus, using an experimentally verified set of BGCs as a query provides a promising direction to infer specific SMs using bioinformatics tools. In this study, we present TaxiBGC, a computational pipeline for the prediction of known BGCs from shotgun metagenomes. TaxiBGC employs three main steps for predicting microbial BGCs: i) identification of BGC-producing species, ii) initial prediction of experimentally verified BGCs based upon the presence of those identified BGC-producing species, and iii) *in silico* confirmation of the predicted BGCs by mapping their genes in the metagenome. We evaluated the prediction accuracy of the TaxiBGC pipeline on simulated metagenomes constructed using varied combinations of species identities and sequencing library sizes. We showed that BGC prediction using our taxonomy-guided approach was superior to directly predicting BGC genes from the metagenomes (average F_1 scores of 0.70–0.82 and 0.26–0.57, respectively). Next, to

identify BGCs (and thereby their corresponding SMs) associated with human health and disease, we demonstrated our TaxiBGC pipeline on metagenomes sampled from different body sites; and on gut metagenomes collected from patients with various diseases (IBD (CD and UC)], CRC, ASD, RA, and schizophrenia) and their controls. Our results show that human body sites harbor distinct BGC signatures in their microbiomes. Furthermore, by characterizing the potential chemical milieu of the gut environment for these diseases, we provide candidate stool-borne SM biomarkers. Importantly, TaxiBGC is not limited to the use in only human metagenomes, but can easily be applied to environmental metagenome samples for BGC identification. To the best of our knowledge, our novel pipeline to date is the most rapid, accurate, and comprehensive method for identifying BGCs from shotgun metagenomes of microbial communities. In sum, we anticipate that TaxiBGC can facilitate the large-scale identification of experimentally verified BGCs and their annotated SMs from microbiome datasets; and assist in the computational analyses of metagenomic samples to systematically characterize the chemical ecology of various microbial niches.

Apart from identifying BGCs and their SM products associated with human health and disease, TaxiBGC can also be used to infer SMs that play important roles in interspecies interactions (92–96). For example, within complex microbial communities composed of highly diverse taxa with overlapping metabolic requirements, competition for similar energy resources is one of the most abundant forms of microbial interactions (97–101). Accordingly, microbes evolved the ability to produce SMs with antimicrobial properties (e.g., antibiotics, siderophores, bacteriocins) to limit the number of surrounding species competing for metabolic nutrients (102, 103). Conversely, microbial SMs such as cichofactin, holomycin, and pyrrolnitrin can initiate aggregation of mutually cooperative microorganisms to form biofilms (104–106); in turn, biofilms can bestow competitive advantages to microbial taxa, including protection from the environment, nutrient availability, and metabolic cooperativity, and acquisition of new genetic traits (107–110). Such interspecies competition and cooperation within microbial communities can lead to taxonomic divergence (111–113), which, in turn, leads to the diversity of SM production. The comprehensive prediction of microbial species, their BGCs, and their corresponding SMs provided by TaxiBGC may elucidate novel insights regarding the dynamics and divergence of microbial interactions.

Several limitations of TaxiBGC should be noted when interpreting our results. First, by design, BGC prediction by TaxiBGC depends upon the successful identification of species from the metagenomes. In the case where MetaPhlAn2 does not identify all the BGC-producing species, TaxiBGC would fail to accurately predict the entire breadth of BGCs. Second, predictions regarding the taxonomic origins of BGCs need to be interpreted with caution. This is because, in the ‘mixed bag of genes’ context of metagenomes, gene elements of a particular BGC could originate from other BGCs and/or multiple species. Nevertheless, TaxiBGC can spark new and interesting hypotheses into the connections between microbiome BGCs and their source species, which can be validated in future research. Third, in order to minimize false-positive predictions, TaxiBGC purposefully employs a high stringency cut-off in identifying BGCs from genomes of the GenBank database (i.e., all genes of a BGC need to be detected). The trade-off to having this stringent criterion is that TaxiBGC may miss identifying BGCs of comparably long and intricate gene sequences, thereby leading to false-negative predictions. Fourth, TaxiBGC may have limited utility in metagenomes of relatively small library sizes, as suggested by the drop in predictive accuracy observed in metagenomes of less than 2.5M reads (see **Fig. 3**). Fifth, TaxiBGC provides only the chemical family but not the specific molecule for a small number of SM

products, such as lipopolysaccharide and carotenoid, as that information is not yet available in MIBiG. Finally, experimental validation of whether the small-molecule products of any of the identified BGCs are indeed implicated in specific pathogenic mechanisms is necessary but beyond the scope of this *in silico* study.

Nonetheless, TaxiBGC provides several advantages over currently available tools for BGC identification. First, TaxiBGC employs an assembly-independent, read-based prediction method; this solves a key limitation of BGC prediction tools that assume each BGC is encoded within a single contig in the genome or metagenome assembly, whereas in reality, genes of BGCs can be dispersed throughout several contigs. Second, TaxiBGC can be used to associate BGC presence with species abundance. This can help pinpoint which BGCs originate from which species in different phenotypic conditions. Third, TaxiBGC can simultaneously identify several classes of experimentally verified BGCs in a metagenome sample. This comprehensive detection is essential to deepen our understanding of the diversity of known BGCs and their SMs in a microbiome. Fourth, no parameter selection or optimization is necessary prior to applying TaxiBGC on shotgun metagenomic datasets. The only pre-installed software needed in TaxiBGC are: i) MetaPhlAn2 for species identification, and ii) Bowtie 2 for mapping metagenome reads onto BGC genes. Finally, TaxiBGC can easily be upgraded over time to detect a broader range of BGC classes that are not included in the present version. Notably, we plan to iteratively expand and curate the current TaxiBGC database with newly discovered species and strains, and their corresponding BGC and SM annotations from MIBiG.

DATA AVAILABILITY

The R source code, TaxiBGC database, and MIBiG BGC gene sequences underlying all results presented in this study are publicly available for use at https://github.com/jaeyunsung/TaxiBGC_2021.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

This work was supported in part by the Mayo Clinic Center for Individualized Medicine (to U.B., V.K.G., and J.S.) and Mark E. and Mary A. Davis to Mayo Clinic Center for Individualized Medicine (J.M.D. and J.S.).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank members of the Sung Lab for insightful discussions and editorial assistance on the manuscript.

REFERENCES

1. Baral,B., Akhgari,A. and Metsä-Ketelä,M. (2018) Activation of microbial secondary metabolic pathways: avenues and challenges. *Synth. Syst. Biotechnol.*, 3, 163–178.
2. Singh,B.P., Rateb,M.E., Rodriguez-Couto,S., Polizeli,M. de L.T. de M. and Li,W.-J. (2019) Editorial: microbial secondary metabolites: recent developments and technological challenges. *Front. Microbiol.*, 10, 914.
3. Davies,J. (2013) Specialized microbial metabolites: functions and origins. *J. Antibiot.*, 66, 361–364.
4. Schroeckh,V., Scherlach,K., Nützmann,H.-W., Shelest,E., Schmidt-Heck,W., Schuemann,J., Martin,K., Hertweck,C. and Brakhage,A.A. (2009) Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans. *Proc. Natl. Acad. Sci. U. S. A.*, 106, 14558–14563.
5. Traxler,M.F., Watrous,J.D., Alexandrov,T., Dorrestein,P.C. and Kolter,R. (2013) Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. *MBio*, 4.
6. Visca,P., Imperi,F. and Lamont,I.L. (2007) Pyoverdine siderophores: from biogenesis to biosignificance. *Trends Microbiol.*, 15, 22–30.
7. Glowacki,R.W.P. and Martens,E.C. (2020) In sickness and health: effects of gut microbial metabolites on human physiology. *PLoS Pathog.*, 16, e1008370.
8. Martinez,K.B., Leone,V. and Chang,E.B. (2017) Microbial metabolites in health and disease: navigating the unknown in search of function. *J. Biol. Chem.*, 292, 8553–8559.
9. Donia,M.S. and Fischbach,M.A. (2015) Small molecules from the human microbiota. *Science*, 349**,** 1254766–1254766.
10. Sharon,G., Garg,N., Debelius,J., Knight,R., Dorrestein,P.C. and Mazmanian,S.K. (2014) Specialized metabolites from the microbiome in health and disease. *Cell Metab.*, 20, 719–730.
11. Brown,J.M. and Hazen,S.L. (2017) Targeting of microbe-derived metabolites to improve human health: the next frontier for drug discovery. *J. Biol. Chem.*, 292, 8560–8568.
12. Wyatt,M.A., Wang,W., Roux,C.M., Beasley,F.C., Heinrichs,D.E., Dunman,P.M. and Magarvey,N.A. (2010) Staphylococcus aureus nonribosomal peptide secondary metabolites regulate virulence. *Science*, 329, 294–296.
13. Arthur,J.C. (2020) Microbiota and colorectal cancer: colibactin makes its mark. *Nat. Rev. Gastroenterol. Hepatol.*, 17, 317–318.
14. Nougayrède,J.-P., Homburg,S., Taieb,F., Boury,M., Brzuszkiewicz,E., Gottschalk,G., Buchrieser,C., Hacker,J., Dobrindt,U. and Oswald,E. (2006) *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science*, 313, 848–851.
15. Wilson,M.R., Jiang,Y., Villalta,P.W., Stornetta,A., Boudreau,P.D., Carrá,A., Brennan,C.A., Chun,E., Ngo,L., Samson,L.D., et al. (2019) The human gut bacterial genotoxin colibactin alkylates DNA. *Science*, **363**.
16. Balskus,E.P. (2015) Colibactin: understanding an elusive gut bacterial genotoxin. *Nat. Prod. Rep.*, 32, 1534–1540.
17. Mazmanian,S.K., Round,J.L. and Kasper,D.L. (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453, 620–625.

18. Mazmanian,S.K., Liu,C.H., Tzianabos,A.O. and Kasper,D.L. (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, 122, 107–118.
19. Brown,L.C.W., Wieland Brown,L.C., Penaranda,C., Kashyap,P.C., Williams,B.B., Clardy,J., Kronenberg,M., Sonnenburg,J.L., Comstock,L.E., Bluestone,J.A., et al. (2013) Production of α-Galactosylceramide by a prominent member of the human gut microbiota. *PLoS Biol.*, 11, e1001610.
20. Demain,A.L. (2014) Valuable secondary metabolites from fungi. In Martín,J.-F., García-Estrada,C., Zeilinger,S. (eds), Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites. Fungal Biology. Springer New York, New York, NY, pp. 1–15.
21. Pham,J.V., Yilma,M.A., Feliz,A., Majid,M.T., Maffetone,N., Walker,J.R., Kim,E., Cho,H.J., Reynolds,J.M., Song,M.C., et al. (2019) A review of the microbial production of bioactive natural products and biologics. *Front. Microbiol.*, 10, 1404.
22. Marinelli,F. (2009) Chapter 2. From microbial products to novel drugs that target a multitude of disease indications. *Methods Enzymol.*, 458, 29–58.
23. Donia,M.S., Cimermancic,P., Schulze,C.J., Wieland Brown,L.C., Martin,J., Mitreva,M., Clardy,J., Linington,R.G. and Fischbach,M.A. (2014) A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, 158, 1402–1414.
24. Cohen,L.J., Han,S., Huang,Y.-H. and Brady,S.F. (2018) Identification of the Colicin V bacteriocin gene cluster by functional screening of a human microbiome metagenomic library. *ACS Infect. Dis.*, 4, 27–32.
25. Wilkinson,B. and Micklefield,J. (2007) Mining and engineering natural-product biosynthetic pathways. *Nat. Chem. Biol.*, 3, 379–386.
26. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C., et al. (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, 11, 625–631.
27. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Wieland Brown,L.C., Mavrommatis,K., Pati,A., Godfrey,P.A., Koehrsen,M., Clardy,J., et al. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158, 412–421.
28. Kalkreuter,E., Pan,G., Cepeda,A.J. and Shen,B. (2020) Targeting bacterial genomes for natural product discovery. *Trends Pharmacol. Sci.*, 41, 13–26.
29. Zhou,L., Song,C., Li,Z. and Kuipers,O.P. (2021) Antimicrobial activity screening of rhizosphere soil bacteria from tomato and genome-based analysis of their antimicrobial biosynthetic potential. *BMC Genomics*, 22, 29.
30. Ceniceros,A., Dijkhuizen,L., Petrusma,M. and Medema,M.H. (2017) Genome-based exploration of the specialized metabolic capacities of the genus *Rhodococcus*. *BMC Genomics*, 18, 593.
31. Chen,R., Wong,H.L. and Burns,B.P. (2019) New approaches to detect biosynthetic gene clusters in the environment. *Medicines*, 6.
32. Grubbs,K.J., Bleich,R.M., Santa Maria,K.C., Allen,S.E., Farag,S., AgBiome Team, Shank,E.A. and Bowers,A.A. (2017) Large-scale bioinformatics analysis of *Bacillus* genomes uncovers conserved roles of natural products in bacterial physiology. *mSystems*, 2.
33. Wang,B., Guo,F., Huang,C. and Zhao,H. (2020) Unraveling the iterative type I polyketide synthases hidden in *Streptomyces*. *Proc. Natl. Acad. Sci. U. S. A.*, 117, 8449–8454.

34. Sekurova,O.N., Schneider,O. and Zotchev,S.B. (2019) Novel bioactive natural products from bacteria via bioprospecting, genome mining and metabolic engineering. *Microb. Biotechnol.*, 12, 828–844.
35. Sugimoto,Y., Camacho,F.R., Wang,S., Chankhamjon,P., Odabas,A., Biswas,A., Jeffrey,P.D. and Donia,M.S. (2019) A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science*, 366.
36. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, 39, W339–46.
37. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, 140, 13–17.
38. Skinnider,M.A., Dejong,C.A., Rees,P.N., Johnston,C.W., Li,H., Webster,A.L.H., Wyatt,M.A. and Magarvey,N.A. (2015) Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.*, 43, 9645–9662.
39. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D., et al. (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, 47, e110.
40. Navarro-Muñoz,J.C., Selem-Mojica,N., Mullowney,M.W., Kautsar,S.A., Tryon,J.H., Parkinson,E.I., De Los Santos,E.L.C., Yeong,M., Cruz-Morales,P., Abubucker,S., et al. (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, 16, 60–68.
41. Pascal Andreu,V., Roel-Touris,J., Dodd,D., Fischbach,M.A. and Medema,M.H. (2021) The gutSMASH web server: automated identification of primary metabolic gene clusters from the gut microbiota. *Nucleic Acids Res.*, 49, W263–W270.
42. Meleshko,D., Mohimani,H., Tracanna,V., Hajirasouliha,I., Medema,M.H., Korobeynikov,A. and Pevzner,P.A. (2019) BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.*, 29, 1352–1362.
43. Lee,N., Hwang,S., Kim,J., Cho,S., Palsson,B. and Cho,B.-K. (2020) Mini review: genome mining approaches for the identification of secondary metabolite biosynthetic gene clusters in *Streptomyces*. *Comput. Struct. Biotechnol. J.*, 18, 1548–1556.
44. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, 12, 902–903.
45. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, 40, D48–D53.
46. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., de los Santos,E.L.C., Kim,H.U., Nave,M., et al. (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, 45, W36–W41.
47. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., van der Hooft,J.J.J., van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Andreu,V.P., et al. (2019) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, 48, D454–D458.
48. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and

1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

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

50. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, 486, 207–214.

51. Lloyd-Price,J., Mahurkar,A., Rahnavard,G., Crabtree,J., Orvis,J., Hall,A.B., Brady,A., Creasy,H.H., McCracken,C., Giglio,M.G., et al. (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, 550, 61–66.

52. Qin,J., Li,R., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T., *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464, 59–65.

53. Hall,A.B., Yassour,M., Sauk,J., Garner,A., Jiang,X., Arthur,T., Lagoudas,G.K., Vatanen,T., Fornelos,N., Wilson,R., et al. (2017) A novel *Ruminococcus gnavus* clade enriched in inflammatory bowel disease patients. *Genome Med.*, 9, 103.

54. Schirmer,M., Franzosa,E.A., Lloyd-Price,J., McIver,L.J., Schwager,R., Poon,T.W., Ananthakrishnan,A.N., Andrews,E., Barron,G., Lake,K., et al. (2018) Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat. Microbiol.*, 3, 337–346.

55. Franzosa,E.A., Sirota-Madi,A., Avila-Pacheco,J., Fornelos,N., Haiser,H.J., Reinker,S., Vatanen,T., Hall,A.B., Mallick,H., McIver,L.J., et al. (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.*, 4, 293–305.

56. Nielsen,H.B., Bjørn Nielsen,H., MetaHIT Consortium, Almeida,M., Juncker,A.S., Rasmussen,S., Li,J., Sunagawa,S., Plichta,D.R., Gautier,L., et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, 32, 822–828.

57. Lloyd-Price,J., Arze,C., Ananthakrishnan,A.N., Schirmer,M., Avila-Pacheco,J., Poon,T.W., Andrews,E., Ajami,N.J., Bonham,K.S., Brislawn,C.J., et al. (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569, 655–662.

58. He,Q., Gao,Y., Jie,Z., Yu,X., Laursen,J.M., Xiao,L., Li,Y., Li,L., Zhang,F., Feng,Q., et al. (2017) Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience*, 6, 1–11.

59. Feng,Q., Liang,S., Jia,H., Stadlmayr,A., Tang,L., Lan,Z., Zhang,D., Xia,H., Xu,X., Jie,Z., et al. (2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.*, 6, 6528.

60. Yu,J., Feng,Q., Wong,S.H., Zhang,D., Liang,Q.Y., Qin,Y., Tang,L., Zhao,H., Stenvang,J., Li,Y., et al. (2017) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66, 70–78.

61. Zeller,G., Tap,J., Voigt,A.Y., Sunagawa,S., Kultima,J.R., Costea,P.I., Amiot,A., Böhm,J., Brunetti,F., Habermann,N., et al. (2014) Potential of fecal microbiota for early□stage detection of colorectal cancer. *Mol. Syst. Biol.*, 10, 766.

62. Zhang,X., Zhang,D., Jia,H., Feng,Q., Wang,D., Liang,D., Wu,X., Li,J., Tang,L., Li,Y., et al. (2015) The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.*, 21, 895–905.

63. Wang,M., Wan,J., Rong,H., He,F., Wang,H., Zhou,J., Cai,C., Wang,Y., Xu,R., Yin,Z., et al. (2019) Alterations in gut glutamate metabolism associated with changes in gut microbiota composition in children with autism spectrum disorder. *mSystems*, 4.
64. Zhu,F., Ju,Y., Wang,W., Wang,Q., Guo,R., Ma,Q., Sun,Q., Fan,Y., Xie,Y., Yang,Z., et al. (2020) Metagenome-wide association of gut microbiome features for schizophrenia. *Nat. Commun.*, 11, 1612.
65. Nayfach,S. and Pollard,K.S. (2016) Toward accurate and quantitative comparative metagenomics. *Cell*, 166, 1103–1116.
66. Quince,C., Walker,A.W., Simpson,J.T., Loman,N.J. and Segata,N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35, 833–844.
67. Rodriguez-R,L.M. and Konstantinidis,K.T. (2014) Estimating coverage in metagenomic data sets and why it matters. *ISME J.*, 8, 2349–2351.
68. Zaheer,R., Noyes,N., Ortega Polo,R., Cook,S.R., Marinier,E., Van Domselaar,G., Belk,K.E., Morley,P.S. and McAllister,T.A. (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.*, 8, 5890.
69. Gweon,H.S., Shaw,L.P., Swann,J., De Maio,N., AbuOun,M., Niehus,R., Hubbard,A.T.M., Bowes,M.J., Bailey,M.J., Peto,T.E.A., et al. (2019) The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environ. Microbiome*, 14, 7.
70. Sassone-Corsi,M., Nuccio,S.-P., Liu,H., Hernandez,D., Vu,C.T., Takahashi,A.A., Edwards,R.A. and Raffatellu,M. (2016) Microcins mediate competition among *Enterobacteriaceae* in the inflamed gut. *Nature*, 540, 280–283.
71. Parrot,M., Caufield,P.W. and Lavoie,M.C. (1990) Preliminary characterization of four bacteriocins from *Streptococcus mutans*. *Can. J. Microbiol.*, 36, 123–130.
72. Qi,F., Chen,P. and Caufield,P.W. (1999) Purification of mutacin III from group III *Streptococcus mutans* UA787 and genetic analyses of mutacin III biosynthesis genes. *Appl. Environ. Microbiol.*, 65, 3880–3887.
73. Merritt,J. and Qi,F. (2012) The mutacins of *Streptococcus mutans*: regulation and ecology. *Mol. Oral Microbiol.*, 27, 57–69.
74. Zipperer,A., Konnerth,M.C., Laux,C., Berscheid,A., Janek,D., Weidenmaier,C., Burian,M., Schilling,N.A., Slavetinsky,C., Marschal,M., et al. (2016) Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*, 535, 511–516.
75. Bitschar,K., Sauer,B., Focken,J., Dehmer,H., Moos,S., Konnerth,M., Schilling,N.A., Grond,S., Kalbacher,H., Kurschus,F.C., et al. (2019) Lugdunin amplifies innate immune responses in the skin in synergy with host- and microbiota-derived factors. *Nat. Commun.*, 10, 2730.
76. Hyink,O., Balakrishnan,M. and Tagg,J.R. (2005) *Streptococcus rattusstrain* BHT produces both a class I two-component lantibiotic and a class II bacteriocin. *FEMS Microbiol. Lett.*, 252, 235–241.
77. Yonezawa,H. and Kuramitsu,H.K. (2005) Genetic analysis of a unique bacteriocin, Smb, produced by *Streptococcus mutans* GS5. *Antimicrob. Agents Chemother.*, 49, 541–548.
78. Hamada,S. and Slade,H.D. (1980) Biology, immunology, and cariogenicity of *Streptococcus mutans*. *Microbiol. Rev.*, 44, 331–384.
79. Hugon,P., Mishra,A.K., Robert,C., Raoult,D. and Fournier,P.-E. (2012) Non-contiguous finished genome sequence and description of *Anaerococcus vaginalis*. *Stand. Genomic Sci.*, 6, 356–365.

80. Bian,X., Fu,J., Plaza,A., Herrmann,J., Pistorius,D., Stewart,A.F., Zhang,Y. and Müller,R. (2013) In vivo evidence for a prodrug activation mechanism during colibactin maturation. *Chembiochem*, **14**, 1194–1197.
81. Dubinsky,V., Dotan,I. and Gophna,U. (2020) Carriage of Colibactin-producing bacteria and colorectal cancer risk. *Trends Microbiol.*, 28, 874–876.
82. Secher,T., Payros,D., Brehin,C., Boury,M., Watrin,C., Gillet,M., Bernard-Cadenat,I., Menard,S., Theodorou,V., Saoudi,A., et al. (2015) Oral tolerance failure upon neonatal gut colonization with *Escherichia coli* producing the genotoxin colibactin. *Infect. Immun.*, 83, 2420–2429.
83. Fasano,A. and Shea-Donohue,T. (2005) Mechanisms of disease: the role of intestinal barrier function in the pathogenesis of gastrointestinal autoimmune diseases. *Nat. Clin. Pract. Gastroenterol. Hepatol.*, 2, 416–422.
84. Pasternak,B.A., D'Mello,S., Jurickova,I.I., Han,X., Willson,T., Flick,L., Petiniot,L., Uozumi,N., Divanovic,S., Traurnicht,A., et al. (2010) Lipopolysaccharide exposure is linked to activation of the acute phase response and growth failure in pediatric Crohn's disease and murine colitis. *Inflamm. Bowel Dis.*, 16, 856–869.
85. Caradonna,L., Amati,L., Magrone,T., Pellegrino,N.M., Jirillo,E. and Caccavo,D. (2000) Enteric bacteria, lipopolysaccharides and related cytokines in inflammatory bowel disease: biological and clinical significance. *J. Endotoxin Res.*, 6, 205–214.
86. de Waal,G.M., de Villiers,W.J.S., Forgan,T., Roberts,T. and Pretorius,E. (2020) Colorectal cancer is associated with increased circulating lipopolysaccharide, inflammation and hypercoagulability. *Sci. Rep.*, 10, 8777.
87. Lorenz,W., Buhrmann,C., Mobasheri,A., Lueders,C. and Shakibaei,M. (2013) Bacterial lipopolysaccharides form procollagen-endotoxin complexes that trigger cartilage inflammation and degeneration: implications for the development of rheumatoid arthritis. *Arthritis Res. Ther.*, 15, R111.
88. Arakawa,K., Kawai,Y., Iioka,H., Tanioka,M., Nishimura,J., Kitazawa,H., Tsurumi,K. and Saito,T. (2009) Effects of gassericins A and T, bacteriocins produced by *Lactobacillus gasseri*, with glycine on custard cream preservation. *J. Dairy Sci.*, 92, 2365–2372.
89. Coburn,P.S. and Gilmore,M.S. (2003) The *Enterococcus faecalis* cytolysin: a novel toxin active against eukaryotic and prokaryotic cells. *Cell. Microbiol.*, 5, 661–669.
90. Sung,J., Wang,Y., Chandrasekaran,S., Witten,D.M. and Price,N.D. (2012) Molecular signatures from omics data: from chaos to consensus. *Biotechnol. J.*, 7, 946–957.
91. Pahl,H.L., Krauss,B., Schulze-Osthoff,K., Decker,T., Traenckner,E.B., Vogt,M., Myers,C., Parks,T., Warring,P., Mühlbacher,A., et al. (1996) The immunosuppressive fungal metabolite gliotoxin specifically inhibits transcription factor NF-kappaB. *J. Exp. Med.*, 183, 1829–1840.
92. Tyc,O., de Jager,V.C.L., van den Berg,M., Gerards,S., Janssens,T.K.S., Zaagman,N., Kai,M., Svatos,A., Zweers,H., Hordijk,C., et al. (2017) Exploring bacterial interspecific interactions for discovery of novel antimicrobial compounds. *Microb. Biotechnol.*, 10, 910–925.
93. Deveau,A., Gross,H., Palin,B., Mehnaz,S., Schnepf,M., Leblond,P., Dorrestein,P.C. and Aigle,B. (2016) Role of secondary metabolites in the interaction between Pseudomonas fluorescens and soil microorganisms under iron-limited conditions. *FEMS Microbiol. Ecol.*, 92.
94. Yan,Q., Lopes,L.D., Shaffer,B.T., Kidarsa,T.A., Vining,O., Philmus,B., Song,C., Stockwell,V.O., Raaijmakers,J.M., McPhail,K.L., et al. (2018) Secondary metabolism and interspecific competition affect accumulation of spontaneous mutants in the GacS-GacA regulatory system in *Pseudomonas protegens*. *MBio*,

9.
95. Stubbendieck,R.M., Vargas-Bautista,C. and Straight,P.D. (2016) Bacterial communities: interactions to scale. *Front. Microbiol.*, 7, 1234.
96. O'Brien,J. and Wright,G.D. (2011) An ecological perspective of microbial secondary metabolism. *Curr. Opin. Biotechnol.*, 22, 552–558.
97. Levy,R. and Borenstein,E. (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci. U. S. A.*, 110, 12804–12809.
98. Sung,J., Kim,S., Cabatbat,J.J.T., Jang,S., Jin,Y.-S., Jung,G.Y., Chia,N. and Kim,P.-J. (2017) Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.*, 8, 15393.
99. Hibbing,M.E., Fuqua,C., Parsek,M.R. and Peterson,S.B. (2010) Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.*, 8, 15–25.
100. Bauer,M.A., Kainz,K., Carmona-Gutierrez,D. and Madeo,F. (2018) Microbial wars: competition in ecological niches and within the microbiome. *Microb. Cell Fact.*, 5, 215–219.
101. Seth,E.C. and Taga,M.E. (2014) Nutrient cross-feeding in the microbial world. *Front. Microbiol.*, 5, 350.
102. Losada,L., Ajayi,O., Frisvad,J.C., Yu,J. and Nierman,W.C. (2009) Effect of competition on the production and activity of secondary metabolites in *Aspergillus* species. *Med. Mycol.*, 47, S88–96.
103. Challis,G.L. and Hopwood,D.A. (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 14555–14561.
104. Pauwelyn,E., Huang,C.-J., Ongena,M., Leclère,V., Jacques,P., Bleyaert,P., Budzikiewicz,H., Schäfer,M. and Höfte,M. (2013) New linear lipopeptides produced by *Pseudomonas cichorii* SF1-54 are involved in virulence, swarming motility, and biofilm formation. *Mol. Plant. Microbe. Interact.*, 26, 585–598.
105. Zhang,S.-D., Isbrandt,T., Lindqvist,L.L., Larsen,T.O. and Gram,L. (2021) Holomycin, an antibiotic secondary metabolite, is required for biofilm formation by the native producer *Photobacterium galatheae* S2753. *Appl. Environ. Microbiol.*, 87.
106. Selin,C., Habibian,R., Poritsanos,N., Athukorala,S.N.P., Fernando,D. and de Kievit,T.R. (2010) Phenazines are not essential for *Pseudomonas chlororaphis* PA23 biocontrol of *Sclerotinia sclerotiorum*, but do play a role in biofilm formation. *FEMS Microbiol. Ecol.*, 71, 73–83.
107. Flemming,H.-C., Wingender,J., Szewzyk,U., Steinberg,P., Rice,S.A. and Kjelleberg,S. (2016) Biofilms: an emergent form of bacterial life. *Nat. Rev. Microbiol.*, 14, 563–575.
108. Xavier,J.B. and Foster,K.R. (2007) Cooperation and conflict in microbial biofilms. *Proc. Natl. Acad. Sci. U. S. A.*, 104, 876–881.
109. Davey,M.E. and O'toole,G.A. (2000) Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.*, 64, 847–867.
110. Nadell,C.D., Drescher,K. and Foster,K.R. (2016) Spatial structure, cooperation and competition in biofilms. *Nat. Rev. Microbiol.*, 14, 589–600.
111. Ren,D., Madsen,J.S., Sørensen,S.J. and Burmølle,M. (2015) High prevalence of biofilm synergy among bacterial soil isolates in cocultures indicates bacterial interspecific cooperation. *ISME J.*, 9, 81–89.
112. Baishya,J. and Wakeman,C.A. (2019) Selective pressures during chronic infection drive microbial

competition and cooperation. *NPJ Biofilms Microbiomes*, 5, 16.

113. Koeppel,A.F. and Wu,M. (2014) Species matter: the role of competition in the assembly of congeneric bacteria. *ISME J.*, 8, 531–540.