Discovery of antimicrobial peptides in the global microbiome with machine learning

Graphical abstract

Highlights
- Machine learning predicts nearly 1 million new antibiotics in the global microbiome
- Out of 100 tested peptides, 79 were active in vitro; 63 of these targeted pathogens
- Some peptides may originate from longer sequences through genomic fragmentation
- The AMPSphere is an open-access resource to accelerate antibiotic discovery

Authors
Célio Dias Santos-Júnior, Marcelo D.T. Torres, Yiqian Duan, ..., Jaime Huerta-Cepas, Cesar de la Fuente-Nunez, Luis Pedro Coelho

Correspondence
cfuente@upenn.edu (C.d.I.F.-N.), luispedro@big-data-biology.org (L.P.C.)

In brief
A machine-learning-based approach predicts nearly one million new antibiotics from the global microbiome, with 79 out of 100 tested peptides being active in vitro and several showing efficacy comparable to a clinical antibiotic in a mouse preclinical model of infection.
Discovery of antimicrobial peptides in the global microbiome with machine learning

Célio Dias Santos-Júnior, Marcelo D.T. Torres, Yiqian Duan, Álvaro Rodríguez del Río, Thomas S.B. Schmidt, Huí Chong, Anthony Fullam, Michael Kuhn, Chengkai Zhu, Amy Houseman, Jelena Somborski, Anna Vines, Xing-Ming Zhao, Peer Bork, Jaime Huerta-Cepas, Cesar de la Fuente-Núñez, and Luis Pedro Coelho

SUMMARY

Novel antibiotics are urgently needed to combat the antibiotic-resistance crisis. We present a machine-learning-based approach to predict antimicrobial peptides (AMPs) within the global microbiome and leverage a vast dataset of 63,410 metagenomes and 87,920 prokaryotic genomes from environmental and host-associated habitats to create the AMPSphere, a comprehensive catalog comprising 863,498 non-redundant peptides, few of which match existing databases. AMPSphere provides insights into the evolutionary origins of peptides, including by duplication or gene truncation of longer sequences, and we observed that AMP production varies by habitat. To validate our predictions, we synthesized and tested 100 AMPs against clinically relevant drug-resistant pathogens and human gut commensals both in vitro and in vivo. A total of 79 peptides were active, with 63 targeting pathogens. These active AMPs exhibited antibacterial activity by disrupting bacterial membranes. In conclusion, our approach identified nearly one million prokaryotic AMP sequences, an open-access resource for antibiotic discovery.

INTRODUCTION

Antibiotic-resistant infections are becoming increasingly difficult to treat with conventional therapies. Indeed, such infections currently kill 1.27 million people per year. Therefore, there is an urgent need for novel methods for antibiotic discovery. Computational approaches have recently been developed to accelerate our ability to identify novel antibiotics, including antimicrobial peptides (AMPs). Recently, proteome mining approaches have even been developed to identify antimicrobial agents in extinct organisms in an attempt to further expand our repertoire of known antimicrobials.
AMPs, found in all domains of life, are short sequences (operationally defined here as 10–100 amino acid residues) capable of disturbing microbial growth. AMPs most commonly interfere with cell wall integrity and cause cell lysis. Natural AMPs can originate by proteolysis, by non-ribosomal synthesis, or, as we focus on in the present study, they can be encoded within the genome.

Bacteria live in an intricate balance of antagonism and mutualism in natural habitats. AMPs play an important role in modulating such microbial interactions and can displace competitor strains, facilitating cooperation. For instance, pathogens such as Shigella spp., Staphylococcus spp., Vibrio cholerae, and Listeria spp. produce AMPs that eliminate competitors (sometimes from the same species), allowing them to occupy their niche.

AMPs hold promise as potential therapeutics and have already been used clinically as antimicrobial agents (e.g., enfuvirtide and telaprevir). AMPs that exhibit immunomodulatory properties are currently undergoing clinical trials. As are peptides that may be used to address yeast and bacterial infections (e.g., pexiganan, LL-37, and PAC-113). Although most AMPs display broad-spectrum activity, some are only active against closely related members of the same species or genus. Such AMPs are more targeted agents than conventional broad-spectrum antibiotics. Furthermore, contrary to conventional antibiotics, the evolution of resistance to many AMPs occurs at low rates and is not related to cross-resistance to other classes of widely used antibiotics.

The application of metagenomic analyses to the study of AMPs has been limited due to technical constraints, primarily stemming from the challenge of distinguishing genuine protein-coding sequences from false positives. Therefore, the significance of small open reading frames (smORFs) has been historically overlooked in (meta)genomic analyses. In recent years, significant progress has been made in metagenomic analyses of human-associated smORFs. These advancements have incorporated machine learning (ML) techniques to identify smORFs encoding proteins belonging to specific functional categories. Notably, a recent study used predicted smORFs to uncover approximately 2,000 AMPs from metagenomic samples of human gut microbiomes. Nevertheless, it is important to note that the human gut represents only a fraction of the overall microbial diversity, suggesting that there remains an immense potential for the discovery of AMPs from prokaryotes in the diverse range of habitats across the globe.

In this study, we employed ML to predict and catalog AMPs from the global microbiome as currently represented in public databases. By computationally exploring 63,410 globally distributed publicly available metagenomes, we uncovered a vast array of AMP diversity. This resulted in the creation of the AMPSphere, a collection of 863,498 non-redundant peptide sequences, encompassing candidate AMPs (c-AMPs) derived from (meta)genomic data. Remarkably, the majority of these c-AMP sequences had not been previously described. Our analysis revealed that these c-AMPs were specific to particular habitats and were predominantly not core genes in the pan-genome.

Moreover, we synthesized 100 c-AMPs from AMPSphere and found that 79 were active, with 63 exhibiting antimicrobial activity in vitro against clinically significant ESKAPEE pathogens, which are recognized as public health concerns. These peptides were further compared to encrypted peptides (EPs), which are peptide sequences hidden in protein sequences and mined computationally, and demonstrated their ability to target bacterial membranes and their propensity to adopt a-helical and b-structures. Notably, the leading candidates displayed promising anti-infective activity in a preclinical animal model. Together, our work demonstrates the ability of ML approaches to identify functional AMPs from the global microbiome.

RESULTS

AMPSphere comprises almost 1 million c-AMPs from several habitats

AMPSphere incorporates c-AMPs predicted with ML using Macrel, a pipeline that uses random forests to predict AMPs from large peptide datasets with an emphasis on precision over recall. It was applied to 63,410 globally distributed publicly available metagenomes (Figure 1A; Table S1) and 87,920 high-quality bacterial and archaeal genomes. Sequences present in a single sample were removed, except when they had a significant match (defined as amino acid identity ≥75% and E-value ≤10⁻⁵) to a sequence in the AMP-dedicated database Data Repository of Antimicrobial Peptides (DRAMP) version 3.0. This resulted in 5,518,294 genes, 0.1% of the total predicted smORFs, coding for 863,498 non-redundant c-AMPs (on average 37 ± 8 residues long; Figures 1A and S1). Similar to validated sequences with antimicrobial activity, c-AMPs from AMPSphere present a positive charge (4.7 ± 2.6), high isoelectric point (10.9 ± 1.2), amphipathic (hydrophobic moment, 0.6 ± 0.1), and a potential to bind to membranes or other proteins. As expected, in general, the distribution of physicochemical properties of peptides from AMPSphere, DRAMP version 3.0, and the positive training dataset used in Macrel are more similar to each other than to the negative training set (assumed to not be AMPs). Nonetheless, c-AMPs from AMPSphere are on average longer (37 ± 8 residues) than those in DRAMP version 3.0 (28 ± 22 residues), and we observed differences in the distribution of other features (e.g., charge, aliphaticity, amphipathicity, and isoelectric point; Figure S1).

We subsequently estimated the quality of the smORF predictions and detected 20% (172,840) of the c-AMP sequences in independent publicly available metaproteomes or metascriptomes (Figures 2 and S2A; see STAR Methods section “Quality control of c-AMPs”) belonging to several habitats included in the AMPSphere, such as the human gut, plants, and others (Table S6). We then subjected all c-AMPs to a bundle of in silico quality tests (see STAR Methods section “Quality control of c-AMPs”). A subset of c-AMPs (9.2% or 80,213 c-AMPs) passed all of them, and this subset is hereafter designated as high-quality. Testing with other AMP prediction systems (AMPScanner v2, the model for mature peptides in ampir, ampEPPpy, APIN, AI4AMP, and AMPLib 57), we observed that 98.4% (849,703 peptides) of AMPSphere c-AMPs were also predicted as AMPs by at least one other AMP prediction system. Approximately 15% (132,440 out of 863,498 peptides) of AMPSphere c-AMPs were co-predicted by all methods used.
Only 0.7% of the identified c_AMPs (6,339 peptides) are homologous (operationally defined as amino acid identity ≥ 75% and E-value ≤ 10⁻⁵) to experimentally validated AMP sequences in DRAMP version 3.0. Moreover, most c_AMPs were also absent from protein databases not specific to AMPs (Figure 1B), such as the Small Proteins database (SmProt2), bioactive peptide databases (DRAMP version 3.0, starPepDB 45k), and general protein datasets (GMGCv1; see also Figure S2B). Also shown is the number of homologs in the AMPSphere in each database as well as the total. The number of homologs passing all of our quality tests regardless of their experimental evidence of translation/transcription is also shown along with the percentage it represents in the homologs identified. Note that some peptides have homologs in multiple databases and thus the total count is not the sum of the individual databases.

(C) Shown are rarefaction curves showing how AMP discovery is impacted by sampling, with most of the habitats presenting steep sampling curves.

(D) Sharing of c_AMPs between habitats is limited. The width of the ribbons represents the proportion of the shared c_AMPs in the habitat on the left. See also Figures S2C and S2D and Tables S1 and S2.

Only 9% of c_AMPs have detectable homologs in other small protein databases (SmProt2, STsORFs), bioactive peptide databases (DRAMP version 3.0, starPepDB 45k), and general protein datasets (GMGCv1; see also Figure S2B). Also shown is the number of homologs in the AMPSphere in each database as well as the total. The number of homologs passing all of our quality tests regardless of their experimental evidence of translation/transcription is also shown along with the percentage it represents in the homologs identified. Note that some peptides have homologs in multiple databases and thus the total count is not the sum of the individual databases.

(C) Shown are rarefaction curves showing how AMP discovery is impacted by sampling, with most of the habitats presenting steep sampling curves.

(D) Sharing of c_AMPs between habitats is limited. The width of the ribbons represents the proportion of the shared c_AMPs in the habitat on the left. See also Figures S2C and S2D and Tables S1 and S2.
peptide sequence space that is not present in these other databases. In total, we could find only 73,774 (8.5%) c_AMPs with homologs in any of the databases we considered. High-quality c_AMPs were detected in public databases at a higher frequency than general c_AMPs (2.5-fold, \( P_{\text{hypergeometric}} = 4.2 \times 10^{-250}\). Figure 1B), with 23,012 out of the 80,213 high-quality c_AMPs having a match in another database. However, it is notable that 76.4% (4,843 peptides out of 6,339) of those c_AMPs that have a homolog in DRAMP\(^{46}\) version 3.0 (and, therefore, are highly likely to be functional) are not high-quality c_AMPs. Thus, while our quality tests do enrich for validated sequences, a failure to pass the tests is not a sufficient reason to conclude that the sequence is not active.

To put c_AMPs in an evolutionary context, we hierarchically clustered peptides using a reduced amino acid alphabet of 8 letters.\(^5^8\) The three sequence clustering levels adopted identity cutoffs of 100%, 85%, and 75% (Figure S3). At the 75% identity level, we obtained 521,760 protein clusters, of which 405,547 were singletons, corresponding to 47% of all c_AMPs from AMPSphere. A total of 78,481 (19.3%) of these singletons were detected in metatranscriptomes or metaproteomes from various habitats; Figure 1D). This fraction is 7.25-fold smaller than would be expected by a random assignment of habitats to samples (\( P_{\text{Permutation}} < 10^{-50}\); see STAR Methods section “Testing c_AMPs overlap across habitats”; Figure 1D).

Mutations in larger genes generate c_AMPs as independent genomic entities

Many c_AMPs are rare and habitat-specific

The AMPSphere spans 72 different habitats, which were classified into eight high-level habitat groups, e.g., soil/plant (36.6% of c_AMPs in AMPSphere), aquatic (24.8%), and human gut (13%; Figure 1A; Table S2). Most of the habitats, except for the human gut, appear to be far from saturated in terms of discovered c_AMPs (Figure 1C). In fact, most AMPs are rare (median number of detections is 99, or 0.17% of the dataset; when restricted to high-quality c_AMPs, the median number of detections is 81, or 0.14% of the dataset), with 83.97% being observed in <1% of samples (Figure S2). Only 10.8% (93,280) of c_AMPs were detected in more than one high-level habitat group (henceforth termed “multi-habitat c_AMPs”); this fraction is 2.4–192-fold less than expected by chance (2.4–192-fold less, \( P_{\text{Permutation}} < 5.4 \times 10^{-50}\); see STAR Methods section “Testing c_AMPs overlap across habitats”; Figure 1D).

Figure 2. Quality control of AMPSphere candidates

(A) The number of AMPSphere candidates passing each of the tests proposed for quality is shown. The high-quality set is composed of 7.3% of candidates without experimental evidence and 2% of candidates with evidence of their translation or transcription, as well as the number of homologs found in the high-quality set of AMP candidates. Although the high-quality set displays some overlap with the homologs, most of the homologs are not found in the high-quality set.

(B) The number of AMP candidates co-predicted by AMP prediction systems beyond Macrel (AMPS-canner v2,\(^{53}\) ampir\(^{40}\) with the model for mature peptides, amPEPpy\(^{34}\) APIN\(^{55}\) with their proposed model, A4AMP,\(^{32}\) and AMPlyt\(^{51}\) ). Only a small portion of AMPSphere (<2%) cannot be co-predicted by any system other than Macrel.\(^5^2\)

All the c_AMPs predicted here can be accessed at https://ampsphere.big-data-biology.org/\. Users can retrieve the peptide sequences, ORFs, and predicted biochemical properties of each c_AMP (e.g., molecular weight, isoelectric point, and net charge at pH 7.0). We also provide the distribution across geographical regions, habitats, and microbial species for each c_AMP.
Figure 3. Mutations in genes encoding large proteins generate c_AMPs as independent genomic entities

(A) The distribution of positions (as a percentage of the length of the larger protein) from which the AMP homologs start their alignment is shown. About 7% of c_AMPs are homologous to proteins from GMGCv1.52 with approximately one-fourth of the hits having the same start position as the larger protein.

(B) As an illustrative example of an AMP homologous to a full-length protein, AMP10.271_016 was recovered from three samples of human saliva from the same donor.59 AMP10.271_016 is predicted to be produced by Prevotella jejuni, sharing the start codon (bolded) of an NAD(P)-dependent dehydrogenase gene (WP_089365220.1), the transcription of which was stopped by a mutation (in red; TGG > TGA).

(C) The distribution of AMPs per OG class (left) and their enrichment in comparison to full-length proteins from GMGCv1.52 (right). OGs were classified into subgroups according to the number of c_AMPs they were affiliated with. The OGs of unknown function represent the largest (2,041 out of 3,792 OGs) and most enriched ($p_{Kruskal} = 2.66 \times 10^{-3}$) class with homologs to c_AMPs in GMGCv1.52 Interestingly, when considered individually, the number of c_AMP hits to unknown OGs was the lowest ($p_{Kruskal} = 6 \times 10^{-3}$). These results do not change when underrepresented OGs are excluded by using different thresholds (e.g., at least 10, 20, or 100 homologs per OG). See also Table S3.
To investigate the function of the full-length proteins homologous to AMPs, we mapped the matching proteins from GMGCv1.52 to orthologous groups (OGs) from eggNOG 5.0.60 We identified 3,792 (out of 43,789) OGs significantly enriched (Hypergeom. < 0.05, after multiple hypothesis corrections with the Holm-Sidak method) among the hits from AMPSphere. Although OGs of unknown function comprise 53.8% of all identified OGs, when considered individually, these OGs are on average smaller than OGs in other categories. Thus, despite each OG having a relatively small number of c_AMP hits, when compared to the background distribution of the OGs in GMGCv1.52 OGs of unknown function were the most enriched among the c_AMP hits, with an average enrichment of 10,857-fold ($p_{\text{mann}} \leq 3.9 \times 10^{-16}$; Figure 3C; Table S3).

**c_AMP genes may arise after gene duplication events**

We next raised the question of whether c_AMPs would be predominantly present in specific genomic contexts. To investigate the functions of the neighboring genes of the c_AMPs, we mapped them against 169,484 genomes included in a recent study.61 A total of 38.9% (21,465 out of 55,191) of c_AMPs with more than two homologs in different genomes in the database showed phylogenetically conserved genomic context with genes of known function (see STAR Methods section “Genomic context conservation analysis”). This holds true for curated versions of the catalog: 35.32% of high-quality c_AMPs and 32.06% of high-quality c_AMPs with experimental evidence show conserved genomic neighbors. These conservation values are similar to that of 3,899,674 gene families with more than two homologs calculated de novo on the gene catalog (34.4%), indicating that the genomic location of c_AMPs is not random.

Despite being involved in similar processes, c_AMPs were generally depleted from conserved genomic contexts involving known systems of antibiotic synthesis and resistance, even when compared to small protein families (Figure 4). Instead, we found that c_AMPs are encoded in conserved genomic contexts with ribosomal genes (23.6%) at a higher frequency than other gene families (4.75%; Figure 4A; Table S4).

Most of the c_AMPs (2,201 out of 2,642) in a conserved context with ribosomal subunits are homologous to ribosomal proteins (Figure 4D), congruent with the observation that in some species, ribosomal proteins have antimicrobial properties.62 Seventy-seven c_AMPs homologous to ribosomal proteins were also homologous to a ribosomal gene in their immediate vicinity (up to 1 gene up/downstream). This phenomenon is not exclusive to ribosomal proteins: 1,951 c_AMPs can be annotated to the same KEGG Orthologous Group (KO) as some of their immediate neighbors and may have originated from gene duplication events. This shared annotation was interpreted in this context as evidence for a common evolutionary origin and not as a functional prediction for the c_AMPs. These duplications may have arisen by recombination of flanking homologous sequences, which can happen during cell division.53-65 Interestingly, 1,635 (83.8%) of these c_AMPs are located upstream of the neighbor with the same KO annotation. Different permeases and transposases are the most common KOs assigned to c_AMPs and their neighbors (400 and 125 c_AMPs, respectively; see Table S5).

**Most c_AMPs are members of the accessory pangenome**

We observed that only a small portion (5.9%, $p_{\text{Permutation}} = 4.8 \times 10^{-3}$, $N_{\text{species}} = 416$) of c_AMP families present in Pro-Genomes19-21 are contained in ≥95% of genomes from the same species (Figure 5), here referred to as “core.”56 This is consistent with previous work, in which AMP production was observed to be strain-specific.63 In contrast, a high proportion (circa 68.8%) of full-length protein families are core in Pro-Genomes19-21 species. There is a 1.9-fold greater chance ($p_{\text{Fisher}} = 2.2 \times 10^{-56}$) that a pair of genomes from the same species share at least one c_AMP when they belong to the same strain (99.5% ≤ ANI <99.99%).

One example of this strain-specific behavior is AMP10.018_194, the only c_AMP found in Mycoplasma pneumoniae genomes. M. pneumoniae strains are traditionally classified into two groups based on their P1 adhesin gene.70 Of the 76 M. pneumoniae genomes present in our study, 29 were classified as type-1, 29 were classified as type-2, and the remaining 18 were undetermined in this classification system71 (see STAR Methods section “Determination of accessory AMPs”). Twenty-six of the 29 type-2 genomes contain AMP10.018_194, as did 2 undetermined type genomes, but none of the type-1 genomes contain this AMP.

**More transmissible species have lower c_AMP density**

We investigated the taxonomic composition of AMPSphere by annotating contigs with the Genome Taxonomy Database (GTDB) taxonomy68,69 (see STAR Methods section “c_AMP density in microbial species”), which resulted in 570,187 c_AMPs being annotated to a genus or species. The genera contributing the most c_AMPs to AMPSphere were Prevotella (18,593 c_AMPs), Bradyrhizobium (11,846 c_AMPs), Pelagibacter (6,675 c_AMPs), Faecalibacterium (5,917 c_AMPs), and CAG-110 (5,254 c_AMPs; see Figure 5). This distribution reflects the fact that these genera are among those that contribute the most assembled sequences in our dataset (all occupying percentiles above 99.75% among the assembled genera). Therefore, we calculated the c_AMP density ($\rho_{\text{AMP}}$) by determining the number of c_AMP genes per megabase pairs of assembled sequence. To avoid bias due to the unequal sampling of habitats, we included all the sequences predicted by Macrel62 in each sample, including singleton sequences that were subsequently removed and are not part of AMPSphere.

To further explore the importance of AMP production in ecological processes, we investigated the role of AMPs in the mother-to-child transmissibility of bacterial species in a recently published paper72 by correlating the $\rho_{\text{AMP}}$ for each bacterial species to the published measures of microbial transmission. Human gut bacteria showed increased transmissibility at lower AMP densities (RSpearman = −0.42, $p_{\text{Holm-Sidak}} = 3.4 \times 10^{-2}$, $N_{\text{species}} = 43$). Similarly, in human oral microbiome bacterial species, transmissibility from mother to offspring is consistently inversely correlated with their $\rho_{\text{AMP}}$ for the first year (RSpearman = −0.55, $p_{\text{Holm-Sidak}} = 1.4 \times 10^{-3}$, $N_{\text{species}} = 41$). This suggests that human gut bacteria and oral microbiome bacterial species show increased transmissibility at lower $\rho_{\text{AMP}}$. Moreover, it highlights the potential influence of $\rho_{\text{AMP}}$ on the transmissibility of gut and
oral microbiota, suggesting a link between AMPs and the transmission success rates of microbial species.

Physicochemical features and secondary structure of AMPs

To investigate the properties and structure of the synthesized peptides, we first compared their amino acid composition to AMPs from available databases of experimentally verified sequences (DRAMP46 version 3.0, Database of Antimicrobial Activity and Structure of Peptides [DBAASP], 73 and Antimicrobial Peptides Database [APD] 74 version 3). Overall, the composition was similar, as was expected, given that Macrel’s ML model was trained using known AMPs.42 Notably, AMPSphere sequences displayed a slightly higher abundance of aliphatic amino acid residues, specifically alanine and valine. However, these AMPSphere sequences consistently differed (Figure 6A) from EPs. 4,10,33 The resemblances in amino acid composition between the identified c_AMPs and known AMPs suggested similar physicochemical characteristics and secondary structures, both of which are recognized for their influence on antimicrobial activity.16 The c_AMPs exhibited comparable hydrophobicity, net charge, and amphiphilicity to AMPs sourced from databases (Figure S1). Furthermore, they displayed a slight propensity for disordered conformations (Figure 6B) and had a lower net positive charge compared to other EPs (Figure 6A).

To evaluate the structural and antimicrobial properties of c_AMPs from AMPSphere, we first filtered the AMPSphere for peptides that were predicted as suitable for in vitro assays due to their solubility in aqueous solution and ease of chemical synthesis. We chose a set of high-quality AMPs with 50 peptide sequences based on their prevalence and taxonomic diversity (see STAR Methods section “Peptide selection for synthesis and testing”). Additionally, to provide an unbiased evaluation of the peptides we report here, we first excluded any peptides with a homolog in one of the published databases and then randomly selected 50 additional peptides from the AMPSphere, including 25 peptides with AMP probabilities of at least 0.6 (as reported by Macrel42) and 25 peptides with lower probabilities (0.5–0.6).

Subsequently, we conducted experimental assessments of the secondary structure of the active c_AMPs using circular dichroism (Figures 6B and S4). Similar to AMPs documented in databases, peptides derived from AMPSphere exhibited different
propensities for adopting α-helical structures; also, some of them were unstructured or adopted β-antiparallel conformations in all media analyzed. Notably, they also displayed an unusually high content of β-antiparallel structures in both water and methanol/water mixtures (Figure 6B) despite their amino acid composition similarities to AMPs and EPs. We attribute these findings to the slightly elevated occurrence of alanine and valine residues, which are known to favor β-like structures with a preference for β-antiparallel conformation.76

Validation of c_AMPs as potent antimicrobials through in vitro assays
Next, we tested the 100 synthesized peptides against 11 clinically relevant pathogenic strains encompassing Acinetobacter baumannii, Escherichia coli (including one colistin-resistant strain), Klebsiella pneumoniae, Pseudomonas aeruginosa, Staphylococcus aureus (including one methicillin-resistant strain), vancomycin-resistant Enterococcus faecalis, and vancomycin-resistant Enterococcus faecium. Our initial screening revealed that 63 AMPs (out of 100 synthesized) completely eradicated the growth of at least one of the pathogens tested (Figure 6C). Remarkably, in some cases, the AMPs were active at concentrations as low as 1 μmol L⁻¹, close to the peptide antibiotic polymyxin B and the antibiotic levofloxacin that were used as positive controls in all experiments (Figure S4A). The Gram-negative bacteria A. baumannii and E. coli, as well as the Gram-positive vancomycin-resistant strains E. faecalis and E. faecium, displayed higher susceptibility to the AMPs, with 39, 24, 21, and 26 peptide hits, respectively. However, none of the tested AMPs affected methicillin-resistant S. aureus (MRSA) (Figure 6C). We also synthesized and tested the scrambled versions of five of the most active peptides from the high-quality group for antimicrobial activity (i.e., actinomycin-1, enterococcin-1, lachnospirin-1, proteobactin-1, and synechocucin-1). All scrambled versions were inactive except for lachnospirin-1_scrambled, which presented modest activity against A. baumannii at 32 μmol L⁻¹ (16 times higher concentration compared to its parent peptide lachnospirin-1; Figure S5A). These results underscore the importance of the specific sequence of these peptides to exert their antimicrobial activity. To further explore the influence of sequence on
Figure 6. Amino acid composition, structure, antimicrobial activity, and mechanism of action of c_AMPs
(A) Amino acid frequency in c_AMPs from AMPSphere, AMPs from databases (DRAMP version 0, APD3, 0, and DBAASP 0), and encrypted peptides (EPs) from the human proteome.

(B) Heatmap with the percentage of secondary structure found for each peptide in three different solvents: water, 60% trifluoroethanol (TFE) in water, and 50% methanol (MeOH) in water. Secondary structure was calculated using BeStSel server. 75

(C) Activity of c_AMPs assessed against ESKAPEE pathogens and human gut commensal strains. Briefly, 10^6 CFU mL^{-1} was exposed to c_AMPs 2-fold serially diluted ranging from 64 to 1 μmol L^{-1} in 96-well plates and incubated at 37 °C for one day. After the exposure period, the absorbance of each well was measured at 600 nm. Untreated solutions were used as controls, and minimal concentration values for complete inhibition were presented as a heatmap of antimicrobial activities (μmol L^{-1}) against 11 pathogenic and eight human gut commensal bacterial strains. All the assays were performed in three independent replicates, and the heatmap shows the mode obtained within the 2-fold dilution concentration range studied. Gram-positive (+) and Gram-negative (−) bacteria are indicated as such (top).

(legend continued on next page)
structure, we assessed the secondary structure tendency of the scrambled peptides using circular dichroism. We noticed a decrease in helical fraction for sequences with higher helical content (enterococcin-1, lachnospirin-1, and synechocucin-1), while the predominately random coiled sequences actinomycin-1 and proteobactin-1, as well as their scrambled counterparts, showed similar secondary structural sequences in all media analyzed (Figures S5B–S5E). These results suggest a lack of correlation between secondary structure and antimicrobial activity of the AMPs derived from AMPSphere.

The growth of human gut commensals is impaired by c_AMPs

We screened the AMPs against eight of the most relevant members of the human gut microbiota associated with human health.77–81 We tested commensal bacteria belonging to four phyla (Verrucomicrobiota, Bacteriodota, Actinomycetota, and Bacillota), i.e., Akkermansia muciniphila, Bacteroides fragilis, Bacteroides thetaiotaomicron, Bacteroides uniformis, Phocaeicola vulgaris (formerly Bacteroides vulgaris), Collinsella aerofaciens, Clostridium scindens, and Parabacteroides distasonis.

While it is commonly observed that known natural AMPs do not target microbiome strains,82 our study found that 58 of the synthesized AMPs (58%) demonstrated inhibitory effects on at least one commensal strain at low concentrations (8–16 μmol L⁻¹). Although this concentration range was higher than that observed for the most active peptides against pathogens (1–4 μmol L⁻¹), it still falls within the highly active range of AMPs based on previous studies.83–85 (Figure 6C). Interestingly, all the analyzed gut microbiome strains were susceptible to at least four c_AMPs, with strains of A. muciniphila, B. uniformis, P. vulgaris, C. aerofaciens, C. scindens, and P. distasonis exhibiting the highest susceptibility. In total, 79 AMPs (out of 100 synthesized peptides) demonstrated antimicrobial activity against pathogens and/or commensals. We also screened scrambled sequences of five of the highly active peptides from the high-quality group against gut commensals. Similarly to the results obtained against pathogenic strains (Figure S5), only lachnospirin-1_scrambled was modestly active against C. scindens at 64 μmol L⁻¹ (Figure S5A).

Permeabilization and depolarization of the bacterial membrane by c_AMPs from AMPSphere

To gain insights into the mechanism of action responsible for the antimicrobial activity observed in the peptides derived from AMPSphere (Figure 6C), we conducted experiments to assess their ability to permeabilize and depolarize the outer and cytoplasmic membranes of bacteria at their minimum inhibitory concentrations (MICs). Specifically, we investigated the effects of all 39 peptides that showed activity against A. baumannii (Figures 6D and 6E) and 6 peptides with antimicrobial activity on P. aeruginosa (Figures S6A and S6B). For comparison and as a control, we used polymyxin B, a peptide antibiotic known for its membrane permeabilization and depolarization properties.4

To investigate the potential permeabilization of the outer membranes of Gram-negative bacteria by the selected AMPs, we conducted 1-(N-phenylamino)naphthalene (NPN) uptake assays. NPN is a lipophilic fluorophore that exhibits increased fluorescence in the presence of lips found within bacterial outer membranes. The uptake of NPN indicates membrane permeabilization and damage. Among the 39 peptides evaluated for activity against A. baumannii, 10 peptides caused significant permeabilization of the outer membrane, resulting in fluorescence levels at least 50% higher than that of polymyxin B (Figure 6D) after 45 min of exposure. In the case of P. aeruginosa cells, four out of the six tested peptides showed higher permeabilization than polymyxin B (Figure S6A).

To evaluate the potential membrane depolarization effect of the selected AMPs from AMPSphere, we utilized the fluorescent dye 3,3’-dipropylthiadicarbocyanine iodide (DiSC₃[5]). Among the peptides tested against A. baumannii, bogcin-1 (AMP10.364_543), ampsperhin-2 (AMP10.615_023), and marinobactin-1 (AMP10.321_460) exhibited greater cytoplasmic membrane depolarization than polymyxin B, and among the ones tested against P. aeruginosa, all peptides tested exhibited greater cytoplasmic membrane depolarization than polymyxin B (Figure 6B). Interestingly, all the tested AMPSphere peptides displayed a characteristic crescent-shaped depolarization pattern compared to polymyxin B, with lower levels of depolarization observed during the first 20 min of exposure followed by an increase in depolarization over time (Figures 6E and S6B). Taken together, these results indicate that the kinetics of cytoplasmic membrane depolarization are slower compared to the kinetics of outer membrane permeabilization, which occurs rapidly upon interaction with the bacterial cells.

Our findings indicate that the tested AMPs from AMPSphere primarily exert their effects by permeabilizing the outer membrane rather than depolarizing the cytoplasmic membrane, revealing a similar mechanism of action to that observed for classical AMPs and EPs from the human proteome.3

AMPs exhibit anti-infective efficacy in a mouse model

Next, we tested the anti-infective efficacy of AMPSphere-derived peptides in a skin abscess murine infection model (Figure 7A). Mice were subjected to infection with A. baumannii, a dangerous Gram-negative pathogen known for causing severe infections in various body sites including the bloodstream, lungs, urinary tract, and wounds.86 Ten lead AMPs (A. baumannii: synechocucin-1 (AMP10.000_211, 8 μmol L⁻¹) from Synechococcus sp. (coral-associated, marine microbiome); proteobactin-1 (AMP10.048_551, 16 μmol L⁻¹) from Pseudomonadota (plant and soil microbiome); actinomyacin-1 (AMP10.199_072, 64 μmol L⁻¹) from Actinomycetes (human mouth and saliva

(D) Fluorescence values relative to polymyxin B (PMB, positive control) of the fluorescent probe 1-(N-phenylamino)naphthalene (NPN) that indicate outer membrane permeabilization of A. baumannii ATCC 19606 cells.

(E) Fluorescence values relative to PMB (positive control) of 3,3’-dipropylthiadicarbocyanine iodide (DiSC₃[5]), a hydrophobic fluorescent probe used to indicate cytoplasmic membrane depolarization of A. baumannii ATCC 19606 cells. Depolarization of the cytoplasmic membrane occurred with slow kinetics compared to the permeabilization of the outer membrane and took approximately 20 min to stabilize.
microbiome); lachnospirin-1 (AMP10.015_742, 2 μmol L⁻¹) from Lachnospira sp. (human gut microbiome); enterococcin-1 (AMP10.051_911, 1 μmol L⁻¹) from Enterococcus faecalis (human gut microbiome); alphaprotecin-1 (AMP10.316_798, 1 μmol L⁻¹) from Alphaproteobacteria (aquatic microbiome); oscillospirin (AMP10.771_988, 8 μmol L⁻¹) from Oscillospiraceae (pig gut microbiome); ampsherin-4 (AMP10.466_287, 8 μmol L⁻¹) from an unknown source; methylcellin-1 (AMP10.446_571, 2 μmol L⁻¹) from Methylocella sp. (soil microbiome); and reyranin-1 (AMP10.337_875, 16 μmol L⁻¹) from Reyranella (plant and soil microbiome). The skin abscess infection was established with a bacterial load of 20 mLo of A. baumannii cells at 10⁶ colony-forming units (CFUs) mL⁻¹ onto the wounded area of the dorsal epidermis (Figure 7A). A single dose of each peptide at their respective MIC value obtained in vitro (Figures 6C and S4A) was administered to the infected area. Two days post-infection, synechocycin-1, actynomycin-1, and oscillospirin-1 presented bacteriostatic activity, inhibiting the proliferation of A. baumannii cells, whereas lachnospirin-1, enterococcin-1, ampsherin-4, and reyranin-1 presented bactericidal activity close to that of the antibiotic polymyxin B (at 5 μmol L⁻¹), reducing the CFU counts by 3–4 orders of magnitude (Figure 7B). Four days post-infection, synechocycin-1, lachnospirin-1, enterococcin-1, and ampsherin-4 presented a bacteriostatic effect close to that of the antibiotic polymyxin B, reducing the CFU counts by 2–3 orders of magnitude compared to the untreated control (Figure S6C). These results highlight the anti-infective potential of the tested peptides from AMPSphere as they were administered at a single time immediately after the establishment of the abscess. Mouse weight was monitored as a proxy for toxicity, and no significant changes were observed (Figures 7C and S6D), suggesting that the peptides tested were not toxic.

DISCUSSION

Here, we used ML to identify nearly a million candidate AMPs in the global microbiome. Building on previous studies that focused specifically on the human gut microbiome,6,38,87 we cataloged AMPs from the global microbiome across 63,410 publicly available metagenomes as well as 87,920 high-quality microbial genomes from the ProGenomes2 database.42 This led to the creation of AMPSphere (https://ampsphere.big-data-biology.org/), an open-access and publicly available resource encompassing 863,498 non-redundant peptides and 6,499 high-quality AMP families from 72 different habitats, including marine and soil environments and the human gut. Most of the AMPs (91.5%) were previously unknown and lacked detectable homologs in other databases, and about one in five had evidence of translation and/or transcription, as they could be detected in independent publicly available sets of metatranscriptomes or metaproteomes.

We designed a set of tests to capture higher-quality predictions, but many peptides failed these tests despite evidence...
that they were active, including our own in vitro data and the existence of validated homologs in external databases. Low-prevalence peptides will be less likely to pass the tests (RNAcode requires multiple variants), which is independent of their activity and influenced by sampling biases.

Focusing on candidate AMPs that are directly encoded in the genome enabled in vitro and in vivo testing using chemical synthesis without post-translational modifications, but there are other processes that generate active peptides, such as encrypted peptides (EPs), which we used as a comparison point. Notably, the amino acid composition and physicochemical characteristics of the validated AMPs from AMPSphere differed from those of recently identified in EPs. Two evolutionary mechanisms by which AMPs may be generated were explored. First, mutations in genes encoding longer proteins could generate gene fragments via truncation. Among the enriched ortholog groups of proteins from GMGCv1 homologous to c_AMPs, we observed that a majority of groups had unknown function (53.8%), similar to what was reported by Sberro et al. for small proteins from the human gut microbiome. The second mechanism is that a small protein gene could undergo a duplication followed by mutation, which we observed in the case of ribosomal proteins. Ribosomal proteins can harbor antimicrobial activity, possibly due to their amyloidogenic properties. Other origins of AMPs may be horizontal gene transfer or ancestral non-coding sequences.

Nonetheless, the majority of identified AMPs did not have a detectable homolog in other databases. The lack of observed homology may be due to limitations in our ability to robustly detect these homology relationships in small sequences, but there is also the possibility that small proteins, such as AMPs, may be more likely to be generated de novo compared to longer proteins and may have repeatedly evolved in various taxa. This may also be an explanation for the large fraction of c_AMPs in the AMPSphere that do not cluster with any other sequences.

We observed that c_AMPs from AMPSphere were habitat-specific and mostly accessory members of microbial pangenomes. Furthermore, four out of the five genera with the most c_AMPs present in AMPSphere share a host-associated lifestyle, and three of these (Prevotella, Faecalibacterium, and CAG-110) are common in animal hosts (Figure 5).

Valles-Colomer et al., who recently analyzed a large collection of human-associated metagenomes, provide a species-specific index of transmissibility for the seven transmission scenarios they study (e.g., mother to infant). Hypothesizing that AMP production may be related to transmission, we correlated the species-specific $\rho_{\text{AMP}}$ calculated in AMPSphere with transmission scores. In both the human gut and oral microbiomes, species with higher $\rho_{\text{AMP}}$ are less transmissible, possibly because AMPs confer protection against strain replacement. Taken together, these results validate the applicability of AMPSphere in the study of microbial ecology, as they suggest a role for AMPs in determining the transmissibility and colonization ability of microbes, which warrants further investigation and validation in future work.

Finally, we experimentally validated predictions made by our ML model and found that 79 (out of 100) synthesized AMPs displayed antimicrobial activity against either pathogens or commensals. Nonetheless, notably, four peptides (cagicin-1, cagicin-4, and enterococcin-1 against A. baumannii and cagicin-1 and lachnospirin-1 against vancomycin-resistant $E. faecium$) presented MIC values as low as 1 $\mu$mol L$^{-1}$, comparable to the MICs of some of the most potent peptides previously described in the literature.

We show that the tested AMPs from AMPSphere tended to target clinically relevant Gram-negative pathogens and showed activity against vancomycin-resistant $E. faecium$. Although conventional AMPs do not target bacteria from the human gut microbiome, tested AMPs from AMPSphere showed efficacy against commensal bacteria, suggesting potential ecological implications of peptides as protective agents for their producing organisms and their ability to reconfigure microbiome communities.

When assessing their activity in vivo, three peptides exhibited anti-infective efficacy in a murine infection model, with lachnospirin-1 and enterococcin-1 being the most potent, resulting in a reduction of bacterial load by up to three orders of magnitude. The active peptides included those derived from both human-associated and environmental microbiota, validating our approach of investigating the global microbiome. Overall, our findings unveil a wide array of AMP sequences without matches in other databases, highlighting the potential of machine learning in the discovery of much-needed antimicrobials.

**Limitations of the study**

We focused on a particular category of AMPs, namely peptides encoded by their own genes and composed of up to 100 amino acids, which does not cover all active peptides. We explored the global microbiome as represented in public databases, and certain habitats and areas of the globe have been significantly more explored than others. This uneven coverage also impacts our quality estimates, as they depend on data availability. We will, however, continue to update the resource as newer genomes and metagenomes are made available. We report results based on finding homologs to our peptides, but matching small sequences to large databases has a higher rate of errors (particular missed matches) than is the case for longer sequences. Our results on the transmissibility of microbial strains and AMP density were intended to demonstrate the value of AMPSphere as a resource, but a full validation of this link will be the focus of future work. Finally, we tested peptides in vitro and in vivo against a panel of bacteria. Given that we observed species- and even strain-specific responses, it is possible that peptides for which we did not observe any activity would have been active against strains not tested here.

**STAR+METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Bacterial strains and growth conditions
  - Skin abscess infection mouse model
- METHOD DETAILS
SUPPLEMENTAL INFORMATION

SUPPLEMENTAL INFORMATION can be found online at https://doi.org/10.1016/j.cell.2024.05.013.

DECLARATION OF INTERESTS

C.F.-N. provides consulting services to Invaiio Sciences and is a member of the Scientific Advisory Boards of Nowture S.L. and Phare Bio. The de la Fuente Lab has received research funding or in-kind donations from United Therapeutics, Strata Manufacturing PJSC, and Procter & Gamble, none of which were used in support of this work. An invention disclosure associated with this work has been submitted.

Received: June 14, 2023
Revised: April 11, 2024
Accepted: May 6, 2024
Published: June 5, 2024

REFERENCES


ACKNOWLEDGMENTS

We thank Marija Dmitrijeva (University of Zurich) for her helpful comments on a previous version of the manuscript. We thank Kaylyn Toussignant (Queensland University of Technology) for her help editing the manuscript. We thank Georgina H. Joyce (Queensland University of Technology) for her help designing the graphical abstract. We thank members of the Coelho group and the de la Fuente Lab for their support. We thank H. Joyce (Queensland University of Technology) for her help designing the graph.

AUTHOR CONTRIBUTIONS


ADDITIONAL RESOURCES

- Selection of microbial (metagenomes)
- Reads trimming and assembly
- smORF and AMP prediction
- Clustering of AMP families
- Quality control of c_AMPs
- Sample-based c_AMPs accumulation curves
- Multi-habitat and rare c_AMPs
- Testing c_AMPs overlap across habitats
- c_AMP density in microbial species
- c_AMPs and bacterial species transmissibility
- Determination of accessory AMPs
- Annotation of AMPs using different datasets
- Genomic context conservation analysis
- AMPSphere web resource
- Peptide selection for synthesis and testing
- Minimal inhibitory concentration (MIC) determination
- Circular dichroism assays
- Outer membrane permeabilization assays
- Cytoplasmic membrane depolarization assays
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.05.013.
distributed among four halophilic archaeal genera: Natrinema, Halo-


17. Lesiuk, M., Pausdyszyska, M., and Greber, K.E. (2022). Synthetic Anti-
microbial Immunomodulatory Peptides: Ongoing Studies and Clinical Tri-


Nunez, C., and Arciello, A. (2022). Synthetic Antibiotic Derived from Se-


29. Donia, M.S., Cimermancic, P., Schulze, C.J., Wieland Brown, L.C., Mar-


106. The scikit-bio development team (2020). scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. Version 0.5.5.


128. Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Pro-


### STAR METHODS

#### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacterial and virus strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Acinetobacter baumannii</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 19606</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 11775</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td><em>Escherichia coli</em> MG1655 phnE:_2:FRT</td>
<td>AIC221</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td><em>Escherichia coli</em> MG1655 pmrA53 phnE:_2:FRT (polymyxin-resistant; colistin-resistant strain)</td>
<td>AIC222</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 13883</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>N/A</td>
<td>PA01</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 12600</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>American Type Culture Collection</td>
<td>ATCC BAA-1556 (methicillin-resistant strain)</td>
</tr>
<tr>
<td><em>Akkermansia muciniphila</em></td>
<td>American Type Culture Collection</td>
<td>ATCC BAA-635</td>
</tr>
<tr>
<td><em>Bacteroides fragilis</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 25285</td>
</tr>
<tr>
<td><em>Bacteroides thetaiotaomicron</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 29148</td>
</tr>
<tr>
<td><em>Bacteroides uniformis</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 8492</td>
</tr>
<tr>
<td><em>Bacteroides vulgatus (Phocaeicola vulgatus)</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 8482</td>
</tr>
<tr>
<td><em>Collinsella aerofaciens</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 25986</td>
</tr>
<tr>
<td><em>Clostridium scindens</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 35704</td>
</tr>
<tr>
<td><em>Parabacteroides distasonis</em></td>
<td>American Type Culture Collection</td>
<td>ATCC 8503</td>
</tr>
<tr>
<td><strong>Chemicals, peptides, and recombinant proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luria-Bertani broth</td>
<td>BD</td>
<td>244620</td>
</tr>
<tr>
<td>Tryptic soy broth</td>
<td>Sigma</td>
<td>T8907-1KG</td>
</tr>
<tr>
<td>Agar</td>
<td>Sigma</td>
<td>05039</td>
</tr>
<tr>
<td>MacConkey agar</td>
<td>RPI</td>
<td>M42560-500.0</td>
</tr>
<tr>
<td>Phosphate buffer saline</td>
<td>Sigma</td>
<td>P3913-10PAK</td>
</tr>
<tr>
<td>Glucose</td>
<td>Sigma</td>
<td>G5767</td>
</tr>
<tr>
<td>1-(N-phenylamino)naphthalene</td>
<td>Sigma</td>
<td>104043</td>
</tr>
<tr>
<td>3,3’-dipropylthiadicarbocyanine iodide</td>
<td>Sigma</td>
<td>43608</td>
</tr>
<tr>
<td>HEPES</td>
<td>Fisher</td>
<td>BP310-100</td>
</tr>
<tr>
<td>Potassium chloride (KCl)</td>
<td>Sigma</td>
<td>P3911</td>
</tr>
<tr>
<td><strong>Deposited data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code for generation of AMPSphere</td>
<td>This study</td>
<td><a href="https://doi.org/10.5281/zenodo.11055585">https://doi.org/10.5281/zenodo.11055585</a></td>
</tr>
<tr>
<td>AMPSphere database</td>
<td>This study</td>
<td><a href="https://zenodo.org/record/4606582">https://zenodo.org/record/4606582</a></td>
</tr>
<tr>
<td><strong>Experimental models: Organisms/strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse: CD-1</td>
<td>Charles River</td>
<td>18679700-022</td>
</tr>
<tr>
<td><strong>Software and algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NGLess 1.3.0</td>
<td>Coelho et al.</td>
<td><a href="https://github.com/ngless-toolkit/ngless">https://github.com/ngless-toolkit/ngless</a></td>
</tr>
<tr>
<td>JUG 2.1.1</td>
<td>Coelho</td>
<td><a href="https://github.com/luispedro/jug">https://github.com/luispedro/jug</a></td>
</tr>
<tr>
<td>Prodigal 2.6.3</td>
<td>Hyatt et al.</td>
<td><a href="https://github.com/hyattpd/Prodigal">https://github.com/hyattpd/Prodigal</a></td>
</tr>
<tr>
<td>Macrel v.1.0.0</td>
<td>Santos-Júnior et al.</td>
<td><a href="https://github.com/BigDataBiology/macrel">https://github.com/BigDataBiology/macrel</a></td>
</tr>
<tr>
<td>CDHit 4.8.1</td>
<td>Fu et al.</td>
<td><a href="https://github.com/weizhongli/cdhit">https://github.com/weizhongli/cdhit</a></td>
</tr>
<tr>
<td>MMseqs2</td>
<td>Steinegger and Söding</td>
<td><a href="https://github.com/soedinglab/MMseqs2">https://github.com/soedinglab/MMseqs2</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
### REAGENT or RESOURCE SOURCE IDENTIFIER

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>python 3.8.2</td>
<td>Van Rossum</td>
<td><a href="https://www.python.org/">https://www.python.org/</a></td>
</tr>
<tr>
<td>matplotlib 3.4.3</td>
<td>Hunter</td>
<td><a href="https://matplotlib.org/">https://matplotlib.org/</a></td>
</tr>
<tr>
<td>numpy 1.21.2</td>
<td>Harris et al.</td>
<td><a href="https://numpy.org/">https://numpy.org/</a></td>
</tr>
<tr>
<td>pandas 1.3.2</td>
<td>McKinney</td>
<td><a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a></td>
</tr>
<tr>
<td>plotly 5.2.1</td>
<td>Plotly Technologies Inc, 2015</td>
<td><a href="https://plot.ly">https://plot.ly</a></td>
</tr>
<tr>
<td>scipy 1.7.1</td>
<td>Virtanen et al.</td>
<td><a href="https://www.scipy.org">https://www.scipy.org</a></td>
</tr>
<tr>
<td>scikit-learn 0.24</td>
<td>Pedregosa et al.</td>
<td><a href="http://scikit-learn.org/">http://scikit-learn.org/</a></td>
</tr>
<tr>
<td>scikit-bio 0.5.6</td>
<td>The scikit-bio development team, 2020</td>
<td><a href="http://scikit-bio.org/">http://scikit-bio.org/</a></td>
</tr>
<tr>
<td>BioPython 1.7.9</td>
<td>Cock et al.</td>
<td><a href="https://biopython.org/">https://biopython.org/</a></td>
</tr>
<tr>
<td>eggnog-mapper v2</td>
<td>Cantalapiedra et al.</td>
<td><a href="https://github.com/eggnogdb/eggnog-mapper">https://github.com/eggnogdb/eggnog-mapper</a></td>
</tr>
<tr>
<td>HMMER 3.3+dfsg2-1</td>
<td>Eddy</td>
<td><a href="http://hmmer.org/">http://hmmer.org/</a></td>
</tr>
<tr>
<td>FastANI v.1.33</td>
<td>Jain et al.</td>
<td><a href="https://github.com/ParBLISS/FastANI">https://github.com/ParBLISS/FastANI</a></td>
</tr>
<tr>
<td>Megahit 1.2.9</td>
<td>Li et al.</td>
<td><a href="https://voutcn/megahit/">https://voutcn/megahit/</a></td>
</tr>
<tr>
<td>AMPlicify</td>
<td>Li et al.</td>
<td><a href="https://github.com/bcgsc/AMPlicify">https://github.com/bcgsc/AMPlicify</a></td>
</tr>
<tr>
<td>Ampir</td>
<td>Fingerhut et al.</td>
<td><a href="https://github.com/Legana/ampir">https://github.com/Legana/ampir</a></td>
</tr>
<tr>
<td>APIN</td>
<td>Su et al.</td>
<td><a href="https://github.com/zhanglabNKU/APIN">https://github.com/zhanglabNKU/APIN</a></td>
</tr>
<tr>
<td>amPEPpy 1.0</td>
<td>Lawrence et al.</td>
<td><a href="https://github.com/lawrence3/amPEPpy">https://github.com/lawrence3/amPEPpy</a></td>
</tr>
<tr>
<td>RNAcode 0.2-beta</td>
<td>Washietl et al.</td>
<td><a href="https://github.com/ViennaRNA/RNAcode">https://github.com/ViennaRNA/RNAcode</a></td>
</tr>
<tr>
<td>Bwa v.0.7.17</td>
<td>Li et al.</td>
<td><a href="https://github.com/lh3/bwa">https://github.com/lh3/bwa</a></td>
</tr>
<tr>
<td>Statsmodels 0.14.0</td>
<td>Seabold and Perktold</td>
<td><a href="https://www.statsmodels.org">https://www.statsmodels.org</a></td>
</tr>
<tr>
<td>mOTUs2</td>
<td>Milanese et al.</td>
<td><a href="https://github.com/motu-tool/mOTUs">https://github.com/motu-tool/mOTUs</a></td>
</tr>
<tr>
<td>SAMtools 1.18</td>
<td>Li et al.</td>
<td><a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a></td>
</tr>
<tr>
<td>BEDtools v2.31.0</td>
<td>Quinlan and Hall</td>
<td><a href="https://github.com/arksg/bedtools2">https://github.com/arksg/bedtools2</a></td>
</tr>
<tr>
<td>Clustal Omega 1.2.2</td>
<td>Sievers et al.</td>
<td><a href="http://clustal.org/omega/">http://clustal.org/omega/</a></td>
</tr>
<tr>
<td>Diamond v2.1.8</td>
<td>Buchfink et al.</td>
<td><a href="https://github.com/bbuchfink/diamond">https://github.com/bbuchfink/diamond</a></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProGenomes2</td>
<td>Mende et al.</td>
<td><a href="http://progenomes.embl.de/">http://progenomes.embl.de/</a></td>
</tr>
<tr>
<td>DRAMP - Data repository of antimicrobial peptides 3.0</td>
<td>Shi et al.</td>
<td><a href="http://dramp.cpu-bioinfo.org/">http://dramp.cpu-bioinfo.org/</a></td>
</tr>
<tr>
<td>UniprotKB 2021_03</td>
<td>The UniProt Consortium</td>
<td><a href="https://www.uniprot.org/">https://www.uniprot.org/</a></td>
</tr>
<tr>
<td>Eggnog v.5.0</td>
<td>Huerta-Cepas et al.</td>
<td><a href="http://eggnog5.embl.de/">http://eggnog5.embl.de/</a></td>
</tr>
<tr>
<td>SmProt database v2.0</td>
<td>Hao et al.</td>
<td><a href="http://bigdata.ibp.ac.cn/SmProt/index.html">http://bigdata.ibp.ac.cn/SmProt/index.html</a></td>
</tr>
<tr>
<td>PFAM 33.1</td>
<td>Mistry et al.</td>
<td><a href="http://pfam.xfam.org/">http://pfam.xfam.org/</a></td>
</tr>
<tr>
<td>AntiFAM v.7.0</td>
<td>Eberhardt et al.</td>
<td><a href="https://www.ebi.ac.uk/research/batman/software/antifam-tool-identify-sporuous-proteins">https://www.ebi.ac.uk/research/batman/software/antifam-tool-identify-sporuous-proteins</a></td>
</tr>
<tr>
<td>GTDB 07-RS95</td>
<td>Parks et al.</td>
<td><a href="https://gtdb.ecogenomic.org/">https://gtdb.ecogenomic.org/</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Luis Pedro Coelho (luispedro@big-data-biology.org).

Materials availability
This study did not generate new unique reagents.

Data and code availability
- Metagenomes and Genomes data are publicly available at the European Nucleotide Archives (ENA) as of the date of publication. Their accession numbers are listed in Table S1. AMPSphere is available as a public online resource (https://ampsphere.big-data-biology.org/), and its files have been deposited in Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Bacterial strains and growth conditions
The pathogenic strains *Acinetobacter baumannii* ATCC 19606, *Escherichia coli* ATCC 11775, *Escherichia coli* AIC221 [Escherichia coli MG1655 pnhE_2FRT (control strain for AIC 222)], *Escherichia coli* AIC222 [Escherichia coli MG1655 pmrA53 pnhE_2FRT (polymyxin-resistant; colistin-resistant strain)], *Klebsiella pneumoniae* ATCC 13883, *Pseudomonas aeruginosa* PA01, *Pseudomonas aeruginosa* PA14, *Staphylococcus aureus* ATCC 12600, *Staphylococcus aureus* AIC222 [Escherichia coli MG1655 pmrA53 pnhE_2FRT (polymyxin-resistant; colistin-resistant strain)], and *Enterococcus faecalis* ATCC 700802 (vancomycin-resistant strain), and *Enterococcus faecium* ATCC 700221 (vancomycin-resistant strain) were grown and plated on Luria-Bertani (LB) agar plates and incubated overnight at 37°C from frozen stocks. After incubation, one isolated colony was transferred to 6 mL of medium (LB), and cultures were incubated overnight (16 h) at 37°C. The following day, inocula were prepared by diluting the overnight cultures 1:100 in 6 mL of the respective media and incubating them at 37°C until bacteria reached logarithmic phase (OD600 = 0.3–0.5).

The gut commensal strains *Akkermansia muciniphila* ATCC BAA-635, *Bacteroides fragilis* ATCC 25285, *Bacteroides thetaitaomicron* ATCC 29148, *Bacteroides uniformis* ATCC 8492, *Bacteroides vulgatus* ATCC 8426 (Phocaeicola vulgaris), *Collinsella aerofaciens* ATCC 25986, *Clostridium scindens* ATCC 35704, and *Parabacteroides distasonis* ATCC 8503 were grown in brain heart infusion (BHI) agar plates enriched with 0.1% (v/v) vitamin K3 (1 mg mL⁻¹), 1% (v/v) hemin (1 mg mL⁻¹, diluted with 10 mL of 1 N sodium hydroxide), and 10% (v/v) L-cysteine (0.05 mg mL⁻¹), from frozen stocks and incubated overnight at 37°C. Resazurin was used as an oxygen indicator. After the incubation period, a single isolated colony was transferred to 3 mL of BHI broth and incubated overnight (16 h) at 37°C. The following day, inocula were prepared by diluting the overnight cultures 1:100 in 3 mL of the respective media and incubating them at 37°C until bacteria reached logarithmic phase (OD600 = 0.3–0.5).
Skin abscess infection mouse model

To assess the anti-infective efficacy of the peptides against A. baumannii ATCC 19606 in a skin abscess infection mouse model, the bacteria were cultured in tryptic soy broth (TSB) medium until an OD_{600} of 0.5 was reached. Next, the cells were washed twice with sterile PBS (pH 7.4) and suspended to a final concentration of 5 $\times$ 10^9 colony-forming units (CFU) per mL. Six-week-old female CD-1 mice, after being anesthetized with isoflurane, were subjected to a superficial linear skin abrasion on their backs in an area that they could not touch with their mouth or limbs. An aliquot of 20 $\mu$L containing the bacterial load was then administered over the abraded area. A single dose of the peptides diluted in water at their MIC value was administered to the infected area 2 h after the infection. The animals were euthanized two- and four-days post-infection, and the infected area was extracted and homogenized for 20 min using a bead beater (25 Hz) and 10-fold serially diluted for CFU quantification on MacConkey agar plates for easy differentiation of A. baumannii colonies. The experimental groups consisted of 3 mice CD-1 per group ($n$ = 3), all female, and each mouse was infected with an inoculum from a different colony to ensure variability. The animals were single caged to avoid cross-contamination. All the mice were used three days after arrival from the commercial provider. The skin abscess infection mouse model was approved by the University Laboratory Animal Resources (ULAR) from the University of Pennsylvania (Protocol 806763).

METHOD DETAILS

Selection of microbial (meta)genomes

Selection of metagenomes and genomes to compose the AMPSphere was similar to that adopted by Coelho et al. Public metagenomes available on 1 January 2020 produced with Illumina instruments (except for MiSeq, to ensure the consistency and reliability of the meta-analysis findings), with at least 2 million reads and, on average, 75 bp long, were downloaded from the European Nucleotide Archive (ENA). These samples met two criteria: (1) they were tagged with taxonomy ID 408169 (for metagenome) or were a descendant of it in the taxonomic tree; and/or (2) they came from experiments with the library source listed as “METAGENOMIC”. Samples were grouped by project and all projects with at least 20 samples were included for analysis. Additionally, metagenomes deposited by the Integrated Microbial Genomes System (IMG) missing from ENA were also included. Metadata was manually curated from each sample’s describing literature and Biosamples database. For habitat classification groups were created based on the similarity of habitat conditions, such as air, anthropogenic, aquatic, host-associated, ph:alkaline, sediment, terrestrial, and others. The sample origins and information related to host species were obtained using the NCBI taxonomic identification number. High-quality microbial genomes were selected from ProGenomes2 database. The resulting 63,410 publicly available metagenomes and 87,920 high-quality microbial genomes are listed in Table S1.

Reads trimming and assembly

Reads were processed using NGLess, trimming positions with quality lower than 25 and discarding reads shorter than 60 bp post-trimming. Metagenomes obtained from a host-associated microbiome passed through a filtering of reads mapping to the host genome when available. Reads totaling more than 14.7 trillion base pairs of sequenced DNA were assembled with MEGAHIT and the taxonomy of the 16,969,685,977 contigs generated was inferred as previously described, using MMSeqs to map the sequences against the GTDB release 95 and 87,920 high-quality microbial genomes are listed in Table S1.

smORF and AMP prediction

Analogously to Sberro et al., we used a modified version of Prodigal to predict smORFs (33–303 bp) from contigs. The 4,599,187,424 redundant smORFs, most of which (99.25%) originated in metagenomes, were then de-duplicated to optimize the computational resource usage, yielding 2,724,621,233 non-redundant smORFs. Macrel was run on the de-duplicated smORFs to predict c_AMPs. Singleton sequences (those appearing in a single sample or genome) were eliminated, except when they had a significant match (amino acid identity $\geq$ 75% and E-value $\leq$ 10^-5) to a sequence from the Data Repository of Antimicrobial Peptides (DRAMP) version 3.0 using the ‘easy-search’ method from MMSeqs2. In total, AMPSphere encompassed 863,498 non-redundant predicted c_AMPs encoded by 5,518,294 redundant genes. AMP densities were estimated as the number of AMPs per assembled base pairs in a sample or a species.

AMP genes originating from ProGenomes2 had the taxonomy of the original genome assigned to them, whereas AMP genes from metagenomes were assigned the taxonomy predicted for the contig where they were found. Insights about potential structural conformations were obtained using the function secondary_structure_fraction from the ProtParam module implemented in the SeqUtils in Biopython. This function calculates the fraction of amino acids tend to assume conformations of helix [VIYFWL], turn [NPGS], and sheet [EMAL].

Clustering of AMP families

Clustering peptides by sequence identity is only possible at high identities as short low-/medium-identity matches are possible by chance. Therefore, aiming to recover matches where basic features are preserved even if individual amino acids are not identical, we used a reduced amino acids alphabet of 8 letters - [LVIMC], [AG], [ST], [FYW], [EDNQ], [KR], [P], [H]. c_AMPs were hierarchically clustered after alphabet reduction using three sequential identity cutoffs (100%, 85%, and 75%) with CD-Hit.
A cluster was considered an AMP family when it consisted of at least 8 sequences. Representative sequences of peptide clusters were selected according to their length (taking the longest) with ties being broken by their alphabetical order.

To validate this clustering procedure, we used a sample of 3,000 sequences randomly sampled from AMPSphere, excluding cluster representatives. These sequences were aligned against the representative sequence of their cluster using the Smith-Waterman algorithm with the BLOSUM 62 cost matrix, and gap open and extension penalties of −10 and −0.5, respectively. The alignment score was then converted to an E-value according to the model by Karlin and Altschul, which uses the values of $k$ (0.132539) and $l$ (0.313667) constants adjusted to search for a short input sequence as implemented in the BLAST algorithm. Alignments were considered significant if their E-value was less than $10^{-5}$. We found that more than 95.3% of alignments produced in the first two levels (100% and ≥85% of identity) were significant, along with 77.1% of those from the third level (≥75% of identity) – see Figure S3.

**Quality control of c_AMPs**

The c_AMPs in AMPSphere were submitted to another six AMP prediction systems (AMPScanner v2, amPepPy, API55 – with their proposed model, AI4AMP, and AMPify).

The genes of c_AMPs were subjected to five different quality tests to reduce the likelihood that the observed peptides were artifacts or fragments of larger proteins. Initially, the peptides were searched against AntiFam using HMMSearch, which was designed to identify uniquely recurring spurious predicted ORFs, with the option “-cut_ga”. Fewer than 0.05% of c_AMPs had any significant hits.

For each smORF, we searched for an in-frame stop codon upstream of its start codon. When no stop codon is found, we cannot rule out the possibility that the smORF is part of a larger gene which we cannot observe due to fragmented assembly. Most (68.4%) of the c_AMPs are encoded by at least one gene that is not terminally placed. However, the fact that a c_AMP is terminal does not imply that the given c_AMP is an artifact since the AMP genes are short enough to be recovered even in short contigs. For example, 72.9% (4,622/6,339) of homologs to DRAMP as terminal c_AMPs in AMPSphere.

The RNAcode program predicts protein-coding regions based on evolutionary signatures typical for protein genes. This analysis depends on a set of homologous and non-identical genes. Therefore, AMP clusters containing at least three gene variants were aligned. Given that an extensive portion of the AMPSphere candidates (53%; 459,910 out of 863,498) is not part of such a cluster, they could not be tested. Of the tested c_AMPs, 53% (215,421 out of 403,588) were considered genes with evolutionary traits of protein-coding sequences.

We then checked for evidence of transcription and/or translation using 221 publicly available metatranscriptomes, comprising human gut (142), peat (48), plant (13), and symbionts (17); and 109 publicly available metaproteomes from PRIDE database comprising from 37 habitats - Table S6. Using bwa v.0.7.17, reads from the metatranscriptomes were mapped against non-redundant AMP genes, and, using NGLess, we selected genes with at least one read mapped across a minimum of two samples to increase our confidence. This approach is similar to that adopted when predicting AMPs. Using regular expressions implemented in Python 3.8, k-mers of all AMPSphere peptides (with length equal to at least half the length of the sequence) were compared to peptide sequences in metaproteomics data. A perfect match between a k-mer and a metaproteomic peptide was considered additional evidence that this c_AMP is likely to be translated, as described by Ma et al. Briefly, the number of c_AMP peptides mapped against the set of metaproteomic samples was counted, and those c_AMP peptides with at least one match covering more than 50% of the peptide were marked as detected. c_AMPs with experimental evidence in metatranscriptomes and/or metaproteomes accounted for circa 20% of the AMPSphere.

The mapping of c_AMPs was performed without considering genomic context, which may have led to an overestimation of candidates being identified as potentially transcribed. For example, if they are homologous to longer proteins the presence of the longer gene may lead to a false positive detection of the shorter c_AMP. We investigated this using Fisher’s Exact Test to compare the percentage of AMP homologs to the GMGCv1 database with experimental evidence of translation (3.4% - 2,073 out of 61,020 peptides, Odds Ratio = 4.3, $P_{Fisher’s\; exact} < 10^{-305}$) and/or transcription (22.8% - 13,901 out of 61,020 peptides, Odds Ratio = 1.2, $P_{Fisher’s\; exact} = 6.7 \cdot 10^{-108}$). The results suggest that our approach tends to slightly overestimate the potential transcription and translation of candidates with canonical-length homologs.

Given that only a small number of transcriptomic or proteomics data were available and the afore-mentioned limitations in interpreting the mappings, we considered AMPs passing all quality-control tests to be high-quality, regardless of evidence of translation or transcription. We further separated those with experimental evidence of translation/transcription (17,115 c_AMPs, circa 2% of AMPSphere) and those without it (63,098 c_AMPs, circa 7%). For c_AMP families, we considered high-quality those where ≥75% of its c_AMPs pass all quality control tests or those with at least one c_AMP possessing experimental evidence of translation/transcription.

**Sample-based c_AMPs accumulation curves**

To determine the saturation of c_AMP discovery, for each habitat or group of habitats, we computed sample-based accumulation curves by randomly sampling metagenomes in steps of 10 metagenomes. This procedure was repeated 32 times, and the average was taken.
Multi-habitat and rare c_AMPs
We first counted c_AMPs present in ≥2 habitats (“multi-habitat AMPs”). To then test the significance of this value, we opted for a similar approach to that described in Coelho et al.52: habitat labels for each sample were shuffled 100 times and the number of resulting multi-habitat c_AMPs was counted. Shuffling labels resulted in 676,489.7 ± 4,281.8 multi-habitat c_AMPs by chance for high-level habitat groups, and in 685,477.17 ± 4,369.6 multi-habitat c_AMPs by chance when looking at the habitats individually inside the high-level groups. The Shapiro-Wilk test was used to check that the resulting data distribution is normal (p = 0.49, for specific habitats; p = 0.1 for high-level habitats). In the original (non-shuffled data), high-level habitat groups presented 93,280 multi-habitat c_AMPs (136.21 standard deviations below shuffled value), while specific habitats presented 173,955 multi-habitat c_AMPs (117.1 standard deviations below shuffled value).

To determine the rarity of c_AMPs, we adapted the protocol previously established by Coelho et al.52 in which the non-redundant genes in AMPSphere were mapped against the reads of metagenome samples using NGLess.96 We considered only uniquely mapped reads. From the mapping, we computed the c_AMPs detected per sample and the number of detections per c_AMP, considering “rare” c_AMPs as those detected less than the average of the entire AMPSphere (682 detections or 1% of all samples as previously described for species139). This approach was adopted to overcome the high computational costs of a competitive mapping procedure. We expect that our approach overestimates how prevalent c_AMPs are, and because of that, it is a robust way to estimate the rarity of c_AMPs.

As the high-quality designation requires at least 3 gene variants for the RNAcode test to be performed, the rarest genes will not be high-quality. However, for robustness, we quantified this effect by computing the mean and median number of detections in only the high-quality c_AMPs and only non-terminal c_AMPs (a test which does not require a minimum number of genes). The mean number of detections is 682 for the full collection, 789 for high-quality c_AMPs, and 679 for non-terminal ones.

Testing c_AMPs overlap across habitats
Like was done when testing the significance of the number of multi-habitat c_AMPs observed, the number of overlapping c_AMPs was computed for each pair of habitats. We shuffled the sample labels 1,000 times, counting the number of randomly overlapping c_AMPs for each pair of habitats. Then, we estimated the probability of observing the overlap by Chebyshev’s inequality, which does not rely on any assumption regarding the distribution of the data as we observed, using the Shapiro-Wilk’s test, that the shuffled counts do not follow a normal distribution. Chebyshev’s inequality is p ≤ 1/Z², where Z stands for the Z score computed from the average and standard deviations estimated by the shuffling procedure. The p-values were adjusted using Holm-Sidak implemented in multipletests from the statsmodels package,114 and those below 0.05 were considered significant.

c_AMP density in microbial species
The c_AMP density was defined as ρc_AMP = nnc_AMP/L, where nnc_AMP is the number of c_AMP redundant genes and L is the assembled base pairs. We assume, as an approximation, that in a large segment assembled, the start positions of AMP genes are independent and uniformly random. Then, we calculated the standard sample proportion error with the formula: STDERR = √(p(1-p)/L). The standard sample proportion error was used to calculate the margin of error at a 95% confidence interval (Z = 1.96, α = 0.05).

To gain insights about the contributions of different phyla, species, and genera to the AMPSphere, we calculated the c_AMP density for these taxonomy levels using the c_AMPs included within AMPSphere, summing all assembled base pairs for contigs assigned to each taxonomy level in the samples used in AMPSphere. The ρc_AMP of genera, phyla and species within a margin of error superior to 10% of the calculated value were eliminated along with outliers according to Tukey’s fences (k = 1.5). We estimated species’ presence and abundance in each sample using mOTUs2.115 None of the genera with the highest ρc_AMP (Algorimicrobium, TMED78, SFJ001, STGJ01, and CAG-462) were highly prevalent microbes.

c_AMPs and bacterial species transmissibility
We used the species taxonomy and transmissibility indices calculated by Valles-Colomer et al.72 to demonstrate the effect of AMPs on the transmission of bacterial species from mother to children. Only those species overlapping AMPSphere and the datasets from Valles-Colomer et al.72 were used for this analysis, and their AMP densities were calculated as described in the previous section (c_AMP density in microbial species), using all the predicted c_AMPs from metagenomes and genomes we obtained, also including those not in AMPSphere, to avoid sampling bias. The AMP density and the coefficient of transmissibility were correlated using Spearman’s method implemented in the scipy package134; following children’s microbiome after 1, 3, and up to 18 years, as well as, cohabitation and intra-datasets. The ρ-values of correlations were corrected using Holm-Sidak implemented in the multipletests function from the statsmodels package.114

Determination of accessory AMPs
To uncover the prevalence of c_AMPs through the microbial pan-genomes, core, shell, and accessory c_AMP clusters were determined using the subset of c_AMPs obtained from ProGenomes243 because of their high-confidence assigned taxonomies and genomically-defined species (specI140). To increase confidence in our measures, only species containing at least 10 genomes were
used in this analysis. c_AMPs and AMP families, present in fewer than 50% of the genomes from a microbial species were classified as accessory. c_AMPs and families present in 50%–95% of the genomes in the cluster were classified as shell, and those present in >95% of the genomes were classified as core genes.

To determine the propensity of AMPs being shared between genomes belonging to the same strain, we first defined strains within species. For this, we used FastANI v.1.33 to cluster genomes from the same species in ProGenomes2. Genome groups with ANI ≥ 99.99% were considered clonal complexes and only a single representative of each clonal complex was kept for further analyses. Species that had fewer than 10 genomes after this step were not considered further in this analysis. Next, we inferred strains (99.5% ≤ ANI <99.99%) as in Rodriguez et al. We then counted the pairs of genomes from the same species sharing AMPs, stratified by whether the pair originates from the same strain or not, and tested the results with Fisher’s Exact Test implemented in the scapy package.

To determine the proportions of accessory, shell and core full-length proteins in the microbial pangenomes, we also extracted the predicted full-length proteins from the ENA database for each genome and hierarchically clustered them after alphabet reduction in a similar fashion to that described in the topic “AMP families”. Full-length protein clusters with ≥8 sequences for each species were kept. The prevalence of full-length protein families within a species was computed as above and the number of core families was compared to the number of c_AMP core families using the probability, calculated as number of species with proportion of core full-length protein families less or equal to that observed for c_AMPs divided by the total of assessed species.

To determine the genotype of Mycoplasma pneumoniae genomes in ProGenomes2, we extracted the gene coding for P1 adhesin by mapping the reference gene sequence NZ_LR214945.1:c568695-567307 against each genome with bwa v.0.7.17, and later extracted the sequences using with SAMTools 16 and BEDTools. The extracted gene sequences were aligned using Clustal Omega, and a phylogenetic tree was built using the aligned nucleotide sequences and FastTree 2 with the restricted time-reversible substitution model and a bootstrapping procedure with 1,000 pseudo-replicates to determine node support. The tree was used to segregate and classify genomes taking the strain type of reference genomes from Diaz et al.

### Annotation of AMPs using different datasets

To detect homologs to previously published proteins, we aligned AMPSphere candidates against several databases: (i) the small protein sets in SmProt 2, (ii) the bioactive peptides database starPepDB 45k, (iii) the small proteins from the global data-driven census of Salmonella, (iv) the global microbial gene catalog GMGCv1, and (v) and the AMP database DRAMP version 3.0. To strictly avoid any artifacts of assembly for the analysis, only c_AMPs which passed the terminal placement test (i.e., for which there was strong evidence that the ORF is indeed complete) were searched against the GMGCv1. The AMPs were annotated using MMseqs2 with the ‘easy-search’ method, retaining hits with an E-value up to 10^-5. As Macrel removes the starting methionine from the peptides it outputs, hits starting at the second amino acid were treated as if they matched the first one.

We used the hypergeometric test implemented in the scapy package to model the association between c_AMPs and the background distribution of ortholog groups from GMGCv1. The number of genes that were redundant in GMGCv1 for each ortholog group was computed along with the counts for ortholog groups in the top hits to AMPSphere. The enrichment was given as the proportion of hits present in a given ortholog group divided by the proportion of that ortholog group among the redundant sequences in GMGCv1, and results were considered significant if p < 0.05 after correction with the Holm-Sidak method implemented in multipltests from the statsmodels package. When using a robust approach that filters the ortholog groups by the number of c_AMP hits and GMGCv1 hits associated with them, using a minimum of 10, 20, or even 100 proteins, the results were kept similar to those obtained with all data showing that the extension of the ortholog groups in AMPSphere did not affect the enrichment analysis.

To check for genomic entities generated after gene truncation, we screened for c_AMP homologs using the default settings for Blastn against the NCBI database, keeping only significant hits with a maximum E-value of 10^-5. As a case study, we selected the AMP10.271.016, predicted to be produced by Prevotella jejuni, which shares the start codon with the gene coding for a NAD(P)-dependent dehydrogenase (WP_089365220.1). To verify the gene disposition and putative mutations leading to the AMP creation, we used BioPython to codon-align the fragments from metagenomic contigs assembled from samples SAMN09837386, SAMN09837387, and SAMN09837388, and genomic fragments of different strains of Prevotella jejuni CD3:33 (CP023864:1:504836–504949), F1016 (CP072366.1:781389–781502), F0697 (CP072364.1:1466323–1466436), and from Prevotella melaninigenica strains F0106 (CP072366.1:1466323–1466436), and from Prevotella melaninigenica strains F0106 (CP072366.1:1466323–1466436), and from Prevotella melaninigenica strains F0106 (CP072366.1:1466323–1466436), and F0697 (CP072363.1:409769–409769) and compared the segments coding for the AMP and the original full-length protein.

### Genomic context conservation analysis

To gain insights into the gene synteny involving AMP genes, we mapped the 863,498 AMP sequences against a collection of 169,632 reference genomes, metagenome-assembled genomes (MAGs) and single amplified genomes (SAGs) curated elsewhere with DIAMOND in “blastp” mode, as previously reported. Hits with identity >50% (amino acid) and query and target coverage >90% were considered significant. The target coverage threshold avoids hits to larger homologs whose function may be unrelated. This yielded 107,308 AMPs with homologs in at least one genome. We built gene families from the hits of each AMP detected in the prokaryotic genomes and calculated a conservation score based on the functional annotation of the neighboring genes in a window of three genes up and downstream. The vertical conservation score at each position within the window of each c_AMP was

---

**Cell Resource**

---

**Cell Press**

---

**OPEN ACCESS**

---

**Cell 187, 3761–3778.e1–e9, July 11, 2024**

---

**e7**
calculated as the number of genes with a given functional annotation (ortholog group, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, KEGG orthology, KEGG module, 126 PFAM 33.1,122,143 and CARD125; details of annotation and annotated database described previously21), divided by the number of genes in the family. AMPs with more than two hits and a vertical conservation score >0.9 with any functional term were considered to have conserved genomic contexts. Figure 4 shows genomic context conservation of different KEGG pathways.

For testing whether the fraction of AMPs with conserved genomic neighbors is similar to that of other gene families within the 169,632 genomes curated by del Río et al.,11 we calculated genomic context conservation on 3,899,674 gene families calculated de novo with MMSeqs2299 (using a minimal amino acid identity of 30%, coverage of the shorter sequence of at least 50%, and maximum E-value of 10−3). The c_AMPs were also annotated using EggNOG-mapper v2.158 Their KO annotations were compared to that of the immediate neighbors (+/− 1 positions) to identify neighborhoods with the same function. It was possible to annotate 56.1% (60,173 out of 107,308) of c_AMPs with hits to the genomes tested using the EggNOG5 database.60 Of these, 18.1% were assigned to translation-related functions (class J), 14.4% belong to proteins of unknown function (S), 9% were assigned to replication, recombination, and repair (L).

**AMPSphere web resource**

AMPSphere is found at the address https://ampsphere.big-data-biology.org/. The implementation is based on Python100 and Vue Javascript. The database was built with sqlite, and SQLAlchemy was used to map the database to Python objects. Internal and external APIs were built using FastAPI and Gunicorn to serve them. On the front end, Vue 3 was used as the backbone and Quasar was used to generate interactive visualization plots, and Axios to render content seamlessly. LogoJS (https://logos.wenglab.org/app/) was used to generate sequence logos for AMP families; while the helical wheel app (https://github.com/clemlab/helicalwheel) was used to generate AMP helical wheels.

**Peptide selection for synthesis and testing**

We selected two groups of peptides: (i) 50 peptides that were selected as being particularly likely to be active and that were otherwise interesting (as described below), (ii) 50 peptides selected randomly after applying technical exclusions.

For the first group, only high-quality (see the topic “quality control of c_AMPs”) c_AMPs were considered for synthesis. They were further filtered according to six criteria for solubility144 and three criteria for synthesis, as in PepFun.145 We estimated the solubility using the criteria implemented in PepFun,145 observing that 67.4% (581,749 peptides) passed at least half of the solubility criteria evaluated. The subset that is homologous to peptides in DRAMP46 version 3.0 had a slightly lower rate, 44.3% passed half the tests. We then assessed the peptides regarding their ease of synthesis, however, only 21.2% from AMPSphere passed at least 2 out of the 3 criteria established for chemical synthesis.

A peptide approved for at least six of the above-mentioned criteria was then filtered by predicting AMP activity with six methods in addition to Macrel142: AMPScaner v2,53 the mature peptides model in ampir,52 smPEPpy,54 APIN55 – with their proposed model, AI4AMP,56 and AMPLify.57 Peptides predicted to be AMPs by all methods were filtered by length, discarding sequences longer than 40 amino acid residues, for which conventional solid-phase peptide synthesis using Fmoc strategy has lower yields and many recoupling reactions.146–148 Only one peptide was kept from each family or cluster, namely the one with the highest number of observed smORFs. After this process, we obtained 364 candidate AMPs, belonging to 166 families and 198 clusters with <8 c_AMPs. Of these, 30 candidates were homologous to sequences from the databases used in annotation (e.g., SmProt 249). To further filter according to six criteria for solubility 144 and three criteria for synthesis, as in PepFun.145 We estimated the solubility of the peptides that killed 100% of cells after 24 h of incubation at 37°C.

To build the group of randomly selected peptides, we first selected c_AMPs that are not homologous to any other databases tested and that passed the abovementioned synthesis criteria (total of 768,061 out of 863,498 peptides). We further divided this group into subgroups: (i) those with Macrel-assigned probability >0.6 (271,555 c_AMPs) and (ii) those in the range 0.5–0.6 (496,506 c_AMPs; note that all c_AMPs in AMPSphere have a Macrel-assigned probability ≥ 0.5). We randomly sampled 25 peptides from each group.

**Minimal inhibitory concentration (MIC) determination**

The 100 AMPs were tested for antimicrobial activity using the broth microdilution method.150 MIC values were considered as the concentration of the peptides that killed 100% of cells after 24 h of incubation at 37°C. First, peptides diluted in water were added to untreated flat-bottom polystyrene microwell 96-well plates in 2-fold dilutions ranging from 64 to 1 µmol L−1, and then peptides were exposed to an inoculum of 2·10⁶ cells in LB or BHI broth, for pathogens and gut commensals, respectively. After the incubation time, the absorbance of each well representing each of the conditions was analyzed using a spectrophotometer at 600 nm. The assays were conducted in three biological replicates to ensure statistical reliability.

**Circular dichroism assays**

Circular dichroism experiments were conducted using a J1500 circular dichroism spectropolarimeter (Jasco) at the Biological Chemistry Resource Center (BCRC) of the University of Pennsylvania. The experiments were carried out at a temperature of 25°C. Circular
dichroism spectra were obtained by averaging three accumulations using a quartz cuvette with an optical path length of 1.0 mm. The spectra were recorded in the wavelength range from 260 to 190 nm at a scanning rate of 50 nm min⁻¹ with a bandwidth of 0.5 nm. The peptides were tested at a concentration of 50 μmol L⁻¹. Measurements were performed in water, a mixture of water and trifluoroethanol (TFE) in a ratio of 3:2, and a mixture of water and methanol in a ratio of 1:1. Baseline measurements were recorded prior to each measurement. To minimize background effects, a Fourier transform filter was applied. The helical fraction values were calculated using the single spectra analysis tool available on the BeStSel server.

**Outer membrane permeabilization assays**

Membrane permeability was analyzed using the 1-(N-phenylamino)naphthalene (NPN) uptake assay. NPN demonstrates weak fluorescence in an extracellular environment but displays strong fluorescence when in contact with lipids from the bacterial outer membrane. Thus, NPN will show increased fluorescence when the integrity of the outer membrane is compromised. *A. baumannii* ATCC 19606 and *P. aeruginosa* PA01 were cultured until cell numbers reached an OD₆₀₀ of 0.4, followed by centrifugation (10,000 rpm at 4 °C for 3 min), washing, and resuspension in buffer (5 mmol L⁻¹ HEPES, 5 mmol L⁻¹ glucose, pH 7.4). Subsequently, 4 μL of NPN solution (working concentration of 0.5 mmol L⁻¹) was added to 100 μL of bacterial solution in a white flat bottom 96-well plate. The fluorescence was monitored at λₑₓ = 350 nm and λₑₘ = 420 nm. The peptide solutions in water (100 μL solution at their MIC values) were introduced into each well, and fluorescence was monitored as a function of time until no further increase in fluorescence was observed (30 min). The relative fluorescence was calculated using a non-linear fit. The positive control (antibiotic polymyxin B) was used as baseline. The following equation was applied to reflect % of difference between the baseline (polymyxin B) and the sample:

\[
\text{Relative fluorescence} = \frac{100 \times (\text{fluorescence}_{\text{sample}} - \text{fluorescence}_{\text{polymyxin B}})}{\text{fluorescence}_{\text{polymyxin B}}}
\]

**Cytoplasmic membrane depolarization assays**

The ability of the peptides to depolarize the cytoplasmic membrane was assessed by measuring the fluorescence of the membrane potential-sensitive dye 3,3′-dipropylthiadicarbocyanine iodide [DiSC₃-(5)]. This potentiometric fluorophore fluoresces upon release from the interior of the cytoplasmic membrane in response to an imbalance of its transmembrane potential. *A. baumannii* ATCC 19606 and *P. aeruginosa* PA01 cells were grown with agitation at 37 °C until they reached mid-log phase (OD₆₀₀ = 0.5). The cells were then centrifuged and washed twice with washing buffer (20 mmol L⁻¹ glucose, 5 mmol L⁻¹ HEPES, pH 7.2) and re-suspended to an OD₆₀₀ of 0.05 in 20 mmol L⁻¹ glucose, 5 mmol L⁻¹ HEPES, 0.1 mol L⁻¹ KCl, pH 7.2. An aliquot of 100 μL of bacterial cells was added to a black flat bottom 96-well plate and incubated with 20 mmol L⁻¹ of DiSC₃-(5) for 15 min until the fluorescence stabilized, indicating the incorporation of the dye into the cytoplasmic membrane. The membrane depolarization was monitored by observing the change in the fluorescence emission intensity of the dye (λₑₓ = 622 nm, λₑₘ = 670 nm), after the addition of the peptides (100 μL solution at their MIC values). The relative fluorescence was calculated using a non-linear fit. The positive control (antibiotic polymyxin B) was used as baseline. We estimated the % of difference between the baseline (polymyxin B) and the sample using the same mathematical approach as in the “Outer membrane permeabilization assays”.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Graphs for the experimental results were created and statistical tests conducted in GraphPad Prism v.9.5.1 (GraphPad Software, San Diego, California USA).

**ADDITIONAL RESOURCES**

AMPSphere is freely available for download in Zenodo and as a web server (https://ampsphere.big-data-biology.org/).
Figure S1. General physical-chemical features of c_AMPs in AMPSphere and validated databases of antimicrobial peptides, related to Figure 1

Shown are density curves; the arbitrary density units are not shown, as all curves are independently normalized so the area under the curve is one. For each dataset and feature, the top 1% and bottom 1% of values were considered outliers and are not shown in the plot. Proportions of residues with small side chains [A, C, D, G, N, P, S, T, V] per c_AMP along with the proportions of basic residues [H, R, K] per c_AMP were also shown. The distributions of each feature were compared among the datasets using the Mann-Whitney test with multiple hypothesis testing corrected using Holm-Sidak. Almost all differences are significant (adjusted p value < 0.05). The exceptions are: aliphatic index did not differ between the peptides from DRAMP version 3 and the ones present in the positive training set used in Macrel; AMPSphere peptides did not differ from the positive training set used in Macrel in the fraction of aromatic (pMann = 0.58), non-polar (pMann = 0.97), polar (pMann = 0.97), and acidic (pMann = 0.69) residues; the instability index (pMann = 0.58) and the hydrophobicity (pMann = 0.31) of AMPSphere peptides also were not different from the positive training set used in Macrel.12

---

12 Cramer et al. (2018)
Figure S2. c_AMP quality and habitat distribution, related to Figures 1 and 2

(A) Quality assessment of AMPSphere revealed most of the peptides passed at least one of the tests. The RNAcode test depends on gene diversity, which is very low for AMPSphere, which led to a low rate of positives among our candidates.

(B) c_AMPs homologous to databases of validated bioactive peptides also showed a higher average quality of these datasets.

(C) The limited overlap of c_AMPs among habitats argues in favor of using habitat groups to gain resolution. Note that the group of habitats with the highest paired overlaps belongs to human body sites and samples from human guts and non-human mammalian guts. Only habitats with at least 100 samples were shown.

(D) We observed a large proportion of rare genes in AMPSphere from different habitat groups.
Figure S3. Clustering validation of families, related to STAR Methods section “Clustering of AMP families”

To validate the clustering procedure using a reduced amino acid alphabet, samples of 1,000 peptides were randomly drawn from AMPSphere (excluding representative sequences) and aligned against their cluster representatives. Three different levels (I, II, and III) of clustering were tested. The E-values were computed per alignment and plotted against the corresponding alignment identity. The averaged proportion of significant alignments is shown in each graph above.
Figure S4. Antimicrobial activity of polymyxin B and levofloxacin and circular dichroism spectra of the c_AMPs, related to STAR Methods section “Circular dichroism assays”

(A) Minimal inhibitory concentration values for polymyxin B, a peptide antibiotic, and levofloxacin against all the strains tested. Polymyxin B and levofloxacin were used as positive controls in all antimicrobial assays.

(B–D) The c_AMPs' secondary structural tendency was analyzed using three different solvents: (B) water, (C) trifluoroethanol (TFE) and water mixture (3:2, V:V), and (D) methanol (MeOH) and water mixture (1:1, V:V). The experiments were carried out at 25°C, and the circular dichroism spectra shown are an average of three accumulations obtained using a quartz cuvette with an optical path length of 1.0 mm, ranging from 260 to 190 nm at a rate of 50 nm min⁻¹ and a bandwidth of 0.5 nm. All peptides were tested at a concentration of 50 μmol L⁻¹, with respective baselines recorded prior to measurement. A Fourier transform filter was applied to minimize background effects.
Figure S5. Antimicrobial activity and secondary structure of scrambled versions of some of the lead c_AMPs, related to Figures 6 and 7

(A) MIC values of the scrambled versions of five of the lead c_AMPs from AMPSphere tested against the same 11 pathogenic strains and eight gut commensal strains used to assess the activity of the c_AMPs.

(B–D) The scrambled peptides’ secondary structural tendency was analyzed using three different solvents: (B) water, (C) TFE and water mixture (3:2, V:V), and (D) MeOH and water mixture (1:1, V:V). The experiments were carried out in the same conditions as the ones used for the c_AMPs. A Fourier transform filter was applied to minimize background effects.

(E) Heatmap with the percentage of secondary structure found for each peptide in three different solvents: water, 60% TFE in water, and 50% MeOH in water. Secondary structure was calculated using BeStSel server.11
Figure S6. Mechanism of action of AMPSphere peptides and anti-infective activity of c_AMPs in a preclinical animal model, related to Figures 6 and 7

(A) Fluorescence values relative to polymyxin B (PMB, positive control) of the fluorescent probe 1-(N-phenylamino)naphthalene (NPN) that indicate outer membrane permeabilization of *P. aeruginosa* PAO1 cells.

(B) Fluorescence values relative to PMB (positive control) of 3,3'-dipropylthiadicarbocyanine iodide (DiSC₃[5]), a hydrophobic fluorescent probe used to indicate cytoplasmic membrane depolarization of *P. aeruginosa* PAO1 cells.

(C) Bacterial counts four days post-infection; the c_AMPs were tested at their MIC in a single dose 1 h after the establishment of the infection. Each group consisted of three mice (n = 3), and the bacterial loads used to infect each mouse were derived from a different inoculum.

(D) Mouse weight throughout the experiment (mean ± the standard deviation).

Statistical significance in (C) was determined using one-way ANOVA where all groups were compared to the untreated control group; p values are shown for each of the groups. Features on the violin plots represent median and upper and lower quartiles. Figure created in BioRender.com.