Systematically investigating and identifying unmodified bacteriocins in the human gut microbiome

Dengwei Zhang[1], Yuqi Shi[1], Junliang Zhang[1], Jing Liu[1], Gengfan Wu[1], Jian Zhang[1], Ying Gao[1], Yong-Xin Li[1, *]

[1] Department of Chemistry and The Swire Institute of Marine Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

[*] To whom correspondence may be addressed.

Email: yxpli@hku.hk

Abstract

Unmodified class II bacteriocins are naturally occurring antimicrobial peptides that are synthesized by bacteria. When produced by human gut bacteria, these bacteriocins not only protect against harmful pathogens, but also play a role in regulating host physiology. However, the limitations of current bioinformatic tools impede the comprehensive investigation of their biosynthesis in the human gut microbiome, leaving these natural products largely elusive. Here, we introduce IIBacFinder, an analysis pipeline dedicated to identifying unmodified bacteriocins, which outperforms existing tools. Through large-scale bioinformatic analysis of bacterial genomes and experimental validation of novel bacteriocin sequences, we demonstrate the widespread distribution of unmodified bacteriocins across the bacterial tree of life, which are generally niche-specific and enriched in the human microbiome. Analyzing over 200,000 bacterial reference genomes from the human gut microbiome, we reveal the vast potential of gut bacteria to produce diverse unmodified bacteriocins. Guided by meta-omics analysis, we chemically synthesized 26 hypothetical bacteriocins from primary gut commensal species such as *Bacteroides* and *Prevotella*, with 16 showing inhibitory activities against bacterial pathogens and human gut commensal bacteria. Our study uncovers the underappreciated biosynthetic potential of unmodified bacteriocins in the human gut microbiome, paving the way for elucidating their biological functions and relationship to human health.

Keywords

Bacteriocin; Antimicrobial peptides; Human microbiome; Genome mining; Omics analysis

Introduction

Bacteria are ubiquitously present on Earth, often found in a communal fashion where each species will compete for limited nutrients and space (*1*). These common antagonistic interactions are crucial not only to influence the composition of their members but also to maintain the community stability (*2*, *3*). The interference competition can be governed by various types of molecules, such as polyketides (*4*), protein toxins (*5*), non-ribosomally synthesized peptide (NRP) antimicrobials (*6*), and ribosomally synthesized antimicrobial peptides that are referred to as bacteriocins (*7*, *8*). Despite exhibiting incredible sequence and structural diversity, bacteriocins are usually categorized into two major classes: those that undergo posttranslational modifications by enzymatic tailoring (class I bacteriocins) and unmodified peptides (class II bacteriocins) (*9*). On the basis of modifications, class I bacteriocins can be further grouped into more than 40 subgroups such as lanthipeptide and lasso peptide, whereas class II bacteriocins can be generally divided into four subgroups (class IIa, YGNG-motif containing bacteriocins; class IIb, two-peptide bacteriocins; class IIc, leaderless bacteriocins; class IId, other linear bacteriocins) (*10*).

Due to its antimicrobial nature, bacteriocins have been regarded as promising antibiotic alternatives or adjuncts to contain the antimicrobial resistance threats of current antibiotics (*7*). Compared to class I bacteriocins, unmodified class II bacteriocins are simpler for their production in heterologous systems and chemical synthesis, thereby rendering it more practical for antimicrobial discovery at a large scale. Although a small amount of class II bacteriocins have been identified generally via traditional isolation techniques (Supplementary Table S1), they exhibit a variety of structures and biological activities, such as antifungal and antiviral activities (*11*, *12*). Therefore, class II bacteriocins are promising candidates for the discovery of peptidic antimicrobials. On the other hand, recent studies, although few, have expanded bacteriocin function from an antimicrobial nature to a medium to interplay with the host (*13*–*15*), further emphasizing the significance of studying host-associated class II bacteriocins, particularly those from the human gastrointestinal tract (GIT) where bacteriocins exist as crucial guardians to affect human health (*16*, *17*). For instance, A class IIb bacteriocin, plantaricin EF, can maintain epithelial barrier integrity and thereby alleviate the effects of obesogenic diets in the mouse model with a high-fat diet (*15*). Another case showed the immunomodulatory activity of a lactococcin 972-like class IId bacteriocin from infant gut-derived *Bifidobacterium longum* subsp. *infantis* (*18*). Given that the human microbiome contains 150 times more genes than the human genome (*19*), class II bacteriocins within the human microbiome represent an untapped treasure for antimicrobial discovery and play a crucial role in mediating human health. Although previous investigations on hundreds of human GIT reference genomes or human metagenomic data revealed limited biosynthetic potential of class II bacteriocin in human gut microbiome (*20*, *21*), the overall picture remains substantially incomplete in the context of > 200,000 reference genomes reconstructed from the human gut microbiome (*22*). Hence, there is a pressing need for a thorough investigation into class II bacteriocins in the human gut microbiome.

In the meta-omics era, with tremendous genomic data sets available, bacteriocin discovery has evolved from a traditional activity screening to a more prominent emphasis on genomics-guided discovery via predicting biosynthetic gene clusters (BGCs) (*9*). The BGCs of class I or class II bacteriocins typically include the precursor genes, and other associated genes such as transporter and peptidase (hereinafter referred to as context genes). Advanced by the rapid development of computational techniques, a range of dedicated bioinformatic tools have been in place for bacteriocin mining using distinct algorithms (Table 1). Nevertheless, none has an emphasis on identifying class II bacteriocins. Class II bacteriocins, despite being part of the antimicrobial peptides (AMPs) group that doesn't undergo post-translational modifications, require a distinct approach for their mining compared to the generic AMP prediction(*23*–*25*). This is because, while AMPs constitute the entire small open reading frames (smORFs), most mature class II bacteriocins invariably involve the removal of the leader peptide. Additionally, when constructing machine learning or deep learning models for bacteriocin prediction, some tools, such as NeuBI, mixed up modified and unmodified bacteriocins without considering the post-translational

modification nature of class I bacteriocins (*26–29*). The other two most widely used tools, BAGEL4 and antiSAMSH, also have limitations in identifying class II bacteriocins (*30*, *31*). BAGEL4 can detect class II bacteriocins by aligning query sequences against reference sequences of known class II bacteriocins using BLAST algorithm, which can only identify sequences similar to known ones; meanwhile, antiSAMSH includes few detection rules for class II bacteriocin identification due to its emphasis on detecting overall secondary metabolites (Table 1). Therefore, developing a more dedicated tool for class II bacteriocin mining is a prerequisite for interrogating them in the human gut microbiome.

Here, we commenced with the development of a specialized tool called IIBacFinder (Class II Bacteriocin Finder), which integrates the detection of class II bacteriocins through both precursor genes and context genes. Demonstrated by recovering known class II bacteriocins from the genomes of bacteriocin producers, IIBacFinder outdoes the two wildly used tools, BAGE4 and antiSMASH, in detecting class II bacteriocins. Alongside experimental validation, the systematic mining of ~260,000 bacterial genomes revealed the underappreciated biosynthetic potential of class II bacteriocins in the bacterial kingdom, particularly in Gram-negative bacteria. Those class II bacteriocins exhibited high niche-specificity and were enriched in the human microbiome. Leveraging the comprehensive analysis of ~280,000 reference genomes reconstructed from human gut metagenomes, we gained deep insights into the biosynthetic capacity and diversity of class II bacteriocins. Analyzing metagenomics and metatranscriptomics data of healthy human gut, we prioritized and chemically synthesized 26 predicted class II bacteriocins that were prevalent and abundant in healthy human gut microbiome. They showed diverse activities against common pathogens and commensal strains. Our work demonstrates the previously underappreciated biosynthetic potential of unmodified bacteriocins in the human gut microbiome, which might contribute to maintaining microbiome homeostasis and impacting human health.

Results

The development of IIBacFinder pipeline specializing in detecting unmodified bacteriocins

Several tools have been developed for bacteriocin identification with distinct mining strategies (Table 1), but they underperform in identifying class II bacteriocins. Specifically, some tools [i.e., RMSCNN (*29*), BaPreS (*26*), BPAGS (*27*)] identify motley bacteriocin precursor sequences, rendering it hard to determine modified and unmodified bacteriocins. Other two tools, BAGEL4 (*30*) and PRISM 4 (*32*), detect class II bacteriocins only based on precursor genes, which would constrain detection to similar sequences of known bacteriocins; while another tool, antiSMASH (*31*), includes a few detection rules due to its aims to identify overall secondary metabolites (Table 1). To circumvent these limitations, we set out to develop an improved tool dedicated to identifying class II bacteriocins by integrating the detection of precursor genes and context genes. Given that profile hidden Markov models (pHMMs) offer a straightforward and rapid detection method to identify homologs within a specific protein family (*33*, *34*), we constructed the pHMM library of precursor genes and context genes by curating known sequences and detecting their homologs from the UniRef100 database (Supplementary Figure 1A, Supplementary Table 1, 2). Leveraging the library of 1,153 pHMMs of precursor genes and 37 pHMMs of context genes (Supplementary Tables 3 - 5), we could comprehensively pinpoint class II bacteriocins by either directly detecting precursor genes or indirectly detecting novel bacteriocins genes through anchoring context genes (Figure 1A, Supplementary Figure 1B). We termed this pipeline IIBacFinder (Class II Bacteriocin Finder).

Following the detection of precursor genes, we sought to integrate the prediction of leader which is a marked difference between class II bacteriocins and generic AMPs. Two types of leaders, double-glycine leader and sec-dependent leader, are generally present in class II bacteriocins. To streamline the prediction of double-glycine leader, we adopted and tuned a RiPP leader detection tool, NLPPrecursor (*35*), for the cleavage site prediction. By employing 10-fold cross-validation, we determined that the modified algorithm could attain an accuracy of 97.98% when factoring in cleavage points that are ±3 amino acids (AAs) away from the actual prediction site (Supplementary Figure 2). On the other hand, we applied SignalP 6.0 (*36*) to predict the cleavage site for the sec-dependent leader. Finally, the predicted core peptides were compared to known AMP sequences from four AMP databases (*37–40*), to gain an insight into their sequence novelty (Figure 1A).

Upon the development of IIBacFinder, we next sought to evaluate its performance in identifying class II bacteriocins. We curated 137 bacterial genomes containing 205 complete precursor sequences of known class II bacteriocins, spanning different leader types and subclasses (Supplementary table 6). Among them, 191 sequences were unutilized when constructing the pHMM library of precursor genes, while 12 sequences were not because they were unintentionally excluded when building pHMMs. We then applied three tools (antiSMASH 6, BAGEL4, and IIBacFinder) to these genomes for detecting class II bacteriocins. Reassuringly, IIBacFinder could recover all bacteriocin precursor sequences (205/205, 100%), including 12 sequences that were literally new to IIBacFinder, while BAGEL4 and antiSMASH 6 detected 134 (65.4%) precursor sequences and 112 (54.6%) corresponding bacteriocin gene clusters, respectively (Figure 1B). Notably, their performance in identifying leaderless class IIc bacteriocins was significantly lower. The performance benchmarking demonstrated that IIBacFinder has a superior detection capability for unmodified class II bacteriocins compared to the other two commonly used tools.

Underappreciated biosynthetic capacity of class II bacteriocins in the bacterial kingdom

Having attested to IIBacFinder's remarkable ability to identify bacteriocins, we then applied it to 257,997 bacterial genomes from the NCBI RefSeq database (Supplementary table 6) (*41*), in order to assess the global biosynthetic capacity of unmodified bacteriocins in the bacterial kingdom. These

genomes spanned 56 phyla and 4,259 genera, corresponding to 96,325 Gram-positive and 161,472 Gram-negative bacteria. Using IIBacFinder, a total of 540,587 precursor sequences were identified from 133,806 genomes. Among them, 387,888 (71.75%) and 152,598 (28.23%) precursor sequences were detected from Gram-positive and Gram-negative bacteria, respectively. For comparison, antiSMASH 6 could detect 206,288 RiPP BGCs in 119,697 genomes, consisting of 108,962 (52.84%) and 97,265 (47.16%) BGCs from Gram-positive and Gram-negative bacteria; meanwhile, BAGEL4 detected 137,942 precursor sequences with 93.36% being from Gram-positive genomes, indicating BAGEL4 biased the prediction in Gram-positive bacteria in which most known bacteriocins were discovered (Supplementary Figure 3).

To reduce the probability of false prediction, we only included 481,654 high-confidence sequences for downstream analysis (Supplementary Figure 4). The length of core peptides was approximately 40 amino acids for double-glycine and leaderless bacteriocins, but around 70 amino acids for sec-dependent bacteriocins (Supplementary Figure 5A). When compared to known AMPs deposited in four databases (i.e., APD3, DRAMP, DBAASP, and dbAMP2), they showed less homology, with most being less than 40% identity (Supplementary Figure 5B). Meanwhile, class II bacteriocins had higher median length than AMPs. By analyzing two physicochemical features, which play a pivotal role in elucidating the activity of AMPs, we found that double-glycine or leaderless bacteriocins exhibited a lower number of electronic charges than AMPs (Mann–Whitney U test, $P < 0.001$) but had higher GRAVY index scores, suggesting they are likely to be hydrophobic. In contrast, significantly lower GRAVY scores than AMPs implied that sec-dependent bacteriocins were more likely to be hydrophilic (Supplementary Figure 5C).

To assess the genetically encoded biochemical diversity, we conducted a rarefaction analysis on 8,610 precursor clusters that were categorized from 481,654 sequences using a 50% identity threshold. The curve astonishedly showed that Gram-negative bacteria had a higher biosynthetic potential than Gram-positive bacteria (Figure 2A). We then decorated the GTDB bacterial tree with predicted precursor sequences to reveal the biosynthetic capacity at the phylum level. Through this examination, we found that the bacteriocin biosynthetic capacity stood out in several most sequenced phyla, such as Firmicutes (13,401 unique sequences), Proteobacteria (13,104), Bacteroidota (4,494), and Actinobacteriota (3,801) (Figure 2B). Among them, the bacteria from Firmicutes were the most prolific producers, with the average number of precursors per genome being 4.97, in accordance with the fact that most known bacteriocins have been discovered in this phylum. When down to the genus level, *Streptococcus*, as expected, topped the list of the bacterial tree of life. Other Gram-negative genera, such as the soil bacteria Rhizobium and Bradyrhizobium, also displayed promising biosynthetic capacity (Figure 2C). The overall survey indicated the previously underappreciated bacteriocin biosynthetic capacity in the bacterial kingdom, particularly in Gram-negative bacteria.

To further demonstrate the prediction reliability, we arbitrarily selected and chemically synthesized 28 hypothetical bacteriocins (HBs) with the length of deduced core peptides < 50 amino acids for experimental validation, but 7 failed (Supplementary Table S8). The remaining 21 sequences consisted of 13 double-glycine sequences, 5 sec-dependent sequences, and 3 leaderless sequences from seven phyla (i.e., Firmicutes, Firmicutes_A, Firmicutes_C, Cyanobacteria, Actinobacteria, Bacteroidota, Proteobacteria). Moreover, only HB17 could be detected by BAGEL4, while six HBs with double-glycine precursors could be captured by antiSMASH 6 as RiPP-like BGCs (Figure 2D, Supplementary Figure 6). We examined their antibacterial potential against 19 Gram-positive and 9 Gram-negative indicator strains from four phyla at 100 μg/ml concentration (Figure 2D). The assay showed that 17 of 21 HBs could inhibit the growth of at least one indicator strain (inhibition ratio > 25%, Dunnett's test $P < 0.05$). The active HBs primarily targeted the Gram-positive bacteria from phyla Firmicutes and Actinobacteria and barely inhibited Gram-negative bacteria. They displayed either narrow-spectrum activity (e.g., HB1 and HB2) or broad-spectrum activity (e.g., HB7). Among them, 9 HBs (three from Gram-negative bacteria and six from Gram-positive bacteria) exhibited a robust activity with over 75%

inhibition ratio (Figure 2D). Those results indicated that IIBacFinder could identify authentic novel unmodified bacteriocins the other two tools could not.

Class II bacteriocins exhibit niche-specific enrichment in host-associated bacteria

The global inspection of class II bacteriocins revealed their wide distribution among diverse bacterial phyla, which might account for the previous concept that bacteriocin producers are widespread in natural environments (*42*). To understand their biosynthetic potential in diverse ecosystems, we applied IIBacFinder to ~50,000 metagenome-assembled genomes (MAGs) recovered from metagenomes of diverse habitats, including host-associated environments (primarily from human), aquatic environments, engineered environments, and terrestrial environments (*43*). We identified 4,956 unique high-confidence precursor sequences from the genomic catalog of earth's microbiomes, half of which (2,482 sequences) were from host-associated MAGs, followed by the engineered, aquatic and terrestrial environments (Figure 3A). Notably, the bacteriocins identified from human contributed most to the host-associated bacteriocin (Figure 3B). Interestingly, we observed that class II bacteriocins exhibited high niche-specify, with >90 percent of unmodified bacteriocin sequences found in one specific ecosystem (Figure 3C).

Although the fact that the largest number of bacteriocins were detected in host-associated MAGs is informative, this might be skewed by the uneven number of reconstructed MAGs. Therefore, we then adopted rarefaction analyses to address this bias (Figure 3D). The rarefaction curve still indicated host-associated MAGs presented the greatest biosynthetic potential than other ecosystems, with being around three times over aquatic and engineered environments. The analysis of environmentally diverse bacteriocins implied class II bacteriocins' potential to interact closely with community members in the host. They may also be utilized to interact with the host itself, particularly in the case of human hosts (*18*).

Human gut microbiota has the potential to produce diverse health-beneficial unmodified bacteriocins

Realizing that class II bacteriocins are widely enriched in the human microbiome, we then put our emphasis on the unmodified bacteriocin production in the human gut, where numerous microbes mediate human health through producing various metabolites (*44*). We, therefore, applied IIBacFinder to ~280,000 bacterial genomes recovered from the human gut microbiome (*22*), resulting in 82,806 high-confidence precursor sequences, with 6,238 being unique (100% identity). More than half of them were double-glycine precursors (4,186, 67.10%), followed by sec-dependent leader precursors (765, 12.26%) and leaderless precursors (610, 9.78%). At the class level, Clostridia, Bacilli, and Bacteroidia were the most abundant producers. In accordance with the global bacterial biosynthetic potential, *Streptococcus* could produce the most diverse bacteriocins, although it does not predominate in human gut microbiota. Moreover, other primary human gut bacteria (e.g., *Prevotella*, *Faecalibacterium*, and *Bacteroides*) also have high biosynthetic potential (Figure 4A).

Although this analysis indicated the high biosynthetic potential of human gut bacteria, only a few class II bacteriocins have been identified from gut microbiota yet (*16*). We, thus, wondered whether these predicted unmodified bacteriocins were chemically distinct from known ones. We thus compared the 6,238 unique sequences with 298 known bacteriocin sequences at varying identity thresholds of 50% - 90% (Figure 4B, Supplementary Figure 7). To our surprise, we found 379 (6.07%) and 143 (2.29%) homologous sequences at the identity of 50% and 90%, respectively, indicating that gut microbiota could produce similar bacteriocin products, but the majority remained uncharacterized. Among the most diverse were UviB homologies produced by *Clostridium perfringen* (*45*). Some bacteriocins have established functions in mediating host health, for example, a broad-spectrum class II bacteriocin, ABP118, can protect mice against infection with pathogen *L. monocytogenes* (*46*). These findings

suggest that unmodified bacteriocins with protective functions may have a wide range of health-promoting effects.

Recent advancements in functional metagenomics have facilitated the discovery of bioactive RiPPs in the largely uncharted human microbiota, and have established links to their potential physiological roles (*47*). We next looked at evidence of their transcription in 724 metatranscriptomes of healthy individuals from three cohorts to look for insight into their possible physiological roles. We found 1,966 (31.52%) sequences were actively transcribed in at least one sample, but more sequences could be detected with an increase in sample size as indicated by rarefaction analysis (Figure 4C). Although homologous sequences of known bacteriocins were rarely detected in metatranscriptomes, some novel bacteriocin sequences were widely transcribed in healthy individuals (Figure 4D). Furthermore, multiple distinct bacteriocin sequences were found to be transcribed in individuals, with a median number of 139, 88, and 122 in three cohorts (Figure 4E). Hundreds of transcribed bacteriocins might create a protective barrier for pathogen invasion in a synergistic manner (*48*).

Omics analysis prioritizes putatively bio-functional unmodified bacteriocins in the healthy gut microbiome for experimental characterization

Inspecting the class II bacteriocins from human gut metagenome-assembled genomes (MAGs) allowed us to appreciate their significant biosynthetic potential in the human gut. Next, we set out to experimentally characterize the bioactive class II bacteriocins that might play a crucial role in maintaining healthy human gut homeostasis. As the species abundance and sequence depth would skew the genome assembly from metagenomics data, we prioritized putatively bioactive bacteriocin using mapping-based analysis of metagenomic and metatranscriptomic data instead of those that were the most abundant in gut MAGs. To do so, we first created a reference database of class II bacteriocin by applying IIBacFinder to ~700,000 bacterial genomes from multiple sources, including RefSeq (*41*), human bodies (*22*, *49*, *50*), earth environment (*43*), ocean (*51*), mouse and ruminant gastrointestinal tract (*52*, *53*). Following the construction of this database encompassing 32,735 non-redundant high-confidence class II bacteriocin sequences, we examined their profiles by re-visiting 1,901 metagenomic and 724 metatranscriptomic datasets from healthy individuals (Figure 5A, Supplementary Table 9). To remove the redundancy, we computed their abundance at the cluster level (80% identity and 80% coverage) rather than at the individual level. Given the fact that bioactive bacteriocin could inhibit the growth of other bacteria and thus might regulate the overall bacterial community, we ranked them by considering (1) their prevalence and (2) abundance, (3) negative association with bacterial species, and (4) association with bacterial community diversity assessed by Shannon index (Figure 5A). This analysis allowed us to prioritize bioactive bacteriocins in the human gut microbiome (Supplementary Table 10).

Finally, on the basis of the ranking analysis, we scrutinized and selected 40 clusters of human gut microbiota-derived hypothetical bacteriocins (HGM-HBs) with a length of core peptide < 50 amino acids for experimental validation. We attempted to chemically synthesize the core peptides from their representative sequences; however, 14 of these attempts were unsuccessful. Consequently, we had 26 HGM-HBs available for examination (Supplementary Table 11). All of them were identified from human gut-derived bacterial genomes. None was detected by BAGEL 4, while eight were identified by antiSMASH 6 (seven within RiPP-like BGCs and one within NRPS BGC). Meanwhile, HGM-HP11 was identical to bacteroidetocin A, but we did not include the intrinsic disulfide bonds for chemical synthesis as the putative disulfide bonds in other HGM-HPs were not predictable. Likewise, we initially examined their antibacterial activity against 21 indicator strains encompassing 13 clinical pathogens and 8 gut commensal bacteria at a concentration of 100 μg/ml. Under a cutoff of inhibition ratio > 25% and $P < 0.05$, 16 peptides (70%) were found to be active against, generally, Gram-positive bacteria. Notably, 8 HGM-HBs, including HGM-HP11, effectively inhibited at least one indicator strain with a >

75% inhibition ratio. The assay indicated that the human gut microbiota-derived class II bacteriocin could inhibit the growth of clinical pathogens and gut commensals.

Discussion

The human gut microbiota is a rich source of bacteriocins that might combat pathogenic bacteria invasion and participate in human physiology (*16*, *17*). Although a few efforts have been made to decipher their distribution in human gut microbiome (*20*, *21*), the unmodified bacteriocins remain incompletely characterized. In this study, we conducted the first large-scale survey, to our best knowledge, on unmodified class II bacteriocins in the bacterial kingdom and human gut microbiome by applying an in-house bioinformatic tool. Through systematic investigation in combination with experimental validation, we demonstrated novel class II bacteriocins throughout the tree of bacterial life as well as in the human gut microbiome. These investigations will not only disclose that the class II bacteriocins are potentially an unprecedented resource for antimicrobial discovery but also act as a crucial pathway for unraveling bacteriocins-mediated interactions between the microbiome and human host.

To enhance the detection of class II bacteriocins, we have developed a specialized software pipeline that identifies them by scanning pHMMs of signature genes. While this tool outperforms existing software and is capable of identifying novel bacteriocins, it may still produce false predictions, particularly when detecting new bacteriocins based on context genes. Given that only a small number of class II bacteriocins have been characterized, our current limited understanding can make it challenging to distinguish genuine bacteriocin sequences from false predictions. Additionally, while we can identify novel bacteriocins through anchoring context genes, we may still miss other unknown bacteriocins that fall outside our current understanding of bacteriocin biosynthetic logic. Recent studies have discovered and experimentally validated numerous AMPs from the human gut microbiome (*24*, *25*), implying the underappreciated potential of antagonistic AMPs released by the human gut microbiota. Although whether these AMPs are authentically produced and whether they undergo post-translational modification or leader removal as does bacteriocin biosynthesis remain questioned, they could be potential novel bacteriocins. The availability of tremendous genome data in the meta-omics era has unprecedentedly advanced bacteriocin mining in an untargeted manner, leading to biological characterization lagging far behind (*9*). Rather than sole prediction, bacteriocin discovery should combine ‘wet-laboratory’ demonstration in future studies. Particularly for unmodified bacteriocins, a large-scale identification could be achieved by cell-free protein synthesis systems or robotic automation systems; two techniques have been successfully applied to AMP identification and modified bacteriocin characterization (*54*–*56*). Hence, we believe that future advances in techniques can accelerate the discovery of bacteriocin, deepening our accurate understanding of their biological mechanism and functions.

In this study, we chemically synthesized predicted mature peptides of unmodified bacteriocins, which, however, might not represent their mature form in nature. Although not underwent post-translational modifications, unmodified bacteriocins might incorporate other intramolecular features. The most common is intramolecular disulfide bonds, which are prevalent in various unmodified bacteriocins such as class IIa bacteriocin and defensin-like bacteriocins. Nevertheless, the disulfide bonds seem to be crucial in stabilizing three-dimensional structure rather than function (*10*). For leaderless bacteriocins, the presence of formylated methionine at the N-terminal seems a common feature, which might not influence their activities either (*58*). However, as exemplified by Lactococcin B (class IId), the reduced state of cysteine residue could determine the bacteriocin activity (*59*). Additionally, like AMPs (*48*), synergies of bacteriocins might be frequent in the human gut, considering the fact that hundreds of bacteriocins were found to be actively transcribed in individuals and many bacteriocin gene clusters often harbor more than one precursor gene. Consequently, the in vivo functions of synthetic bacteriocins might be distinct from our in vitro characterization. Despite this limitation, our study found that the gut microbiota could produce various class II bacteriocins that might interact with the microbial community or human host.

Although we have partially determined the antibacterial activities of synthetic bacteriocins from the human gut microbiome, their functions in vivo might be beyond the antibacterial nature, because unmodified bacteriocins can exhibit other antimicrobial activities, which include but are not limited to antibiofilm, antifungal, and antiviral activities. For example, enterocin A, a class IIa bacteriocin, can inhibit the biofilm formation of three pathogens (*Staphylococcus aureus*, *Listeria monocytogenes* and *Acinetobacter baumannii*) (*60*). This feature can conduce to the decolonization of pathogens in the human gut. Sharma and Srivastava reported an antifungal of two-peptides class IIb bacteriocins, plantaricins EF and JK (*11*). Given that the fungal community represents a crucial component of human health (*61*), these antifungal bacteriocins might regulate the fungal community to impact human health indirectly. Another incredible activity is antiviral, exemplified by enterocin DD14 which can combat the infection of eukaryotic cells with enveloped viruses such as SARS-CoV-2 (*12*), despite the mechanism remaining elusive. Another class IIa bacteriocin, enterocin CRL35, is also active against herpesviruses (*62*), indicating antiviral activity of class II bacteriocins is not orphan.

On top of antimicrobial activities, unmodified bacteriocins can also directly interplay with the human host, exhibiting various activities such as anticancer and immunomodulation. Along with the antibiofilm activities of enterocin A, Ankaiah and colleagues also demonstrated its anticancer activity against human colon, gastric, and cervical cancer cells while maintaining no cytotoxicity toward normal intestinal epithelial cells (*60*). Inspired by this case, we can legitimately argue that those unmodified bacteriocins in the human gut might also play essential roles in controlling colon cancer development. The immunomodulatory activity has also been found in various class II bacteriocins, such as enterocin DD14, BacSp22 and acidocin A (*63–65*). More intriguingly, the two-peptide bacteriocin, plantaricin EF, could increase the production of tight junction protein in intestinal epithelial, maintaining epithelial barrier integrity (*15*). Although most of these bacteriocins are not produced by human gut bacteria, our systematic analysis reveals that human gut bacteria contain homologous precursor genes of these bacteriocins, indicating biosynthetic potential (Figure 4, Supplementary Figure 7). These findings imply possible biological roles of unmodified bacteriocins in promoting human health.

In conclusion, we provide a comprehensive overview of unmodified bacteriocins in the human gut microbiome. However, our analysis may not be exhaustive due to the current limitations in our understanding of bacteriocin biosynthetic logic. The precise biological functions of unmodified bacteriocins in the human gut microbiome remain under-characterized and warrant further investigation in human gut or rodent models. Despite these limitations, our study reveals the chemical repertoire of unmodified bacteriocins encoded by the human gut microbiome, paving the way for antimicrobial discovery from the untapped human microbiome and deepening our understanding of its role in human health and disease.

Methods

Development of IIBacFinder

Profile hidden Markov models (pHMMs) construction

A total of 286 non-redundant class II bacteriocin sequences (Supplementary Table S1) were curated from previous studies. To comprehensively identify their homologous sequences, we used jackhammer (*66*) to iteratively align them against the UniProt UniRef100 database (December 2022, comprising 323,519,324 sequences) with the parameters "-N 5 -T 20 --domT 20 --incT 20 --incdomT 20". Putative homologous sequences longer than 150 amino acids were considered false positives and discarded. We manually collated the homologous sequences, removing those with specific functions unrelated to antimicrobial activity or those with unknown functions, resulting in 297,986 remaining sequences. Combined with the 286 known bacteriocin sequences, these class II bacteriocin sequences were deduplicated to avoid significant similarity when constructing pHMMs using CD-HIT with the parameters "-n 2 -p 1 -c 0.8 -d 200 -M 5000 -aL 0.6 -g 1" (*67*). The deduplicated sequences were then clustered at a 50% identity threshold using CD-HIT, resulting in 2,884 clusters with low sequence similarity among them. One cluster was removed due to containing a sequence with ambiguous amino acids. We used MAFFT (*68*) for multiple sequence alignment within each cluster. Following this, we used trimAl (*69*) to automatically remove poorly aligned regions with the model "-automated1". All aligned sequences were then used to build the pHMM for each cluster using hmmbuild (*66*). This process resulted in 2,883 class II bacteriocin-related pHMMs. Additionally, we collected pHMMs from the PFAM (*70*) and NCBI databases, obtaining 16 and 14 pHMMs, respectively. To avoid redundancy between the 30 available pHMMs and the 2,883 self-constructed pHMMs, we used hmmscan (*66*) to annotate the seed sequences of the 2,883 clusters against the PFAM and NCBI databases. Clusters whose seed sequences were entirely annotated by either PFAM or NCBI domains were discarded, retaining a total of 1,123 clusters. Consequently, we obtained 1,153 precursor-related pHMMs (Supplementary Table S3).

Subsequently, we curated 29 immunity protein sequences (Supplementary Table S2) associated with class II bacteriocins from previous reports. To build the pHMMs for these immunity proteins, we first aligned the sequences against the UniRef100 database using Diamond with the parameter "--id 40 --masking 0", retaining homologous sequences with identities greater than 40%. These homologous sequences were processed similarly to the precursor sequences, resulting in 29 pHMMs for class II bacteriocin immunity proteins. Additionally, we included eight publicly available pHMMs related to context genes, including protease genes, inducer genes, and accessory genes. Finally, we obtained 37 context gene-related pHMMs (Supplementary Table S4).

Precursor gene identification

IIBacFinder takes a DNA sequence in FASTA format as input, which is subjected to open reading frame (ORF) annotation using Prodigal-short (*71*). All ORFs are queried against a library of 1,153 precursor-related pHMMs and 37 context gene-related pHMMs using hmmsearch (*66*). Putative precursor genes are detected using both precursor-based and context gene-based rules in parallel. For the precursor-based rule, significant hits identified by the 1,153 precursor-related pHMMs are initially considered putative precursor genes if the hit sequence is less than 150 amino acids in length. The region encompassing ±8 genes around the identified precursor genes is defined as the biosynthetic gene cluster (BGC). For the context gene-based rule, significant hits identified by the 37 context gene-related pHMMs are considered associated with class II bacteriocin biosynthesis. The eight genes upstream and downstream of the identified context genes are included as the putative BGCs. BGC regions lacking a transporter are discarded. Subsequently, the BGC region is examined for the presence of proteins containing peptidase_C39 (PF03412) and LANC_like (PF05147) domains. BGCs containing both

domains are discarded as they are likely to be modified lantibiotics. All small ORFs within the BGC, with lengths less than 150 AA, are retrieved and subjected to antimicrobial peptide annotation using both amPEP (*72*) and ampir (*72*). Small ORFs predicted as antimicrobial peptides (AMPs) with a probability threshold of 0.8 by either tool are considered putative precursor genes. The putative precursor genes detected by either the precursor-based or context gene-based rules are then queried against the Pfam and NCBI databases using hmmscan (*66*). Sequences containing domains unrelated to class II bacteriocins are removed (Supplementary Table 5). The remaining sequences are regarded as putative precursors of class II bacteriocins. Finally, BGC regions detected by both the precursor-based and context gene-based rules are combined.

Leader type prediction

The leader type of precursor genes detected by the precursor-based rule is considered identical to that of known homologous bacteriocins (Supplementary Table 3). For precursor genes detected by the context gene-based rule, we annotate the leader types based on the context genes within the BGC region (Supplementary Table 4). For example, if the BGC contains an ORF with a peptidase_C39 domain (PF03412), the leader type of the precursor genes is considered to be double-glycine.

Leader peptide prediction

To predict the cleavage site of the double-glycine leader, we tuned a deep learning model, NLPPrecursor, specifically designed for the prediction of RiPP precursor cleavage sites (*35*). The protein sequences were tokenized using a vocabulary of 24 tokens, including a start and end tag, twenty proteinogenic amino acids, an ambiguous amino acid, and a padding token. Unlike RiPP precursors, which may contain leader and follower sequences at the N-terminal and C-terminal, respectively, double-glycine bacteriocin precursors only have a leader sequence. Therefore, we modified the sequence labels in the model. The allowed transitions of protein sequences within the model were: (start-before), (before-before), (before-propeptide), (propeptide-propeptide), (propeptide-stop), (stop-pad), and (pad-pad). All sequences were unified to a final length of 150 by adding pad tokens at the end.

To train the deep learning model, we created a dataset of double-glycine class II bacteriocins. Since only 142 known class II bacteriocin precursors contain a double-glycine leader, we collected their homologous sequences from the UniProt UniRef100 database to expand the training dataset. DIAMOND (*35*) was used to query the 142 sequences against the UniRef100 database with the parameters "--id 50 --query-cover 80 --masking 0", resulting in 1,224 homologs. To reduce sequence bias in training the prediction model, the 1,336 double-glycine sequences were deduplicated at a sequence identity threshold of 90% using CD-HIT (*67*). This deduplication process yielded 439 non-redundant sequences. Subsequently, we used MAFFT (*68*) with the parameters "--maxiterate 1000 --genafpair" to conduct multiple sequence alignments and manually check whether the homologous sequences contained a conserved double-glycine cleavage site. Finally, 412 class II bacteriocin precursor sequences with a double-glycine leader were retained for model training with 100 epochs (Supplementary Figure 2).

For predicting the Sec-dependent leader, we adopted the well-established tool SignalP 6.0 (*36*), which predicts the signal peptides of proteins secreted by the secretory (Sec) pathway, using default parameters.

Physicochemical property calculation

Following the prediction of the leader peptide, we compute common descriptors for the predicted core peptide using a pure-Python package, peptides.py (https://peptides.readthedocs.io/en/stable/index.html). The physicochemical properties include the aliphatic index, Boman index, isoelectric point, molecular weight, instability index, and net charge at a pH of 7.

Comparison with antimicrobial peptides (AMPs)

To gain insight into sequence novelty, predicted core peptides of class II bacteriocins were compared with known AMPs. AMP sequences were retrieved from four publicly available databases in April 2024: APD3 (*37*), dbAMP (*38*), DRAMP (*39*), and DBAASP (*40*). Sequences less than 10 amino acids or containing ambiguous amino acids were discarded, resulting in 51,692 AMP sequences, of which 21,403 were unique (100% identity). The predicted core peptides were queried against the AMP dataset using DIAMOND (*40*) with the parameters "--ultra-sensitive --max-target-seqs 1 --id 20 --query-cover 20 --masking 0". The best alignment hit was retained.

Prediction confidence assignment

Given that bioinformatic predictions can inevitably yield false positives, we added a confidence label to indicate whether the predicted class II bacteriocins are likely authentic or potentially false (Supplementary Figure 4). We considered several factors in determining this confidence level: the presence of a ribosome binding site (RBS), the completeness of the predicted ORF of the precursor gene, the detection rules (i.e., bacteriocin detection based on one rule or both rules), the leader peptide prediction, and the length of the predicted core peptide. The decision tree in Supplementary Figure 4 illustrates the classification into high-confidence or low-confidence categories. It is important to note that the inferred confidence level does not confirm whether the predicted bacteriocins are authentic or false without experimental validation; it only suggests a likelihood of the prediction's accuracy.

Performance benchmarking on identifying class II bacteriocins

To evaluate the performance of IIBacFinder, BAGEL4 (*40*), and antiSMASH 6 (*31*) in detecting class II bacteriocins, we applied these tools to the genomes of known bacteriocin producers. Since most producers' genomes have not been documented, we first attempted to locate known class II bacteriocins in sequenced bacterial genomes. The ORFs of 257,997 bacterial genomes from the RefSeq database were identified using Prodigal-short (*71*) with default parameters. These ORFs were then searched against 286 non-redundant known precursor sequences using MMseqs2 with the parameter "easy-search --max-seqs 1 --min-seq-id 0.95" (*73*). A best hit with an identity > 95% and coverage > 95% was considered a known precursor sequence. Consequently, we obtained 137 genomes containing 205 complete and non-redundant precursor sequences of known class II bacteriocins.

We then applied the three prediction tools to these genomes using default parameters to assess their ability to identify class II bacteriocins. Notably, both IIBacFinder and BAGEL4 can pinpoint the precursor sequences, while antiSMASH 6 preferentially detects the biosynthetic region rather than the precursor sequences. Accordingly, if the precursor sequence identified by either IIBacFinder or BAGEL4 was within the BGC region detected by antiSMASH 6, we considered it detected by antiSMASH.

Phylogenetic tree construction

The multiple sequence alignment of 286 non-redundant known precursor sequences was achieved by MAFFT v7.508 (*68*) with the parameter of "--maxiterate 1000 --localpair", which was further used to infer the phylogenetic tree. IQ-TREE version 2.2.0.3 (*74*) was adapted for constructing maximum likelihood (ML) phylogenetic trees, with 1,000 ultrafast bootstrap replicates. The best-fit model was chosen as VT+F+I+I+R4 identified by the in-built ModelFinder (*75*). The inferred phylogeny was visualized and annotated using iTOL (*76*).

Analysis of class II bacteriocins in bacterial genomes from multiple datasets

We used IIBacFinder to predict the biosynthetic potential of class II bacteriocins in bacterial genomes from eight datasets, consisting of the RefSeq database (RefSeq, n = 257,997), human microbiome (SGB,

n = 126,702), the Unified Human Gastrointestinal Genome (UHGG, n = 285,835), the Human Reference Gut Microbiome (HRGM, n = 29,035), global earth microbiome (GEM, n = 49,466), ocean microbiome (Ocean, n = 21,592), mouse (MGBC, n = 26,640) and ruminant gut microbiome (RGM, n = 10,211) (Supplementary Table S7). To confirm and unify the taxonomical annotation, GTDB-Tk v2.1.1 (*77*) was adopted for taxonomical annotation against GTDB-Tk reference data version r207 (*78*). Genomes not belonging to bacteria were discarded.

Rarefaction analysis

To assess the bacteriocin biosynthetic potential of bacterial genomes from the RefSeq database, R package vegan (*79*) was used. CD-HIT (*67*) was used to group 481,654 high-confidence sequences from RefSeq genomes into 8,610 clusters at a 50% identity threshold The accumulations of bacteriocin clusters detected in Gram-positive and Gram-negative bacteria were computed using *specaccum* function with "random" method.

For the biosynthetic potential in Earth's microbiome, rarefaction analyses were conducted using the iNEXT R package (*80*). We adopted CD-HIT (*67*) for deduplicating 10,581 high-confidence sequences from Earth's microbiome to 4,956 unique sequences at a 100% identity threshold. A unique sequence presence/absence table (sequence versus strain matrix) was constructed for each group considered. The type of input data was set as 'incidence-raw' data in the iNEXT main function. Besides, knots of 2,000 and an endpoint of 50,000 were adopted. The number of bootstrap replications is 50 by default.

Metagenomics and metatranscriptomics analysis

The raw metagenomic sequencing reads from five cohorts and the raw metatranscriptomic data from three cohorts were acquired from NCBI SRA (Sequence Read Archive) under respective project accession numbers (Supplementary Table 9). Only the sequencing data of healthy individuals was retained for analysis. Fastp v0.21.1 (*81*) with default parameters was used to detect and remove low-quality sequencing reads. The sequencing reads belonging to the human host were detected and discarded by kneaddata (https://github.com/biobakery/kneaddata), through searching against the human reference genome (GRCh38.p13) from GENCODE (*82*). The high-quality sequencing reads retained were subjected to downstream analysis.

Bacteriocin precursor profiling

Before profiling bacteriocin precursors in metagenomic and metatranscriptomic data, we constructed a reference dataset of class II bacteriocin precursor sequences. IIBacFinder identified 645,318 high-confidence precursor sequences in bacterial genomes from eight databases. To remove duplicates and reduce computational load, we deduplicated these sequences at a threshold of 95% identity and 95% coverage using CD-HIT (*67*), resulting in a reference precursor dataset of 32,735 sequences. This reference dataset was then queried by high-quality sequencing reads from metagenomic and metatranscriptomic data using PALADIN software (*83*) to profile the abundance of bacteriocin precursors in each sample. We used SAMtools (*84*) to calculate their abundance, normalizing the results using transcripts per kilobase million (TPM). To reduce inaccuracies in abundance calculation caused by highly similar sequences in the reference dataset, we grouped the sequences into clusters with a threshold of 80% identity and 80% coverage using CD-HIT (*67*), and computed their abundance at the cluster level rather than the individual sequence level.

Taxonomic profiling, community diversity, and correlation analysis

Metagenomic taxonomic profiling was conducted using MetaPhlAn version 4.0.6 (*85*). The relative abundance of species was used to represent the abundance of bacterial species in the community. The alpha diversity (Shannon index) of each sample was calculated with the R package vegan v2.6.4 (*79*).

Given that bioactive bacteriocins could inhibit the growth of other bacteria and thus regulate the overall bacterial community, we computed the Spearman correlation between taxonomic composition (at the species level) or Shannon diversity and precursor clusters using the corAndPvalue function from the R package WGCNA (*86*). The p-values were then adjusted using the "BH" method to control the false discovery rate in multiple comparisons (*87*). Considering that bioactive bacteriocins would inhibit the growth of specific bacterial species, we retained only the significantly negative associations (*rho* < -0.3 and adjusted $P < 0.05$).

Bacteriocin precursor prioritization in metagenomic and metatranscriptomic data

To infer the importance of class II bacteriocins in metagenomic and metatranscriptomic data, we contrived a scheme to assign ranking scores for each profiled precursor clusters, through considering the ecological importance (prevalence and abundance of bacteriocins) and functional importance (association between bacterial species and bacteriocins and between Shannon diversity and bacteriocins). Mathematically, the ecological importance is scored as:

$$RS_eco = \frac{1}{\frac{1}{R_p} \times 0.5 + \frac{1}{R_a} \times 0.5}$$

Where RS_eco indicates the ranking score (ecological importance) of a particular bacteriocin within a cohort. R_p and R_a represent the ranking percentiles based on prevalence and mean abundance (largest to smallest), respectively. A bacteriocin with high prevalence and high abundance would give rise to a larger RS_eco.

The functional importance is mathematically scored as:

$$RS_fun = \frac{RS_{cor} + RS_{reg}}{2}$$

Where RS_fun is the ranking score (functional importance) of a particular bacteriocin within a cohort. RS_{cor} indicates the ranking percentiles based on the negative correlation between bacteriocins and bacterial species, while RS_{reg} is the ranking percentiles based on the correlation between bacteriocins and the Shannon diversity index.

The overall ranking score of a bacteriocin in metagenomic and metatranscriptomic data is the average ranking score of ecological and functional importance in all cohorts, mathematically represented as:

$$RS = \frac{\sum_{i=1}^{n} RS_eco + \sum_{i=1}^{n} RS_fun}{n}$$

Where n represents the number of metagenomic and metatranscriptomic datasets. A high score of RS implies the great importance of the bacteriocin in the gut microbiome.

After prioritizing the bacteriocins in the human gut microbiome, we selected bacteriocin clusters to validate their antibacterial activities based on the ranking scores (RS) and the following criteria: (1) The hypothetical bacteriocin cluster should include $\geqslant$ 5 precursor sequences, as sporadically identified bacteriocins are likely to be false predictions. (2) The cleavage site for double-glycine bacteriocins should be a typical "GG" rather than alternative sites such as "GA" or "GS" to reduce false positives. (3) The length of the deduced core peptide should be less than 50 amino acids due to cost considerations and the feasibility of chemical synthesis. (4) The bacteriocin sequences should not be annotated as peptides with functions other than antimicrobials when queried against the NCBI database. Based on these criteria, we chose to synthesize and test the representative sequences, assigned by CD-HIT (*67*), from 40 bacteriocin clusters.

Three-dimensional peptide structure prediction

ColabFold was used to predict the three-dimensional structure of peptides, with a parameter of "--amber --templates --num-recycle 3 --use-gpu-relax --model-type auto" (*88*). The model with the highest relaxed rank was retained and visualized using PyMol (*89*).

Peptide synthesis

The deducted core peptides of predicted bacteriocins were chemically synthesized by solid-phase peptide synthesis by Sangon Biotech (Shanghai, China). Their molecular weights were confirmed by mass spectrometry, and their required purity was greater than 90% determined by high-performance liquid chromatography. The synthesized peptide powder was stored at − 80 °C and dissolved in Dimethyl sulfoxide (DMSO) to 50 mg/mL upon use.

Microbial strains and culturing

A total of 43 bacterial strains are included in this study and are detailed in Supplementary Table 12 regarding the strains and growth conditions. The following media were used: Brain heart infusion medium (BHI); Tryptic Soy Broth with supplement of 0.3% yeast extract (TSB-YE); De Man–Rogosa–Sharpe medium (MRS); Peptone Yeast Glucose medium (PYG); Reinforced Clostridium Medium (RCM);Gifu Anaerobic Broth (GAM); Bifidobacterium Medium (BM); M17 broth. Bacterial strains were grown and streaked on the agar plate, and incubated under respective optimal growth conditions. Individual colonies were then inoculated into 5 ml corresponding culture broth. Once the bacterial culture reached sufficient density, 500 µl culture bacterial suspension was supplemented with 500 µl 50% glycerol to generate liquid glycerol stocks and stored at −80 °C.

Antibacterial activity screening

The bacterial inhibition experiment was adopted from a previous study (*24*). All bacterial indicators were revived on agar medium under respective growth conditions (Supplementary Table 12). The individual colonies were picked into broth medium and incubated for one or two days. The bacterial suspension was diluted to a particular concentration [optical density at 600 nm (OD_{600}) = 0.1] and then diluted 100 or 1,000 times for the antibacterial activity screening (Supplementary Table 12). Three experimental groups were set: (1) blank control group, 100 µL of broth culture; (2) bacterial control group, 50 µL of broth culture and 50 µL of bacterial solution with DMSO (equivalent amount to the solvent of peptides); (3) experimental group, 50 µL of broth culture and 50 µL of bacterial solution with 100 µg/mL peptide. Experiments were conducted on flat bottom 96-well plates. The bacterial growth was measured with OD_{600} value after statically incubating for 24h or 48h. The inhibition ratio was calculated as:

$$inhibition\ ratio = \frac{OD_{experiment} - OD_{blank}}{OD_{control} - OD_{blank}}$$

Where OD_{blank}, $OD_{control}$, and $OD_{experiment}$ represent the OD_{600} of the blank control group, bacterial control group, and experimental group, respectively.

All experiments were conducted with three independent replicates. We performed Dunnett's test for computing the statistical difference between control and experiment groups. The peptide with a criterion of inhibition ratio ≥ 25% and P value < 0.05 was considered effective against the tested indicator strain. We categorized the activity results into four groups: (1) no activity, inhibition ratio < 25% or P value ≥ 0.05; (2) low activity, 25% ≤ inhibition ratio < 50% and P value < 0.05; (3) medium activity, 50% ≤ inhibition ratio < 75% and P value < 0.05; (4) high activity, inhibition ratio ≥ 75% and P value < 0.05. Peptide with inhibition ratio > 90% was further examined the MIC against this indicator strain.

Determination of minimum inhibitory concentrations (MICs)

The minimum inhibitory concentrations (MICs) of synthetic peptides were performed by broth microdilution in 96-well flat bottom plates as previously reported (*90*). Vancomycin was used as positive controls. Peptides were added to the plate containing 100 μL aliquots of bacterial suspensions and were two-fold serially diluted (ranging from 100 μg/mL to 0.195 μg/mL). After incubating for 24h or 48h, bacterial growth was assessed by determining OD_{600}. The MIC value was defined as the lowest concentration of the peptides where no bacterial or fungal growth was detected. The assays were conducted in triplicate.

Statistical analysis and visualization

The intersection of identified bacteriocins from three tools was visualized using the R package UpSetR v1.4.0 (*91*). Dunnett's test (two sided) was done by function *DunnettTest* in R package DescTools v0.99.49 (*92*). The sankey plot showing the biosynthetic potential of bacteriocins in the human gut microbiome was achieved by the package ggsankey v0.0.99999 (*93*). The heat maps were plotted using the package ComplexHeatmap v2.14.0 (*94*). The similarity network was visualized using Cytoscape 3.9.0 (*95*). The organization of gene clusters was visualized by R package gggenes v0.5.0 (*96*). Without a specific statement, other figures were generated using ggplot2 v3.4.4 (*97*), and all statistical analyses were finished in R v4.2.3.

Data availability

The bacterial genomes were collected from eight publicly available datasets, including the NCBI Assembly RefSeq database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/), human microbiome (http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html), UHGG database (, https://www.ebi.ac.uk/ena/browser/view/PRJEB33885), HRGM database (https://www.mbiomenet.org/HRGM/), Earth's microbiome (https://portal.nersc.gov/GEM/genomes/), ocean microbiome (https://www.ebi.ac.uk/ena/browser/view/PRJEB45951), mouse (https://zenodo.org/records/4840600) and ruminant gut microbiome (https://www.ebi.ac.uk/ena/browser/view/PRJNA657473). The raw data for metagenomes and metatranscriptomes are publicly available in NCBI-SRA under the BioProject IDs: PRJNA422434, PRJNA319574, PRJEB4336, PRJEB1220, PRJDB4176, PRJNA354235, PRJEB51894, and PRJNA398089. Publicly available AMP data can be accessed from four databases – APD3: https://aps.unmc.edu/downloads; dbAMP v2.0: https://awi.cuhk.edu.cn/dbAMP/download.php; DRAMP v3.0: http://dramp.cpu-bioinfor.org/downloads/; DBAASP v3.0: https://dbaasp.org/. Protein family models can be found in the Pfam database (version 34.0, http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam34.0) and NCBI database (version 11.0, https://ftp.ncbi.nlm.nih.gov/hmm/11.0/). Other source data required to reanalyze the data reported in this paper can be found in supplementary materials or the Zenodo repository (https://zenodo.org/records/10996042).

Code availability

The IIBacFinder codes can be found at https://github.com/ZhangDengwei/IIBacFinder.

Acknowledgments

The author would like to thank Prof. Wang (University of Nebraska Medical Center) for graciously providing the AMP sequences from the latest APD3 database in response to an email request.

Funding

This work is partially funded by three Hong Kong Research Grants Council General research grants (HKU27107320, HKU17115322, and HKU17102123).

Author contributions

Conceptualization: Y.L. and D.Z. Methodology: D.Z., Y.S., J.Z., J.L., G.W., J.Z., Y.G. Investigation: Y.L. Visualization: Y.L. and D.Z. Supervision: Y.L. Writing—original draft: D.Z. Writing—review & editing: Y.L. and D.Z.

Competing interests

Authors declare that they have no competing interests.

References

1. M. Ghoul, S. Mitri, The Ecology and Evolution of Microbial Competition. *Trends Microbiol.* **24**, 833–845 (2016).

2. L. García-Bayona, L. E. Comstock, Bacterial antagonism in host-associated microbial communities. *Science (80-.).* **361** (2018).

3. J. D. Palmer, K. R. Foster, Bacterial species rarely work together. *Science (80-.).* **376**, 581–582 (2022).

4. M. Özçam, J. H. Oh, R. Tocmo, D. Acharya, S. Zhang, T. J. Astmann, M. Heggen, S. Ruiz-Ramírez, F. Li, C. C. Cheng, E. Vivas, F. E. Rey, J. Claesen, T. S. Bugni, J. Walter, J. P. van Pijkeren, A secondary metabolite drives intraspecies antagonism in a gut symbiont that is inhibited by cell-wall acetylation. *Cell Host Microbe* **30**, 824-835.e6 (2022).

5. K. G. Roelofs, M. J. Coyne, R. R. Gentyala, M. Chatzidaki-Livanis, L. E. Comstock, Bacteroidales Secreted Antimicrobial Proteins Target Surface Molecules Necessary for Gut Colonization and Mediate Competition In Vivo. *MBio* **7**, 1–10 (2016).

6. P. Piewngam, Y. Zheng, T. H. Nguyen, S. W. Dickey, H. S. Joo, A. E. Villaruz, K. A. Glose, E. L. Fisher, R. L. Hunt, B. Li, J. Chiou, S. Pharkjaksu, S. Khongthong, G. Y. C. Cheung, P. Kiratisin, M. Otto, Pathogen elimination by probiotic Bacillus via signalling interference. *Nature* **562**, 532–537 (2018).

7. P. D. Cotter, R. P. Ross, C. Hill, Bacteriocins-a viable alternative to antibiotics? *Nat. Rev. Microbiol.* **11**, 95–105 (2013).

8. S. Heilbronner, B. Krismer, H. Brötz-Oesterhelt, A. Peschel, The microbiome-shaping roles of bacteriocins. Nature Publishing Group (2021). https://doi.org/10.1038/s41579-021-00569-w.

9. I. Sugrue, R. P. Ross, C. Hill, Bacteriocin diversity, function, discovery and application as antimicrobials. *Nat. Rev. Microbiol.*, doi: 10.1038/s41579-024-01045-x (2024).

10. J. Z. Acedo, S. Chiorean, J. C. Vederas, M. J. van Belkum, The expanding structural variety among bacteriocins from Gram-positive bacteria. Oxford University Press (2018). https://doi.org/10.1093/femsre/fuy033.

11. A. Sharma, S. Srivastava, Anti-Candida activity of two-peptide bacteriocins, plantaricins (Pln E/F and J/K) and their mode of action. *Fungal Biol.* **118**, 264–275 (2014).

12. R. Teiar, F. Sane, I. Erol, M. P. Nekoua, D. Lecouturier, R. Boukherroub, S. Durdağı, D. Hober, D. Drider, Enterocin DD14 can inhibit the infection of eukaryotic cells with enveloped viruses. *Arch. Microbiol.* **206**, 269 (2024).

13. M. Nazari, I. Yaghoubian, D. L. Smith, The stimulatory effect of Thuricin 17, a PGPR-produced bacteriocin, on canola (Brassica, napus L.) germination and vegetative growth under stressful temperatures. *Front. Plant Sci.* **13**, 1–17 (2022).

14. J. Hu, L. Ma, Y. Nie, J. Chen, W. Zheng, X. Wang, C. Xie, Z. Zheng, Z. Wang, T. Yang, M. Shi, L. Chen, Q. Hou, Y. Niu, X. Xu, Y. Zhu, Y. Zhang, H. Wei, X. Yan, A Microbiota-Derived Bacteriocin Targets the Host to Confer Diarrhea Resistance in Early-Weaned Piglets. *Cell Host Microbe* **24**, 817-832.e8 (2018).

15. D. D. Heeney, Z. Zhai, Z. Bendiks, J. Barouei, A. Martinic, C. Slupsky, M. L. Marco, Lactobacillus plantarum bacteriocin is associated with intestinal and systemic improvements in diet-induced obese mice and maintains epithelial barrier integrity in vitro. *Gut Microbes* **10**, 382–397 (2019).

16. E. Garcia-Gutierrez, M. J. Mayer, P. D. Cotter, A. Narbad, Gut microbiota as a source of novel antimicrobials. *Gut Microbes* **10**, 1–21 (2019).

17. Q. Gu, J. Yan, Y. Lou, Z. Zhang, Y. Li, Z. Zhu, M. Liu, D. Wu, Y. Liang, J. Pu, X. Zhao, H. Xiao, P. Li, Bacteriocins: Curial guardians of gastrointestinal tract. *Compr. Rev. Food Sci. Food Saf.* **23**, 1–27 (2024).

18. S. G. Javvadi, M. Kujawska, D. Papp, A. Jordan, M. A. E. Lawson, P. Clarke, N. Beraza, L. J. Hall, G. Microbes, I. Microbiome, U. Kingdom, U. Kingdom, A novel bacteriocin produced by Bifidobacterium longum subsp. infantis has dual antimicrobial and immunomodulatory activity. *bioRxiv*, doi: 10.1101/2022.01.27.477972 (2022).

19. B. Zhu, X. Wang, L. Li, Human gut microbiome: The second genome of human body. *Protein Cell* **1**, 718–725 (2010).

20. C. J. Walsh, C. M. Guinane, C. Hill, R. P. Ross, P. W. O'Toole, P. D. Cotter, In silico identification of bacteriocin gene clusters in the gastrointestinal tract, based on the Human Microbiome Project's reference genome database. *BMC Microbiol.* **15** (2015).

21. J. Zheng, M. G. Gänzle, X. B. Lin, L. Ruan, M. Sun, Diversity and dynamics of bacteriocins from human microbiome. *Environ. Microbiol.* **17**, 2133–2143 (2015).

22. A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).

23. C. D. Santos-Júnior, M. D. T. Torres, Y. Duan, Á. Rodríguez del Río, T. S. B. Schmidt, H. Chong, A. Fullam, M. Kuhn, C. Zhu, A. Houseman, J. Somborski, A. Vines, X.-M. Zhao, P. Bork, J. Huerta-Cepas, C. de la Fuente-Nunez, L. P. Coelho, Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell*, 1–18 (2024).

24. Y. Ma, Z. Guo, B. Xia, Y. Zhang, X. Liu, Y. Yu, N. Tang, X. Tong, M. Wang, X. Ye, J. Feng, Y. Chen, J. Wang, Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).

25. M. D. T. Torres, E. Brooks, A. Cesaro, H. Sberro, C. Nicolaou, A. S. Bhatt, C. de la Fuente-Nunez, Human gut metagenomic mining reveals an untapped source of peptide antibiotics. *bioRxiv*, 2023.08.31.555711 (2023).

26. S. Akhter, J. H. Miller, BaPreS: a software tool for predicting bacteriocins using an optimal set of features. *BMC Bioinformatics* **24**, 313 (2023).

27. S. Akhter, J. H. Miller, BPAGS: a web application for bacteriocin prediction via feature evaluation using alternating decision tree, genetic algorithm, and linear support vector classifier. *Front. Bioinforma.* **3**, 1–11 (2024).

28. M. N. Hamid, I. Friedberg, Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* **35**, 2009–2016 (2019).

29. Z. Cui, Z. H. Chen, Q. H. Zhang, V. Gribova, V. F. Filaretov, D. S. Huang, RMSCNN: A Random Multi-Scale Convolutional Neural Network for Marine Microbial Bacteriocins Identification. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **19**, 3663–3672 (2022).

30. A. J. van Heel, A. de Jong, C. Song, J. H. Viel, J. Kok, O. P. Kuipers, BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.* **46**, W278–W281 (2018).

31. K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Wezel, M. H. Medema, T. Weber, antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).

32. M. A. Skinnider, C. W. Johnston, M. Gunabalasingam, N. J. Merwin, A. M. Kieliszek, R. J.

MacLellan, H. Li, M. R. M. Ranieri, A. L. H. Webster, M. P. T. Cao, A. Pfeifle, N. Spencer, Q. H. To, D. P. Wallace, C. A. Dejong, N. A. Magarvey, Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).

33. P. Baldi, Y. Chauvin, T. Hunkapiller, M. A. Mcclure, Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1059–1063 (1994).

34. S. R. Eddy, Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

35. N. J. Merwin, W. K. Mousa, C. A. Dejong, M. A. Skinnider, M. J. Cannon, H. Li, K. Dial, M. Gunabalasingam, C. Johnston, N. A. Magarvey, DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 371–380 (2020).

36. F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, M. H. Gíslason, S. I. Pihl, K. D. Tsirigos, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).

37. G. Wang, X. Li, Z. Wang, APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).

38. J. H. Jhong, L. Yao, Y. Pang, Z. Li, C. R. Chung, R. Wang, S. Li, W. Li, M. Luo, R. Ma, Y. Huang, X. Zhu, J. Zhang, H. Feng, Q. Cheng, C. Wang, K. Xi, L. C. Wu, T. H. Chang, J. T. Horng, L. Zhu, Y. C. Chiang, Z. Wang, T. Y. Lee, dbAMP 2.0: Updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Res.* **50**, D460–D470 (2022).

39. G. Shi, X. Kang, F. Dong, Y. Liu, N. Zhu, Y. Hu, H. Xu, X. Lao, H. Zheng, DRAMP 3.0: An enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Res.* **50**, D488–D496 (2022).

40. M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt, M. Tartakovsky, DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).

41. D. H. Haft, M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, M. Gwadz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, C. Zheng, F. Thibaud-Nissen, L. Y. Geer, A. Marchler-Bauer, K. D. Pruitt, RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).

42. S. Arbulu, M. Kjos, Revisiting the Multifaceted Roles of Bacteriocins: The Multifaceted Roles of Bacteriocins. *Microb. Ecol.* **87** (2024).

43. S. Nayfach, S. Roux, R. Seshadri, D. Udwary, N. Varghese, F. Schulz, D. Wu, D. Paez-Espino, I.-M. Chen, M. Huntemann, K. Palaniappan, J. Ladau, S. Mukherjee, T. B. K. Reddy, T. Nielsen, E. Kirton, J. P. Faria, J. N. Edirisinghe, C. S. Henry, S. P. Jungbluth, D. Chivian, P. Dehal, E. M. Wood-Charlson, A. P. Arkin, S. G. Tringe, A. Visel, H. Abreu, S. G. Acinas, E. Allen, M. A. Allen, L. V. Alteio, G. Andersen, A. M. Anesio, G. Attwood, V. Avila-Magaña, Y. Badis, J. Bailey, B. Baker, P. Baldrian, H. A. Barton, D. A. C. Beck, E. D. Becraft, H. R. Beller, J. M. Beman, R. Bernier-Latmani, T. D. Berry, A. Bertagnolli, S. Bertilsson, J. M. Bhatnagar, J. T. Bird, J. L. Blanchard, S. E. Blumer-Schuette, B. Bohannan, M. A. Borton, A. Brady, S. H. Brawley, J. Brodie, S. Brown, J. R. Brum, A. Brune, D. A. Bryant, A. Buchan, D. H. Buckley, J. Buongiorno, H. Cadillo-Quiroz, S. M. Caffrey, A. N. Campbell, B. Campbell, S. Carr, J. Carroll, S. C. Cary, A. M. Cates, R. A. Cattolico, R. Cavicchioli, L. Chistoserdova, M. L. Coleman, P. Constant, J. M. Conway, W. P. Mac Cormack, S. Crowe, B. Crump, C. Currie, R. Daly, K. M. DeAngelis, V. Denef, S. E. Denman, A. Desta, H. Dionisi, J. Dodsworth, N.

Dombrowski, T. Donohue, M. Dopson, T. Driscoll, P. Dunfield, C. L. Dupont, K. A. Dynarski, V. Edgcomb, E. A. Edwards, M. S. Elshahed, I. Figueroa, B. Flood, N. Fortney, C. S. Fortunato, C. Francis, C. M. M. Gachon, S. L. Garcia, M. C. Gazitua, T. Gentry, L. Gerwick, J. Gharechahi, P. Girguis, J. Gladden, M. Gradoville, S. E. Grasby, K. Gravuer, C. L. Grettenberger, R. J. Gruninger, J. Guo, M. Y. Habteselassie, S. J. Hallam, R. Hatzenpichler, B. Hausmann, T. C. Hazen, B. Hedlund, C. Henny, L. Herfort, M. Hernandez, O. S. Hershey, M. Hess, E. B. Hollister, L. A. Hug, D. Hunt, J. Jansson, J. Jarett, V. V. Kadnikov, C. Kelly, R. Kelly, W. Kelly, C. A. Kerfeld, J. Kimbrel, J. L. Klassen, K. T. Konstantinidis, L. L. Lee, W.-J. Li, A. J. Loder, A. Loy, M. Lozada, B. MacGregor, C. Magnabosco, A. Maria da Silva, R. M. McKay, K. McMahon, C. S. McSweeney, M. Medina, L. Meredith, J. Mizzi, T. Mock, L. Momper, M. A. Moran, C. Morgan-Lang, D. Moser, G. Muyzer, D. Myrold, M. Nash, C. L. Nesbø, A. P. Neumann, R. B. Neumann, D. Noguera, T. Northen, J. Norton, B. Nowinski, K. Nüsslein, M. A. O'Malley, R. S. Oliveira, V. Maia de Oliveira, T. Onstott, J. Osvatic, Y. Ouyang, M. Pachiadaki, J. Parnell, L. P. Partida-Martinez, K. G. Peay, D. Pelletier, X. Peng, M. Pester, J. Pett-Ridge, S. Peura, P. Pjevac, A. M. Plominsky, A. Poehlein, P. B. Pope, N. Ravin, M. C. Redmond, R. Reiss, V. Rich, C. Rinke, J. L. M. Rodrigues, W. Rodriguez-Reillo, K. Rossmassler, J. Sackett, G. H. Salekdeh, S. Saleska, M. Scarborough, D. Schachtman, C. W. Schadt, M. Schrenk, A. Sczyrba, A. Sengupta, J. C. Setubal, A. Shade, C. Sharp, D. H. Sherman, O. V. Shubenkova, I. N. Sierra-Garcia, R. Simister, H. Simon, S. Sjöling, J. Slonczewski, R. S. Correa de Souza, J. R. Spear, J. C. Stegen, R. Stepanauskas, F. Stewart, G. Suen, M. Sullivan, D. Sumner, B. K. Swan, W. Swingley, J. Tarn, G. T. Taylor, H. Teeling, M. Tekere, A. Teske, T. Thomas, C. Thrash, J. Tiedje, C. S. Ting, B. Tully, G. Tyson, O. Ulloa, D. L. Valentine, M. W. Van Goethem, J. VanderGheynst, T. J. Verbeke, J. Vollmers, A. Vuillemin, N. B. Waldo, D. A. Walsh, B. C. Weimer, T. Whitman, P. van der Wielen, M. Wilkins, T. J. Williams, B. Woodcroft, J. Woolet, K. Wrighton, J. Ye, E. B. Young, N. H. Youssef, F. B. Yu, T. I. Zemskaya, R. Ziels, T. Woyke, N. J. Mouncey, N. N. Ivanova, N. C. Kyrpides, E. A. Eloe-Fadrosh, A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).

44. Y. Fan, O. Pedersen, Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).

45. T. Gamier, S. T. Cole, Studies of UV-inducible promoters from Clostridium perfringens in vivo and in vitro. *Mol. Microbiol.* **2**, 607–614 (1988).

46. S. C. Corr, Y. Li, C. U. Riedel, P. W. O'Toole, C. Hill, C. G. M. Gahan, Bacteriocin production as a mechanism for the antiinfective activity of Lactobacillus salivarius UCC118. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7617–7621 (2007).

47. J. Li, J. Jin, S. Li, Y. Zhong, Y. Jin, X. Zhang, B. Xia, Y. Zhu, R. Guo, X. Sun, J. Guo, F. Hu, W. Xiao, F. Huang, H. Ye, R. Li, Y. Zhou, X. Xiang, H. Yao, Q. Yan, L. Su, L. Wu, T. Luo, Y. Liu, X. Guo, J. Qin, H. Qi, J. He, J. Wang, Z. Li, Tonsillar Microbiome-Derived Lantibiotics Induce Structural Changes of IL-6 and IL-21 Receptors and Modulate Host Immunity. *Adv. Sci.* **9**, 2202706 (2022).

48. B. P. Lazzaro, M. Zasloff, J. Rolff, Antimicrobial peptides: Application informed by evolution. American Association for the Advancement of Science (2020). https://doi.org/10.1126/science.aau5480.

49. E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, N. Segata, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).

50. C. Y. Kim, M. Lee, S. Yang, K. Kim, D. Yong, H. R. Kim, I. Lee, Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med.* **13** (2021).

51. L. Paoli, H.-J. Ruscheweyh, C. C. Forneris, F. Hubrich, S. Kautsar, A. Bhushan, A. Lotti, Q. Clayssen, G. Salazar, A. Milanese, C. I. Carlström, C. Papadopoulou, D. Gehrig, M. Karasikov, H. Mustafa, M. Larralde, L. M. Carroll, P. Sánchez, A. A. Zayed, D. R. Cronin, S. G. Acinas, P. Bork, C. Bowler, T. O. Delmont, J. M. Gasol, A. D. Gossert, A. Kahles, M. B. Sullivan, P. Wincker, G. Zeller, S. L. Robinson, J. Piel, S. Sunagawa, Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).

52. B. S. Beresford-Jones, S. C. Forster, M. D. Stares, G. Notley, E. Viciani, H. P. Browne, D. J. Boehmler, A. T. Soderholm, N. Kumar, K. Vervier, J. R. Cross, A. Almeida, T. D. Lawley, V. A. Pedicord, The Mouse Gastrointestinal Bacteria Catalogue enables translation between the mouse and human gut microbiotas via functional mapping. *Cell Host Microbe* **30**, 124-138.e8 (2022).

53. F. Xie, W. Jin, H. Si, Y. Yuan, Y. Tao, J. Liu, X. Wang, C. Yang, Q. Li, X. Yan, L. Lin, Q. Jiang, L. Zhang, C. Guo, C. Greening, R. Heller, L. L. Guan, P. B. Pope, Z. Tan, W. Zhu, M. Wang, Q. Qiu, Z. Li, S. Mao, An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* **9**, 1–20 (2021).

54. A. Pandi, D. Adam, A. Zare, V. T. Trinh, S. L. Schaefer, M. Burt, B. Klabunde, E. Bobkova, M. Kushwaha, Y. Foroughijabbari, P. Braun, C. Spahn, C. Preußer, E. Pogge von Strandmann, H. B. Bode, H. von Buttlar, W. Bertrams, A. L. Jung, F. Abendroth, B. Schmeck, G. Hummer, O. Vázquez, T. J. Erb, Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nat. Commun.* **14**, 7197 (2023).

55. R. S. Ayikpoe, C. Shi, A. J. Battiste, S. M. Eslami, S. Ramesh, M. A. Simon, I. R. Bothwell, H. Lee, A. J. Rice, H. Ren, Q. Tian, L. A. Harris, R. Sarksian, L. Zhu, A. M. Frerk, T. W. Precord, W. A. van der Donk, D. A. Mitchell, H. Zhao, A scalable platform to discover antimicrobials of ribosomal origin. *Nat. Commun.* **13**, 1–15 (2022).

56. W.-Q. Liu, X. Ji, F. Ba, Y. Zhang, H. Xu, S. Huang, X. Zheng, Y. Liu, S. Ling, M. C. Jewett, J. Li, Cell-free biosynthesis and engineering of ribosomally synthesized lanthipeptides. *Nat. Commun.* **15**, 4336 (2024).

57. C. S. Sit, C. T. Lohans, M. J. van Belkum, C. D. Campbell, M. Miskolzie, J. C. Vederas, Substitution of a Conserved Disulfide in the Type IIa Bacteriocin, Leucocin A, with L-Leucine and L-Serine Residues: Effects on Activity and Three-Dimensional Structure. *ChemBioChem* **13**, 35–38 (2012).

58. R. H. Perez, T. Zendo, K. Sonomoto, Circular and leaderless bacteriocins: Biosynthesis, mode of action, applications, and prospects. Frontiers Media S.A. (2018). https://doi.org/10.3389/fmicb.2018.02085.

59. K. Venema, T. Abee, A. J. Haandrikman, K. J. Leenhouts, J. Kok, W. N. Konings, G. Venema, Mode of Action of Lactococcin B, a Thiol-Activated Bacteriocin from Lactococcus lactis. *Appl. Environ. Microbiol.* **59**, 1041–1048 (1993).

60. D. Ankaiah, P. Esakkiraj, V. Perumal, R. Ayyanna, A. Venkatesan, Probiotic characterization of Enterococcus faecium por1: Cloning, over expression of Enterocin-A and evaluation of antibacterial, anti-cancer properties. *J. Funct. Foods* **38**, 280–292 (2017).

61. Q. Yan, S. Li, Q. Yan, X. Huo, C. Wang, X. Wang, Y. Sun, W. Zhao, Z. Yu, Y. Zhang, R. Guo, Q. Lv, X. He, C. Yao, Z. Li, F. Chen, Q. Ji, A. Zhang, H. Jin, G. Wang, X. Feng, L. Feng, F. Wu, J. Ning, S. Deng, Y. An, D. Guo, F. M. Martin, X. Ma, A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. *Cell*, 1–21 (2024).

62. M. Ó. B. Wachsman, M. E. Farías, E. Takeda, F. Sesma, A. P. De Ruiz Holgado, R. A. De Torres, C. E. Coto, Antiviral activity of enterocin CRL35 against herpesviruses. *Int. J. Antimicrob.*

Agents **12**, 293–299 (1999).

63. R. Teiar, A. Pérez-Ramos, H. Zgheib, B. Cudennec, Y. Belguesmia, D. Drider, Anti-adhesion and Anti-inflammatory Potential of the Leaderless Class IIb Bacteriocin Enterocin DD14. *Probiotics Antimicrob. Proteins* **14**, 613–619 (2022).

64. J. Śmiałek, P. Mak, M. Bzowska, A. Hinz, R. Mężyk-Kopeć, K. Sołtys, Bacteriocin BacSp222 and Its Succinylated Forms Exhibit Proinflammatory Activities Toward Innate Immune Cells. *J. Inflamm. Res.* **15**, 4601–4621 (2022).

65. D. V. Antoshina, S. V. Balandin, I. V. Bogdanov, M. A. Vershinina, E. V. Sheremeteva, I. Y. Toropygin, E. I. Finkina, T. V. Ovchinnikova, Antimicrobial Activity and Immunomodulatory Properties of Acidocin A, the Pediocin-like Bacteriocin with the Non-Canonical Structure. *Membranes (Basel).* **12** (2022).

66. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

67. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

68. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

69. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

70. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

71. J. Santos-Aberturas, G. Chandra, L. Frattaruolo, R. Lacret, T. H. Pham, N. M. Vior, T. H. Eyles, A. W. Truman, Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool. *Nucleic Acids Res.* **47**, 4624–4637 (2019).

72. T. J. Lawrence, D. L. Carper, M. K. Spangler, A. A. Carrell, T. A. Rush, S. J. Minter, D. J. Weston, J. L. Labbé, amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. *Bioinformatics* **37**, 2058–2060 (2021).

73. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

74. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, R. Lanfear, E. Teeling, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

75. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

76. I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, 256–259 (2019).

77. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

78. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

79. J. Oksanen, Vegan: ecological diversity. (2017).

80. T. C. Hsieh, K. H. Ma, A. Chao, iNEXT: an R package for rarefaction and extrapolation of

species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).

81. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

82. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

83. A. Westbrook, J. Ramsdell, T. Schuelke, L. Normington, R. D. Bergeron, W. K. Thomas, M. D. MacManes, PALADIN: Protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics* **33**, 1473–1478 (2017).

84. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

85. A. Blanco-Míguez, F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, L. Dubois, K. D. Huang, A. M. Thomas, W. A. Nickols, G. Piccinno, E. Piperni, M. Punčochář, M. Valles-Colomer, A. Tett, F. Giordano, R. Davies, J. Wolf, S. E. Berry, T. D. Spector, E. A. Franzosa, E. Pasolli, F. Asnicar, C. Huttenhower, N. Segata, Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).

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

87. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).

88. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

89. W. L. DeLano, others, Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40**, 82–92 (2002).

90. Z. Cheng, B. B. He, K. Lei, Y. Gao, Y. Shi, Z. Zhong, H. Liu, R. Liu, H. Zhang, S. Wu, W. Zhang, X. Tang, Y. X. Li, Rule-based omics mining reveals antimicrobial macrocyclic peptides against drug-resistant clinical isolates. *Nat. Commun.* **15** (2024).

91. J. R. Conway, A. Lex, N. Gehlenborg, UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

92. A. Signorell, DescTools: Tools for Descriptive Statistics. (2024).

93. D. Sjoberg, ggsankey: Sankey, Alluvial and Sankey Bump Plots. (2024).

94. Z. Gu, Complex heatmap visualization. *iMeta* **1**, 1–15 (2022).

95. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

96. D. Wilkins, gggenes: Draw Gene Arrow Maps in “ggplot2.” (2023).

97. R. A. M. Villanueva, Z. J. Chen, ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Meas. Interdiscip. Res. Perspect.* **17**, 160–167 (2019).

Figure legends:

Figure 1. Overview of IIBacFinder specializing in detecting unmodified bacteriocins. (A) Schematic overview of the genome mining procedure in IIBacFinder. IIBacFinder detects potential unmodified class II bacteriocins from nucleotide sequences in FASTA format, by leveraging the library of 1,153 pHMMs of precursor genes and 37 pHMMs of context genes. Putative core peptides of hypothetical bacteriocins are be deduced and compared against known AMPs from four databases (i.e., APD3, dbAMP v2.0, DRAMP v3.0, and DBAASP v3.0). (B) Performance of three tools in recovering 205 known class II bacteriocins from 137 bacterial genomes. The maximum likelihood phylogenetic tree was inferred based on full bacteriocin sequences. Symbols at the leaf tips represent the subclass of class II bacteriocins, and the black triangles indicate the 12 bacteriocin sequences excluded from the pHMM model construction. Numbers in brackets indicate the number of bacteriocins detected by each prediction tool.

Figure 2. Biosynthetic potential of class II bacteriocins in the bacterial kingdom. (A) Rarefaction curves illustrating how the number of detected bacteriocins increases with the inclusion of more bacterial genomes. (B) The GTDB bacterial tree, decorated with the number of genomes, detected bacteriocin sequences, unique bacteriocin sequences, and bacteriocin sequences per genome, demonstrates the bacteriocin biosynthetic potential across different bacterial phyla. (C) The top 20 bacterial genera containing the most abundant class II bacteriocins. (D) Antibacterial activity was examined for 21 synthetic hypothetical bacteriocins (HBs) against both Gram-positive (+) and Gram-negative (-) bacteria. A red cross signifies that HBs cannot be identified by the prediction tool, while a green tick indicates successful detection. The activity screening was conducted at a concentration of 100 μg/ml, with activity being categorized into four groups (no activity, < 25% inhibitory ratio or *P* value > 0.05; low activity, 25% ≤ inhibitory ratio < 50% & *P* value < 0.05; medium activity, 50% ≤ inhibitory ratio < 75% & *P* value < 0.05; high activity, inhibitory ratio > 75% & *P* value < 0.05). HBs with high activity are subjected to MIC determination. All assays are performed in three independent replicates.

Figure 3. Class II bacteriocins exhibit niche-specific enrichment in host-associated bacteria. (A) Bar plots illustrating the number of genomes analyzed for class II bacteriocin identification and the corresponding number of non-redundant bacteriocin sequences detected. (B) Bar plot showing the number of non-redundant bacteriocin sequences detected from bacterial genomes across different habitats. (C) Venn diagram depicting the overlap of bacteriocin sequences detected from four distinct habitats, with pie charts indicating the proportion of niche-specific sequences, labeled with percentages. (D) Accumulation curves of bacteriocin sequences detected from the four habitats, with solid lines representing interpolated and actual data, and dashed lines representing extrapolated data.

Figure 4. Human gut bacteria have significant potential to produce a diverse array of unmodified bacteriocins. (A) Sankey plot showing the bacteriocin biosynthetic potential of 285,835 human gut bacteria across different taxonomic levels. Numbers in brackets indicate the number of non-redundant class II bacteriocin sequences detected. (B) Heatmap illustrating the number of gut microbiota-derived class II bacteriocins homologous to specific (left panel) and all (right panel) known class II bacteriocins at varying identity thresholds (50%-90%). Only the top 20 most abundant known class II bacteriocins are shown. (C) Rarefaction curves showing the increase in detected bacteriocins with the inclusion of metatranscriptomes. The pie chart indicates the percentage of class II bacteriocins detected (red) and not detected (grey) in 724 metatranscriptomes. (D) Scatter plot illustrating the prevalence of 6,238 putative bacteriocins in metatranscriptomes. (E) Violin plot depicting the distribution of the number of class II bacteriocins actively transcribed in individuals from three cohorts.

Figure 5. Prioritization and experimental validation of class II bacteriocins from human gut microbiota. (A) Schematic overview of prioritizing class II bacteriocins using metagenomics and metatrancriptomics analysis. Class II bacteriocins are detected from >700,000 bacterial genomes, which

are subsequently surveyed in human gut metagenomes and metatranscriptomes. The detected bacteriocins are ranked based on a meta-ranking scheme consisting of four criteria (i.e., prevalence, abundance, negative association with bacterial species, and correlation with community diversity). (B) Antibacterial activity was determined for 26 human gut microbiota-derived hypothetical bacteriocins (HGM-HBs). A red cross signifies that HGM-HBs cannot be identified by the prediction tool, while a green tick indicates successful detection. The heatmap in the middle panel depicts the prevalence and abundance of class II bacteriocins in different cohorts. The ones in the right panel show the activity screening for 26 HGM-HBs at a concentration of 100 μg/ml and MIC determination for HGM-HBs with high activity. All assays are performed in three independent replicates.

Figure 1

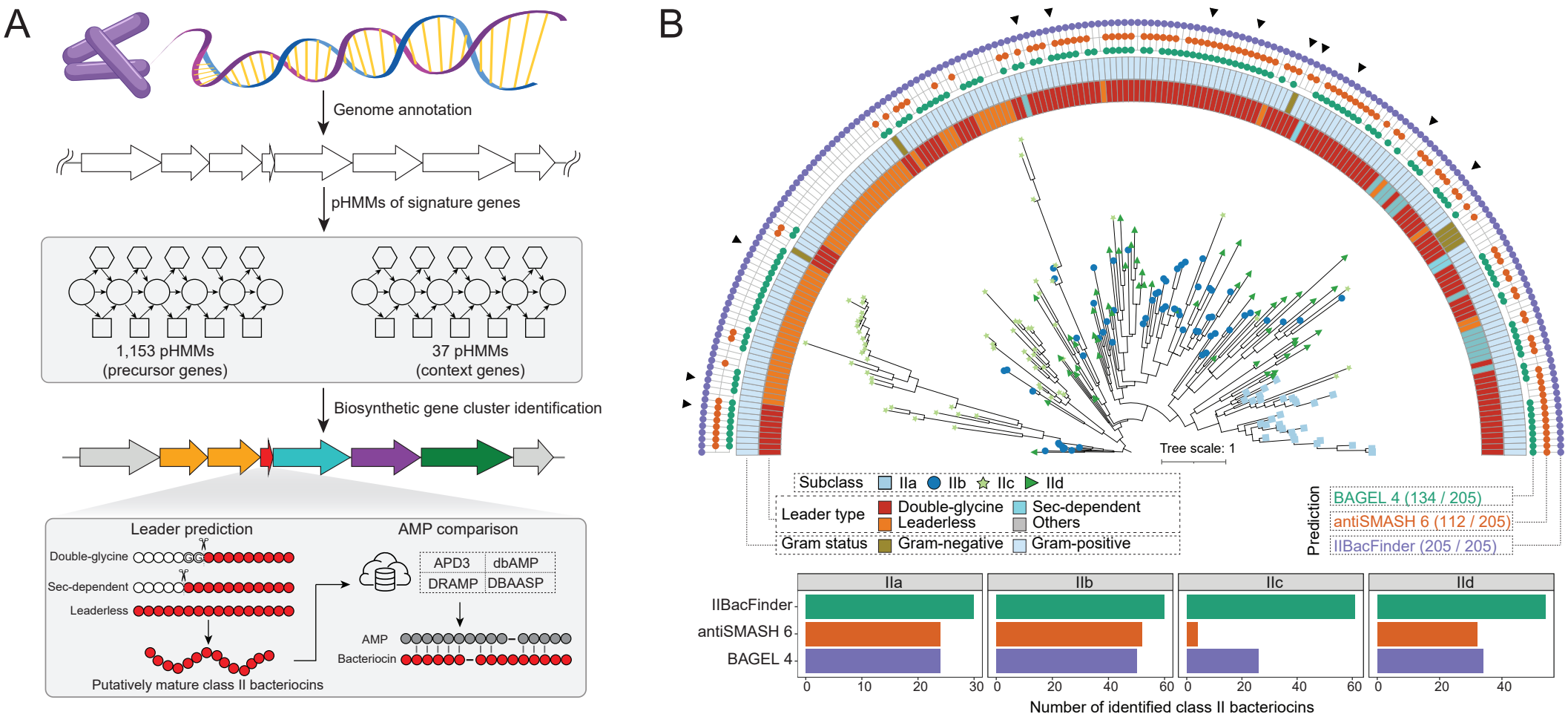

A
Genome annotation
pHMMs of signature genes
1,153 pHMMs
(precursor genes)
37 pHMMs
(context genes)
Biosynthetic gene cluster identification
Leader prediction
AMP comparison
Double-glycine
Sec-dependent
Leaderless
APD3
dbAMP
DRAMP
DBAASP
AMP
Bacteriocin
Putatively mature class II bacteriocins
B
Tree scale: 1
Subclass
IIa
IIb
IIc
IId
Leader type
Double-glycine
Leaderless
Sec-dependent
Others
Gram status
Gram-negative
Gram-positive
Prediction
BAGEL 4 (134 / 205)
antiSMASH 6 (112 / 205)
IIBacFinder (205 / 205)
IIBacFinder
antiSMASH 6
BAGEL 4
Number of identified class II bacteriocins

Figure 2

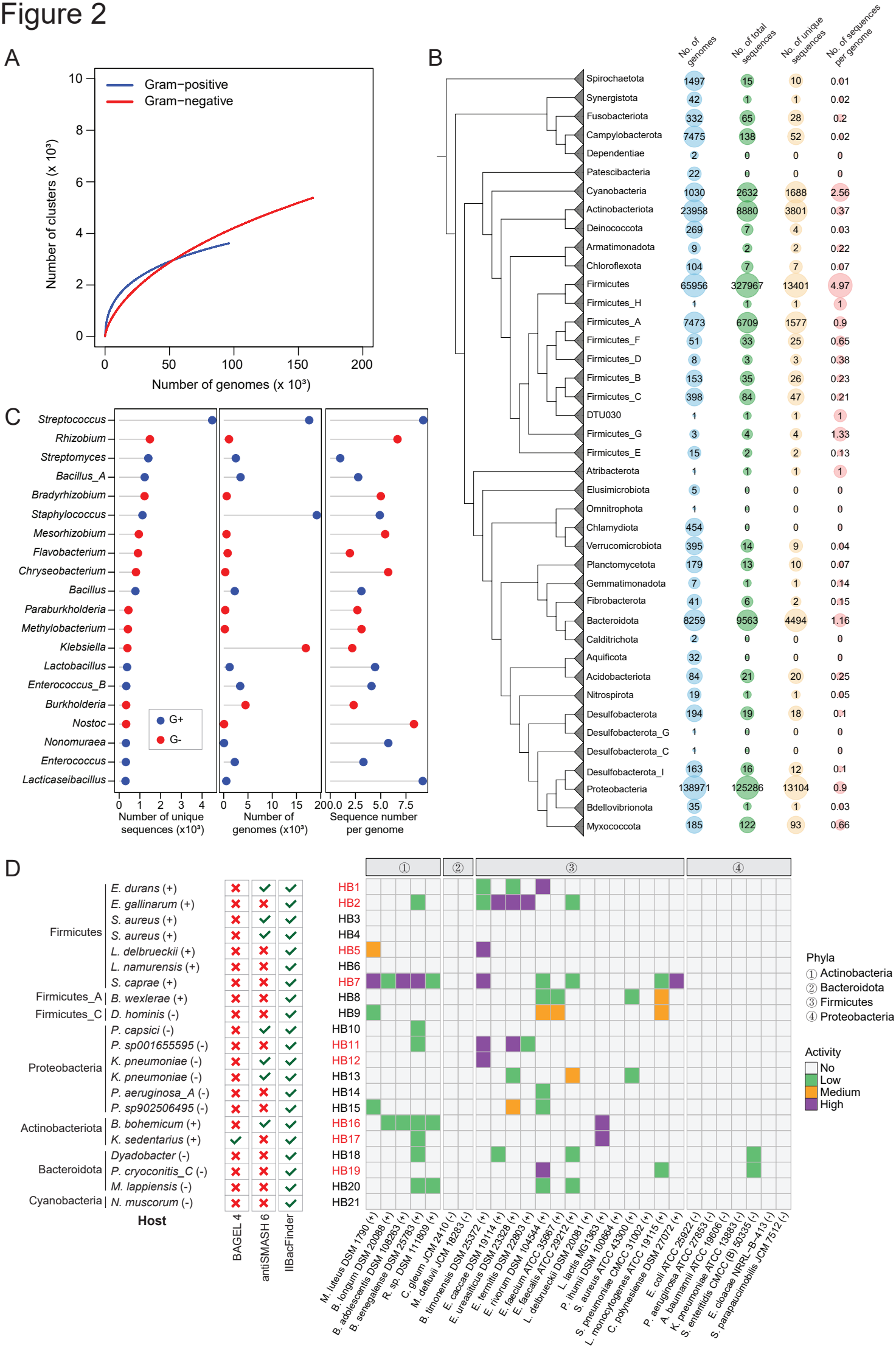

A
Gram−positive
Gram−negative
Number of clusters (x 10³)
Number of genomes (x 10³)
B
No. of genomes
No. of total sequences
No. of unique sequences
No. of sequences per genome
Spirochaetota 1497 15 10 0.01
Synergistota 42 1 1 0.02
Fusobacteriota 332 65 28 0.2
Campylobacterota 7475 138 52 0.02
Dependentiae 2 0 0 0
Patescibacteria 22 0 0 0
Cyanobacteria 1030 2632 1688 2.56
Actinobacteriota 23958 8880 3801 0.37
Deinococcota 269 7 4 0.03
Armatimonadota 9 2 2 0.22
Chloroflexota 104 7 7 0.07
Firmicutes 65956 327967 13401 4.97
Firmicutes_H 1 1 1 1
Firmicutes_A 7473 6709 1577 0.9
Firmicutes_F 51 33 25 0.65
Firmicutes_D 8 3 3 0.38
Firmicutes_B 153 35 26 0.23
Firmicutes_C 398 84 47 0.21
DTU030 1 1 1 1
Firmicutes_G 3 4 4 1.33
Firmicutes_E 15 2 2 0.13
Atribacterota 1 1 1 1
Elusimicrobiota 5 0 0 0
Omnitrophota 1 0 0 0
Chlamydiota 454 0 0 0
Verrucomicrobiota 395 14 9 0.04
Planctomycetota 179 13 10 0.07
Gemmatimonadota 7 1 1 0.14
Fibrobacterota 41 6 2 0.15
Bacteroidota 8259 9563 4494 1.16
Calditrichota 2 0 0 0
Aquificota 32 0 0 0
Acidobacteriota 84 21 20 0.25
Nitrospirota 19 1 1 0.05
Desulfobacterota 194 19 18 0.1
Desulfobacterota_G 1 0 0 0
Desulfobacterota_C 1 0 0 0
Desulfobacterota_I 163 16 12 0.1
Proteobacteria 138971 125286 13104 0.9
Bdellovibrionota 35 1 1 0.03
Myxococcota 185 122 93 0.66
C
Streptococcus
Rhizobium
Streptomyces
Bacillus_A
Bradyrhizobium
Staphylococcus
Mesorhizobium
Flavobacterium
Chryseobacterium
Bacillus
Paraburkholderia
Methylobacterium
Klebsiella
Lactobacillus
Enterococcus_B
Burkholderia
Nostoc
Nonomuraea
Enterococcus
Lacticaseibacillus
G+
G-
Number of unique sequences (x10³)
Number of genomes (x10³)
Sequence number per genome
D
Firmicutes
E. durans (+)
E. gallinarum (+)
S. aureus (+)
S. aureus (+)
L. delbrueckii (+)
L. namurensis (+)
S. caprae (+)
Firmicutes_A
B. wexlerae (+)
Firmicutes_C
D. hominis (-)
Proteobacteria
P. capsici (-)
P. sp001655595 (-)
K. pneumoniae (-)
K. pneumoniae (-)
P. aeruginosa_A (-)
P. sp902506495 (-)
Actinobacteriota
B. bohemicum (+)
K. sedentarius (+)
Bacteroidota
Dyadobacter (-)
P. cryoconitis_C (-)
M. lappiensis (-)
Cyanobacteria
N. muscorum (-)
Host
BAGEL 4
antiSMASH 6
IIBacFinder
HB1
HB2
HB3
HB4
HB5
HB6
HB7
HB8
HB9
HB10
HB11
HB12
HB13
HB14
HB15
HB16
HB17
HB18
HB19
HB20
HB21
M. luteus DSM 1790 (+)
B. longum DSM 20088 (+)
B. adolescentis DSM 108263 (+)
B. senegalense DSM 25783 (+)
R. sp. DSM 111809 (+)
C. gleum JCM 2410 (-)
M. defluvii JCM 18283 (-)
B. timonensis DSM 25372 (+)
E. caecae DSM 19114 (+)
E. ureasiticus DSM 23328 (+)
E. termitis DSM 22803 (+)
E. rivorum DSM 104544 (+)
E. faecium ATCC 35667 (+)
E. faecalis ATCC 29212 (+)
L. delbrueckii DSM 20081 (+)
L. lactis MG1363 (+)
P. ihumii DSM 100664 (+)
S. aureus ATCC 43300 (+)
S. pneumoniae CMCC 31002 (+)
L. monocytogenes ATCC 19115 (+)
C. polynesiense DSM 27072 (+)
E. coli ATCC 25922 (-)
P. aeruginosa ATCC 27853 (-)
A. baumannii ATCC 19606 (-)
K. pneumoniae ATCC 13883 (-)
S. enteritidis CMCC (B) 50335 (-)
E. cloacae NRRL-B-413 (-)
S. parapaucimobilis JCM 7512 (-)
Phyla
① Actinobacteria
② Bacteroidota
③ Firmicutes
④ Proteobacteria
Activity
No
Low
Medium
High

Figure 3

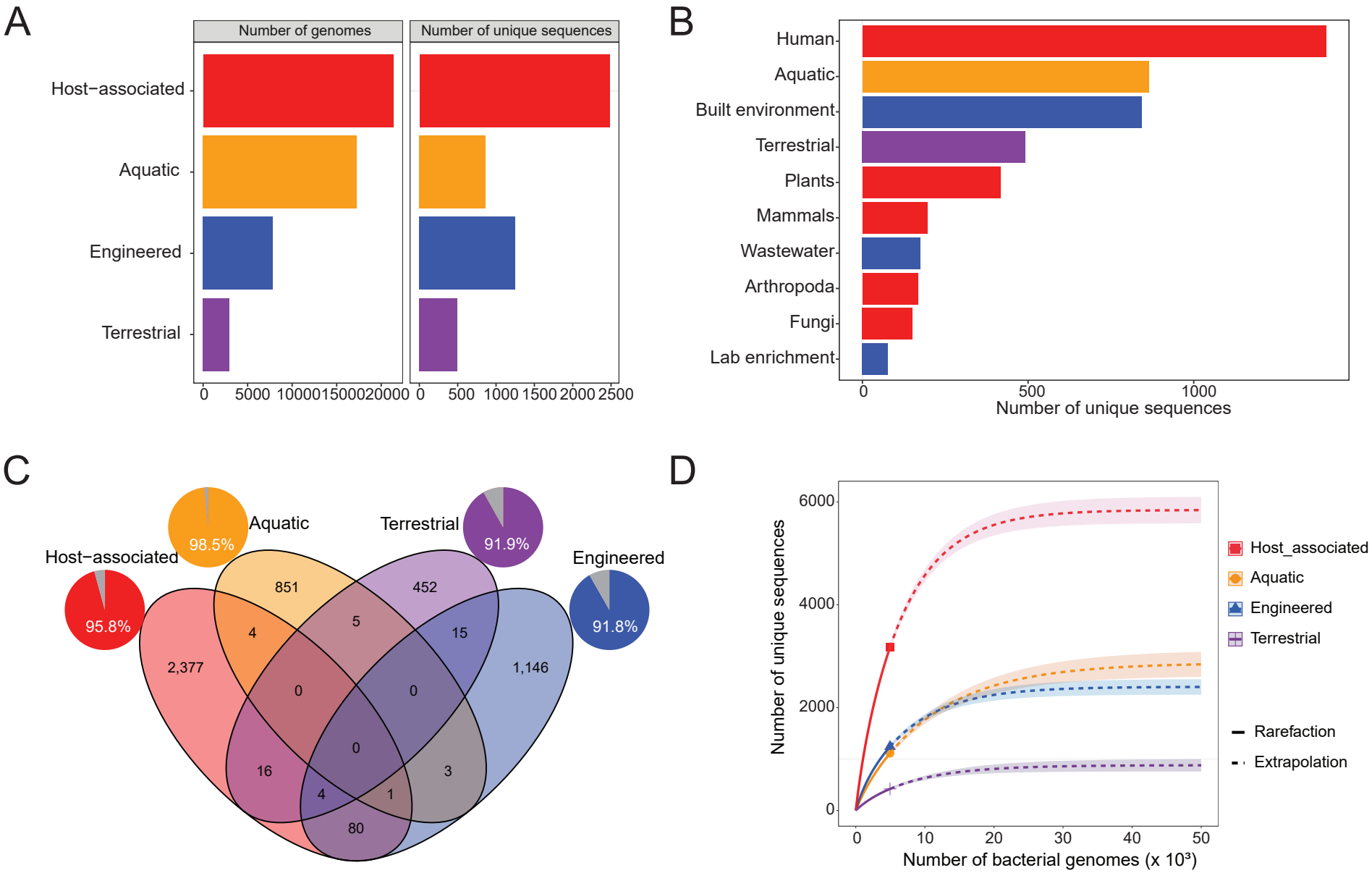
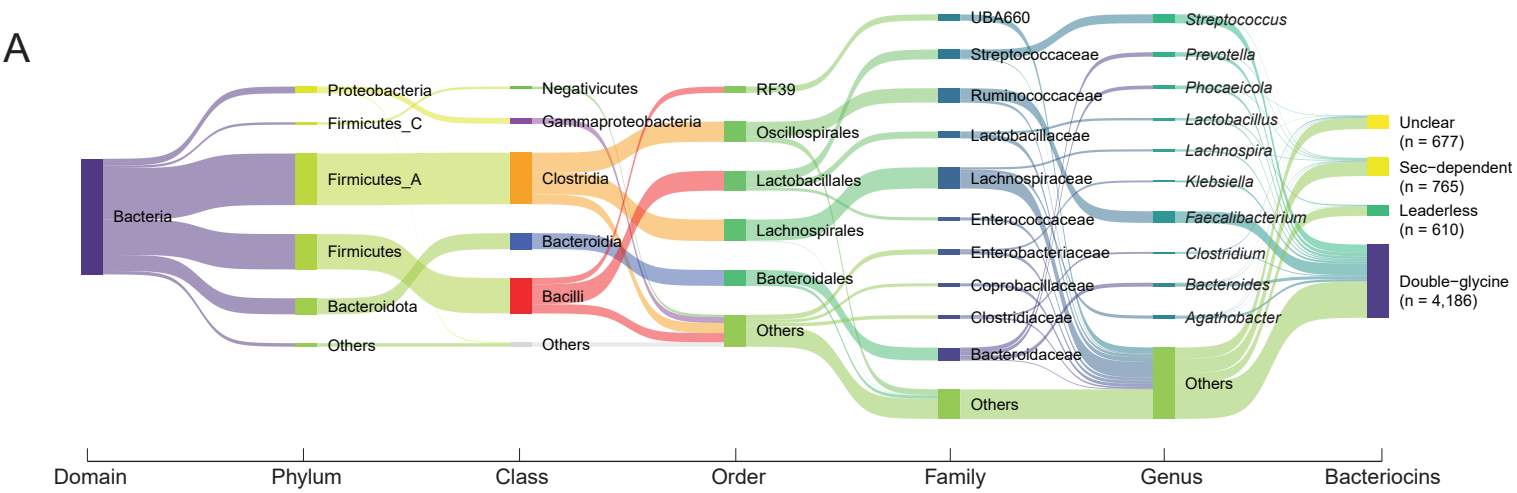

A
Number of genomes
Number of unique sequences
Host−associated
Aquatic
Engineered
Terrestrial
0
5000
10000
15000
20000
0
500
1000
1500
2000
2500
B
Human
Aquatic
Built environment
Terrestrial
Plants
Mammals
Wastewater
Arthropoda
Fungi
Lab enrichment
0
500
1000
Number of unique sequences
C
Host−associated
Aquatic
Terrestrial
Engineered
95.8%
98.5%
91.9%
91.8%
2,377
851
452
1,146
4
5
15
0
0
0
16
3
4
1
80
D
Number of unique sequences
6000
4000
2000
0
0
10
20
30
40
50
Number of bacterial genomes (x 10³)
Host_associated
Aquatic
Engineered
Terrestrial
Rarefaction
Extrapolation

Figure 4



B

	UviB	Mutacin VI	BM1829	Thermophilin 9	Bacteriocin LS2	Enterocin T	SprG1 (short)	Thermophilin 13	Lactiscin	Enterocin P−like	Garvicin C	Bavaricin MN	Pediocin PA−1	coagulin	Ubericin A	Leucocin C	ABP−118	lactacin F	Gallocin A	SprG1 (long)	Number of homologous sequences
Identity > 50%	42	28	25	23	21	17	16	13	13	11	11	10	9	9	8	8	7	7	7	7	379
Identity > 60%	19	22	4	6	3	10	0	1	13	5	4	0	0	0	8	0	7	6	3	0	242
Identity > 70%	7	9	1	1	1	3	0	1	13	5	0	0	0	0	1	0	7	1	2	0	177
Identity > 80%	6	6	1	1	1	0	0	1	6	4	0	0	0	0	0	0	6	1	0	0	154
Identity > 90%	4	4	1	1	1	0	0	1	6	1	0	0	0	0	0	0	6	1	0	0	143

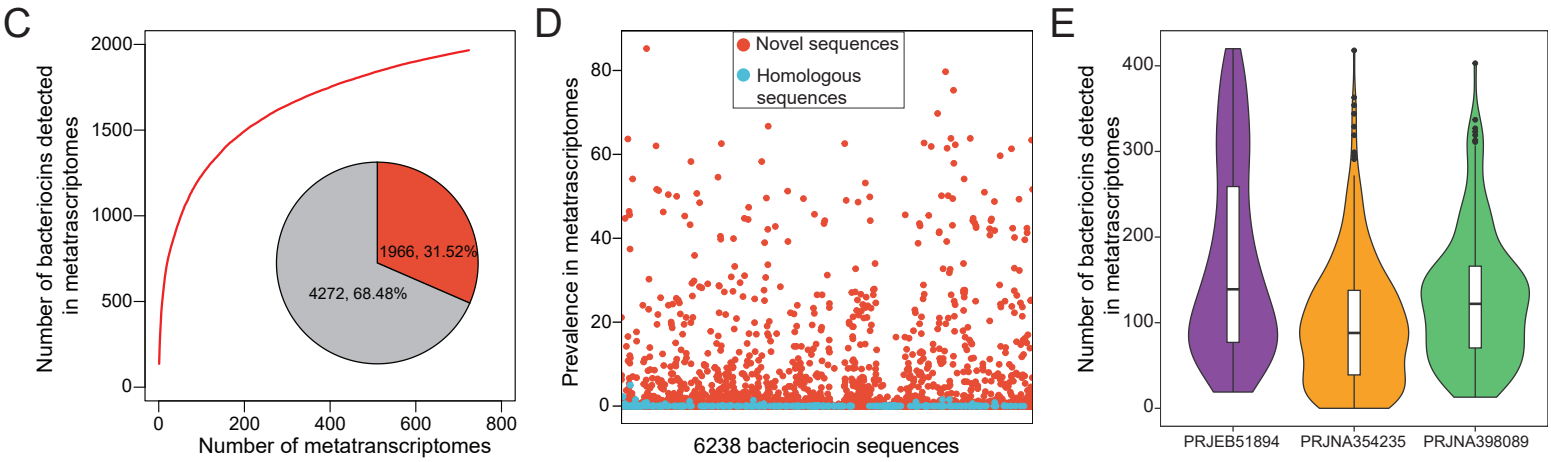
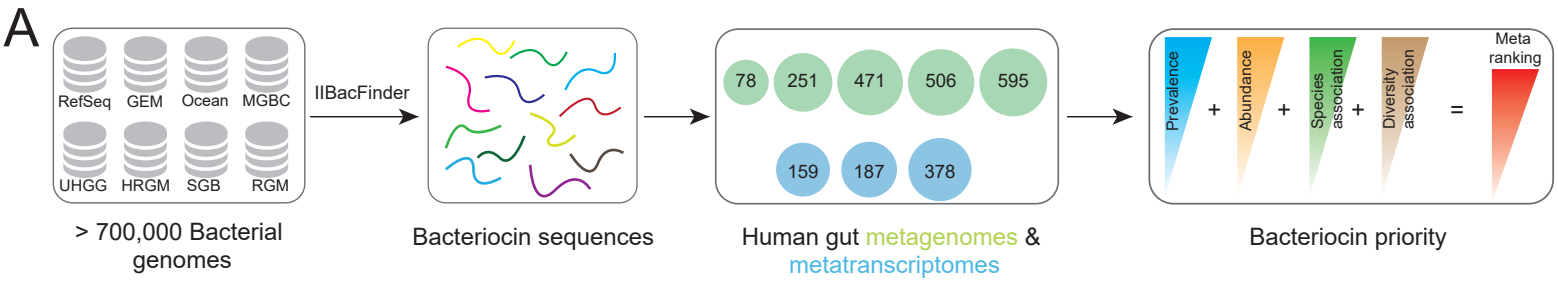
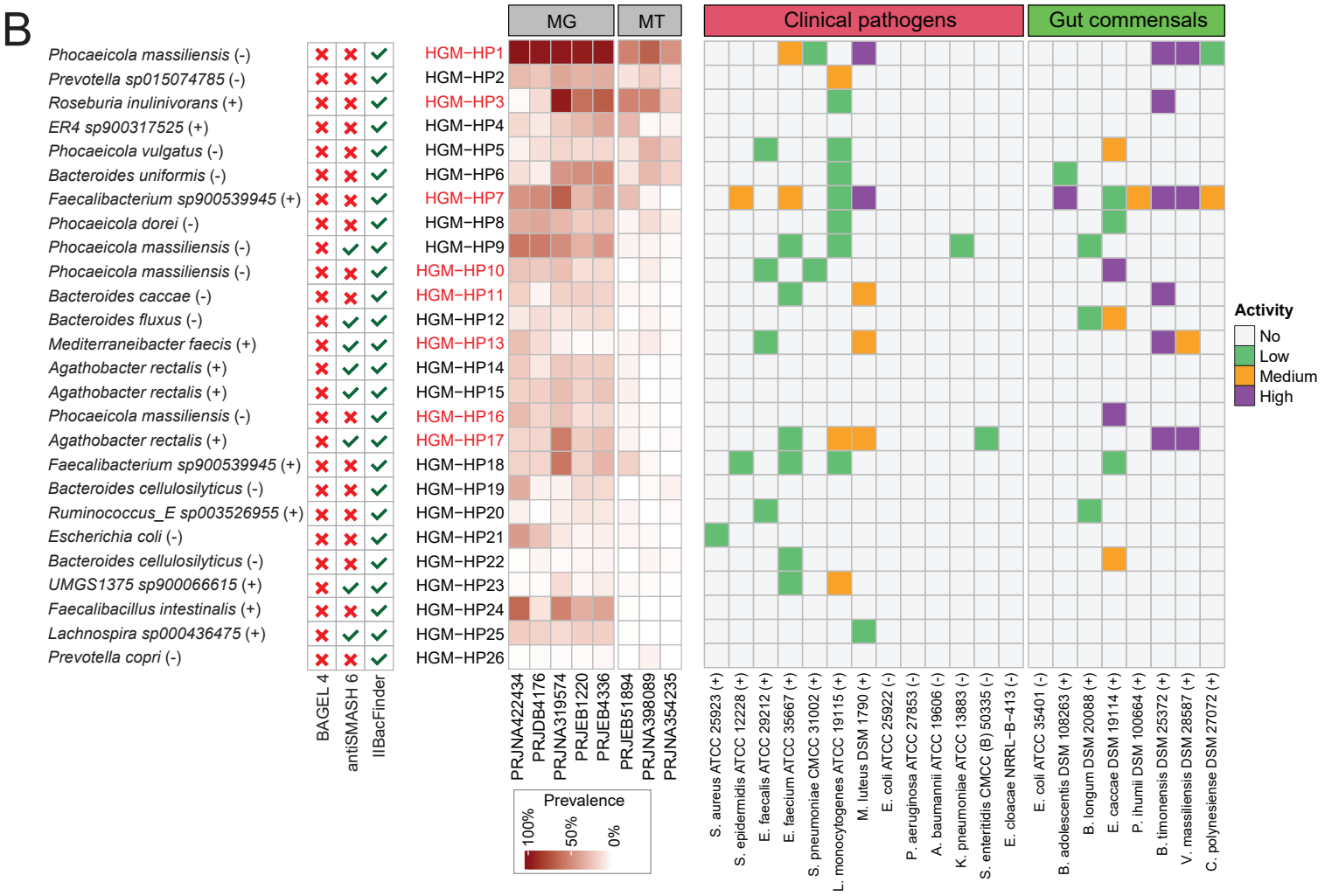


Figure 5


A
RefSeq
GEM
Ocean
MGBC
UHGG
HRGM
SGB
RGM
IIBacFinder
> 700,000 Bacterial genomes
Bacteriocin sequences
78
251
471
506
595
159
187
378
Human gut metagenomes & metatranscriptomes
Prevalence
+
Abundance
+
Species association
+
Diversity association
=
Meta ranking
Bacteriocin priority


B
MG
MT
Clinical pathogens
Gut commensals
Phocaeicola massiliensis (-)
Prevotella sp015074785 (-)
Roseburia inulinivorans (+)
ER4 sp900317525 (+)
Phocaeicola vulgatus (-)
Bacteroides uniformis (-)
Faecalibacterium sp900539945 (+)
Phocaeicola dorei (-)
Phocaeicola massiliensis (-)
Phocaeicola massiliensis (-)
Bacteroides caccae (-)
Bacteroides fluxus (-)
Mediterraneibacter faecis (+)
Agathobacter rectalis (+)
Agathobacter rectalis (+)
Phocaeicola massiliensis (-)
Agathobacter rectalis (+)
Faecalibacterium sp900539945 (+)
Bacteroides cellulosilyticus (-)
Ruminococcus_E sp003526955 (+)
Escherichia coli (-)
Bacteroides cellulosilyticus (-)
UMGS1375 sp900066615 (+)
Faecalibacillus intestinalis (+)
Lachnospira sp000436475 (+)
Prevotella copri (-)
BAGEL 4
antiSMASH 6
IIBacFinder
HGM-HP1
HGM-HP2
HGM-HP3
HGM-HP4
HGM-HP5
HGM-HP6
HGM-HP7
HGM-HP8
HGM-HP9
HGM-HP10
HGM-HP11
HGM-HP12
HGM-HP13
HGM-HP14
HGM-HP15
HGM-HP16
HGM-HP17
HGM-HP18
HGM-HP19
HGM-HP20
HGM-HP21
HGM-HP22
HGM-HP23
HGM-HP24
HGM-HP25
HGM-HP26
PRJNA422434
PRJDB4176
PRJNA319574
PRJEB1220
PRJEB4336
PRJEB51894
PRJNA398089
PRJNA354235
Prevalence
100%
50%
0%
S. aureus ATCC 25923 (+)
S. epidermidis ATCC 12228 (+)
E. faecalis ATCC 29212 (+)
E. faecium ATCC 35667 (+)
S. pneumoniae CMCC 31002 (+)
L. monocytogenes ATCC 19115 (+)
M. luteus DSM 1790 (+)
E. coli ATCC 25922 (-)
P. aeruginosa ATCC 27853 (-)
A. baumannii ATCC 19606 (-)
K. pneumoniae ATCC 13883 (-)
S. enteritidis CMCC (B) 50335 (-)
E. cloacae NRRL-B-413 (-)
E. coli ATCC 35401 (-)
B. adolescentis DSM 108263 (+)
B. longum DSM 20088 (+)
E. caccae DSM 19114 (+)
P. ihumii DSM 100664 (+)
B. timonensis DSM 25372 (+)
V. massiliensis DSM 28587 (+)
C. polynesiense DSM 27072 (+)
Activity
No
Low
Medium
High