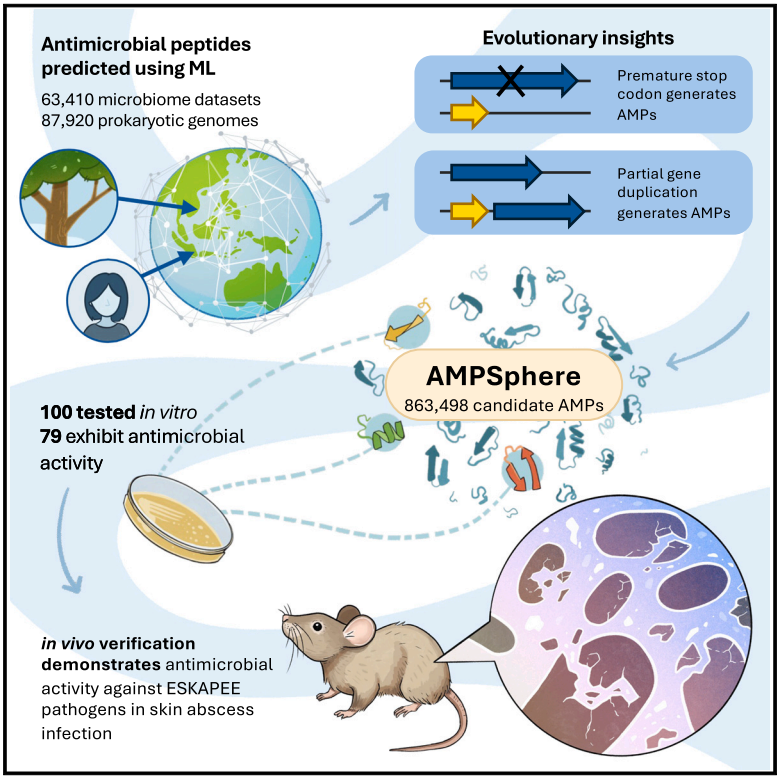
Cell

Resource

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.l.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.

Santos-Júnior et al., 2024, Cell *187*, 3761–3778
July 11, 2024
https://doi.org/10.1016/j.cell.2024.05.013

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Discovery of antimicrobial peptides in the global microbiome with machine learning

Célio Dias Santos-Júnior,[1,2,16] Marcelo D.T. Torres,[3,4,5,6,16] Yiqian Duan,[1] Álvaro Rodríguez del Río,[7] Thomas S.B. Schmidt,[8,9] Hui Chong,[1] Anthony Fullam,[8] Michael Kuhn,[8] Chengkai Zhu,[1] Amy Houseman,[1] Jelena Somborski,[1] Anna Vines,[1] Xing-Ming Zhao,[1,12,13,14] Peer Bork,[8,10,11] Jaime Huerta-Cepas,[7] Cesar de la Fuente-Nunez,[3,4,5,6,*] and Luis Pedro Coelho[1,15,17,*]

[1]Institute of Science and Technology for Brain-Inspired Intelligence - ISTBI, Fudan University, Shanghai 200433, China
[2]Laboratory of Microbial Processes & Biodiversity - LMPB, Department of Hydrobiology, Universidade Federal de São Carlos – UFSCar, São Carlos, São Paulo 13565-905, Brazil
[3]Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA
[5]Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
[6]Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA, USA
[7]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Campus de Montegancedo-UPM, Pozuelo de Alarcón, 28223 Madrid, Spain
[8]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[9]APC Microbiome & School of Medicine, University College Cork, Cork, Ireland
[10]Max Delbrück Centre for Molecular Medicine, Berlin, Germany
[11]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany
[12]Department of Neurology, Zhongshan Hospital, Fudan University, Shanghai, China
[13]State Key Laboratory of Medical Neurobiology, Institutes of Brain Science, Fudan University, Shanghai, China
[14]MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China
[15]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, QLD, Australia
[16]These authors contributed equally
[17]Lead contact
*Correspondence: cfuente@upenn.edu (C.d.l.F.-N.), luispedro@big-data-biology.org (L.P.C.)
https://doi.org/10.1016/j.cell.2024.05.013

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.[1] Indeed, such infections currently kill 1.27 million people per year.[2] 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).[3–9] 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.[10]

AMPs, found in all domains of life,[11–14] are short sequences (operationally defined here as 10–100 amino acid residues[15]) capable of disturbing microbial growth.[12,15] AMPs most commonly interfere with cell wall integrity and cause cell lysis.[12,16] Natural AMPs can originate by proteolysis,[4,17] by non-ribosomal synthesis,[18] or, as we focus on in the present study, they can be encoded within the genome.[19]

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.[20] For instance, pathogens such as *Shigella* spp.,[21] *Staphylococcus* spp.,[22] *Vibrio cholerae*,[23] and *Listeria* spp.[24,25] 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 antiviral drugs (e.g., enfuvirtide and telaprevir[26]). AMPs that exhibit immunomodulatory properties are currently undergoing clinical trials,[27] as are peptides that may be used to address yeast and bacterial infections[28] (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.[29] Such AMPs are more targeted agents than conventional broad-spectrum antibiotics.[30,31] 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.[4,32,33]

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.[34] Therefore, the significance of small open reading frames (smORFs) has been historically overlooked in (meta)genomic analyses.[35–37] In recent years, significant progress has been made in metagenomic analyses of human-associated smORFs.[6,38] These advancements have incorporated machine learning (ML) techniques to identify smORFs encoding proteins belonging to specific functional categories.[39–42] Notably, a recent study used predicted smORFs to uncover approximately 2,000 AMPs from metagenomic samples of human gut microbiomes.[6] 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 publicly available metagenomes and 87,920 high-quality microbial genomes,[43] 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 pangenome.

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.[44,45] These peptides were further compared to encrypted peptides (EPs), which are peptide sequences hidden in protein sequences and mined computationally,[4,10] and demonstrated their ability to target bacterial membranes and their propensity to adopt α-helical and β-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,[42] 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.[43] Sequences present in a single sample were removed,[42] except when they had a significant match (defined as amino acid identity $\geq$75% and E-value $\leq 10^{-5}$) to a sequence in the AMP-dedicated database Data Repository of Antimicrobial Peptides (DRAMP) version 3.0.[46] 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 $\pm$ 8 residues long; Figures 1A and S1). Similar to validated sequences with antimicrobial activity,[42,47,48] c_AMPs from AMPSphere present a positive charge (4.7 $\pm$ 2.6), high isoelectric point (10.9 $\pm$ 1.2), amphiphilicity (hydrophobic moment, 0.6 $\pm$ 0.1), and a potential to bind to membranes or other proteins (Boman index, 1.14 $\pm$ 1.1). As expected, in general, the distribution of physicochemical properties of peptides from AMPSphere, DRAMP[46] version 3.0, and the positive training dataset used in Macrel[42] 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 $\pm$ 8 residues) than those in DRAMP[46] version 3.0 (28 $\pm$ 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 metatranscriptomes (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,[53] the model for mature peptides in ampir,[40] amPEPpy,[54] APIN,[55] AI4AMP,[56] and AMPLify[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.

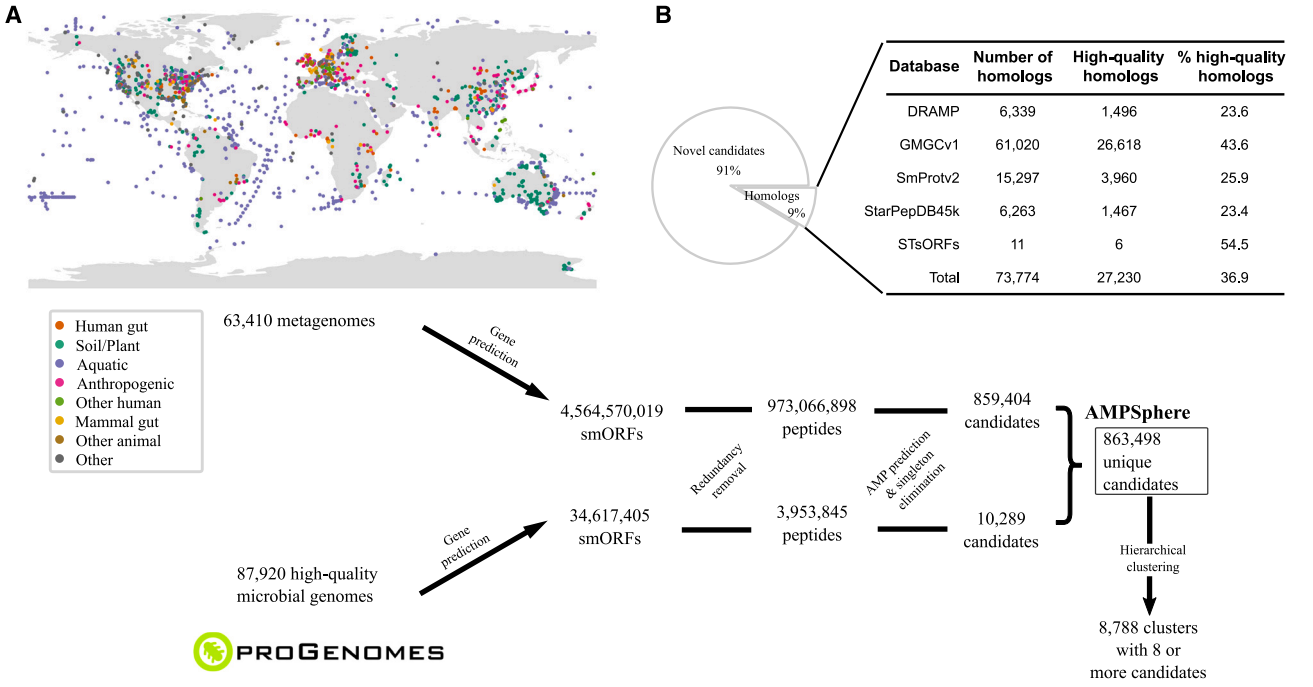


Figure 1. AMPSphere comprises 836,498 non-redundant c_AMPs from thousands of metagenomes and high-quality microbial genomes

(A) To build the AMPSphere, we first assembled 63,410 publicly available metagenomes from diverse habitats. A modified version of Prodigal,[34] which can also predict smORFs (30–300 bp), was used to predict genes on the resulting metagenomic contigs as well as on 87,920 microbial genomes from ProGenomes2.[43] Macrel[42] was applied to the 4,599,187,424 predicted smORFs to obtain 863,498 non-redundant c_AMPs (see also Figure S1). c_AMPs were then hierarchically clustered in a reduced amino acid alphabet using 100%, 85%, and 75% identity cutoffs. We observed 118,051 non-singleton clusters at 75% of identity, and 8,788 of them were considered families (≥8 c_AMPs).

(B) Only 9% of c_AMPs have detectable homologs in other small protein databases (SmProt 2,[49] STsORFs[50]), bioactive peptide databases (DRAMP[46] version 3.0, starPepDB 45k[51]), and general protein datasets (GMGCv1[52]; 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 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 0.7% of the identified c_AMPs (6,339 peptides) are homologous (operationally defined as amino acid identity ≥75% and E-value $\leq 10^{-5}$) to experimentally validated AMP sequences in DRAMP version 3.0.[46] Moreover, most c_AMPs were also absent from protein databases not specific to AMPs (Figure 1B), such as the Small Proteins database (SmProt2)[49] or the Global Microbiome Gene Catalog of canonical-length proteins (GMGCv1),[52] suggesting that c_AMPs represent a region of

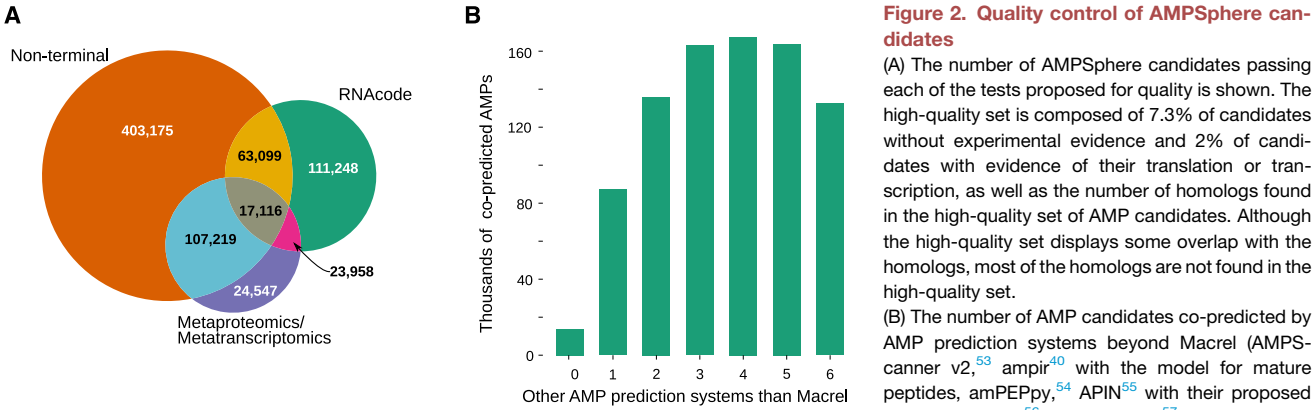


Figure 2. Quality control of AMPSphere candidates

(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 (AMPScanner v2,[53] ampir[40] with the model for mature peptides, amPEPpy,[54] APIN[55] with their proposed model, AI4AMP,[56] and AMPLify[57]). Only a small portion of AMPSphere (<2%) cannot be co-predicted by any system other than Macrel.[42]

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, $p_{Hypergom.} = 4.2 \times 10^{-250}$; 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.[58] 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 sources, indicating that they were not artifacts. The large number of singletons suggests that most c_AMPs originated from processes other than diversification within families, which is the opposite of the hypothesized origin of full-length proteins, in which singleton families are rare.[52] The 8,788 clusters with $\geq$8 peptides obtained at 75% of identity are hereafter named "families," as in Sberro et al.[38] Among them, we considered 6,499 as high-quality families because they contained evidence of translation or transcription or because $\geq$75% of their sequences pass all *in silico* quality tests, regardless of whether experimental evidence is available (see STAR Methods section "AMP families"). These high-quality families span 15.4% of the AMPSphere (133,309 peptides).

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.

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 7.25-fold smaller than would be expected by a random assignment of habitats to samples ($p_{Permutation} < 10^{-300}$; see STAR Methods section "Multi-habitat and rare c_AMPs"). Even within high-level habitat groups, c_AMPs overlap between habitats much less frequently than expected by chance (2.4–192-fold less, $p_{Permutation} < 5.4 \times 10^{-50}$; see STAR Methods section "Testing c_AMPs overlap across habitats"; Figure 1D).

Mutations in larger genes generate c_AMPs as independent genomic entities

Many AMPs are generated post-translationally by the fragmentation of larger proteins.[17] For example, EPs are computationally detected fragments from protein sequences within the human proteome and other proteomes that have been shown to be highly active.[4,10] EPs present diverse secondary structures and act on the membrane of bacterial cells similarly to known natural AMPs but have different physicochemical features compared to known AMPs.[4,33] AMPSphere only considered peptides encoded by dedicated genes. Nonetheless, we hypothesized that some of these have originated from larger proteins by fragmentation at the genomic level. To explore this, we aligned the AMPSphere c_AMPs to the full-length proteins in GMGCv1[52] and observed that about 7% (61,020) of them are homologous to a canonical-length protein (Figure 1B), with 27% of these hits sharing the start codon with the longer protein. This suggests early termination of full-length proteins as one mechanism for generating novel c_AMPs (Figures 3A and 3B).

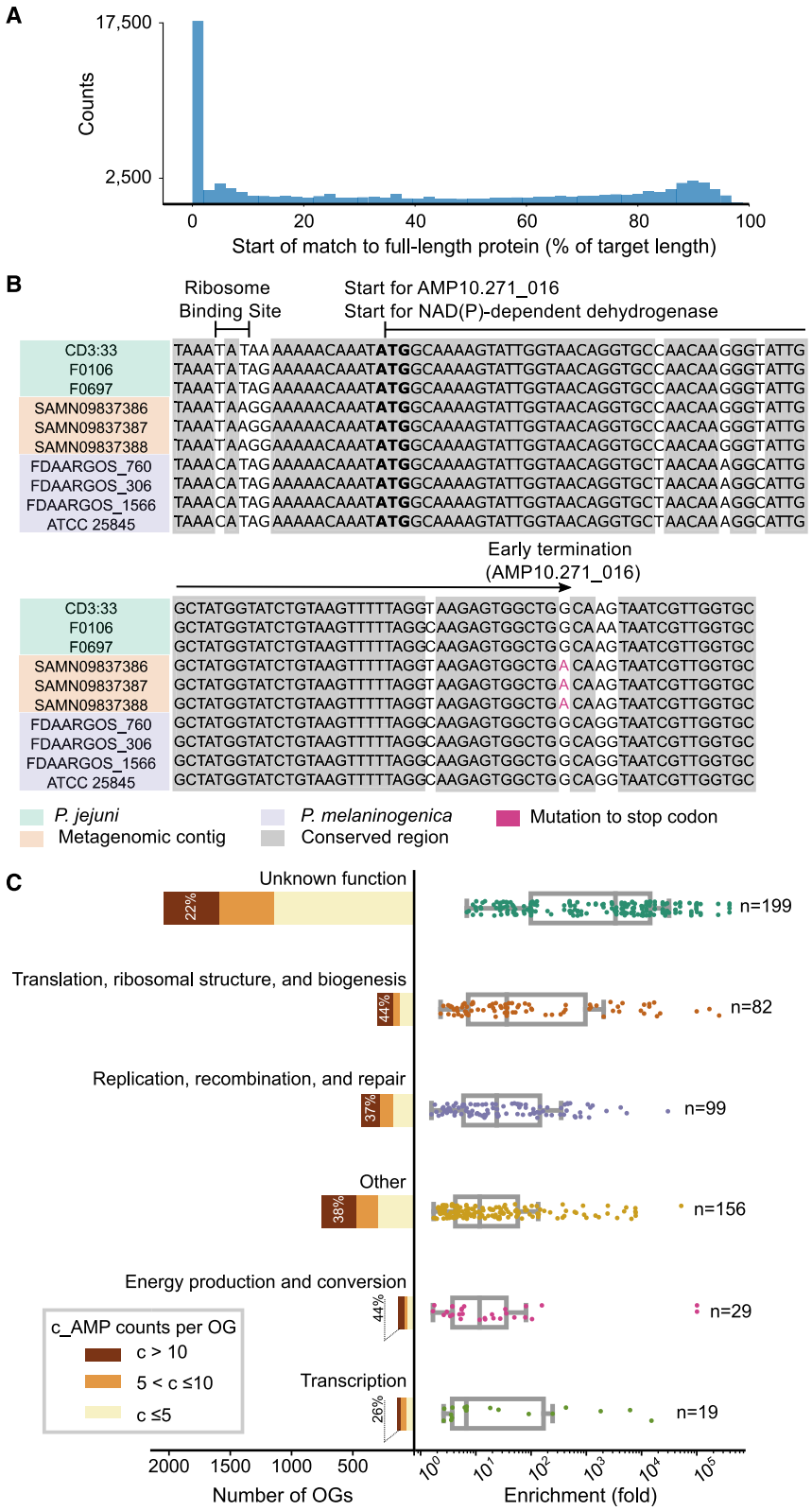


Figure 3. Mutations in genes encoding large proteins generate c_AMPs as independent genomic entities

(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 × 10^{-39}) 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 × 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 ($p_{Hypergeom.} < 0.05$, 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 ($p_{Mann} \leq 3.9 \times 10^{-4}$; 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.[63–65] 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%, $p_{Permutation} = 4.8 \times 10^{-3}$, $N_{Species} = 416$) of c_AMP families present in ProGenomes2[43] are contained in ≥95% of genomes from the same species (Figure 5), here referred to as "core."[66] This is consistent with previous work, in which AMP production was observed to be strain-specific.[67] In contrast, a high proportion (circa 68.8%) of full-length protein families are core in ProGenomes2[43] species. There is a 1.9-fold greater chance ($p_{Fisher} = 2.2 \times 10^{-92}$) 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 system[71] (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) taxonomy[68,69] (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 (ρ_{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 Macrel[42] 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 paper[72] by correlating the ρ_{AMP} for each bacterial species to the published measures of microbial transmission. Human gut bacteria showed increased transmissibility at lower AMP densities ($R_{Spearman} = -0.42$, $p_{Holm\text{-}Sidak} = 3.4 \times 10^{-2}$, $N_{Species} = 43$). Similarly, in human oral microbiome bacterial species, transmissibility from mother to offspring is consistently inversely correlated with their ρ_{AMP} for the first year ($R_{Spearman} = -0.55$, $p_{Holm\text{-}Sidak} = 1.4 \times 10^{-3}$, $N_{Species} = 41$). This suggests that human gut bacteria and oral microbiome bacterial species show increased transmissibility at lower ρ_{AMP}. Moreover, it highlights the potential influence of ρ_{AMP} on the transmissibility of gut and

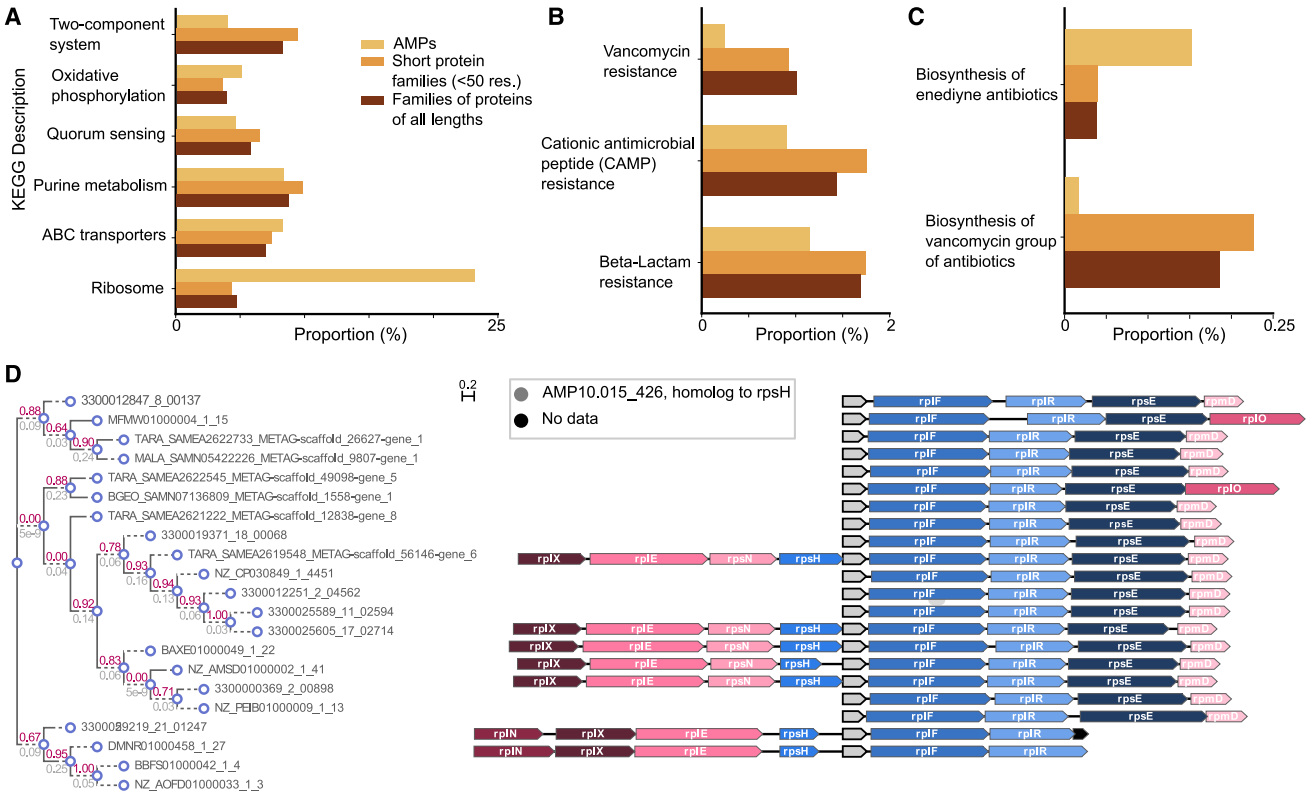


Figure 4. The genome context of c_AMPs shows a preference for neighborhoods containing ribosome assembly proteins

(A) Compared to other proteins, c_AMPs in conserved genomic architectures tend to be closer to ribosomal-machinery-related genes than families of proteins with different sizes (all length and small proteins with ≤50 amino acids).

(B) The proportion of c_AMPs in a genome context involving antibiotic resistance genes is lower than in other gene families.

(C) The proportion of c_AMPs in neighborhoods with antibiotic-synthesis-related genes is very small (<0.25%).

(D) The conserved genomic context of the gene encoding AMP10.015_426 is shown in different genomes (the tree on the left depicts the phylogenetic relationship of the genes homologous to it). This c_AMP is homologous to the ribosomal protein rpsH and is found in the context of *rpsH* and other ribosomal protein genes. See also Table S4.

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 (DRAMP[46] 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 Macrel[42]) 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

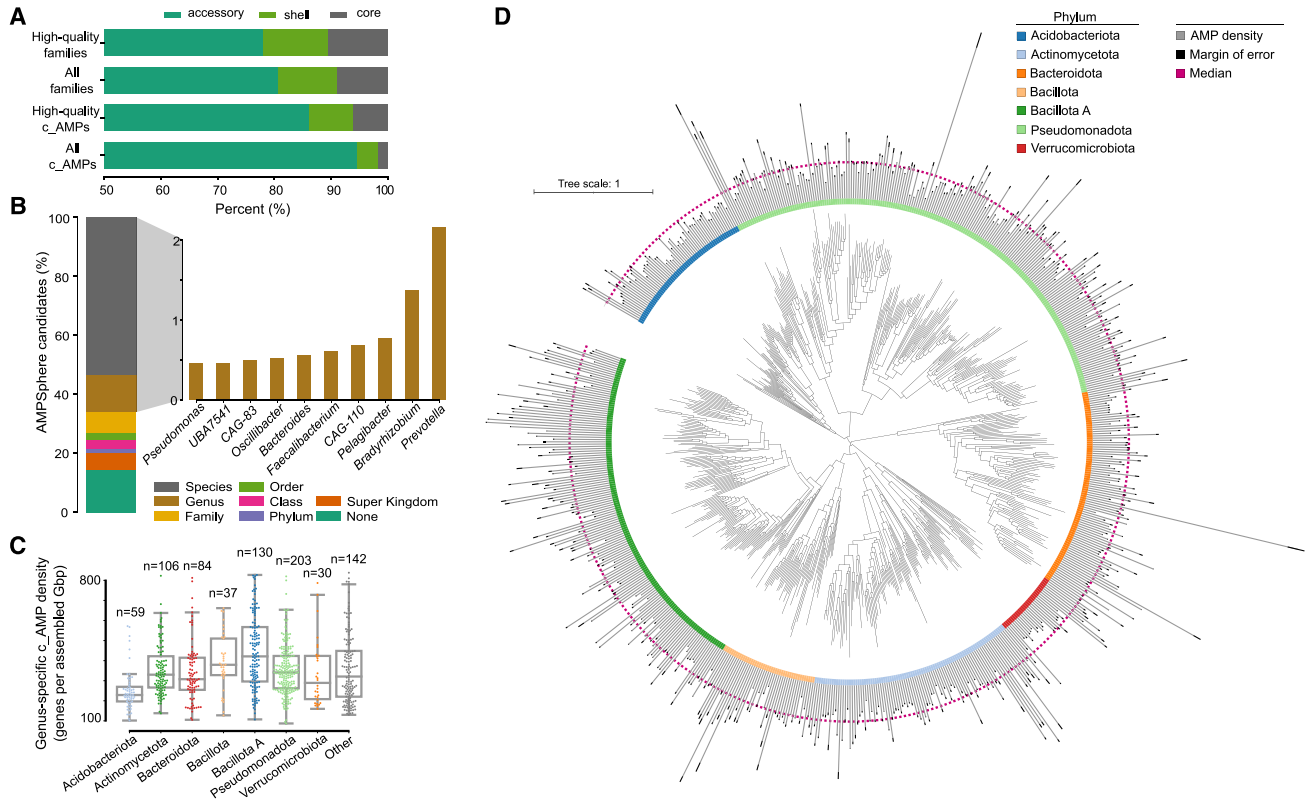


Figure 5. AMP variation in AMPSphere database is taxonomy-dependent

(A) Shown are the fractions of AMPs (or AMP families) that are accessory (present in <50% of genomes from same species), shell (50%–95%), or core (≥95%).

(B) Distribution of the lowest taxonomic level at which c_AMPs were annotated. In detail (right) are the top 10 genera with the highest numbers of c_AMPs included in AMPSphere. Animal-associated genera (e.g., *Prevotella*, *Faecalibacterium*, and *CAG-110*) contribute the most c_AMPs, possibly reflecting data sampling.

(C) Using the ρ_{AMP} per genus (calculated with c_AMPs in AMPSphere), we observed the distribution of c_AMPs per phylum, with Bacillota A as the densest (the number of samples used to build the graph is shown above each box).

(D) Taxonomy of the detected taxa in AMPSphere is shown using the GTDB[68,69] reference tree. The gray bars show ρ_{AMP} distribution with respect to taxonomy, with black bars representing the confidence interval of 95%. Bacillota A, Actinomycetota, and Pseudomonadota are the densest phyla in c_AMPs. As a reference, the median of ρ_{AMP} for the presented genera is indicated by a magenta dashed line.

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, proteobacticin-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

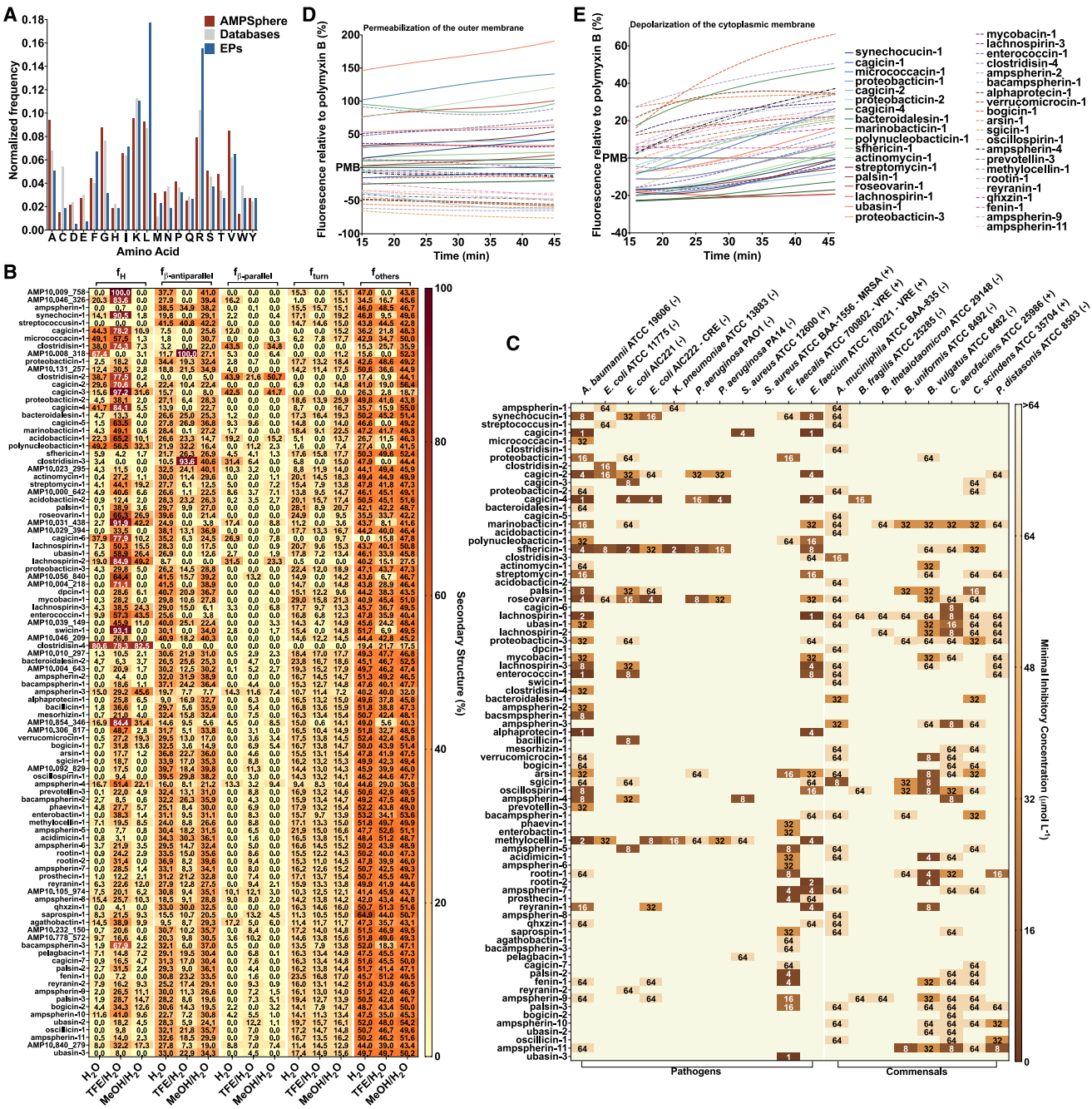


Figure 6. Amino acid composition, structure, antimicrobial activity, and mechanism of action of c_AMPs

(A) Amino acid frequency in c_AMPs from AMPSphere, AMPs from databases (DRAMP[46] version 3, APD3,[74] and DBAASP[73]), and encrypted peptides[4] (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, Bacteroidota, Actinomycetota, and Bacillota), i.e., *Akkermansia muciniphila*, *Bacteroides fragilis*, *Bacteroides thetaiotaomicron*, *Bacteroides uniformis*, *Phocaeicola vulgatus (*formerly *Bacteroides vulgatus)*, *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. vulgatus*, *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 lipids 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 ($DiSC_3$-[5]). Among the peptides tested against *A. baumannii*, bogicin-1 (AMP10.364_543), ampspherin-2 (AMP10.615_023), and marinobacticin-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.[4]

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 from different sources displayed potent *in vitro* activity against *A. baumannii*: synechocucin-1 (AMP10.000_211, 8 μmol L^{-1}) from *Synechococcus* sp. (coral-associated, marine microbiome); proteobacticin-1 (AMP10.048_551, 16 μmol L^{-1}) from Pseudomonadota (plant and soil microbiome); actynomycin-1 (AMP10.199_072, 64 μmol L^{-1}) from *Actinomyces* (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 ($DiSC_3$-[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.

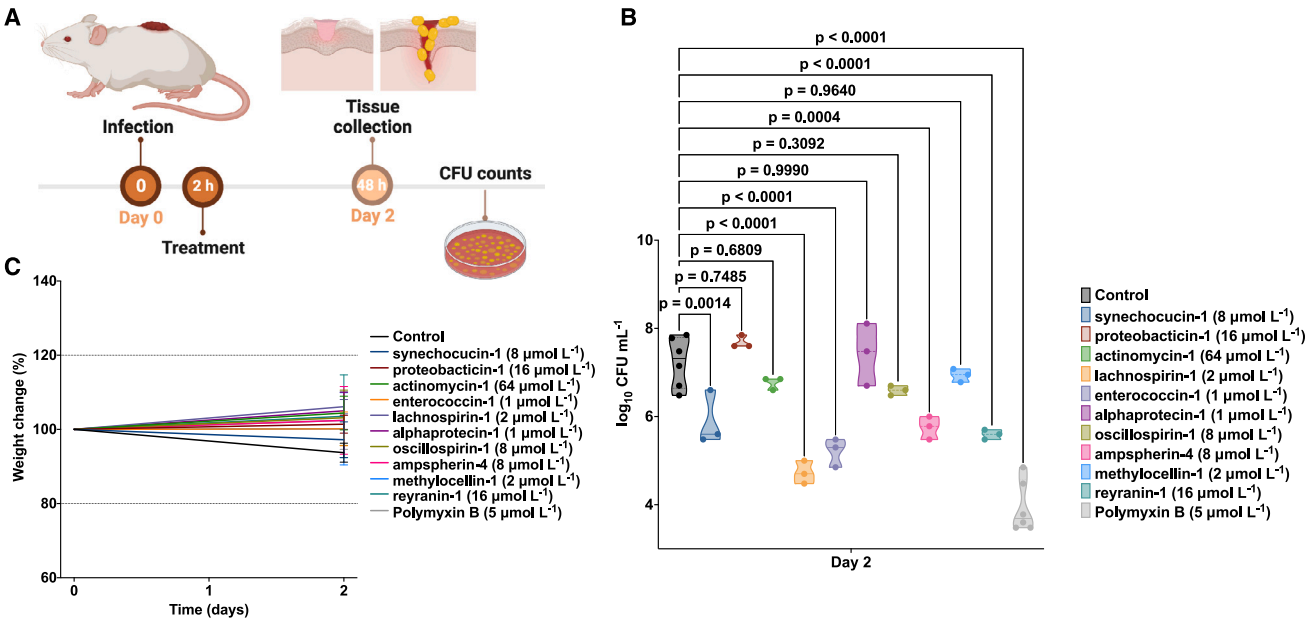


Figure 7. Anti-infective activity of AMPs in preclinical animal model

(A) Schematic of the skin abscess mouse model used to assess the anti-infective activity of the peptides against *A. baumannii* cells.

(B) Peptides were tested at their MIC in a single dose 2 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.

(C) To rule out toxic effects of the peptides, mouse weight was monitored throughout the experiment.

Statistical significance in (B) 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. Data in (C) are the mean ± the standard deviation. Figure created in BioRender.com.

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); ampspherin-4 (AMP10.466_287, 8 μmol L^{-1}) from an unknown source; methylocellin-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 μL of *A. baumannii* cells at 10^6 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, synechocucin-1, actynomycin-1, and oscillosporin-1 presented bacteriostatic activity, inhibiting the proliferation of *A. baumannii* cells, whereas lachnospirin-1, enterococcin-1, ampspherin-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, synechocucin-1, lachnospirin-1, enterococcin-1, and ampspherin-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 c_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[88] 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),[4] 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.[4] 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[55] 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.[38] 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,[62] possibly due to their amyloidogenic properties.[89] Other origins of AMPs may be horizontal gene transfer[90] or ancestral non-coding sequences.[91]

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.[92] 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[93–95] (Figure 5).

Valles-Colomer et al.,[72] who recently analyzed a large collection of human-associated metagenomes, provide a species-specific index of transmissibility for the several transmission scenarios they study (e.g., mother to infant). Hypothesizing that AMP production may be related to transmission, we correlated the species-specific ρ_{AMP} calculated in AMPSphere with transmission scores. In both the human gut and oral microbiomes, species with higher ρ_{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[42] 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 μmol L^{-1}, comparable to the MICs of some of the most potent peptides previously described in the literature.[84,85]

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,[82] 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 (particularly 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.

ACKNOWLEDGMENTS

We thank Marija Dmitrijeva (University of Zurich) for her helpful comments on a previous version of the manuscript. We thank Kaylyn Tousignant (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 insightful discussions. C.F.-N. holds a Presidential Professorship at the University of Pennsylvania and acknowledges funding from the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, the Penn Health-Tech Accelerator Award, Defense Threat Reduction Agency grants HDTRA11810041 and HDTRA1-23-1-0001, and the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania. We thank Dr. Mark Goulian for kindly donating the strains *Escherichia coli* AIC221 (*Escherichia coli* MG1655 phnE_2:FRT [control strain for AIC 222]) and *Escherichia coli* AIC222 (*Escherichia coli* MG1655 pmrA53 phnE_2:FRT [polymyxin-resistant]). This work was partly funded by the EMBL and the following grants: National Natural Science Foundation of China grants T2225015 and 61932008 (L.P.C. and X.-M.Z.); Shanghai Science and Technology Commission Program grant 23JS1410100 (L.P.C. and X.-M.Z.); National Key R&D Program of China grants 2023YFF1204800 and 2020YFA0712403 (L.P.C. and X.-M.Z.); Shanghai Municipal Science and Technology Major Project grant 2018SHZDZX01 (L.P.C. and X.-M.Z.); Lingang Laboratory and National Key Laboratory of Human Factors Engineering Joint Grant LG-TKN-202203-01 (X.-M.Z.); The Science and Technology Commission of Shanghai Municipality grant 22JC1410900 (L.P.C.); Australian Research Council grant FT230100724 (L.P.C.); the Langer Prize from the AIChE Foundation (C.F.-N.); National Institutes of Health grant R35GM138201 (C.F.-N.); Defense Threat Reduction Agency grant HDTRA1-21-1-0014 (C.F.-N.); PID2021-127210NB-I00, MCIN/AEI/10.13039/501100011033/FEDER, UE (J.H.-C.); 'la Caixa' Foundation ID 100010434, fellowship code LCF/BQ/DI18/11660009 (A.R.d.R.); and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 713673 (A.R.d.R.).

AUTHOR CONTRIBUTIONS

Conceptualization, C.D.S.-J., L.P.C., M.D.T.T., and C.F.-N.; Data curation, C.D.S.-J., Y.D., T.S.B.S., M.K., A.F., L.P.C., M.D.T.T., and C.F.-N.; Formal analysis, C.D.S.-J., L.P.C., and M.D.T.T.; Funding acquisition, L.P.C., X.-M.Z., and C.F.-N.; Investigation, C.D.S.-J., L.P.C., M.D.T.T., and C.F.-N.; Methodology, C.D.S.-J., Y.D., J.H.-C., A.R.d.R., L.P.C., M.D.T.T., and C.F.-N.; Project administration, L.P.C., M.K., X.-M.Z., P.B., and C.F.-N.; Resources, L.P.C., X.-M.Z., and C.F.-N.; Supervision, L.P.C. and C.F.-N.; Visualization, C.D.S.-J., J.H.-C., J.S., A.V., A.H., C.Z., L.P.C., and M.D.T.T.; Writing – original draft, C.D.S.-J., M.D.T.T., C.F.-N., and L.P.C.; Writing – review & editing, C.D.S.-J., Y.D., J.H.-C., A.R.d.R., T.S.B.S., A.F., P.B., X.-M.Z., L.P.C., M.D.T.T., and C.F.-N.

DECLARATION OF INTERESTS

C.F.-N. provides consulting services to Invaio 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

1. de la Fuente-Nunez, C., Torres, M.D., Mojica, F.J., and Lu, T.K. (2017). Next-generation precision antimicrobials: towards personalized treatment of infectious diseases. Curr. Opin. Microbiol. *37*, 95–102. https://doi.org/10.1016/j.mib.2017.05.014.
2. Antimicrobial Resistance Collaborators (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet *399*, 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0.
3. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. Cell *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.
4. Torres, M.D.T., Melo, M.C.R., Flowers, L., Crescenzi, O., Notomista, E., and de la Fuente-Nunez, C. (2022). Mining for encrypted peptide antibiotics in the human proteome. Nat. Biomed. Eng. *6*, 67–75. https://doi.org/10.1038/s41551-021-00801-1.
5. Porto, W.F., Irazazabal, L., Alves, E.S.F., Ribeiro, S.M., Matos, C.O., Pires, Á.S., Fensterseifer, I.C.M., Miranda, V.J., Haney, E.F., Humblot, V., et al. (2018). In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. Nat. Commun. *9*, 1490. https://doi.org/10.1038/s41467-018-03746-3.
6. Ma, Y., Guo, Z., Xia, B., Zhang, Y., Liu, X., Yu, Y., Tang, N., Tong, X., Wang, M., Ye, X., et al. (2022). Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat. Biotechnol. *40*, 921–931. https://doi.org/10.1038/s41587-022-01226-0.
7. Wong, F., de la Fuente-Nunez, C., and Collins, J.J. (2023). Leveraging artificial intelligence in the fight against infectious diseases. Science *381*, 164–170. https://doi.org/10.1126/science.adh1114.
8. Cesaro, A., Bagheri, M., Torres, M., Wan, F., and de la Fuente-Nunez, C. (2023). Deep learning tools to accelerate antibiotic discovery. Expert Opin. Drug Discov. *18*, 1245–1257. https://doi.org/10.1080/17460441.2023.2250721.
9. Torres, M.D.T., and de la Fuente-Nunez, C. (2019). Toward computer-made artificial antibiotics. Curr. Opin. Microbiol. *51*, 30–38. https://doi.org/10.1016/j.mib.2019.03.004.
10. Maasch, J.R.M.A., Torres, M.D.T., Melo, M.C.R., and de la Fuente-Nunez, C. (2023). Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. Cell Host Microbe *31*, 1260–1274.e6. https://doi.org/10.1016/j.chom.2023.07.001.
11. Besse, A., Vandervennet, M., Goulard, C., Peduzzi, J., Isaac, S., Rebuffat, S., and Carré-Mlouka, A. (2017). Halocin C8: an antimicrobial peptide

distributed among four halophilic archaeal genera: Natrinema, Haloterrigena, Haloferax, and Halobacterium. Extremophiles *21*, 623–638. https://doi.org/10.1007/s00792-017-0931-5.

12. Cotter, P.D., Ross, R.P., and Hill, C. (2013). Bacteriocins — a viable alternative to antibiotics? Nat. Rev. Microbiol. *11*, 95–105. https://doi.org/10.1038/nrmicro2937.
13. Wang, S., Zheng, Z., Zou, H., Li, N., and Wu, M. (2019). Characterization of the secondary metabolite biosynthetic gene clusters in archaea. Comput. Biol. Chem. *78*, 165–169. https://doi.org/10.1016/j.compbiolchem.2018.11.019.
14. Zasloff, M. (2019). Antimicrobial Peptides of Multicellular Organisms: My Perspective. In Antimicrobial Peptides: Basics for Clinical Application, K. Matsuzaki, ed. (Springer Singapore), pp. 3–6. https://doi.org/10.1007/978-981-13-3588-4_1.
15. Huang, K.-Y., Chang, T.-H., Jhong, J.-H., Chi, Y.-H., Li, W.-C., Chan, C.-L., Robert Lai, K., and Lee, T.-Y. (2017). Identification of natural antimicrobial peptides from bacteria through metagenomic and metatranscriptomic analysis of high-throughput transcriptome data of Taiwanese oolong teas. BMC Syst. Biol. *11*, 131. https://doi.org/10.1186/s12918-017-0503-4.
16. Torres, M.D.T., Sothiselvam, S., Lu, T.K., and de la Fuente-Nunez, C. (2019). Peptide Design Principles for Antimicrobial Applications. J. Mol. Biol. *431*, 3547–3567. https://doi.org/10.1016/j.jmb.2018.12.015.
17. Pizzo, E., Cafaro, V., Di Donato, A., and Notomista, E. (2018). Cryptic Antimicrobial Peptides: Identification Methods and Current Knowledge of their Immunomodulatory Properties. Curr. Pharm. Des. *24*, 1054–1066. https://doi.org/10.2174/1381612824666180327165012.
18. Nolan, E.M., and Walsh, C.T. (2009). How nature morphs peptide scaffolds into antibiotics. Chembiochem *10*, 34–53. https://doi.org/10.1002/cbic.200800438.
19. Singh, N., and Abraham, J. (2014). Ribosomally synthesized peptides from natural sources. J. Antibiot. *67*, 277–289. https://doi.org/10.1038/ja.2013.138.
20. García-Bayona, L., and Comstock, L.E. (2018). Bacterial antagonism in host-associated microbial communities. Science *361*, eaat2456. https://doi.org/10.1126/science.aat2456.
21. Anderson, M.C., Vonaesch, P., Saffarian, A., Marteyn, B.S., and Sansonetti, P.J. (2017). Shigella sonnei encodes a functional T6SS used for interbacterial competition and niche occupancy. Cell Host Microbe *21*, 769–776.e3. https://doi.org/10.1016/j.chom.2017.05.004.
22. Krismer, B., Weidenmaier, C., Zipperer, A., and Peschel, A. (2017). The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. Nat. Rev. Microbiol. *15*, 675–687. https://doi.org/10.1038/nrmicro.2017.104.
23. Zhao, W., Caro, F., Robins, W., and Mekalanos, J.J. (2018). Antagonism toward the intestinal microbiota and its effect on Vibrio cholerae virulence. Science *359*, 210–213. https://doi.org/10.1126/science.aap8775.
24. Quereda, J.J., Nahori, M.A., Meza-Torres, J., Sachse, M., Titos-Jiménez, P., Gomez-Laguna, J., Dussurget, O., Cossart, P., and Pizarro-Cerdá, J. (2017). Listeriolysin S is a streptolysin s-like virulence factor that targets exclusively prokaryotic cells in vivo. mBio *8*, e00259-17. https://doi.org/10.1128/mBio.00259-17.
25. Quereda, J.J., Dussurget, O., Nahori, M.A., Ghozlane, A., Volant, S., Dillies, M.A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B., et al. (2016). Bacteriocin from epidemic Listeria strains alters the host intestinal microbiota to favor infection. Proc. Natl. Acad. Sci. USA *113*, 5706–5711. https://doi.org/10.1073/pnas.1523899113.
26. Gomes, B., Augusto, M.T., Felício, M.R., Hollmann, A., Franco, O.L., Gonçalves, S., and Santos, N.C. (2018). Designing improved active peptides for therapeutic approaches against infectious diseases. Biotechnol. Adv. *36*, 415–429. https://doi.org/10.1016/j.biotechadv.2018.01.004.
27. Lesiuk, M., Paduszyńska, M., and Greber, K.E. (2022). Synthetic Antimicrobial Immunomodulatory Peptides: Ongoing Studies and Clinical Trials. Antibiotics (Basel) *11*, 1062. https://doi.org/10.3390/antibiotics11081062.
28. Mahlapuu, M., Håkansson, J., Ringstad, L., and Björn, C. (2016). Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. Front. Cell. Infect. Microbiol. *6*, 235805.
29. Baquero, F., Lanza, V.F., Baquero, M.R., Del Campo, R., and Bravo-Vázquez, D.A. (2019). Microcins in Enterobacteriaceae: peptide antimicrobials in the eco-active intestinal chemosphere. Front. Microbiol. *10*, 2261. https://doi.org/10.3389/fmicb.2019.02261.
30. Kim, S.G., Becattini, S., Moody, T.U., Shliaha, P.V., Littmann, E.R., Seok, R., Gjonbalaj, M., Eaton, V., Fontana, E., Amoretti, L., et al. (2019). Microbiota-derived lantibiotic restores resistance against vancomycin-resistant Enterococcus. Nature *572*, 665–669. https://doi.org/10.1038/s41586-019-1501-z.
31. Nakatsuji, T., Hata, T.R., Tong, Y., Cheng, J.Y., Shafiq, F., Butcher, A.M., Salem, S.S., Brinton, S.L., Rudman Spergel, A.K., Johnson, K., et al. (2021). Development of a human skin commensal microbe for bacteriotherapy of atopic dermatitis and use in a phase 1 randomized clinical trial. Nat. Med. *27*, 700–709. https://doi.org/10.1038/s41591-021-01256-2.
32. Spohn, R., Daruka, L., Lázár, V., Martins, A., Vidovics, F., Grézal, G., Méhi, O., Kintses, B., Számel, M., Jangir, P.K., et al. (2019). Integrated evolutionary analysis reveals antimicrobial peptides with limited resistance. Nat. Commun. *10*, 4538. https://doi.org/10.1038/s41467-019-12364-6.
33. Cesaro, A., Torres, M.D.T., Gaglione, R., Dell'Olmo, E., Di Girolamo, R., Bosso, A., Pizzo, E., Haagsman, H.P., Veldhuizen, E.J.A., de la Fuente-Nunez, C., and Arciello, A. (2022). Synthetic Antibiotic Derived from Sequences Encrypted in a Protein from Human Plasma. ACS Nano *16*, 1880–1895. https://doi.org/10.1021/acsnano.1c04496.
34. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.
35. Ahrens, C.H., Wade, J.T., Champion, M.M., and Langer, J.D. (2022). A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. J. Bacteriol. *204*, e0035321. https://doi.org/10.1128/JB.00353-21.
36. Storz, G., Wolf, Y.I., and Ramamurthi, K.S. (2014). Small Proteins Can No Longer Be Ignored. Annu. Rev. Biochem. *83*, 753–777. https://doi.org/10.1146/annurev-biochem-070611-102400.
37. Su, M., Ling, Y., Yu, J., Wu, J., and Xiao, J. (2013). Small proteins: untapped area of potential biological importance. Front. Genet. *4*, 286. https://doi.org/10.3389/fgene.2013.00286.
38. Sberro, H., Fremin, B.J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M.P., Pavlopoulos, G.A., Kyrpides, N.C., and Bhatt, A.S. (2019). Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell *178*, 1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.
39. Donia, M.S., Cimermancic, P., Schulze, C.J., Wieland Brown, L.C., Martin, J., Mitreva, M., Clardy, J., Linington, R.G., and Fischbach, M.A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell *158*, 1402–1414. https://doi.org/10.1016/j.cell.2014.08.032.
40. Fingerhut, L.C.H.W., Miller, D.J., Strugnell, J.M., Daly, N.L., and Cooke, I.R. (2021). ampir: an R package for fast genome-wide prediction of antimicrobial peptides. Bioinformatics *36*, 5262–5263. https://doi.org/10.1093/bioinformatics/btaa653.
41. Sugimoto, Y., Camacho, F.R., Wang, S., Chankhamjon, P., Odabas, A., Biswas, A., Jeffrey, P.D., and Donia, M.S. (2019). A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science *366*, eaax9176. https://doi.org/10.1126/science.aax9176.

42. Santos-Júnior, C.D., Pan, S., Zhao, X.-M., and Coelho, L.P. (2020). Macrel: antimicrobial peptide screening in genomes and metagenomes. PeerJ *8*, e10555. https://doi.org/10.7717/peerj.10555.

43. Mende, D.R., Letunic, I., Maistrenko, O.M., Schmidt, T.S.B., Milanese, A., Paoli, L., Hernández-Plaza, A., Orakov, A.N., Forslund, S.K., Sunagawa, S., et al. (2020). proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res. *48*, D621–D625. https://doi.org/10.1093/nar/gkz1002.

44. Navidinia, M. (2016). The clinical importance of emerging ESKAPE pathogens in nosocomial infections. Archives of Advances in Biosciences *7*, 43–57. https://doi.org/10.22037/jps.v7i3.12584.

45. Mulani, M.S., Kamble, E.E., Kumkar, S.N., Tawre, M.S., and Pardesi, K.R. (2019). Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. Front. Microbiol. *10*, 539. https://doi.org/10.3389/fmicb.2019.00539.

46. Shi, G., Kang, X., Dong, F., Liu, Y., Zhu, N., Hu, Y., Xu, H., Lao, X., and Zheng, H. (2022). DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res. *50*, D488–D496. https://doi.org/10.1093/nar/gkab651.

47. Zhang, L.-J., and Gallo, R.L. (2016). Antimicrobial peptides. Curr. Biol. *26*, R14–R19. https://doi.org/10.1016/j.cub.2015.11.017.

48. Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S.W.I. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci. Rep. *8*, 1697. https://doi.org/10.1038/s41598-018-19752-w.

49. Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., and Chen, R. (2018). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief. Bioinform. *19*, 636–643. https://doi.org/10.1093/bib/bbx005.

50. Venturini, E., Svensson, S.L., Maaß, S., Gelhausen, R., Eggenhofer, F., Li, L., Cain, A.K., Parkhill, J., Becher, D., Backofen, R., et al. (2020). A global data-driven census of Salmonella small proteins and their potential functions in bacterial virulence. microLife *1*, uqaa002. https://doi.org/10.1093/femsml/uqaa002.

51. Aguilera-Mendoza, L., Marrero-Ponce, Y., Beltran, J.A., Tellez Ibarra, R., Guillen-Ramirez, H.A., and Brizuela, C.A. (2019). Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. Bioinformatics *35*, 4739–4747. https://doi.org/10.1093/bioinformatics/btz260.

52. Coelho, L.P., Alves, R., Del Río, Á.R., Myers, P.N., Cantalapiedra, C.P., Giner-Lamia, J., Schmidt, T.S., Mende, D.R., Orakov, A., Letunic, I., et al. (2022). Towards the biogeography of prokaryotic genes. Nature *601*, 252–256. https://doi.org/10.1038/s41586-021-04233-4.

53. Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. Bioinformatics *34*, 2740–2747. https://doi.org/10.1093/bioinformatics/bty179.

54. Lawrence, T.J., Carper, D.L., Spangler, M.K., Carrell, A.A., Rush, T.A., Minter, S.J., Weston, D.J., and Labbé, J.L. (2021). amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. Bioinformatics *37*, 2058–2060. https://doi.org/10.1093/bioinformatics/btaa917.

55. Su, X., Xu, J., Yin, Y., Quan, X., and Zhang, H. (2019). Antimicrobial peptide identification using multi-scale convolutional network. BMC Bioinf. *20*, 730. https://doi.org/10.1186/s12859-019-3327-y.

56. Lin, T.-T., Yang, L.-Y., Lu, I.-H., Cheng, W.-C., Hsu, Z.-R., Chen, S.-H., and Lin, C.-Y. (2021). AI4AMP: an Antimicrobial Peptide Predictor Using Physicochemical Property-Based Encoding Method and Deep Learning. mSystems *6*, e0029921. https://doi.org/10.1128/mSystems.00299-21.

57. Li, C., Sutherland, D., Hammond, S.A., Yang, C., Taho, F., Bergman, L., Houston, S., Warren, R.L., Wong, T., Hoang, L.M.N., et al. (2022). AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against whom priority pathogens. BMC Genom. *23*, 77. https://doi.org/10.1186/s12864-022-08310-4.

58. Murphy, L.R., Wallqvist, A., and Levy, R.M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. *13*, 149–152. https://doi.org/10.1093/protein/13.3.149.

59. Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., de Beaufort, C., and Wilmes, P. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat. Microbiol. *2*, 16180. https://doi.org/10.1038/nmicrobiol.2016.180.

60. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

61. Rodríguez del Río, Á., Giner-Lamia, J., Cantalapiedra, C.P., Botas, J., Deng, Z., Hernández-Plaza, A., Munar-Palmer, M., Santamaría-Hernando, S., Rodríguez-Herva, J.J., Ruscheweyh, H.-J., et al. (2023). Functional and evolutionary significance of unknown genes from uncultivated taxa. Nature, 1–3. https://doi.org/10.1038/s41586-023-06955-z.

62. Hurtado-Rios, J.J., Carrasco-Navarro, U., Almanza-Pérez, J.C., and Ponce-Alquicira, E. (2022). Ribosomes: The New Role of Ribosomal Proteins as Natural Antimicrobials. Int. J. Mol. Sci. *23*, 9123. https://doi.org/10.3390/ijms23169123.

63. Shoja, V., and Zhang, L. (2006). A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat. Mol. Biol. Evol. *23*, 2134–2141. https://doi.org/10.1093/molbev/msl085.

64. Sukhodolets, V.V. (2006). Unequal crossing-over in Escherichia coli. Russ. J. Genet. *42*, 1285–1293. https://doi.org/10.1134/S102279540611010X.

65. Kim, M.K., Kang, T.H., Kim, J., Kim, H., and Yun, H.D. (2012). Evidence Showing Duplication and Recombination of cel Genes in Tandem from Hyperthermophilic Thermotoga sp. Appl. Biochem. Biotechnol. *168*, 1834–1848. https://doi.org/10.1007/s12010-012-9901-7.

66. Blaustein, R.A., McFarland, A.G., Ben Maamar, S., Lopez, A., Castro-Wallace, S., and Hartmann, E.M. (2019). Pangenomic Approach To Understanding Microbial Adaptations within a Model Built Environment, the International Space Station, Relative to Human Hosts and Soil. mSystems *4*, e00281-18. https://doi.org/10.1128/mSystems.00281-18.

67. Collins, F.W.J., Mesa-Pereira, B., O'Connor, P.M., Rea, M.C., Hill, C., and Ross, R.P. (2018). Reincarnation of Bacteriocins From the Lactobacillus Pangenomic Graveyard. Front. Microbiol. *9*, 1298. https://doi.org/10.3389/fmicb.2018.01298.

68. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7.

69. Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. *38*, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8.

70. Simmons, W.L., Daubenspeck, J.M., Osborne, J.D., Balish, M.F., Waites, K.B., and Dybvig, K. (2013). Type 1 and type 2 strains of Mycoplasma pneumoniae form different biofilms. Microbiology (Read.) *159*, 737–747. https://doi.org/10.1099/mic.0.064782-0.

71. Diaz, M.H., Desai, H.P., Morrison, S.S., Benitez, A.J., Wolff, B.J., Caravas, J., Read, T.D., Dean, D., and Winchell, J.M. (2017). Comprehensive bioinformatics analysis of Mycoplasma pneumoniae genomes to investigate underlying population structure and type-specific determinants. PLoS One *12*, e0174701. https://doi.org/10.1371/journal.pone.0174701.

72. Valles-Colomer, M., Blanco-Míguez, A., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., Armanini, F., Cumbo, F., Huang, K.D., Manara, S., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. Nature *614*, 125–135. https://doi.org/10.1038/s41586-022-05620-1.

73. Pirtskhalava, M., Amstrong, A.A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D.E., and Tartakovsky, M. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Res. *49*, D288–D297. https://doi.org/10.1093/nar/gkaa991.

74. Wang, G., Li, X., and Wang, Z. (2016). APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. *44*, D1087–D1093. https://doi.org/10.1093/nar/gkv1278.

75. Micsonai, A., Moussong, É., Wien, F., Boros, E., Vadászi, H., Murvai, N., Lee, Y.-H., Molnár, T., Réfrégiers, M., Goto, Y., et al. (2022). BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Res. *50*, W90–W98. https://doi.org/10.1093/nar/gkac345.

76. Lifson, S., and Sander, C. (1979). Antiparallel and parallel β-strands differ in amino acid residue preferences. Nature *282*, 109–111. https://doi.org/10.1038/282109a0.

77. Derrien, M., Collado, M.C., Ben-Amor, K., Salminen, S., and de Vos, W.M. (2008). The Mucin Degrader Akkermansia muciniphila Is an Abundant Resident of the Human Intestinal Tract. Appl. Environ. Microbiol. *74*, 1646–1648. https://doi.org/10.1128/AEM.01226-07.

78. Earley, H., Lennon, G., Balfe, Á., Coffey, J.C., Winter, D.C., and O'Connell, P.R. (2019). The abundance of Akkermansia muciniphila and its relationship with sulphated colonic mucins in health and ulcerative colitis. Sci. Rep. *9*, 15683. https://doi.org/10.1038/s41598-019-51878-3.

79. Daquigan, N., Seekatz, A.M., Greathouse, K.L., Young, V.B., and White, J.R. (2017). High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. npj Biofilms Microbiomes *3*, 35. https://doi.org/10.1038/s41522-017-0043-0.

80. Saenz, C., Fang, Q., Gnanasekaran, T., Trammell, S.A.J., Buijink, J.A., Pisano, P., Wierer, M., Moens, F., Lengger, B., Brejnrod, A., and Arumugam, M. (2023). Clostridium scindens secretome suppresses virulence gene expression of Clostridioides difficile in a bile acid-independent manner. Microbiol. Spectr. *11*, e0393322. https://doi.org/10.1128/spectrum.03933-22.

81. Geerlings, S.Y., Kostopoulos, I., De Vos, W.M., and Belzer, C. (2018). Akkermansia muciniphila in the Human Gastrointestinal Tract: When, Where, and How? Microorganisms *6*, 75. https://doi.org/10.3390/microorganisms6030075.

82. Cullen, T.W., Schofield, W.B., Barry, N.A., Putnam, E.E., Rundell, E.A., Trent, M.S., Degnan, P.H., Booth, C.J., Yu, H., and Goodman, A.L. (2015). Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. Science *347*, 170–175. https://doi.org/10.1126/science.1260580.

83. Torres, M.D.T., Pedron, C.N., Araújo, I., Silva, P.I., Silva, F.D., and Oliveira, V.X. (2017). Decoralin Analogs with Increased Resistance to Degradation and Lower Hemolytic Activity. ChemistrySelect *2*, 18–23. https://doi.org/10.1002/slct.201601590.

84. Torres, M.D.T., Pedron, C.N., Higashikuni, Y., Kramer, R.M., Cardoso, M.H., Oshiro, K.G.N., Franco, O.L., Silva Junior, P.I., Silva, F.D., Oliveira Junior, V.X., et al. (2018). Structure-function-guided exploration of the antimicrobial peptide polybia-CP identifies activity determinants and generates synthetic therapeutic candidates. Commun. Biol. *1*, 221. https://doi.org/10.1038/s42003-018-0224-2.

85. Silva, O.N., Torres, M.D.T., Cao, J., Alves, E.S.F., Rodrigues, L.V., Resende, J.M., Lião, L.M., Porto, W.F., Fensterseifer, I.C.M., Lu, T.K., et al. (2020). Repurposing a peptide toxin from wasp venom into antiinfectives with dual antimicrobial and immunomodulatory properties. Proc. Natl. Acad. Sci. USA *117*, 26936–26945. https://doi.org/10.1073/pnas.2012379117.

86. Morris, F.C., Dexter, C., Kostoulias, X., Uddin, M.I., and Peleg, A.Y. (2019). The Mechanisms of Disease Caused by Acinetobacter baumannii. Front. Microbiol. *10*, 1601.

87. Petruschke, H., Schori, C., Canzler, S., Riesbeck, S., Poehlein, A., Daniel, R., Frei, D., Segessemann, T., Zimmerman, J., Marinos, G., et al. (2021). Discovery of novel community-relevant small proteins in a simplified human intestinal microbiome. Microbiome *9*, 55. https://doi.org/10.1186/s40168-020-00981-z.

88. Washietl, S., Findeiß, S., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. RNA *17*, 578–594. https://doi.org/10.1261/rna.2536111.

89. Galzitskaya, O.V. (2021). Exploring Amyloidogenicity of Peptides From Ribosomal S1 Protein to Develop Novel AMPs. Front. Mol. Biosci. *8*, 705069. https://doi.org/10.3389/fmolb.2021.705069.

90. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304. https://doi.org/10.1038/35012500.

91. Zheng, D., and Gerstein, M.B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet. *23*, 219–224. https://doi.org/10.1016/j.tig.2007.03.003.

92. Lazzaro, B.P., Zasloff, M., and Rolff, J. (2020). Antimicrobial peptides: Application informed by evolution. Science *368*, eaau5480. https://doi.org/10.1126/science.aau5480.

93. Sun, S., Wang, H., Howard, A.G., Zhang, J., Su, C., Wang, Z., Du, S., Fodor, A.A., Gordon-Larsen, P., and Zhang, B. (2022). Loss of Novel Diversity in Human Gut Microbiota Associated with Ongoing Urbanization in China. mSystems *7*, e0020022. https://doi.org/10.1128/msystems.00200-22.

94. Piquer-Esteban, S., Ruiz-Ruiz, S., Arnau, V., Diaz, W., and Moya, A. (2022). Exploring the universal healthy human gut microbiota around the World. Comput. Struct. Biotechnol. J. *20*, 421–433. https://doi.org/10.1016/j.csbj.2021.12.035.

95. Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K.R., and Sharma, V.K. (2019). The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience *8*, giz004. https://doi.org/10.1093/gigascience/giz004.

96. Coelho, L.P., Alves, R., Monteiro, P., Huerta-Cepas, J., Freitas, A.T., and Bork, P. (2019). NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. Microbiome *7*, 84. https://doi.org/10.1186/s40168-019-0684-8.

97. Coelho, L.P. (2017). Jug: Software for Parallel Reproducible Computation in Python. J. Open Res. Softw. *5*, 30. https://doi.org/10.5334/jors.161.

98. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

99. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

100. Van Rossum, G. (2020). Python Release Python 3.8.2. Python.org. https://www.python.org/downloads/release/python-382/.

101. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

102. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

103. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

104. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

105. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Machine Learning In Python *12*, 2825–2830.

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

107. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

108. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

109. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

110. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

111. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114. https://doi.org/10.1038/s41467-018-07641-9.

112. Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., Yamashita, H., and Lam, T.W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.

113. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

114. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference, pp. 92–96. https://doi.org/10.25080/Majora-92bf1922-011.

115. Milanese, A., Mende, D.R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., Alves, R., Costea, P.I., Coelho, L.P., et al. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. Nat. Commun. *10*, 1014. https://doi.org/10.1038/s41467-019-08844-4.

116. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

117. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

118. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539. https://doi.org/10.1038/msb.2011.75.

119. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

120. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

121. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

122. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

123. Eberhardt, R.Y., Haft, D.H., Punta, M., Martin, M., O'Donovan, C., and Bateman, A. (2012). AntiFam: a tool to help identify spurious ORFs in protein annotation. Database *2012*, bas003. https://doi.org/10.1093/database/bas003.

124. NCBI Resource Coordinators (2015). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *43*, D6–D17. https://doi.org/10.1093/nar/gku1130.

125. Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L.V., Cheng, A.A., Liu, S., et al. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. *48*, D517–D525. https://doi.org/10.1093/nar/gkz935.

126. Kanehisa, M., and Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. *29*, 28–35. https://doi.org/10.1002/pro.3711.

127. Courtot, M., Cherubin, L., Faulconbridge, A., Vaughan, D., Green, M., Richardson, D., Harrison, P., Whetzel, P.L., Parkinson, H., and Burdett, T. (2019). BioSamples database: an updated sample metadata hub. Nucleic Acids Res. *47*, D1172–D1178. https://doi.org/10.1093/nar/gky1061.

128. Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., et al. (2021). The European Nucleotide Archive in 2020. Nucleic Acids Res. *49*, D82–D85. https://doi.org/10.1093/nar/gkaa1028.

129. Jones, P., Côté, R.G., Martens, L., Quinn, A.F., Taylor, C.F., Derache, W., Hermjakob, H., and Apweiler, R. (2006). PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res. *34*, D659–D663. https://doi.org/10.1093/nar/gkj138.

130. Schmidt, T.S.B., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O.M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., Sunagawa, S., et al. (2024). SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res. *52*, D777–D783. https://doi.org/10.1093/nar/gkad943.

131. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., and Levy Karin, E. (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics *37*, 3029–3031. https://doi.org/10.1093/bioinformatics/btab184.

132. Oren, A., Arahal, D.R., Rosselló-Móra, R., Sutcliffe, I.C., and Moore, E.R.B. (2021). Emendation of Rules 5b, 8, 15 and 22 of the International Code of Nomenclature of Prokaryotes to include the rank of phylum. Int. J. Syst. Evol. Microbiol. *71*. https://doi.org/10.1099/ijsem.0.004851.

133. Oren, A., and Garrity, G.M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. Int. J. Syst. Evol. Microbiol. *71*. https://doi.org/10.1099/ijsem.0.005056.

134. Solis, A.D. (2015). Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. Proteins *83*, 2198–2216. https://doi.org/10.1002/prot.24936.

135. Peterson, E.L., Kondev, J., Theriot, J.A., and Phillips, R. (2009). Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics *25*, 1356–1362. https://doi.org/10.1093/bioinformatics/btp164.

136. Smith, T.F., and Waterman, M.S. (1981). Identification of Common Molecular Subsequences. J. Mol. Biol. *147*, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

137. Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA *87*, 2264–2268. https://doi.org/10.1073/pnas.87.6.2264.

138. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

139. Cena, J.A. de, Zhang, J., Deng, D., Damé-Teixeira, N., and Do, T. (2021). Low-Abundant Microorganisms: The Human Microbiome's Dark Matter, a Scoping Review. Front. Cell. Infect. Microbiol. *11*, 689197.

140. Mende, D.R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat. Methods *10*, 881–884. https://doi.org/10.1038/nmeth.2575.

141. Sélem-Mojica, N., Aguilar, C., Gutiérrez-García, K., Martínez-Guerrero, C.E., and Barona-Gómez, F. (2019). EvoMining reveals the origin and fate of natural product biosynthetic enzymes. Microb. Genom. *5*, e000260. https://doi.org/10.1099/mgen.0.000260.

142. Rodriguez-R, L.M., Conrad, R.E., Viver, T., Feistel, D.J., Lindner, B.G., Venter, S.N., Orellana, L.H., Amann, R., Rossello-Mora, R., and Konstantinidis, K.T. (2024). An ANI gap within bacterial species that advances the definitions of intra-species units. mBio *15*, e02696-23. https://doi.org/10.1128/mbio.02696-23.

143. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285. https://doi.org/10.1093/nar/gkv1344.

144. SolyPep: a fast generator of soluble peptides https://bioserv.rpbs.univ-paris-diderot.fr/services/SolyPep/

145. Ochoa, R., and Cossio, P. (2021). PepFun: Open Source Protocols for Peptide-Related Computational Analysis. Molecules *26*, 1664. https://doi.org/10.3390/molecules26061664.

146. Kochendoerfer, G.G., and Kent, S.B. (1999). Chemical protein synthesis. Curr. Opin. Chem. Biol. *3*, 665–671. https://doi.org/10.1016/s1367-5931(99)00024-1.

147. Sheppard, R. (2003). The fluorenylmethoxycarbonyl group in solid phase synthesis. J. Pept. Sci. *9*, 545–552. https://doi.org/10.1002/psc.479.

148. Palomo, J.M. (2014). Solid-phase peptide synthesis: an overview focused on the preparation of biologically relevant peptides. RSC Adv. *4*, 32658–32672. https://doi.org/10.1039/C4RA02458C.

149. Schmidt, T.S.B., Li, S.S., Maistrenko, O.M., Akanni, W., Coelho, L.P., Dolai, S., Fullam, A., Glazek, A.M., Hercog, R., Herrema, H., et al. (2022). Drivers and determinants of strain dynamics following fecal microbiota transplantation. Nat. Med. *28*, 1902–1912. https://doi.org/10.1038/s41591-022-01913-0.

150. Wiegand, I., Hilpert, K., and Hancock, R.E.W. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. Nat. Protoc. *3*, 163–175. https://doi.org/10.1038/nprot.2007.521.

151. Santos-Júnior, C.D., Schmidt, T.S.B., Fullam, A., Duan, Y., Bork, P., Zhao, X.-M., and Coelho, L.P. (2021). AMPSphere : The Worldwide Survey of Prokaryotic Antimicrobial Peptides (Zenodo) https://doi.org/10.5281/zenodo.4606582.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Acinetobacter baumannii	American Type Culture Collection	ATCC 19606
Escherichia coli	American Type Culture Collection	ATCC 11775
Escherichia coli	*Escherichia coli* MG1655 phnE_2:FRT	AIC221
Escherichia coli	*Escherichia coli* MG1655 pmrA53 phnE_2:FRT (polymyxin-resistant; colistin-resistant strain)	AIC222
Klebsiella pneumoniae	American Type Culture Collection	ATCC 13883
Pseudomonas aeruginosa	N/A	PAO1
Pseudomonas aeruginosa	N/A	PA14
Staphylococcus aureus	American Type Culture Collection	ATCC 12600
Staphylococcus aureus	American Type Culture Collection	ATCC BAA-1556 (methicillin-resistant strain)
Akkermansia muciniphila	American Type Culture Collection	ATCC BAA-635
Bacteroides fragilis	American Type Culture Collection	ATCC 25285
Bacteroides thetaiotaomicron	American Type Culture Collection	ATCC 29148
Bacteroides uniformis	American Type Culture Collection	ATCC 8492
Bacteroides vulgatus (Phocaeicola vulgatus)	American Type Culture Collection	ATCC 8482
Collinsella aerofaciens	American Type Culture Collection	ATCC 25986
Clostridium scindens	American Type Culture Collection	ATCC 35704
Parabacteroides distasonis	American Type Culture Collection	ATCC 8503
Chemicals, peptides, and recombinant proteins		
Luria-Bertani broth	BD	244620
Tryptic soy broth	Sigma	T8907-1KG
Agar	Sigma	05039
MacConkey agar	RPI	M42560–500.0
Phosphate buffer saline	Sigma	P3913-10PAK
Glucose	Sigma	G5767
1-(N-phenylamino)naphthalene	Sigma	104043
3,3′-dipropylthiadicarbocyanine iodide	Sigma	43608
HEPES	Fisher	BP310-100
Potassium chloride (KCl)	Sigma	P3911
Deposited data		
Code for generation of AMPSphere	This study	https://doi.org/10.5281/zenodo.11055585
AMPSphere database	This study	https://zenodo.org/record/4606582
Experimental models: Organisms/strains		
Mouse: CD-1	Charles River	18679700–022
Software and algorithms		
NGLess 1.3.0	Coelho et al.[96]	https://github.com/ngless-toolkit/ngless
JUG 2.1.1	Coelho[97]	https://github.com/luispedro/jug
Prodigal 2.6.3	Hyatt et al.[34]	https://github.com/hyattpd/Prodigal
Macrel v.1.0.0	Santos-Júnior et al.[42]	https://github.com/BigDataBiology/macrel
CDHit 4.8.1	Fu et al.[98]	https://github.com/weizhongli/cdhit
MMseqs2	Steinegger and Söding[99]	https://github.com/soedinglab/MMseqs2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
python 3.8.2	Van Rossum[100]	https://www.python.org/
matplotlib 3.4.3	Hunter[101]	https://matplotlib.org/
numpy 1.21.2	Harris et al.[102]	https://numpy.org/
pandas 1.3.2	McKinney[103]	https://pandas.pydata.org/
plotly 5.2.1	Plotly Technologies Inc, 2015	https://plot.ly
scipy 1.7.1	Virtanen et al.[104]	https://www.scipy.org
scikit-learn 0.24	Pedregosa et al.[105]	https://scikit-learn.org/
scikit-bio 0.5.6	The scikit-bio development team, 2020[106]	http://scikit-bio.org/
BioPython 1.7.9	Cock et al.[107]	https://biopython.org/
eggnog-mapper v2	Cantalapiedra et al.[108]	https://github.com/eggnogdb/eggnog-mapper
HMMer 3.3+dfsg2-1	Eddy[109]	http://hmmer.org/
FastTree 2.1	Price et al.[110]	http://www.microbesonline.org/fasttree/
FastANI v.1.33	Jain et al.[111]	https://github.com/ParBLiSS/FastANI
Megahit 1.2.9	Li et al.[112]	https://github.com/voutcn/megahit/
AMPlify	Li et al.[57]	https://github.com/bcgsc/AMPlify
Ampir	Fingerhut et al.[40]	https://github.com/Legana/ampir
AMPScanner v2	Veltri et al.[53]	https://www.dveltri.com/ascan/v2/ascan.html
APIN	Su et al.[55]	https://github.com/zhanglabNKU/APIN
amPEPpy 1.0	Lawrence et al.[54]	https://github.com/tlawrence3/amPEPpy
AI4AMP	Lin et al.[56]	https://github.com/LinTzuTang/AI4AMP_predictor
RNAcode 0.2-beta	Washietl et al.[88]	https://github.com/ViennaRNA/RNAcode
Bwa v.0.7.17	Li et al.[113]	https://github.com/lh3/bwa
Statsmodels 0.14.0	Seabold and Perktold[114]	https://www.statsmodels.org
mOTUs2	Milanese et al.[115]	https://github.com/motu-tool/mOTUs
SAMtools 1.18	Li et al.[116]	https://github.com/samtools/samtools
BEDtools v2.31.0	Quinlan and Hall[117]	https://github.com/arq5x/bedtools2
Clustal Omega 1.2.2	Sievers et al.[118]	http://clustal.org/omega/
Diamond v2.1.8	Buchfink et al.[119]	https://github.com/bbuchfink/diamond
Blast+ 2.13.0	Camacho et al.[120]	https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html
Other		
ProGenomes2	Mende et al.[43]	http://progenomes.embl.de/
DRAMP - Data repository of antimicrobial peptides 3.0	Shi et al.[46]	http://dramp.cpu-bioinfor.org/
UniprotKB 2021_03	The UniProt Consortium[121]	https://www.uniprot.org/
Eggnog v.5.0	Huerta-Cepas et al.[60]	http://eggnog5.embl.de/
SmProt database v.2.0	Hao et al.[49]	http://bigdata.ibp.ac.cn/SmProt/index.html
StarPep45k	Aguilera-Mendoza et al.[51]	http://mobiosd-hub.com/starpep
PFAM 33.1.	Mistry et al.[122]	http://pfam.xfam.org/
AntiFAM v.7.0	Eberhardt et al.[123]	https://www.ebi.ac.uk/research/bateman/software/antifam-tool-identify-spurious-proteins
GTDB 07-RS95	Parks et al.[68,69]	https://gtdb.ecogenomic.org/
NCBI release 207	NCBI Resource Coordinators[124]	https://ftp.ncbi.nih.gov/refseq/release/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Database of Antimicrobial Activity and Structure of Peptides - DBAASP	Pirtskhalava et al.[73]	https://dbaasp.org/home
Antimicrobial peptides database - APD3	Wang and Wang[74]	https://aps.unmc.edu/
Salmonella Typhimurium small ORFs - STsORFs	Venturini et al.[50]	https://academic.oup.com/microlife/article/1/1/uqaa002/5928550#supplementary-data
CARD - Comprehensive Antibiotic Resistance Database	Alcock et al.[125]	https://card.mcmaster.ca/
Kyoto Encyclopedia of Genes and Genomes (KEGG) release 102	Kanehisa et al.[126]	https://www.genome.jp/kegg/
Biosamples database	Courtot et al.[127]	http://www.ebi.ac.uk/biosamples
European Nucleotide Archive - ENA	Harrison et al.[128]	https://www.ebi.ac.uk/ena
Proteomics Identification Database - PRIDE	Jones et al.[129]	https://www.ebi.ac.uk/pride/

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 phnE_2FRT (control strain for AIC 222)], *Escherichia coli* AIC222 [*Escherichia coli* MG1655 pmrA53 phnE_2FRT (polymyxin-resistant; colistin-resistant strain)], *Klebsiella pneumoniae* ATCC 13883, *Pseudomonas aeruginosa* PAO1, *Pseudomonas aeruginosa* PA14, *Staphylococcus aureus* ATCC 12600, *Staphylococcus aureus* ATCC BAA-1556 (methicillin-resistant strain), *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 (OD_{600} = 0.3–0.5).

The gut commensal strains *Akkermansia muciniphila* ATCC BAA-635, *Bacteroides fragilis* ATCC 25285, *Bacteroides thetaiotaomicron* ATCC 29148, *Bacteroides uniformis* ATCC 8492, *Bacteroides vulgatus* ATCC 8482 (*Phocaeicola vulgatus*), *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 at 37°C. The next day, inocula were prepared by diluting the bacterial overnight cultures 1:100 in 3 mL of BHI broth and incubated at 37°C until cells reached the logarithmic phase (OD_{600} = 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 $5 \cdot 10^6$ colony-forming units (CFU) per mL^{-1}. 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 μL 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.[52,130] 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.[127] 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.[43] 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,[96] 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 1.2.9[112] and the taxonomy of the 16,969,685,977 contigs generated was inferred as previously described,[131] using MMSeqs2[99] to map the sequences against the GTDB release 95.[68,69] Mapped taxonomy lineages were then manually curated to conform to the International Code of Nomenclature of Prokaryotes.[132,133]

smORF and AMP prediction

Analogously to Sberro et al.,[38] we used a modified version of Prodigal[34] 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[42] 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)[46] version 3.0 using the 'easy-search' method from MMSeqs2.[99] 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[43] 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.[107] 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,[134,135] we used a reduced amino acids alphabet of 8 letters[58] - [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.[98]

A cluster was considered an AMP family when it consisted of at least 8 sequences.[38] 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[136] 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,[137] which uses the values of κ (0.132539) and λ (0.313667) constants adjusted to search for a short input sequence as implemented in the BLAST algorithm.[120,138] 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 $\geq$85% of identity) were significant, along with 77.1% of those from the third level ($\geq$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,[53] ampir[40] - with the model for mature peptides, amPEPpy,[54] APIN[55] – with their proposed model, AI4AMP,[56] and AMPLify[57]).

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 v.7.0[123] using HMMSearch,[109] which was designed to identify commonly recurring spuriously 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[46] version 3.0 were found as terminal c_AMPs in AMPSphere.

The RNAcode[88] 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[129] database comprising from 37 habitats - Table S6. Using bwa v.0.7.17,[113] reads from the metatranscriptomes were mapped against non-redundant AMP genes, and, using NGLess,[96] 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.[42] Using regular expressions implemented in Python 3.8,[100] 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.[6] 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[52] database with experimental evidence of translation (3.4% - 2,073 out of 61,020 peptides, Odds Ratio = 4.3, $P_{\text{Fisher's exact}} < 10^{-300}$) and/or transcription (22.8% - 13,901 out of 61,020 peptides, Odds Ratio = 1.2, $P_{\text{Fisher's exact}} = 6.7 \cdot 10^{-108}$). 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 dataset 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 $\geq$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-Wilks 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 species[139]). 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 $p \leq \frac{1}{Z^2}$, 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 $\rho_{AMP} = \frac{n_{c_{AMPs}}}{L}$, where $n_{c_{AMPs}}$ 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: $STDerr = \sqrt{\frac{\rho*(1-\rho)}{L}}$. The standard sample proportion error was used to calculate the margin of error at a 95% confidence interval ($Z = 1.96, \alpha = 0.05$).

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 ρ_{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 ρ_{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 package[104]: following children's microbiome after 1, 3, and up to 18 years, as well as, cohabitation and intra-datasets. The p-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 pangenomes, core, shell, and accessory c_AMP clusters were determined using the subset of c_AMPs obtained from ProGenomes2[43] because of their high-confidence assigned taxonomies and genomically-defined species (specI[140]). 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,[141] and those present in >95% of the genomes were classified as core genes.[66]

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[111] to cluster genomes from the same species in ProGenomes2.[43] Genome groups with ANI $\geq$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% $\leq$ ANI <99.99%) as in Rodriguez et al.[142] 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 scipy package.[104]

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 $\geq$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,[43] we extracted the gene coding for P1 adhesin[70] by mapping the reference gene sequence NZ_LR214945.1:c568695-567307 against each genome with bwa v.0.7.17[113], and later extracted the sequences using with SAMtools[116] and BEDtools.[117] The extracted gene sequences were aligned using Clustal Omega,[118] and a phylogenetic tree was built using the aligned nucleotide sequences and FastTree 2[110] 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.[71]

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,[49] (ii) the bioactive peptides database starPepDB 45k,[51] (iii) the small proteins from the global data-driven census of *Salmonella*,[50] (iv) the global microbial gene catalog GMGCv1,[52] (v) and the AMP database DRAMP[46] 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.[52] The AMPs were annotated using MMseqs2[99] with the 'easy-search' method, retaining hits with an E-value up to 10^{-5}. As Macrel[42] 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 scipy package[104] to model the association between c_AMPs and the background distribution of ortholog groups from GMGCv1.[52] The number of genes that were redundant in GMGCv1[52] 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,[52] and results were considered significant if $p < 0.05$ after correction with the Holm-Sidak method implemented in multipletests from the statsmodels package.[114] When using a robust approach that filters the ortholog groups by the number of c_AMP hits and GMGCv1[52] 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[120] against the NCBI database,[124] 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[107] 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), F0106 (CP072366.1:781389–781502), F0697 (CP072364.1:1466323–1466436), and from *Prevotella melaninogenica* strains FDAARGOS_760 (CP054010.1:157726–157839), FDAARGOS_306 (CP022041.2:943522–943635), FDAARGOS_1566 (CP085943.1:1102942–1103055), and ATCC 25845 (CP002123.1:409656–409769) 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[61] with DIAMOND[119] in "blastp" mode, as previously reported.[61] 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

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 CARD[125]; details of annotation and annotated database described previously[61]). 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.,[61] we calculated genomic context conservation on 3,899,674 gene families calculated *de novo* with MMSeqs2[99] (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.[108] 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 Python[100] 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 built the layout. Plotly was used to generate interactive visualization plots, and Axios to render content seamlessly. LogoJS (https://logojs.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 solubility[144] 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 DRAMP[46] 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 Macrel[42]: AMPScanner v2,[53] the mature peptides model in ampir,[40] amPEPpy,[54] APIN[55] – 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 2[49]). To compose the list of 50 high-likelihood candidates: (i) we selected 34 of the most prevalent peptides; (ii) we randomly selected 14 c_AMPs (30% of our set) with homologs to the GMGCv1[52] and one that matched SmProt 2[49]; and (iii) we included one peptide that was found in the MAGs binned from stool samples used to investigate fecal transplantations.[149] We also included scrambled sequences made using five of the most active peptide sequences to verify the potency of randomly generated sequences.

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 microtiter 96-well plates in 2-fold dilutions ranging from 64 to 1 μmol L^{-1}, and then peptides were exposed to an inoculum of $2 \cdot 10^6$ 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.[75]

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 OD_{600} 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 λ_{ex} = 350 nm and λ_{em} = 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:

$$Relative\ fluorescence = \frac{100 \times \left(fluorescence_{sample} - fluorescence_{polymyxinB}\right)}{fluorescence_{polymyxinB}}$$

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 [$DiSC_3$-(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 (OD_{600} = 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 OD_{600} 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 nmol L^{-1} of $DiSC_3$-(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 (λ_{ex} = 622 nm, λ_{em} = 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[151] and as a web server (https://ampsphere.big-data-biology.org/).

CellPress
OPEN ACCESS

Supplemental figures

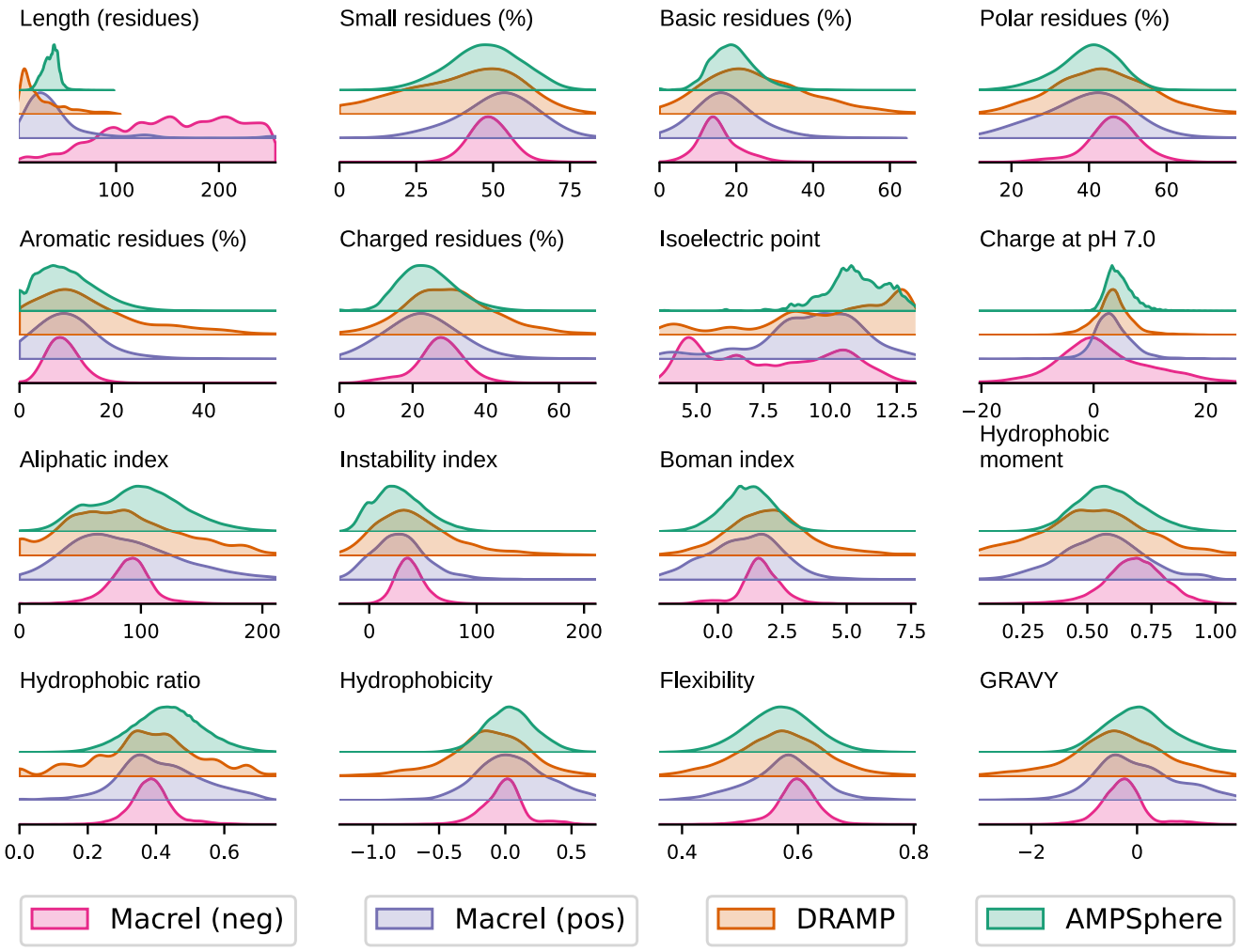


Figure S1. General physical-chemical features of c_AMPs in AMPSphere and validated databases of antimicrobial peptides, related to Figure 1

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[46] and the ones present in the positive training set used in Macrel[42] (p_{Mann} = 0.71); AMPSphere peptides did not differ from the positive training set used in Macrel[42] in the fraction of aromatic (p_{Mann} = 0.58), non-polar (p_{Mann} = 0.97), polar (p_{Mann} = 0.97), and acidic (p_{Mann} = 0.69) residues; the instability index (p_{Mann} = 0.58) and the hydrophobicity (p_{Mann} = 0.31) of AMPSphere peptides also were not different from the positive training set used in Macrel.[42]

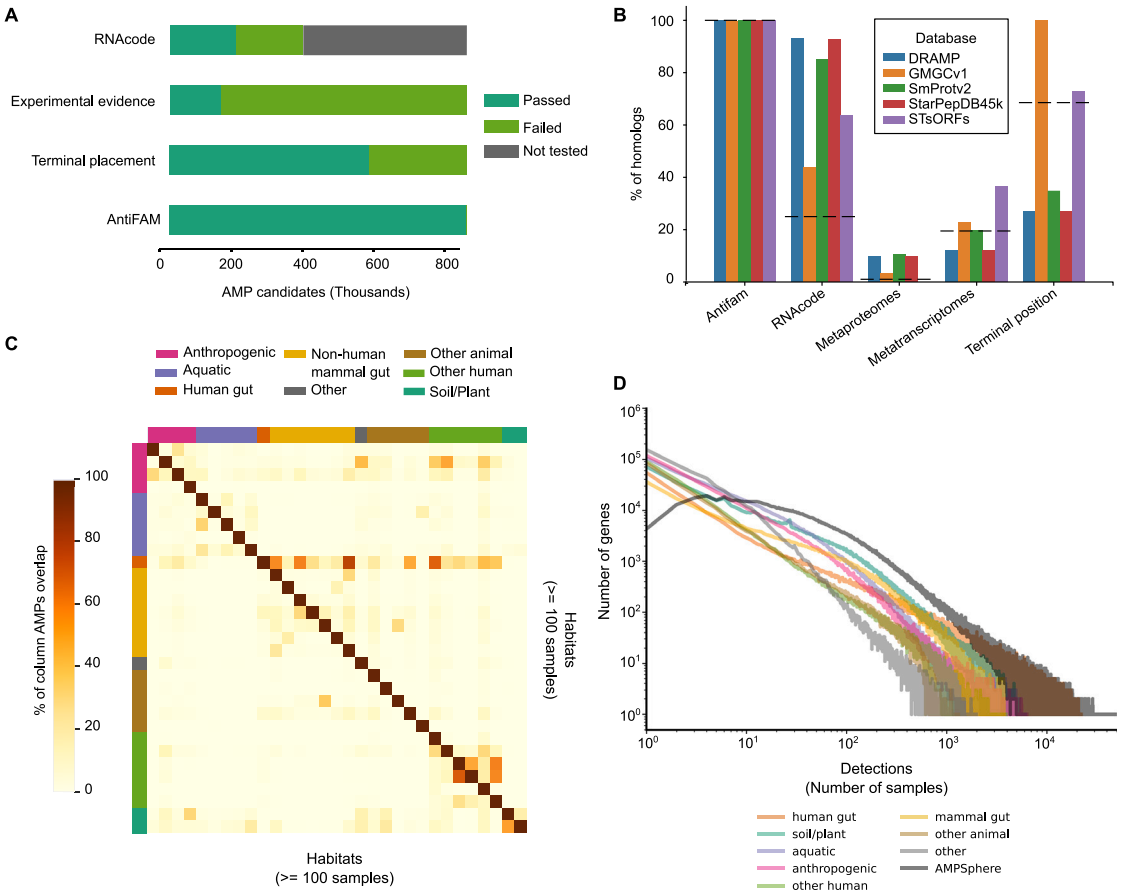


Figure S2. c_AMP quality and habitat distribution, related to Figures 1 and 2

(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.

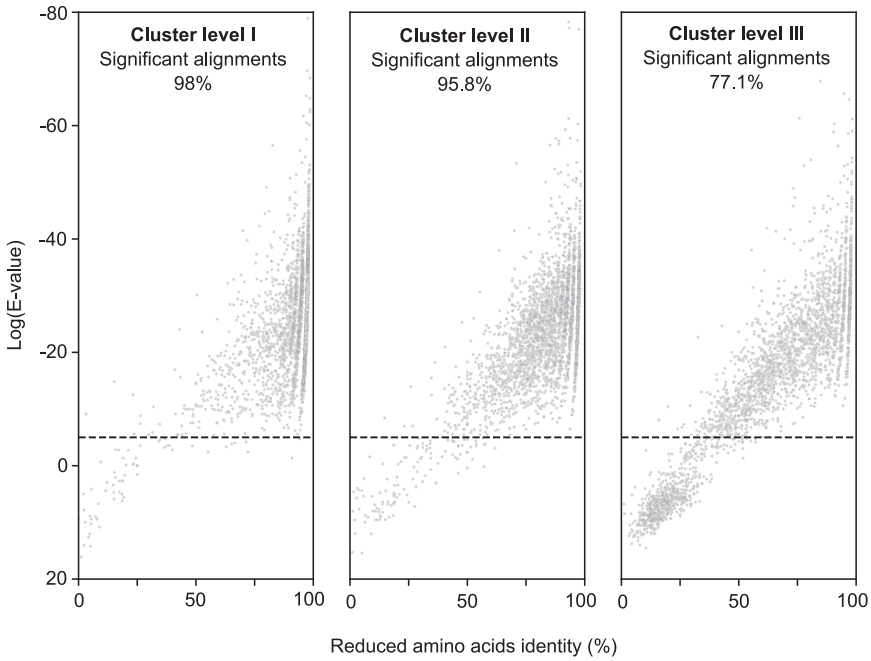


Figure S3. Clustering validation of families, related to STAR Methods section "Clustering of AMP families"

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.

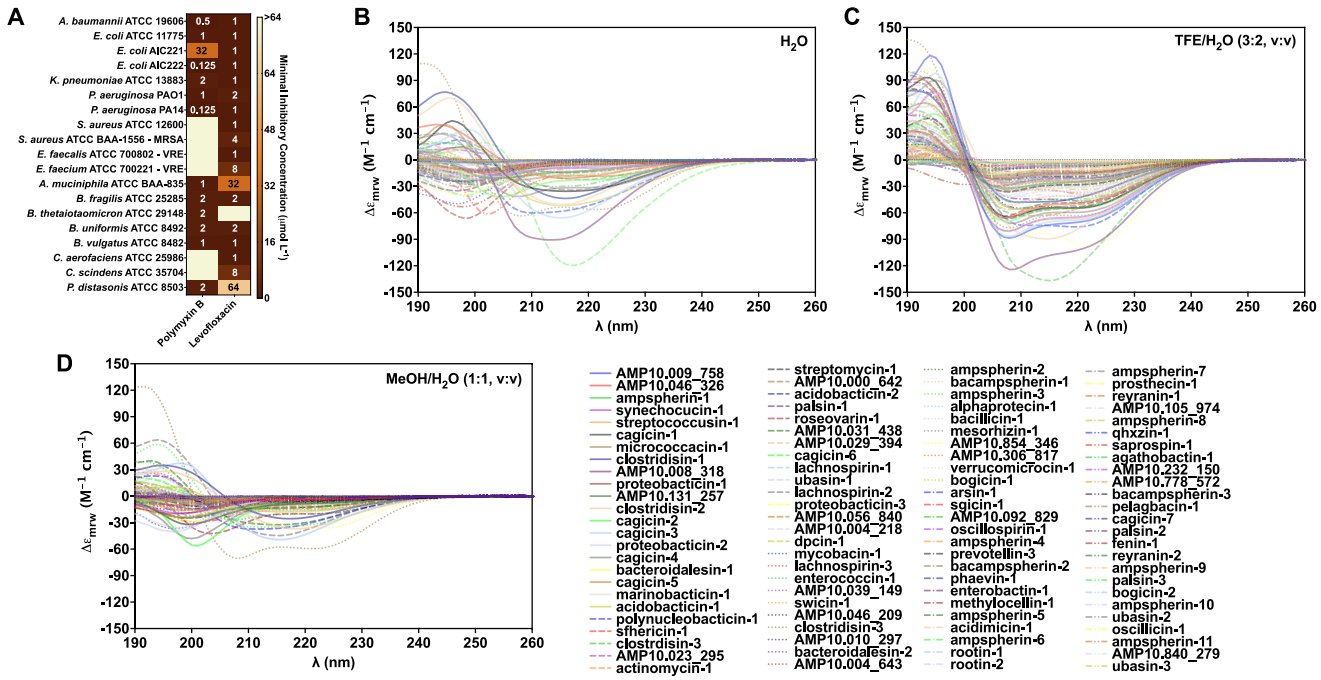
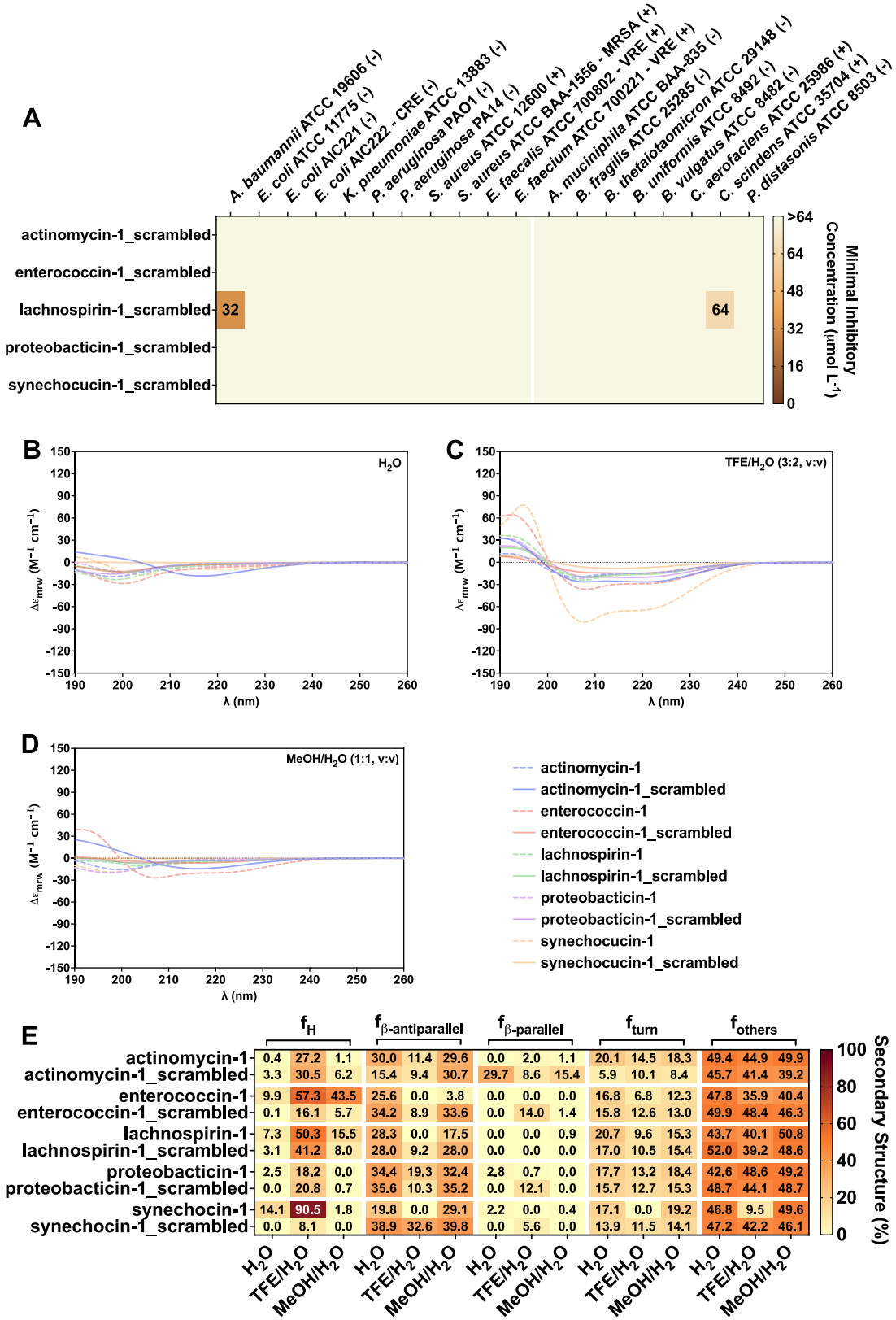


Figure S4. Antimicrobial activity of polymyxin B and levofloxacin and circular dichroism spectra of the c_AMPs, related to STAR Methods section "Circular dichroism assays"

(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.



(legend on next page)

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.[75]

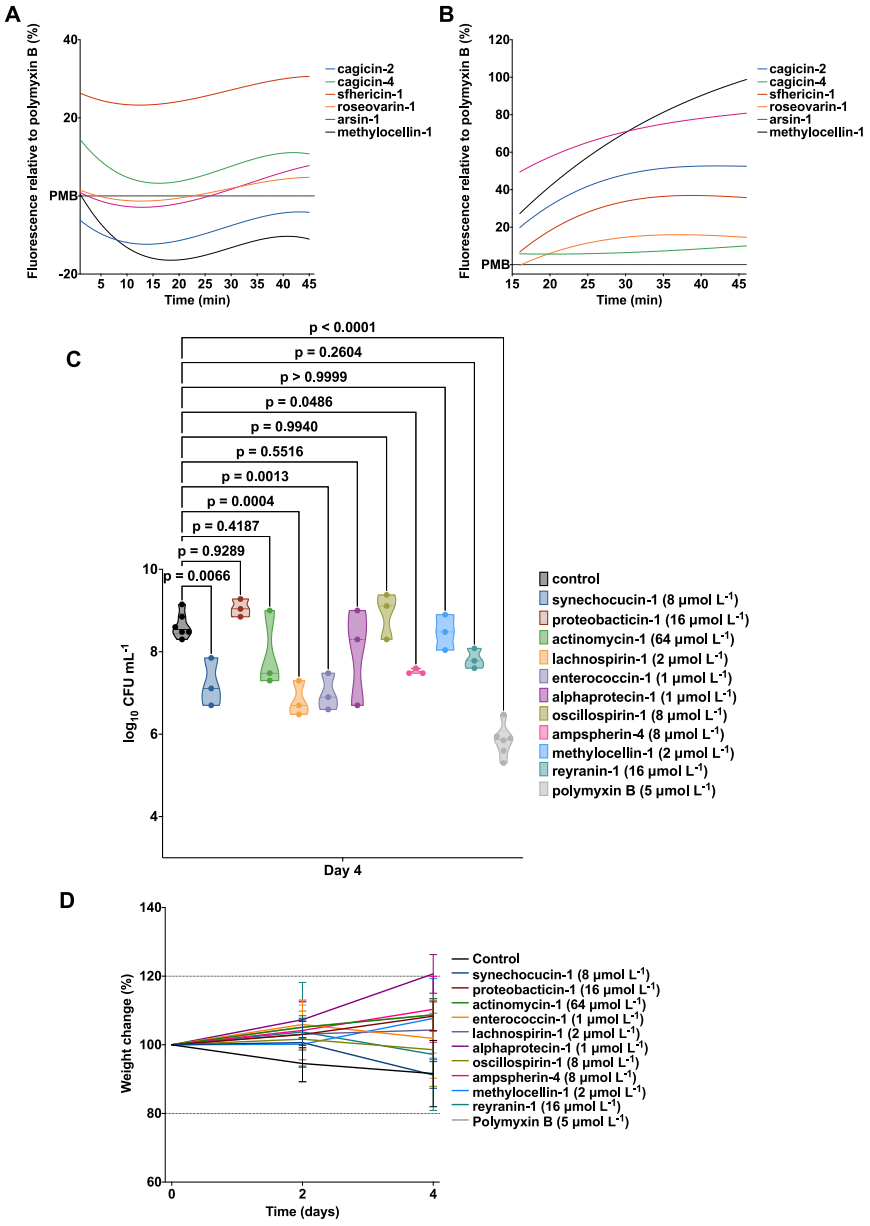


Figure S6. Mechanism of action of AMPSphere peptides and anti-infective activity of c_AMPs in a preclinical animal model, related to Figures 6 and 7

(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 ($DiSC_3$-[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.