TaxiBGC: a Taxonomy-guided Approach for Profiling Experimentally Characterized Microbial Biosynthetic Gene Clusters and Secondary Metabolite Production Potential in Metagenomes

Vinod K. Gupta[a,b], Utpal Bakshi[c], Daniel Chang[d], Aileen R. Lee[e], John M. Davis III[f], Sriram Chandrasekaran[g,h,i], Yong-Su Jin[j,k], Michael F. Freeman[e], and Jaeyun Sung[a,b,f,#]

[a]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA

[b]Division of Surgery Research, Department of Surgery, Mayo Clinic, Rochester, MN 55905, USA

[c]Institute of Health Sciences, Presidency University, Kolkata 700073, India

[d]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA

[e]Department of Biochemistry, Molecular Biology, and Biophysics and BioTechnology Institute, University of Minnesota-Twin Cities, St. Paul, MN 55108, USA

[f]Division of Rheumatology, Department of Medicine, Mayo Clinic, Rochester, MN 55905, USA

[g]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, 48109, USA

[h]Program in Chemical Biology, University of Michigan, Ann Arbor, MI, 48109, USA

[i]Center for Bioinformatics and Computational Medicine, University of Michigan, Ann Arbor, MI, 48109, USA

[j]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[k]Department of Food Science and Human Nutrition, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[#]To whom correspondence should be addressed: Jaeyun Sung, Sung.Jaeyun@mayo.edu.

Running title: Predicting Biosynthetic Gene Clusters in Metagenomes

ABSTRACT

Biosynthetic gene clusters (BGCs) in microbial genomes encode for bioactive secondary metabolites (SMs), which can play important roles in microbe-microbe and host-microbe interactions. Given the biological significance of SMs and the current profound interest in the metabolic functions of microbiomes, the unbiased identification of BGCs from high-throughput metagenomic data could offer novel insights into the complex chemical ecology of microbial communities. Currently available tools for predicting BGCs from shotgun metagenomes have several limitations, including the need for computationally demanding read-assembly; predicting a narrow breadth of BGC classes; and not providing the SM product. To overcome these limitations, we developed TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters), a command-line tool for predicting experimentally characterized BGCs (and inferring their known SMs) in metagenomes by first pinpointing the microbial species likely to harbor them. We benchmarked TaxiBGC on various simulated metagenomes, showing that our taxonomy-guided approach could predict BGCs with much-improved performance (mean F_1 score: 0.56, mean PPV score: 0.80) compared to directly identifying BGCs by mapping sequencing reads onto the BGC genes (mean F_1 score: 0.49, mean PPV score: 0.41). Next, by applying TaxiBGC on 2,650 metagenomes from the Human Microbiome Project and various case-control gut microbiome studies, we were able to associate BGCs (and their SMs) with different human body sites and with multiple gastrointestinal diseases, including Crohn's disease and liver cirrhosis. In all, TaxiBGC provides an *in silico* platform to predict experimentally characterized BGCs and their SM production potential in metagenomic data while demonstrating important advantages over existing techniques.

IMPORTANCE

Currently available bioinformatics tools to identify BGCs from metagenomic sequencing data are limited in their predictive capability or ease of use to even computationally-oriented researchers. We present an automated computational pipeline called TaxiBGC, which predicts experimentally characterized BGCs (and infers their known SMs) in shotgun metagenomes by first considering the microbial species source. Through rigorous benchmarking techniques on simulated metagenomes, we show that TaxiBGC provides a significant advantage over existing methods. When demonstrating TaxiBGC on thousands of human microbiome samples, we associate BGCs encoding for bacteriocins with different human body sites and diseases, thereby elucidating a possible novel role of this antibiotic class in maintaining the stability of microbial ecosystems throughout the human body. Furthermore, we report for the first time gut microbial BGC associations shared amongst multiple pathologies. Ultimately, we expect our tool to facilitate future investigations into the chemical ecology of microbial communities across diverse niches and pathologies.

Keywords: Biosynthetic Gene Cluster, Natural Product, Secondary Metabolite, Metagenomics, Microbiome, Biomarkers, Bacteriocin

INTRODUCTION

Microbial secondary metabolites (SMs) (also called natural products) are a group of low molecular-weight, structurally diverse, and bioactive chemical compounds (1–3). Although not essential for the growth, development, and reproduction of microorganisms, these compounds—which include bacterial and fungal cell-to-cell signaling molecules, pigments, bacteriocins, and siderophores—are known to primarily mediate key interactions within microbial ecosystems (4–7). Interestingly, these microbial SMs have recently been found to substantially regulate many aspects of human health and disease (8–12). Furthermore, microbial SMs are a significant source of antibiotics, antifungals, anticancer agents, immunosuppressants, and other pharmaceutical drugs (13–18). For example, alkaloids were found to have cytotoxic activity against human cancer cell lines (13) and to inhibit biofilm formation of Methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococci* (VRE) (19).

The genes mainly responsible for synthesizing, modifying, and exporting SMs in microbes are often found in biosynthetic gene clusters (BGCs). In brief, BGCs are physically clustered groups of genes in a microbial genome that encode a biosynthetic pathway for producing an SM and its chemical variants (20, 21). With the advent of large-scale genomic studies and the ensuing big data revolution in natural products research (22), researchers have been able to apply computational genome mining to discover novel microbial BGCs and SMs (1, 10, 20, 23), including those from human microbiomes (16, 24). Hence, bioinformatics methods hold strong promise for detecting BGCs and inferring their SM products.

Early computational tools to detect BGCs from microbial genomes (e.g., antiSMASH (25), CLUSEAN (26), PRISM (27)) use heuristic, rule-based algorithms based on the compositional similarity of gene/protein domains to annotated reference BGCs. Other tools that utilize machine learning models have lately emerged, demonstrating a greater ability to discover novel BGCs. One such widely used method is ClusterFinder (21), which employs Hidden Markov Models for BGC detection. Additionally, DeepBGC (28) implements recurrent neural networks for BGC identification. Moreover, Navarro-Muñoz *et al*. have provided an integrated computational workflow consisting of two software tools (BiG-SCAPE and CORASON) for identifying novel gene cluster families and their phylogenetic relationships from microbial genomes (29). Furthermore, the primary specialized metabolic gene clusters (MGCs) of anaerobic bacteria were extensively studied using gutSMASH (30); this web server determines the metabolic potential of anaerobic bacteria by predicting both known and putative MGCs using the same Pfam domain detection rules used by antiSMASH. Notably, gutSMASH is specifically designed to identify MGCs from the gut microbiome. Lastly, Andreu *et al.* have recently introduced BiG-MAP, an automated command-line tool for profiling the abundance and expression of gene clusters in metagenomic and metatranscriptomic samples, respectively (31).

Approaches to predict BGCs have expanded beyond assembled or complete genomes to enable BGC prediction from culture-independent microbial communities using shotgun metagenomic sequencing. One such method is BiosyntheticSPAdes (32), which utilizes assembly graphs to detect clusters encoding for nonribosomal peptide synthetases

(NRPSs) and polyketide synthases (PKSs) from whole-genome shotgun sequences and low-complexity metagenomes. In addition, MetaBGC (24) demonstrated the characterization of BGCs directly from metagenomic reads. This method first detects BGC reads from shotgun metagenome sequences based on sequence-scoring models; afterward, the identified reads are binned for targeted assembly to reconstruct 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 (25), CLUSEAN (26), and PRISM (27) are designed to predict BGCs of known pathway classes from only individual microbial genomes. Machine learning-based methods can predict unknown BGCs, but these methods can generate more false positive predictions than rule-based approaches (33). Moreover, these rule-based and machine learning-based approaches require assembled genome (or metagenome) sequences to detect BGCs, limiting the scalability of their application toward metagenomic datasets. Among BGC detection strategies for metagenomes, BiosyntheticSPAdes is limited by the requirement of read-assembly before BGC prediction and thus requires considerable computational time and memory consumption. MetaBGC, which is an assembly-independent method, can be used directly on metagenomic reads for BGC prediction; however, as mentioned in its original publication (24), the requirement of laborious parameter optimization and unavailability of pre-built models for a BGC of interest limits the utility of this tool for a wide range of metagenomic datasets. Furthermore, neither of these two metagenome-based BGC detection methods pinpoint the specific microbes that harbor the predicted BGC genes, leaving researchers uncertain as to where the BGCs could be coming from. Clearly, there is a need to improve current computational tools to enable accurate, rapid, and unbiased predictions of BGCs—and their corresponding SMs—from shotgun metagenomes of microbial communities.

Realizing the limitations mentioned above, we set out to develop a method that accurately and rapidly detects known BGCs from real-world, complex microbiomes. Herein, we introduce TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters), an easy-to-use, command-line tool that identifies experimentally characterized BGCs from shotgun metagenomic data and infers their known SM products. TaxiBGC does not require genome assembly and can be applied to any metagenome. Importantly, our novel approach enables the prediction of experimentally characterized BGCs by first considering the microbial species from which they are derived, thereby allowing the user to trace the predicted BGCs' likely taxonomic origins.

RESULTS

TaxiBGC for profiling experimentally characterized biosynthetic gene clusters in metagenomes

TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters) is a computational strategy for predicting experimentally characterized BGCs and their annotated SMs in shotgun metagenomic sequencing data. The TaxiBGC pipeline consists of three major steps (**Fig. 1A**): the first step of the TaxiBGC pipeline performs species-level, taxonomic

profiling on the metagenome using MetaPhlAn3 (34); the second step performs the first-pass prediction of BGCs through querying these species (identified in the first step) in the TaxiBGC reference database (**Fig. 1B**)—a pre-defined collection of 390 unique species with their experimentally characterized BGCs and known SMs (**Table 1** and **Table S1**); and the last step of the TaxiBGC pipeline performs *in silico* confirmation of the predicted BGCs (from the second step) based on read mapping (i.e., alignment) using BBMap (35). Here, mapping of shotgun sequencing reads onto BGC genes is conducted with a pre-determined pair of minimum BGC gene presence and BGC coverage thresholds. Finally, the SMs corresponding to the confirmed BGCs are retrieved from the TaxiBGC reference database. TaxiBGC is completely open access for anyone to apply to their own metagenomic dataset (see **Data Availability** for links to the TaxiBGC GitHub repository and conda installation).

In our study, the presence of a BGC gene was determined based on the fulfillment of a minimum percentage length of that gene mapped onto by metagenomic reads; and the coverage of a BGC was defined as the proportion of its total number of genes that were found to be present in the metagenome. We determined a set of optimal BGC gene presence and BGC coverage thresholds for the TaxiBGC pipeline based on their overall predictive performance on 125 simulated metagenomes of mock microbial communities (**Fig. 1C**). These simulated metagenomes were constructed using 25 unique sets of varying microbial community sizes and library sizes (**Materials and Methods**). More specifically, analogous to the steps described in **Fig. 1A**, we predicted BGCs in simulated metagenomes accordingly: i) species-level taxonomic profiling; ii) initial prediction of experimentally characterized BGCs; and iii) *in silico* confirmation of BGC predictions wherein BGC genes were searched for in the metagenomes (via read mapping) using a range of BGC gene presence and BGC coverage thresholds. We obtained prediction accuracy (F_1 scores) by comparing the predicted BGCs with the actual BGCs comprising the simulated metagenomes. The pair of thresholds yielding the highest mean F_1 score after testing on all 125 simulated metagenomes was selected as the optimal minimum BGC gene-presence and BGC coverage cut-offs for TaxiBGC.

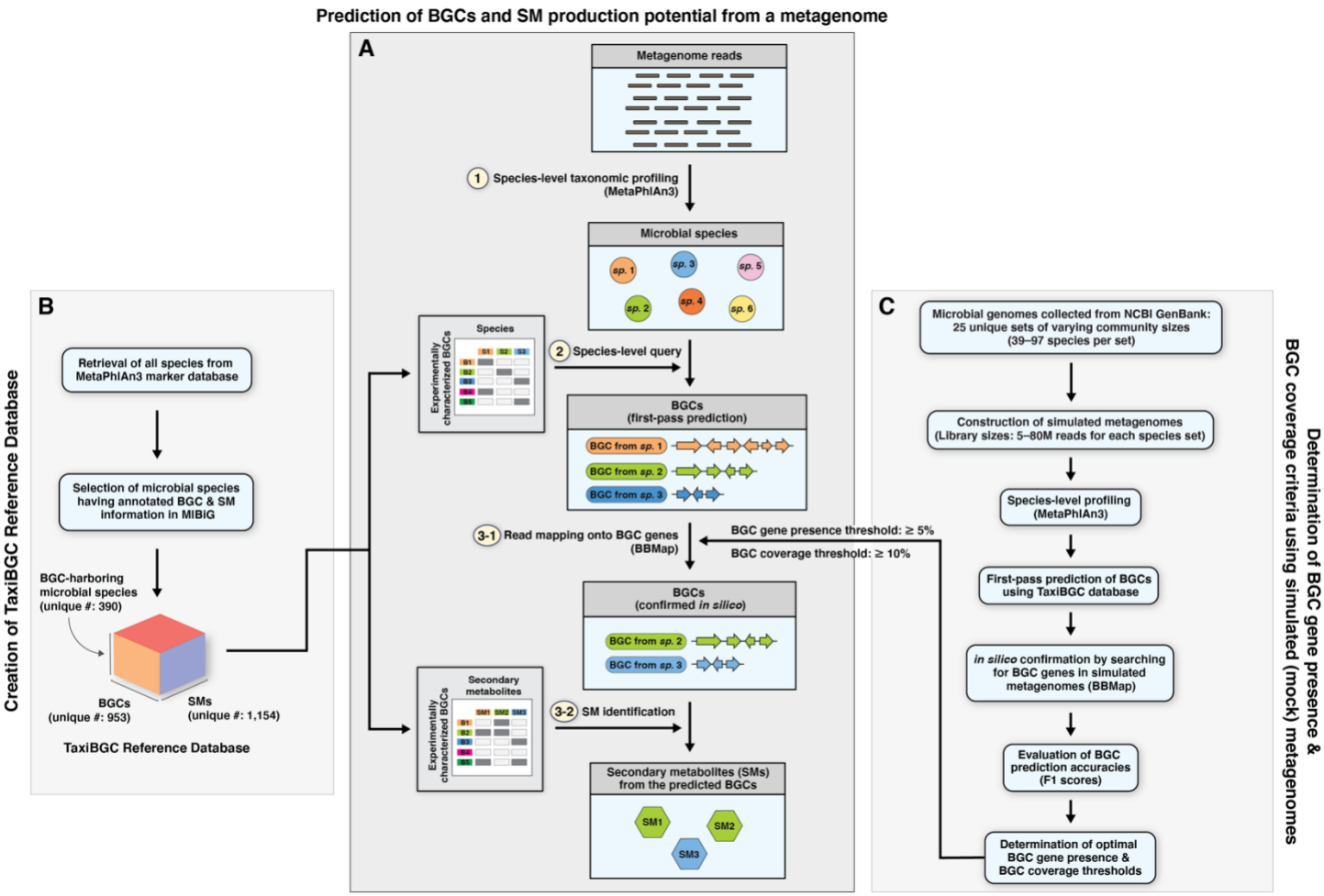

Figure 1. Schematic overview of the TaxiBGC pipeline. A) The prediction of BGCs and their corresponding SMs from a microbiome (i.e., metagenome) sample can be summarized in three main steps: first, species-level taxonomic profiling is performed using MetaPhlAn3; second, all identified species are queried in the TaxiBGC reference database (see **B**) as the first-pass prediction of which experimentally characterized BGCs may be present in the sample; and third, as an *in silico* confirmation of the predicted BGCs, metagenomic reads are mapped (aligned) onto those BGCs using BBMap with a set of pre-determined minimum BGC gene presence and BGC coverage thresholds (see **C**). For the BGCs whose presence are confirmed, their corresponding SM products are retrieved from the TaxiBGC reference database. **B)** The TaxiBGC reference database provides annotated information on microbial species' experimentally characterized BGCs and their SMs. To construct the TaxiBGC reference database, BGC and SM information was retrieved from the MIBiG database (as of January 2022). The TaxiBGC reference database encompasses 390 BGC-encoding microbial species, 953 experimentally characterized BGCs, and 1,154 SMs. **C)** TaxiBGC's BGC gene presence and BGC coverage thresholds were determined based on prediction accuracies on simulated metagenomes of mock microbial communities. Simulated metagenomes were created from 25 unique sets of varying microbial community sizes (39–97 unique species) at five different library sizes (5M, 10M, 20M, 40M, and 80M paired-end reads) (see **Materials and Methods**). Afterward, as described in part by **A**, BGCs were predicted in all combinations of simulated metagenomes to find the optimal pair of thresholds for the TaxiBGC pipeline. Here, the BGC gene presence is defined as the percentage length of a BGC gene covered by metagenomic reads; and BGC coverage is defined as the proportion of the BGC's total number of genes that are mapped onto by metagenomic reads. Prediction accuracy in the form of F_1 scores can be obtained by comparing the predicted and actual BGCs comprising the simulated metagenomes of known compositions. A pair of "BGC gene presence: 5%" and "BGC coverage: 10%" thresholds resulted in the highest mean F_1 score across a total of 125 simulated metagenomes, and were thus selected as the optimal (default) thresholds used in **A**.

Species, BGCs, and SMs of the TaxiBGC reference database

The TaxiBGC reference database essentially provides information linking experimentally characterized BGCs and their corresponding SMs with their source organism (microbial species and/or strains). This database, which is relied on by the TaxiBGC pipeline during BGC prediction (**Fig. 1A**), is pre-defined and not assembled anew for each metagenome sample. For the construction of this database (**Fig. 1B**), we obtained annotated information regarding BGCs, SMs, and their respective source organisms from the MIBiG database (as of January 2022) (36). Since the first step of the TaxiBGC pipeline relies on MetaPhlAn3 for species-level taxonomic profiling, the TaxiBGC reference database takes into consideration only the 390 microbial species in MIBiG (along with their annotated BGC and SM information) found in common with those in the MetaPhlAn3 gene marker database (**Fig. 2A**).

In all, the TaxiBGC reference database encompasses 390 unique BGC-harboring microbial species, 953 unique experimentally characterized BGCs, and 1,169 total (1,154 unique) SMs. Of note, a BGC can be derived from multiple species; a BGC can encode for multiple and/or different SMs; and more than one BGC can encode for the same SM. BGCs encoding for nonribosomal peptides (NRPs) are the most prevalent among the 390 species (165 of 390 species, 42.3%); followed by polyketides (163 of 390 species, 41.8%); ribosomally synthesized and post-translationally modified peptides (RiPPs) (117 of 390 species, 30.0%); others, i.e., not part of the six main categories provided in the MIBiG database and therefore collectively grouped into a single class (80 of 390 species, 20.5%); saccharides (74 of 390 species, 19.0%); terpenes (32 of 390 species, 8.2%); and alkaloids (12 of 390 species, 3.1%). (We followed MIBiG's standard notations for the major SM structure-based classes.) BGCs, along with their classes/subclasses, and the taxa from which they are derived are described in detail in **Table 1** and **Table S1**.

In the TaxiBGC reference database, the number of BGCs a species can harbor varies widely, ranging from one to as many as 21 (*Escherichia coli*), 22 (*Aspergillus nidulans*), or 23 (*Streptomyces griseus*); among all species, 217 (55.6% of 390) harbor only one BGC (**Fig. 2B** and **Table S2**). The number of genes in a BGC also varies widely, ranging from a single gene to 164 genes (BGC0001555 encoding for Colicin V); by and large, however, most of the 953 BGCs have less than 25 biosynthetic genes (**Fig. 2C** and **Table S3**). Among the 953 BGCs, polyketides are the most abundant (338 of 953, 35.5%), followed by NRPs (297 of 953, 31.2%), RiPPs (183 of 953, 19.2%), saccharides (125 of 953, 13.1%), other (110 of 953, 11.5%), terpenes (46 of 953, 4.8%), and alkaloids (15 of 953, 1.6%). (**Fig. 2D**). Although both BGC0001782 and BGC000986 are known to encode for a total of 13 different SMs (highest SM production potential among all BGCs), the majority of BGCs produce only one SM (796 of 953 or 83.5%) (**Fig. 2E** and **Table S4**).

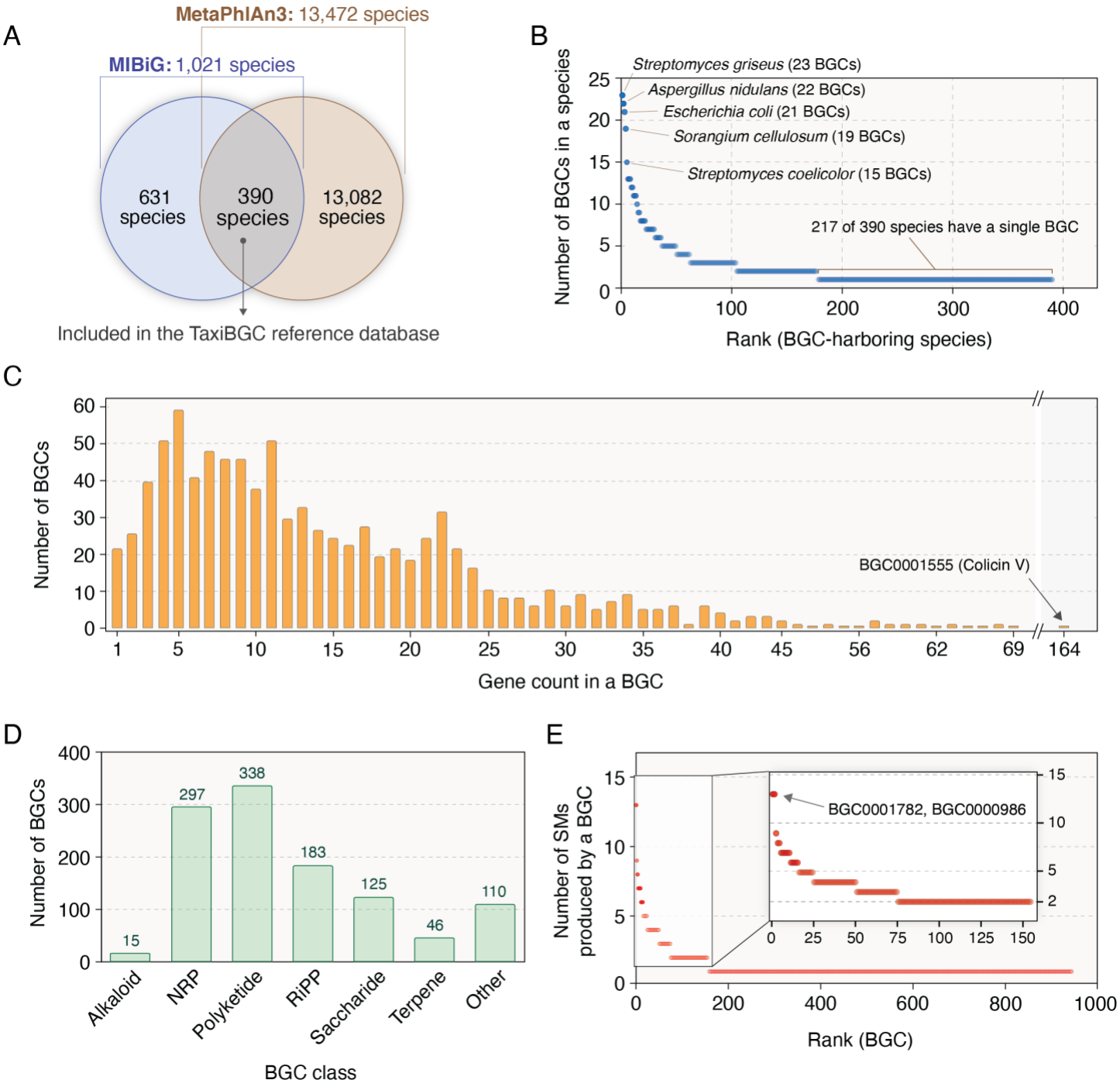

Figure 2. Relationships among microbial species, BGCs, and SMs of the TaxiBGC reference database. The TaxiBGC reference database links 390 microbial species to their 953 BGCs, which encode for 1,154 unique SMs. **A)** 1,021 BGC-harboring species were obtained from the MIBiG database (as of January 2022). Among these species, 390 were common with those in the MetaPhlAn3 gene marker database (as of January 2022). **B)** BGC-harboring microbial species rank-ordered by their number of BGCs. **C)** Number of BGCs with a given gene count. **D)** Number of BGCs in each BGC class. Of note, a BGC can belong to multiple classes (**Table S1**). **E)** BGCs rank-ordered by how many SMs they produce.

Table 1. Taxonomic summary of the BGC-harboring microbial species in the TaxiBGC reference database.

Kingdom	Phylum	Class	# of species	# of strains[a]	# of unique BGCs	BGC class[b]							# of unique SMs[b]
						Alkaloid	NRP	Polyketide	RiPP	Saccharide	Terpene	Other	
Archaea	Euryarchaeota	Methanomicrobia	1	1	1	0	0	0	0	0	0	1	1
Bacteria	Actinobacteria	Actinobacteria	163	209	393	8	92	158	59	45	27	54	515
	Bacteroidetes	Bacteroidia	3	6	8	0	0	0	0	4	0	4	8
		Chitinophagia	2	2	3	0	0	1	1	0	0	1	3
		Flavobacteriia	3	3	3	0	1	1	0	0	0	1	3
	Chloroflexi	Chloroflexia	1	1	2	0	1	0	0	0	1	0	2
	Cyanobacteria	Cyanobacteria	18	32	42	4	20	21	10	1	1	1	56
	Firmicutes	Bacilli	54	71	128	0	20	5	77	22	1	6	129
		Clostridia	8	8	9	0	0	0	8	0	0	1	8
	Proteobacteria	Alphaproteobacteria	19	20	23	0	3	2	6	4	4	5	37
		Betaproteobacteria	20	26	43	1	30	19	3	4	0	2	52
		Deltaproteobacteria	14	20	65	0	48	59	1	0	2	1	82
		Epsilonproteobacteria	2	2	2	0	1	1	0	1	0	0	2
		Gammaproteobacteria	65	81	149	0	52	25	16	44	2	26	161
Eukaryote	Ascomycota	Eurotiomycetes	17	24	82	2	29	46	2	0	8	7	110
Total			390	506	953	15	297	338	183	125	46	110	1169[c]

[a]Strain information from MIBiG (January 2022). [b]According to MIBiG (January 2022). [c]There are a total of 1,169 SMs linked to all strains in the TaxiBGC reference database. However, since strains of different taxonomic classes can have BGCs that encode for the same SM, the number of unique SMs in the reference database is 1,154.

Factors determining the predictive performance of TaxiBGC

In TaxiBGC, identifying whether a BGC is present (or not) in a metagenome is mainly dependent on the following two criteria: regarding a *BGC gene* being present, metagenomic reads should cover at least a minimum sequence length of that gene (i.e., BGC gene presence threshold); and concerning a *BGC* being present, at least a minimum proportion of its constituent genes should be present (i.e., BGC coverage threshold). On simulated metagenomes of which the BGC compositions are known, TaxiBGC prediction performance was evaluated over a range of these thresholds (**Materials and Methods**). More specifically, a total of 400 pairs of thresholds for BGC gene presence (5–100% in increments of 5%) and BGC coverage (5–100% in increments of 5%) were tested on a total of 125 simulated metagenomes. The predicted BGCs were then compared with the actual BGCs to obtain prediction accuracy.

The best overall accuracy (mean F_1 score: 0.56) for predicting BGCs in all simulated metagenomes was achieved with a minimum BGC gene presence threshold of 5% and a BGC coverage threshold of 10% (**Fig. 3A**). (The F_1 score is the harmonic mean of precision and recall.) The F_1 scores when calculated with these optimal thresholds (i.e., BGC gene presence: 5%, BGC coverage: 10%) were found to be very close to the F_1 scores found when using the pair of thresholds most accurate for each metagenome, which can vary sample-to-sample (**Fig. 3B** and **Table S5**). This indicates that the chosen BGC gene presence (5%) and BGC coverage (10%) thresholds are suitable for the TaxiBGC pipeline for predicting BGCs in metagenomes. Thus, these thresholds were set as default parameters in TaxiBGC, although the user is given the option to change them in the command-line arguments based on the desired precision or recall, aspects of the system under study, or the complexity of the dataset.

We next investigated whether TaxiBGC accuracy is influenced by inherent characteristics of the simulated metagenomes, mainly library size and relative abundances of BGC-harboring species. First, we found that F_1 scores for BGC prediction are generally not dependent on library size, although we did observe a slight decrease in F_1 scores with relatively shallow 5M (paired-end) reads compared to those with 40M and 80M reads (**Fig. 3C** and **Table S5**; $P = 0.047$, Mann–Whitney *U* test). Second, to see whether a higher relative abundance of BGC-harboring species is necessary for more favorable predictive performance, we next investigated the relationship between (geometric) mean relative abundance of those species in each sample and the resulting prediction accuracy (**Fig. 3D** and **Table S6**). We found that F_1 scores did not correlate with the mean relative abundances of BGC-harboring species except in the case with 5M reads. In summary, we recommend using TaxiBGC on metagenomes of at least 10M reads, in which case the relative abundances of BGC-harboring species are not a significant contributing factor to prediction accuracy.

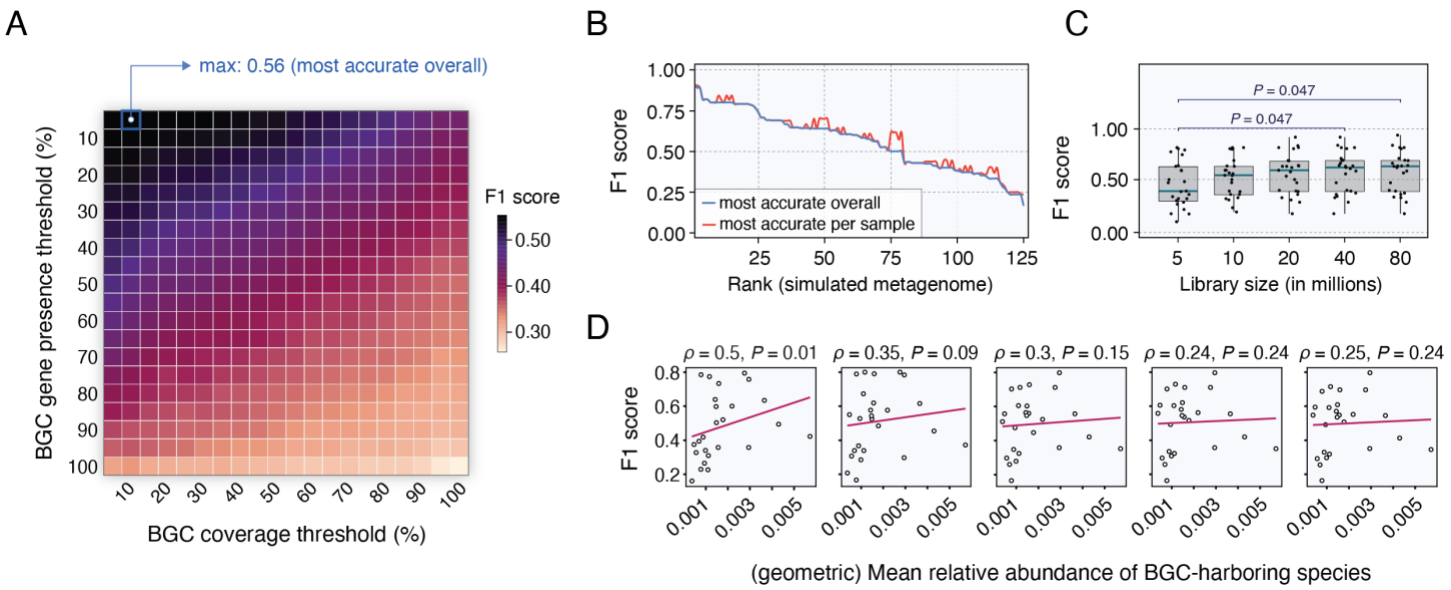

Figure 3. TaxiBGC predictive performance is determined by the BGC gene presence and BGC coverage thresholds but not by metagenome library sizes or BGC-harboring species' mean relative abundances. A) TaxiBGC was used to predict BGCs in all 125 simulated metagenomes of mock microbial communities using 400 total pairwise combinations of BGC gene presence (5–100% with an interval of 5%) and BGC coverage (5–100% with an interval of 5%) thresholds. The best overall accuracy for predicting BGCs in simulated metagenomes was achieved with a minimum BGC gene presence and BGC coverage of 5% and 10%, respectively (mean F_1 score: 0.56). Therefore, these thresholds were set as defaults in TaxiBGC but can be modified by the user. **B)** F_1 scores when using the overall optimal thresholds found in **A** (blue line) resemble those when using the most accurate pair of thresholds for each metagenome sample (red line). **C)** F_1 scores across varying simulated metagenome library sizes when using TaxiBGC with the overall optimal thresholds. Prediction accuracy of TaxiBGC was reasonably consistent in simulated metagenomes with 10M or more paired-end reads. Only the F_1 scores with a library size of 5M reads were statistically different from those with library sizes of 40M and 80M reads ($P = 0.047$, Mann–Whitney U test). Standard box-and-whisker plots (e.g., center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) are used to show distributions of F_1 scores. **D)** No significant correlation was found between F_1 scores and (geometric) mean relative abundances of BGC-harboring species, except for in the case with 5M reads (Spearman's $\rho = 0.5$, $P = 0.01$). Points in **C** and **D** correspond to individual simulated metagenomes.

Taxonomy-guided approach predicts BGCs more accurately than by directly detecting BGC genes

As described above, the TaxiBGC pipeline predicts BGCs in metagenomes by first identifying the microbial species that harbor those BGCs. As an alternative to this taxonomy-guided approach, we can also predict BGCs more directly by mapping metagenomic reads onto known BGC genes (which is the basis of existing state-of-the-art techniques). To demonstrate the predictive capability of this direct BGC detection approach, we again tested a range of BGC gene presence and BGC coverage thresholds (5–100% with intervals of 5%) on the same 125 simulated metagenomes utilized above (**Materials and Methods**). From all possible pairwise combinations of thresholds, we found that a minimum BGC gene presence of 5% and a minimum BGC coverage of 90% led to the best mean F_1 score of 0.49 (**Fig. S1**). Next, the predictive performance of the direct BGC detection approach was compared against that of TaxiBGC using the optimal thresholds found for each respective method. Having tested on simulated metagenomes, we found that TaxiBGC resulted in

significantly higher F_1 scores (**Fig. 4A**; P = 0.001, Mann–Whitney U test). Moreover, positive predictive values (PPVs) were considerably higher for TaxiBGC (**Fig. 4B**; $P < 2 \times 10^{-16}$, Mann–Whitney U test), indicating that the taxonomy-guided method generally results in fewer false positive predictions relative to true positives.

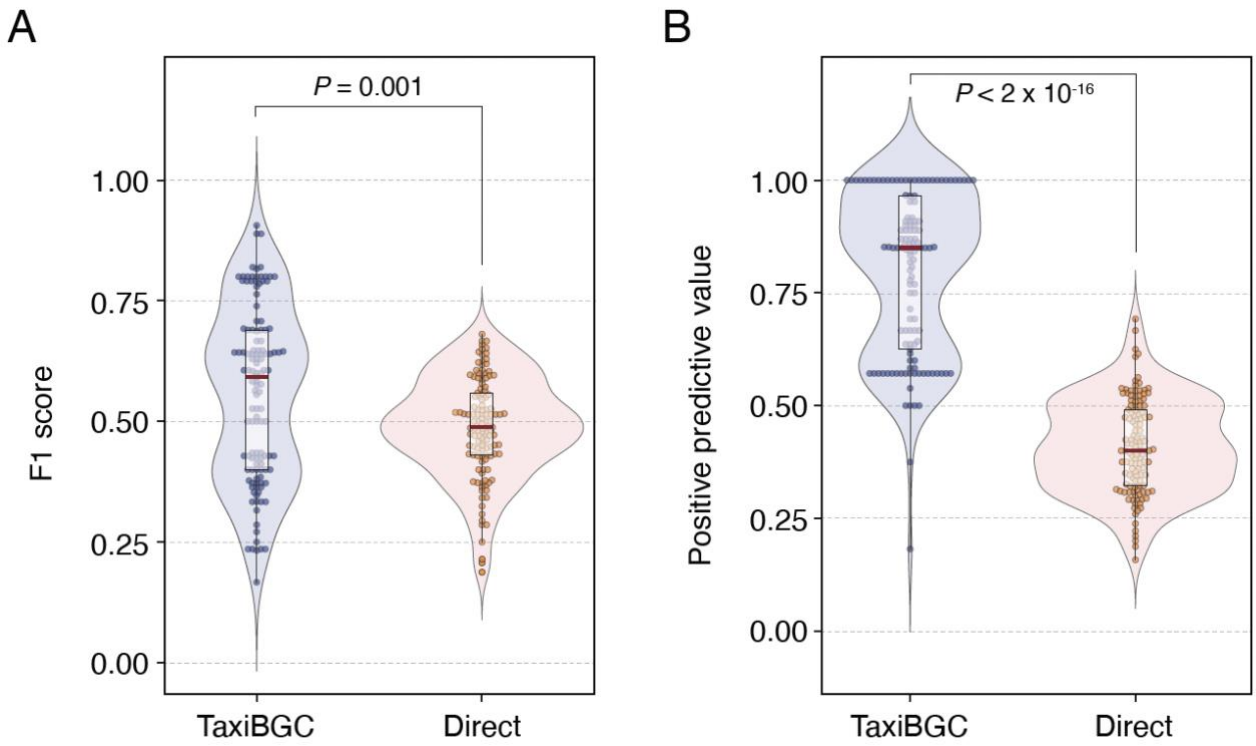

Figure 4. TaxiBGC outperforms the direct BGC detection method in predicting BGCs in metagenomic data. Both **(A)** F_1 scores and **(B)** positive predictive values (PPVs) for BGC predictions in metagenomes were significantly higher when using TaxiBGC (mean F_1 score: 0.56, mean PPV: 0.80) than when directly detecting BGCs (mean F_1 score: 0.49, mean PPV: 0.41). The BGC gene presence (BGC coverage) thresholds for TaxiBGC and the direct BGC detection approach were set as 5% (10%) and 5% (90%), respectively, as these cut-offs led to the best overall prediction accuracies for all 125 simulated metagenomes. The center red line in the box-and-whisker plots denotes the median. Mann–Whitney U test was used to test the level of significance.

BGCs for bacteriocins are prevalent in microbiomes throughout the human body

To demonstrate the utility of TaxiBGC, we applied our novel approach on 1,217 metagenome samples from the Human Microbiome Project (HMP-1-I and HMP-1-II) (37, 38). These metagenomes were sampled from the following five human body sites collected from healthy human individuals: gut, i.e., stool (472 samples), nose, i.e., anterior nares (116 samples), oral (505 samples), skin, i.e., left/right retroauricular crease (26 samples), and vagina (98 samples). (Detailed information regarding the metagenome samples used from these studies is available in **Table S7**; see **Materials and Methods** for sample

exclusion criteria.) We investigated each body site to determine the proportion of its samples predicted to have a certain BGC.

Across all five body sites, we identified a total of 24 unique BGCs present in at least one sample (**Fig. 5** and **Table S7**). Fourteen of these were found at significantly different proportions in one or more body sites ($P < 0.05$, Fisher's exact test). Most (20 of 24) BGCs were exclusive to a particular body site (i.e., absent in metagenomes from other sites); while four BGCs encoding for an exopolysaccharide (BGC0000764), nisin A (BGC0000535), nisin O (BGC0001701), and ruminococcin A (BGC0000545) were found to be present in more than one body site.

The 24 BGCs mentioned above span four main BGC classes: twenty are RiPPs, one is an NRP, one is a saccharide, and two are from the "other" class (**Fig. 5**). The twenty RiPPs, which include gassericins, lacticins, mutacins, nisins, and ruminococcin A, all belong to the bacteriocin group of antibiotics. Bacteriocins are a group of antimicrobial peptides that can kill or inhibit growth in closely-related or non-related bacterial strains (39). Interestingly, there has recently been emerging evidence that supports bacteriocin-producing bacteria as probiotic strains for benefiting human health (40). In our study, mutacin K8 was found only in the oral microbiome ($P < 0.001$, Fisher's exact test), but only for 40 (or 7.9%) of the total 505 samples. Mutacin K8 is a major lantibiotic bacteriocin produced by *Streptococcus mutans*, which is a key bacterium for the formation of dental plaques. It has been shown to have potent activity against other streptococcal species in the oral cavity (41). Additionally, a BGC for nisin O (BGC0001701) was found in both skin and gut microbiomes ($P < 0.001$, Fisher's exact test); this BGC was predicted in 35 (or 7.4%) and 1 (or 3.9%) of the 472 gut microbiome and 26 skin microbiome samples, respectively. Lastly, BGCs for gassericin E (BGC0001388), S (BGC0001601), and T (BGC000619) were predicted exclusively in the vaginal microbiome ($P < 0.001$, Fisher's exact test) for 19 (or 19.4%), 4 (or 4.1%), and 18 (or 18.4%) of the total 98 samples, respectively. A study showed that gassericin E, which was produced by *Lactobacillus gasseri* EV1461 isolated from the vagina of a healthy woman, inhibited the growth of pathogens associated with bacterial vaginosis (42). In all, our findings may motivate future studies on how bacteriocins maintain ecological homeostasis within microbial niches of the human body.

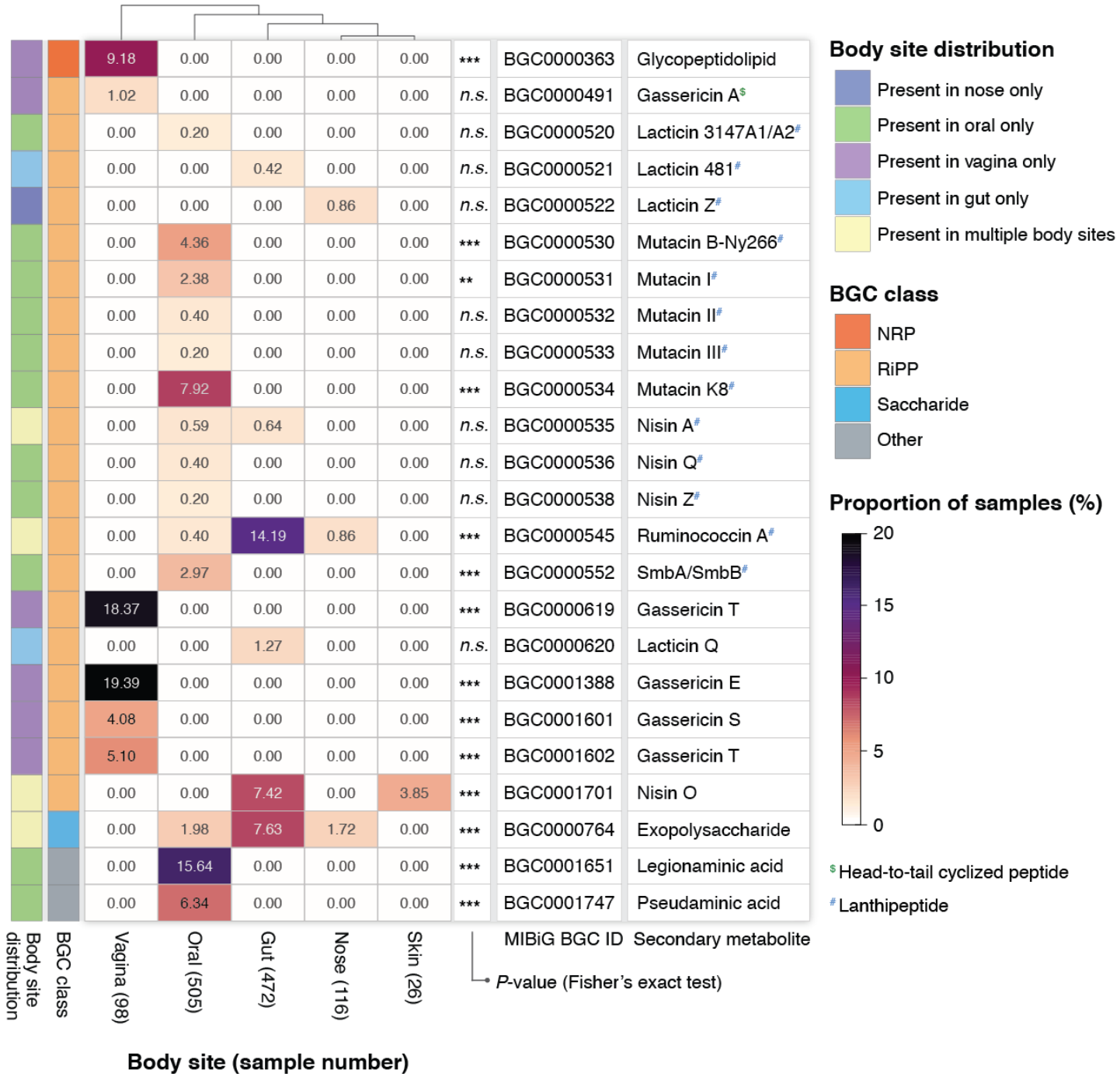

Figure 5. Prevalence of BGCs in metagenomes sampled from five human body sites. TaxiBGC was used to profile metagenomic sequencing data from each body site (provided by the Human Microbiome Project) to determine the proportion of samples in which a BGC was predicted. A total of 24 BGCs were found to be present in at least one sample from all five body sites, of which fourteen had significantly different proportions in one or more body sites ($P < 0.05$, Fisher's exact test). The color intensity in the heatmap reflects the proportion of samples (in a body site) wherein the corresponding BGC was found. Twenty BGCs were predicted in only one body site. The following symbols indicate the level of significance for the two-sided Fisher's exact test: *n.s.* for $P \geq 0.05$, ** for $0.001 \leq P < 0.01$, and *** for $P < 0.001$. RiPP subclasses: $ for Head-to-tail cyclized peptide, and # for Lanthipeptide.

Gut microbial BGCs associate with human diseases

We can also demonstrate TaxiBGC in the context of translational research to find BGCs associated with clinical phenotypes of interest. For this, we applied TaxiBGC on 1,433 gut microbiome (i.e., stool metagenome) samples from the following nine published case-control studies: three Crohn's disease (CD) studies (43–45), four colorectal cancer (CRC) studies (46–49); and one study each for rheumatoid arthritis (RA) (50) and liver cirrhosis (51). (Detailed information regarding the metagenome samples used from these studies is available in **Table S8**; see **Materials and Methods** for sample exclusion criteria.) As shown in **Figure 6**, a total of fourteen unique BGCs were found in gut microbiomes in significantly different proportions between patients and control individuals ($P < 0.05$, Fisher's exact test), thereby associating BGCs and their SM production potential with human diseases.

In the study involving gut microbiomes from patients with LC and controls, ten BGCs were significantly different in proportions between LC patients and controls (**Fig. 6**): nine BGCs (encoding for an exopolysaccharide, gassericin A, gassericin E, gassericin S, gassericin T, mutacin I, mutacin III, macedocin, and ruminococcin A) were found to be present in a significantly higher proportion of LC patients than in controls; on the other hand, only one BGC for nisin O was found more often in controls than in LC patients. Additionally, TaxiBGC found two BGCs encoding for nisin A and nisin Z less frequently in the gut microbiomes of patients with RA compared with those of controls (**Fig. 6**).

Six BGCs were identified in significantly different proportions between patients with CD and controls (**Fig. 6**). Five BGCs encoding for gassericin E (in Franzosa (2018)), gassericin S (in Franzosa (2018)), gassericin T (in Nielsen (2014) and Franzosa (2018)), lacticin Q (in Nielsen (2014)), and ruminococcin A (in Nielsen (2014)) were found more frequently in CD patients in separate individual studies. The one remaining BGC for nisin O (in He (2017) was predicted more often in controls; that these patterns for the nisin O-/gassericin T-encoding BGCs were observed in two different studies may indicate that our findings are robust.

Four BGCs were found to be associated with CRC (**Fig. 6**). Three gut microbial BGCs were found in significantly higher proportions in CRC (compared with controls); these BGCs encode for lactocillin (in Yu (2015)), nisin Z (in Feng (2015)), and ruminococcin A (in Feng (2015) and Wirbel (2019)). Alternatively, as was observed for CD studies, a BGC for nisin O was found in significantly higher proportions in controls than in CRC patients (in Zeller (2014)). Among these four BGCs associated with CRC, three encode for bacteriocins (nisin O, nisin Z, and ruminococcin A), whereas one encodes for lactocillin, which is a thiopeptide antibiotic. Previously, BGCs for thiopeptides were found to be widely distributed in genomes of microbial species inhabiting the human gut, oral, and vaginal microbiota (16).

Interestingly, five of the disease-associated gut microbial BGCs mentioned above were shared across multiple gastrointestinal diseases. For example, BGC0001701 encoding for nisin O was found more often in controls than in LC, CD, and CRC (see pink lines in **Fig. 6**). In contrast, in the same three diseases, BGC0000545 for ruminococcin A was found in higher proportions in cases than in controls (see blue lines in **Fig. 6**). Lastly, three BGCs for gassericin E (BGC0001388;

see green lines in **Fig. 6**), gassericin S (BGC0001601; see gold lines in **Fig. 6**), and gassericin T (BGC0000619; see purple lines in **Fig. 6**) were found in significantly higher proportions in both CD and LC patients than in controls. Furthermore, all five of these BGCs encode for bacteriocins. In summary, building upon previous findings of gut microbiome taxonomies or genes shared amongst different human diseases (52–54), we report for the first time experimentally characterized BGCs associated with the gut microbiome of multiple pathologies.

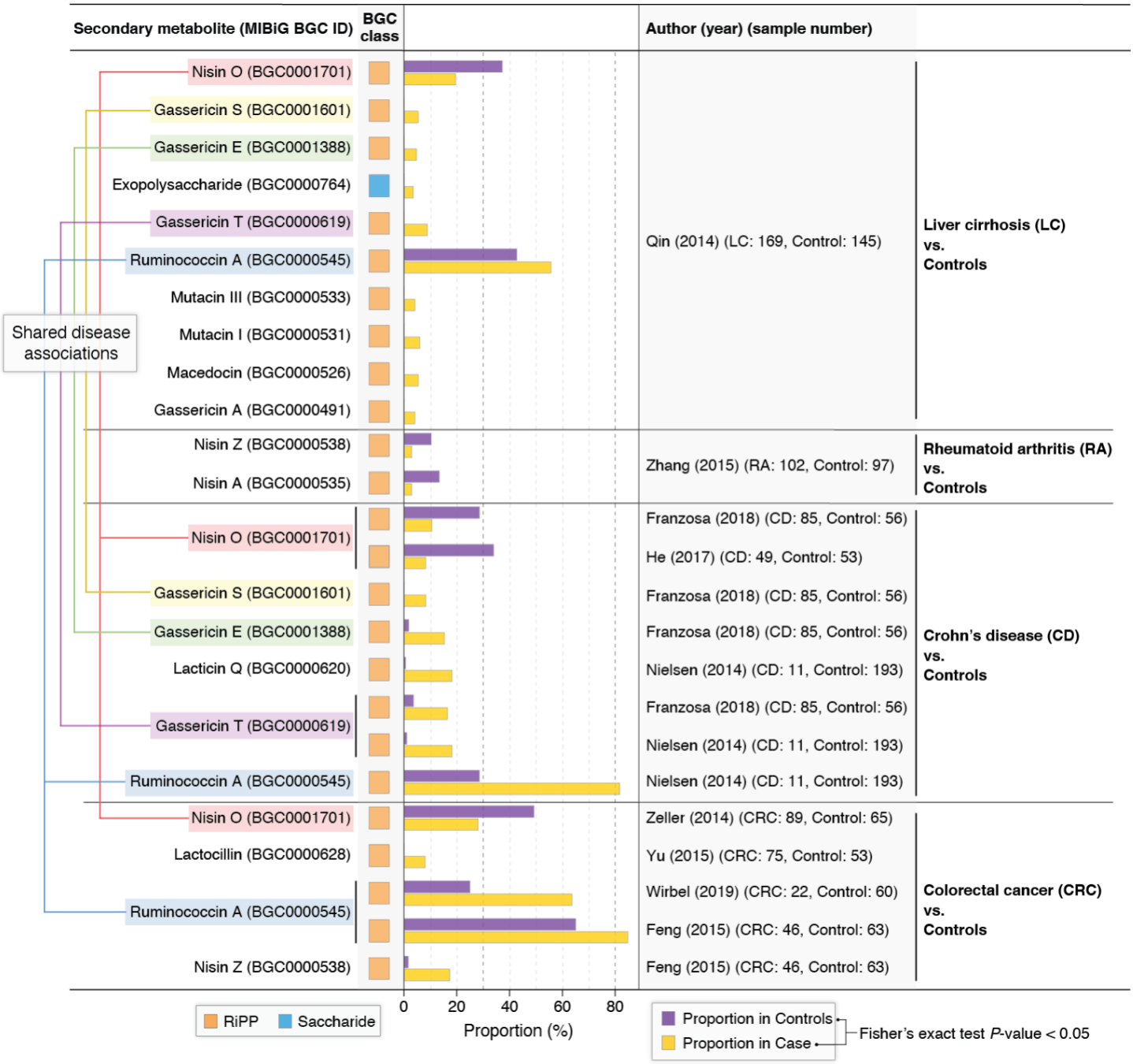

Figure 6. TaxiBGC identifies gut microbial BGCs associated with several human pathologies. TaxiBGC was used to identify BGCs in gut microbiome samples from multiple case-control studies. In liver cirrhosis (LC), BGCs encoding for ten different SMs were found to be associated with either LC or control; in rheumatoid arthritis (RA), two BGCs encoding for nisin A and nisin Z were found in lower proportions in RA than in controls; in Crohn's disease (CD), BGCs encoding for five different SMs (gassericin E, gassericin S, gassericin T, lacticin Q, and ruminococcin A) were found more often in CD. In contrast, a BGC for nisin O was found more frequently in controls; in colorectal cancer (CRC), three BGCs encoding for lactocillin, nisin Z, and ruminococcin A were predicted in higher proportions in CRC. In comparison, one BGC encoding for nisin O was found more frequently in controls. Case-control studies are designated with the lead author's last name and publication year. Sample numbers are shown in parentheses next to their respective cohort. Significance for association was calculated based on a two-sided Fisher's exact test ($P < 0.05$).

DISCUSSION

Accurately predicting the biosynthetic potential of BGCs from uncultivated microbiomes is currently one of the most important challenges in microbial natural products research. Here we present TaxiBGC, a computational pipeline for predicting experimentally characterized, microbial BGCs with their known SMs in shotgun metagenomic datasets. TaxiBGC employs three main steps for predicting BGCs: i) identification of BGC-harboring species, ii) initial prediction of BGCs based on the identified BGC-harboring species, and iii) *in silico* confirmation of the predicted BGCs by mapping metagenomic reads onto the genes. We evaluated the prediction accuracy of TaxiBGC on simulated metagenomes (of mock microbial communities) constructed using varied combinations of species identities, species richness, and sequencing library sizes. Accordingly, we showed that BGC prediction using our taxonomy-guided approach (mean F_1 score: 0.56; and mean PPV score: 0.80) was superior to directly mapping reads onto BGC genes (mean F_1 score: 0.49; and mean PPV score: 0.41). Next, to identify BGCs (and thereby being able to estimate their corresponding SMs) associated with human health and disease, we demonstrated our TaxiBGC pipeline on metagenomes sampled from different body sites; and on gut microbiomes collected from patients with liver cirrhosis, rheumatoid arthritis, Crohn's disease, or colorectal cancer. Our results show that human body sites harbor distinct microbial BGC signatures which mostly encode for bacteriocins. Furthermore, by characterizing the potential chemical milieu of the gut microbial environment for different diseases and their controls, our results provide a starting point for discovering stool-borne SM biomarkers and therapeutic leads.

Several limitations of TaxiBGC should be noted when interpreting our results. First, by design, BGC prediction by TaxiBGC relies on the successful identification of microbial species from metagenomes. In the case where MetaPhlAn3 (or other taxonomy profiling tools in future versions of TaxiBGC) cannot identify all the BGC-harboring species, TaxiBGC would not be able 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 multiple species. Nevertheless, TaxiBGC can spark new and exciting hypotheses regarding the connections between microbiome BGCs and their source species, which can be validated in future empirical studies. Third, although the taxonomy-guided step in our approach can certainly reduce false positive BGC predictions relative to true positives, the trade-off is that we may miss BGCs that are present in the

metagenomes, thereby leading to type II errors (false negatives). Fourth, TaxiBGC provides only the chemical family—but not the specific molecular structure—for a small number of SM products, such as glycopeptidolipid and exopolysaccharide. This is simply because more detailed information is not yet available in MIBiG. Fifth, TaxiBGC is designed to predict only known BGCs whose SMs have been verified by other research groups. Therefore, its predictions do not extend to putative (unvalidated) BGCs that have yet to be isolated and expressed in bacterial hosts. Finally, experimental verification of whether the BGCs (and their SMs) reported herein are indeed implicated in health outcomes, e.g., either maintaining the structure of the so-called "healthy" gut microbiome (53, 55, 56) or eliciting immunopathogenic effects, remains an essential step but beyond the scope of this computational study.

Despite the limitations mentioned above, TaxiBGC provides several advantages over currently available tools for BGC identification. First, TaxiBGC employs an assembly-independent, read-based prediction method; this solves a fundamental limitation of current BGC prediction tools that assume each BGC is encoded within a single contig in the genome or assembled metagenome, 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 characterized BGCs in a metagenome sample. This comprehensive and unbiased 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 before running TaxiBGC. The only pre-installed software needed in TaxiBGC are MetaPhlAn3 for taxonomic species identification; and BBMap v38.90 (or higher) for mapping metagenomic reads onto BGC genes.

In closing, we anticipate that TaxiBGC can facilitate the unbiased identification of experimentally characterized BGCs and their annotated SMs from shotgun metagenomes. It is also worth noting that TaxiBGC is not limited to the use in human microbiomes but can easily be applied to solve questions in other types of microbiomes (e.g., animal models, soil, seawater, plants) for BGC identification. Additionally, we envision that our novel pipeline can contribute to our ongoing efforts to computationally model the wide range of potential metabolite-driven interactions within complex microbial communities (57–59). Finally, to support these efforts and beyond, we plan to continuously expand and curate the current version of TaxiBGC with data newly added to MIBiG, as well as with updates to taxonomic profiling software.

MATERIALS AND METHODS

Constructing simulated metagenomes

Simulated metagenomes were constructed according to the following steps: First, 25 metagenome samples were randomly selected from the Human Microbiome Project (HMP) (37, 38). Second, the species composition of each sample was determined using MetaPhlAn v3.0.13 (34). Third, mock microbial communities were made with BGC-harboring and non-BGC-harboring species while preserving the original species richness and relative abundance quantities found for each

HMP metagenome (**Table S9** and **Fig. S2**). In other words, the actual species in each HMP sample were replaced with BGC-harboring species (from MIBiG) and non-BGC-harboring species (from NCBI GenBank) at random. This uneven distribution of species' counts and abundances allows the generation of simulated metagenomes as close to real metagenomes as possible. In all, 25 mock microbial communities, in which the unique species richness (i.e., count) ranged from ~40 to ~100, were used to construct simulated metagenomes (**Table S10**). Fourth, for each species in a mock microbial community, the genome corresponding to its strain, or that of a randomly selected strain in case there were multiple strains for a particular species, was downloaded from NCBI GenBank. Fifth, the collection of genome sequences for all species of a mock community was converted into paired-end metagenomic reads of 100 bp read length using NeSSM (Next-generation Sequencing Simulator for Metagenomics) (60). In NeSSM, the number of reads generated from each genome is commensurate with the known community composition (thereby considering the species' names, total count, and relative abundances) and the specified library size, resulting in a simulated metagenome for a given mock community. By varying sequencing library sizes, a total of 125 simulated metagenomes ((25 unique sets of microbial species) × (5 library sizes: 5M, 10M, 20M, 40M, and 80M paired-end reads)) were constructed.

Predicting BGCs in simulated metagenomes using TaxiBGC

First, species-level taxonomic profiling was performed on each simulated metagenome using MetaPhlAn v3.0.13 (34) with default parameters. Next, all identified species were queried in the TaxiBGC reference database for the first-pass prediction of BGCs that might be present. Then, as an *in silico* confirmation of those predicted BGCs, genes of those BGCs were searched for in the metagenomes using BBMap v38.90 (35) with minimum BGC gene presence (i.e., percentage length of a BGC gene covered by metagenomic reads) and BGC coverage (i.e., percentage of the total number of genes present for a particular BGC) cut-offs. BGCs that passed these two criteria were considered present in the given metagenome.

Predicting BGCs in simulated metagenomes through direct detection of BGC genes

The direct BGC gene detection strategy was used to predict BGCs in a simulated metagenome without taxonomic profiling. Accounting for all BGCs in the MIBiG database (1,905 BGCs as of January 2022; **Table S1**), this approach searches for BGC genes in simulated metagenomes using BBMap v38.90 with a pair of minimum BGC gene presence and BGC coverage cut-offs. BGCs that passed these two criteria were considered present in the given metagenome.

F_1 scores for evaluating TaxiBGC prediction accuracy

The F_1 score, defined as the harmonic mean of precision and recall, is calculated as follows:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

where:

$$precision = \frac{TP}{TP + FP}$$

and

$$recall = \frac{TP}{TP + FN}$$

Here, for a given metagenome, TP (true positives) is the number of BGCs that were correctly predicted to be present; FP (false positives) is the number of BGCs that were incorrectly predicted to be present; and FN (false negatives) is the number of BGCs that were missed, i.e., incorrectly predicted to be absent by TaxiBGC.

Exclusion criteria for metagenomic datasets used in TaxiBGC demonstration

Metagenome samples from the following were excluded from the current study: i) subjects undergoing dietary/medication interventions; ii) metagenome samples with less than 5M reads; iii) metagenomes samples sequenced as single-end reads; and iv) samples collected from patients with comorbidities. In the case wherein multiple runs were available for a sample, only the sequencing run with the highest read count (of at least 5M reads) was included for analysis. If multiple samples were collected from the same individual across different time points, then only the sample collected at the first recorded time point was considered. In all, the TaxiBGC pipeline was applied on 2,650 publicly available, human-derived metagenomes from eleven independent published studies. Raw sequence files of all metagenomes were downloaded from the NCBI Sequence Read Archive.

DATA AVAILABILITY

TaxiBGC is an open-source command-line tool that is implemented in Python and bash. To avoid dependency conflicts, TaxiBGC can be installed via Anaconda: https://anaconda.org/danielchang2002/TaxiBGC. The source code, MIBiG BGC gene sequences, the TaxiBGC reference database, and complete instructions for installation and usage are freely available at https://github.com/danielchang2002/TaxiBGC_2022.

SUPPLEMENTAL MATERIAL

Figure S1. Predictive performance of the direct BGC detection approach based on two different thresholds.

Figure S2. Relative abundances of strains in each simulated metagenome.

Table S1. BGCs, SMs, and taxonomies in the TaxiBGC reference database.

Table S2. Number of BGCs that each species harbors.

Table S3. Number of genes comprising every BGC in the TaxiBGC reference database.

Table S4. Experimentally characterized BGCs and their SM products in the TaxiBGC reference database.

Table S5. F1 scores of BGC prediction in simulated metagenomes.

Table S6. Mean relative abundances of BGC-harboring species in each simulated metagenome.

Table S7. HMP-1-I (Huttenhower, 2012) and HMP-1-II (Lloyd-Price, 2019) metagenomes used for TaxiBGC demonstration.

Table S8. BGCs predicted by using TaxiBGC on stool metagenomes collected from published case-control studies.

Table S9. Relative abundances of strains (of species) in each mock microbial community used for constructing simulated metagenomes (SMs).

Table S10. Number of BGC-harboring and non-BGC-harboring microbial species in simulated metagenomes.

ACKNOWLEDGMENTS

We thank David N. Rider and Ronald L. Neuharth for their help in downloading all metagenome samples used in our study. This work was supported in part by the Mayo Clinic Center for Individualized Medicine (J.S.) and Mark E. and Mary A. Davis to Mayo Clinic Center for Individualized Medicine (J.M.D. and J.S.). The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

AUTHOR CONTRIBUTIONS

J.S. conceived the problem. V.K.G., U.B., and J.S. designed all analytical methodologies. V.K.G., U.B., and D.C. built the TaxiBGC computational pipeline and software. V.K.G. and U.B. performed all analyses. All authors analyzed the data. V.K.G., U.B., and J.S. wrote the manuscript, with editorial contributions from other authors.

The authors declare no conflict of interest.

REFERENCES

1. Baral B, Akhgari A, Metsä-Ketelä M. 2018. Activation of microbial secondary metabolic pathways: Avenues and challenges. Synth Syst Biotechnol 3:163–178.

2. Singh BP, Rateb ME, Rodriguez-Couto S, Polizeli M de LT de M, 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 (Tokyo) 66:361–364.

4. Schroeckh V, Scherlach K, Nützmann H-W, Shelest E, Schmidt-Heck W, Schuemann J, Martin K, Hertweck C, Brakhage AA. 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 MF, Watrous JD, Alexandrov T, Dorrestein PC, Kolter R. 2013. Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. MBio 4.

6. Visca P, Imperi F, Lamont IL. 2007. Pyoverdine siderophores: from biogenesis to biosignificance. Trends Microbiol 15:22–30.

7. Schmidt R, Jager V de, Zühlke D, Wolff C, Bernhardt J, Cankar K, Beekwilder J, van Ijcken W, Sleutels F, Boer W de, Riedel K, Garbeva P. 2017. Fungal volatile compounds induce production of the secondary metabolite Sodorifen in Serratia plymuthica PRI-2C. Sci Rep 7.

8. Glowacki RWP, Martens EC. 2020. In sickness and health: Effects of gut microbial metabolites on human physiology. PLoS Pathog 16:e1008370.

9. Martinez KB, Leone V, Chang EB. 2017. Microbial metabolites in health and disease: Navigating the unknown in search of function. J Biol Chem 292:8553–8559.

10. Donia MS, Fischbach MA. 2015. HUMAN MICROBIOTA. Small molecules from the human microbiota. Science 349:1254766.

11. Sharon G, Garg N, Debelius J, Knight R, Dorrestein PC, Mazmanian SK. 2014. Specialized metabolites from the microbiome in health and disease. Cell Metab 20:719–730.

12. Brown JM, Hazen SL. 2017. Targeting of microbe-derived metabolites to improve human health: The next frontier for drug discovery. J Biol Chem 292:8560–8568.

13. Demain AL, Vaishnav P. 2011. Natural products for cancer chemotherapy. Microb Biotechnol 4:687–699.

14. Pham JV, Yilma MA, Feliz A, Majid MT, Maffetone N, Walker JR, Kim E, Cho HJ, Reynolds JM, Song MC, Park SR, Yoon YJ. 2019. A review of the microbial production of bioactive natural products and biologics. Front Microbiol 10:1404.

15. Marinelli F. 2009. Chapter 2. From microbial products to novel drugs that target a multitude of disease indications. Methods Enzymol 458:29–58.

16. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG, Fischbach MA. 2014. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158:1402–1414.

17. Mann J. 2001. Natural products as immunosuppressive agents. Nat Prod Rep 18:417–430.

18. Culp EJ, Yim G, Waglechner N, Wang W, Pawlowski AC, Wright GD. 2019. Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. Nat Biotechnol 37:1149–1154.

19. Bessa LJ, Buttachon S, Dethoup T, Martins R, Vasconcelos V, Kijjoa A, Martins da Costa P. 2016. Neofiscalin A and fiscalin C are potential novel indole alkaloid alternatives for the treatment of multidrug-resistant Gram-positive bacterial infections. FEMS Microbiol Lett 363.

20. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Düsterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJN, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, Yim G, Yu F, Xie Y, Aigle B, Apel AK, Balibar CJ, Balskus EP, Barona-Gómez F, Bechthold A, Bode HB, Borriss R, Brady SF, Brakhage AA, Caffrey P, Cheng Y-Q, Clardy J, Cox RJ, De Mot R, Donadio S, Donia MS, van der Donk WA, Dorrestein PC, Doyle S, Driessen AJM, Ehling-Schulz M, Entian K-D, Fischbach MA, Gerwick L, Gerwick WH, Gross H, Gust B, Hertweck C, Höfte M, Jensen SE, Ju J, Katz L, Kaysser L, Klassen JL, Keller NP, Kormanec J, Kuipers OP, Kuzuyama T, Kyrpides NC, Kwon H-J, Lautru S, Lavigne R, Lee CY, Linquan B, Liu X, Liu W, Luzhetskyy A, Mahmud T, Mast Y, Méndez C, Metsä-Ketelä M, Micklefield J, Mitchell DA, Moore BS, Moreira LM, Müller R, Neilan BA, Nett M, Nielsen J, O'Gara F, Oikawa H, Osbourn A, Osburne MS, Ostash B, Payne SM, Pernodet J-L, Petricek M, Piel J, Ploux O, Raaijmakers JM, Salas JA, Schmitt EK, Scott B, Seipke RF, Shen B, Sherman DH, Sivonen K, Smanski MJ, Sosio M, Stegmann E, Süssmuth RD, Tahlan K, Thomas CM, Tang Y, Truman AW, Viaud M, Walton JD, Walsh CT, Weber T, van Wezel GP, Wilkinson B, Willey JM, Wohlleben W, Wright GD, Ziemert N, Zhang C, Zotchev SB, Breitling R, Takano E, Glöckner FO. 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11:625–631.

21. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421.

22. Cech NB, Medema MH, Clardy J. 2021. Benefiting from big data in natural products: importance of preserving foundational skills and prioritizing data quality. Nat Prod Rep 38:1947–1953.

23. Alkreuter E, Pan G, Cepeda AJ, Shen B. 2020. Targeting bacterial genomes for natural product discovery. Trends Pharmacol Sci 41:13–26.

24. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, Jeffrey PD, Donia MS. 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 366:eaax9176.

25. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, 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.

26. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. 2009. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol 140:13–17.

27. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster ALH, Wyatt MA, Magarvey NA. 2015. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 43:9645–9662.

28. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, Durcak J, Wurst M, Kotowski J, Chang D, Wang R, Piizzi G, Temesi G, Hazuda DJ, Woelk CH, Bitton DA. 2019. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res 47:e110.

29. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68.

30. Pascal Andreu V, Roel-Touris J, Dodd D, Fischbach MA, Medema MH. 2021. The gutSMASH web server: automated identification of primary metabolic gene clusters from the gut microbiota. Nucleic Acids Res 49:W263–W270.

31. Pascal Andreu V, Augustijn HE, van den Berg K, van der Hooft JJJ, Fischbach MA, Medema MH. 2021. BiG-MAP: An automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. mSystems 6:e0093721.

32. Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res 29:1352–1362.

33. Lee N, Hwang S, Kim J, Cho S, Palsson B, 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.

34. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10.

35. Bushnell B. 2014. BBMap: A fast, accurate, splice-aware aligner. United States: N. p., 2014. Web.

36. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454–D458.

37. Consortium THMP, The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature https://doi.org/10.1038/nature11234.

38. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61–66.

39. Cotter PD, Hill C, Ross RP. 2005. Bacteriocins: developing innate immunity for food. Nat Rev Microbiol 3:777–788.

40. Dobson A, Cotter PD, Ross RP, Hill C. 2012. Bacteriocin production: a probiotic trait? Appl Environ Microbiol 78:1–6.

41. Watanabe A, Kawada-Matsuo M, Le MN-T, Hisatsune J, Oogai Y, Nakano Y, Nakata M, Miyawaki S, Sugai M, Komatsuzawa H. 2021. Comprehensive analysis of bacteriocins in Streptococcus mutans. Sci Rep 11:12963.

42. Maldonado-Barragán A, Caballero-Guerrero B, Martín V, Ruiz-Barba JL, Rodríguez JM. 2016. Purification and genetic characterization of gassericin E, a novel co-culture inducible bacteriocin from Lactobacillus gasseri EV1461 isolated from the vagina of a healthy woman. BMC Microbiol 16:37.

43. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA, Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ. 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 4:293–305.

44. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD, MetaHIT Consortium. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828.

45. He Q, Gao Y, Jie Z, Yu X, Laursen JM, Xiao L, Li Y, Li L, Zhang F, Feng Q, Li X, Yu J, Liu C, Lan P, Yan T, Liu X, Xu X, Yang H, Wang J, Madsen L, Brix S, Wang J, Kristiansen K, Jia H. 2017. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. Gigascience 6:1–11.

46. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, Su L, Li X, Li X, Li J, Xiao L, Huber-Schönauer U, Niederseer D, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K, Arumugam M, Tilg H, Datz C, Wang J. 2015. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun 6:6528.

47. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Van Nhieu JT, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10:766.

48. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, Wang X, Xu X, Chen N, Wu WKK, Al-Aama J, Nielsen HJ, Kiilerich P, Jensen BAH, Yau TO, Lan Z, Jia H, Li J, Xiao L, Lam TYT, Ng SC, Cheng AS-L, Wong VW-S, Chan FKL, Xu X, Yang H, Madsen L, Datz C, Tilg H, Wang J, Brünner N, Kristiansen K, Arumugam M, Sung JJ-Y, Wang J. 2017. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 66:70–78.

49. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25:679–689.

50. Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, Wu X, Li J, Tang L, Li Y, Lan Z, Chen B, Li Y, Zhong H, Xie H, Jie Z, Chen W, Tang S, Xu X, Wang X, Cai X, Liu S, Xia Y, Li J, Qiao X, Al-Aama JY, Chen H, Wang L, Wu Q-J, Zhang F, Zheng W, Li Y, Zhang M, Luo G, Xue W, Xiao L, Li J, Chen W, Xu X, Yin Y, Yang H, Wang J, Kristiansen K, Liu L, Li T, Huang Q, Li Y, Wang J. 2015. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 21:895–905.

51. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich SD, Zheng S, Li L. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59–64.

52. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications https://doi.org/10.1038/s41467-017-01973-8.

53. Gupta VK, Kim M, Bakshi U, Cunningham KY, Davis JM 3rd, Lazaridis KN, Nelson H, Chia N, Sung J. 2020. A predictive index for health status using species-level gut microbiome profiling. Nat Commun 11:4635.

54. Priya S, Burns MB, Ward T, Mars RAT, Adamowicz B, Lock EF, Kashyap PC, Knights D, Blekhman R. 2022. Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. Nat Microbiol 7:780–795.

55. McBurney MI, Davis C, Fraser CM, Schneeman BO, Huttenhower C, Verbeke K, Walter J, Latulippe ME. 2019. Establishing what constitutes a healthy human gut microbiome: State of the science, regulatory considerations, and future directions. J Nutr 149:1882–1895.

56. Cobo-López S, Gupta VK, Sung J, Guimerà R, Sales-Pardo M. 2022. Stochastic block models reveal a robust nested pattern in healthy human gut microbiomes. PNAS Nexus https://doi.org/10.1093/pnasnexus/pgac055.

57. Sung J, Hale V, Merkel AC, Kim P-J, Chia N. 2016. Metabolic modeling with Big Data and the gut microbiome. Appl Transl Genom 10:10–15.

58. Sung J, Kim S, Cabatbat JJT, Jang S, Jin Y-S, Jung GY, Chia N, Kim P-J. 2017. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nat Commun 8:15393.

59. Kim M, Sung J, Chia N. 2022. Resource-allocation constraint governs structure and function of microbial communities in metabolic modeling. Metab Eng 70:12–22.

60. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. 2013. NeSSM: A Next-generation Sequencing Simulator for metagenomics. PLoS One 8:e75448.