nature biotechnology

Article

https://doi.org/10.1038/s41587-025-02810-w

Bioactive molecules unearthed by terabase-scale long-read sequencing of a soil metagenome

Received: 10 March 2025

Accepted: 14 August 2025

Published online: 12 September 2025

Check for updates

Ján Burian, Robert E. Boer, Yozen Hernandez, Adrian Morales-Amador, Linhai Jiang, Abir Bhattacharjee, Cecilia Panfil, Melinda A. Ternei & Sean F. Brady

Metagenomics provides access to the genetic diversity of uncultured bacteria through analysis of DNA extracted from whole microbial communities. Long-read sequencing is advancing metagenomic discovery by generating larger DNA assemblies than previously possible. However, harnessing the potential of long-read sequencing to access the vast diversity within soil microbiomes is hampered by the challenge of isolating high-quality DNA. Here we introduce a method that can liberate large, high-quality metagenomic DNA fragments from soil bacteria and pair them with optimized nanopore long-read sequencing to generate megabase-sized assemblies. Using this method, we uncover hundreds of complete circular metagenomic genomes from a single soil sample. Through a combination of bioinformatic prediction and chemical synthesis, we convert nonribosomal peptide biosynthetic gene clusters directly into bioactive molecules, identifying antibiotics with rare modes of action and activity against multidrug-resistant pathogens. Our approach advances metagenomic access to the vast genetic diversity of the uncultured bacterial majority and provides a means to convert it to bioactive molecules.

Sequencing has primarily been driven by massively parallel short-read methods, particularly Illumina sequencing. As adoption of this technology rose and costs were driven down, there was an explosion of bacterial isolate whole-genome sequences[1]. The majority of bacteria, however, remain recalcitrant to laboratory culture, preventing direct analysis of their genetic diversity. While short-read-based 16S amplicon profiling of metagenomes (that is, whole-DNA extracts of microbial communities) has been used to provide glimpses at microbial diversity hidden in the uncultured majority[2], a key limitation of direct short-read exploration is the inability to deconvolute complex metagenomes and reconstruct sequences of individual microorganisms within the community. Binning of short-read assemblies by computationally predicted similarity into metagenome-assembled genomes (MAGs) as a means to explore the genetic potential of environmental bacteria has now become routine; however, MAGs tend to be smaller than bacterial genomes, fragmented and often contain sequence contamination through erroneous binning[1,3–5]. Exploration of soil microbiomes presents a particularly difficult challenge because of not only the immense microbial diversity they contain but also the challenges associated with isolating high-quality metagenomic DNA[6,7]. Efficient exploration of environmental microbial genetic potential requires the development of methods that can generate large contiguous assemblies. This is especially important when exploring the biosynthetic potential of the global metagenome, as biosynthetic gene clusters (BGCs) can be tens if not hundreds of kilobases long and are challenging to identify within fragmented assemblies[8]. Long-read sequencing technologies have advanced metagenomic sequencing by producing larger assemblies that can be efficiently interrogated using bioinformatic algorithms

Laboratory of Genetically Encoded Small Molecules, Rockefeller University, New York, NY, USA. e-mail: sbrady@rockefeller.edu

https://doi.org/10.1038/s41587-025-02810-w

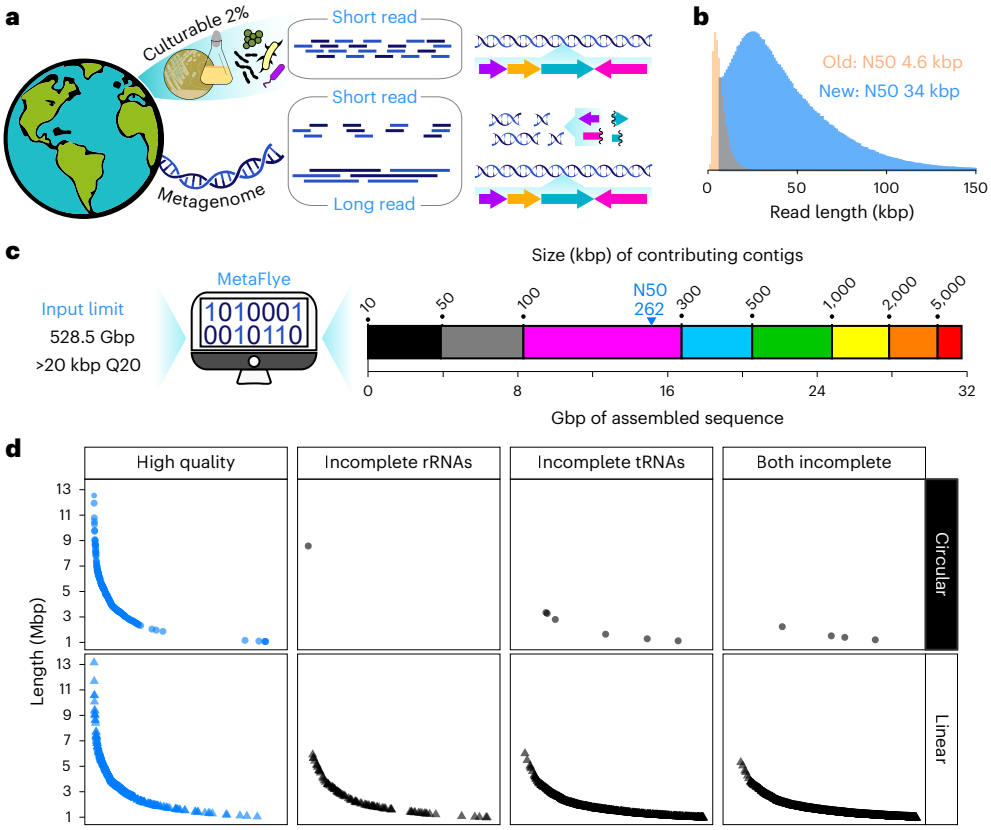

Fig. 1 | Capturing complete genome sequences from a complex soil metagenome. a, The limitations of short-read technology restrict complete genomic sequencing to cultured isolates; emerging long-read sequencing may overcome this challenge. **b**, Nanopore sequencing read-length distribution comparison of soil metagenomic DNA prepared by direct hot CTAB extraction (old) and the newly developed and optimized nycodenz method (new). **c**, Assembly summary of high-quality nanopore soil metagenomic sequencing at metaFlye's input limit. **d**, Assembly length plot of >1 Mbp metaFlye generated contigs with an estimation of genome completeness based on the presence of all rRNA genes (5S, 16S and 23S) and at least 18 unique tRNAs. Plots are separated into circular and linear assemblies.

(Fig. 1a)[9–13] and updated nanopore flow cells (R10.4) and sequencing chemistries (V14) can now generate high-quality assemblies without the need for short-read polishing[14,15].

Numerous methods have been developed to extract metagenomic DNA[16] but the focus has been on DNA quality without the need to optimize these protocols for isolating large DNA fragments. Here, we developed a nanopore long-read sequencing pipeline with size-optimized DNA isolation and preparation that, when applied to a complex soil metagenome, yielded read lengths with N50 > 30 kbp. This is 200 times longer than the commonly used 150-bp short-read technology. The increased read size led to assemblies of contiguous DNA sequences on the order of megabase pairs, including hundreds of complete circular metagenomic genomes. Using a synthetic bioinformatic natural product (synBNP) approach, which couples bioinformatics and total chemical synthesis to abiotically decode BGCs[17], we directly converted unearthed nonribosomal peptide BGCs within the large contiguous assemblies into bioactive molecules. This led to the discovery of a potent cardiolipin-binding broad-spectrum antibiotic and a ClpX-targeting antibiotic.

Scalable access to microbial dark matter

A critical barrier to sequencing soils is the copurification of contaminants when isolating metagenomic DNA[6,7]. To improve soil DNA extraction, we developed a method that first separated bacteria from the soil matrix by nycodenz gradient centrifugation and then used a skim-milk wash to remove impurities from the isolated cells. This process generated a bacterial suspension that appeared similar to a lab-grown culture (Supplementary Fig. 1a). Using standard hot CTAB metagenomic DNA extraction[18] on this cell suspension yielded a cleaner metagenomic DNA extraction with far larger fragments compared to direct soil extraction (Supplementary Fig. 1b). The 'clean' microbial suspension allowed us to test a variety of metagenomic DNA extraction and size-selection approaches to maximize nanopore read length (Supplementary Note 1, Supplementary Fig. 1 and Supplementary Tables 1 and 2). Our final method used commercially available DNA extraction (Monarch's high-molecular-weight (HMW) DNA extraction kit) and size selection (Oxford Nanopore's small fragment eliminator kit), allowing for simple and scalable DNA preparation. We expanded our approach using the original test forest soil sample to generate almost 1 Tbp of sequence data with an N50 of 32.82 kbp. For comparison, a representative nanopore sequencing result using our method and the direct soil hot CTAB extraction method is shown in Fig. 1b. In total, 2.5 Tbp of long-read sequencing data were generated from a single forest soil sample during testing and scaling. All data were pooled for bioinformatic analysis. A summary of all the sequencing runs is given in Supplementary Table 2.

Assembly of large contiguous metagenomic sequences from a soil metagenome

MetaFlye[19], the state-of-the-art algorithm for metagenomic assembly, has an input limit of 2^{39} bp (550 Gbp) and could not process all the data simultaneously because of the high sequencing depth. We, therefore, used two independent approaches to generate assemblies. This included global assembly at metaFlye's input limit using the highest-quality and longest reads and a targeted approach that used extraction and subassembly of reads containing sequences of interest.

Global assembly

To gain broad insight into the metagenomic content of the forest soil, we assembled a 528.5-Gbp subset of our data consisting of the highest-quality and longest reads (Q20+ reads longer than 20,000 bp). Assembly yielded almost 32 Gbp of sequence with an N50 of 262 kbp (Fig. 1c), including over 3,200 contigs > 1 Mbp. As a comparison, previous short-read sequencing explorations of soil metagenomes returned assemblies with an N50 of ~1.6 kbp even with over 3 Tbp of sequencing data[20,21]. In addition, a recent PacBio long-read exploration of a soil metagenome generated read lengths with a N50 of 4 kbp, leading to assemblies with an N50 of 36 kbp (ref. 22). Our optimized nanopore exploration substantially increased assembly length and provided a unique opportunity to explore genome-sized contiguous metagenomic assemblies without the reliance on computationally inferred MAG binning.

Exploration of long contiguous metagenomic assemblies

We focused our exploration on a subset of single contiguous assemblies that likely captured complete or near-complete bacterial genomes. To do this, we determined the subset of >1 Mbp of circular and linear assemblies that contained 5S, 16S and 23S ribosomal RNA (rRNA) genes with at least 18 transfer RNAs (tRNAs) (Fig. 1d). These combined criteria are part of the field benchmark for near-complete MAGs[1] and yielded 563 single contiguous assemblies (Fig. 1d and Supplementary Table 3). These contained 95% (206) of the circular assemblies in the >1-Mbp dataset, which was in line with the expectation that large circular assemblies represent complete contiguous bacterial genomes (Fig. 1d). Among the group of assemblies containing the requisite rRNA and tRNA, the circular subset was on average 4.95 ± 2.05 Mbp in size with the linear assemblies 3.99 ± 1.88 Mbp in size. To put this advance into context, we collated MAGs from soil-linked samples within the Genome Taxonomy Database (GTDB) (Supplementary Table 4)[23], the most up-to-date repository of bacterial sequences. From 5,640 entries across 91 bioprojects, we only identified two single-contig MAGs exceeding 1 Mbp that contained the required repertoire of rRNA and tRNA genes. In fact, only 26 total MAG bins contained an assembly over 1 Mbp, two orders of magnitude lower than the >3,200 in our single assembly.

In addition to the rRNA and tRNA criteria, for a MAG to be considered near complete, it must be rated >90% complete with <5% contamination by CheckM, a software tool designed to assess the quality of genome bins. Irrespective of the rRNA and tRNA content, a MAG is considered of high quality at >70% completeness and medium quality at >50% completeness, both with <10% contamination[1,24]. Of the 206 circular assemblies, only three were below 80% completeness, with 79% (163) matching the near-complete criteria (Supplementary Table 3). The 357 linear assemblies were split into 32%, 29% and 21% near-complete, high-quality and medium-quality MAGs, respectively. Only 15 (2.7%) assemblies within the 563 complete or near-complete dataset exceeded the contamination criteria, with many only by a small margin (Supplementary Table 3). As a comparison, a recent Illumina-based large-scale exploration of soil yielded a total of 679 MAG bins, of which 5% (33) and 14% (92) were >90% and >70% complete, respectively[24]. These high-quality MAGs were on average 3.05 ± 1.25 Mbp in size and composed of 567 ± 417 assemblies with an N50 of 11.8 ± 9.8 kbp. Overall, this analysis continued to demonstrate the advance in metagenomic exploration provided by long-read sequencing, with our optimized protocol greatly increasing the number of complete or near-complete contiguous genomes that can now be assembled from a complex soil metagenome.

16S profiling

To gain an estimate of the species diversity found in each subset of our data, we used 16S profiling. In total, we detected >4,500 unique species within the 2.5-Tbp dataset. Over 90% of these were detected in the global assembly data, 25% were detected in the >1-Mbp dataset and over 12% were detected in the complete or near-complete genome dataset (Supplementary Table 5). Next, we sought to explore the complete or near-complete genome dataset within the comprehensive phylogenetic profiling of the GTDB framework.

Expansion of enigmatic bacterial taxa

Our complete and near-complete metagenome dataset resolved into 16 phyla (Fig. 2a), with only four sequences (0.7%) matched to known species (Supplementary Table 3). To highlight underexplored bacterial sequences within the dataset, we determined which underrepresented and largely uncultured taxa within the GTDB were most expanded by our assembled data (Supplementary Table 6). These taxa included the UBP17 phylum, the Capsulimonadaceae family and the *Lustribacter* and *DAIDGS01* genera from the Eremiobacterota phylum (Fig. 2b). In addition, we also saw a notable expansion of diversity of the *Palsa-744* genus within the well-represented Actinomycetota phylum (Supplementary Fig. 2), the fifth most sequence-rich phylum within the GTDB. Apart from a single completely sequenced cultured representative in the Capsulimonadaceae family, our data added the first complete contiguous genomes for each of the abovementioned taxa. Of these, the most striking was a 12.56-Mbp circular assembly belonging to the UBP17 phylum, which stood out because this mysterious phylum consisted of only ten total MAGs that were on average much smaller and highly fragmented. The theme of highly fragmented MAGs being the only genetic resource for a taxon was common throughout our comparisons (Fig. 2b and Supplementary Fig. 2). Assembly of large contiguous sequences provided not only an unprecedented resolution of a complex microbial community but also an opportunity for the bioinformatic exploration of biosynthetic potential within microbial dark matter.

Biosynthetic potential of complete or near-complete metagenome genomes

Generating complete BGC sequences within fragmented assemblies generated by short-read technology is challenging. Furthermore, the mobile nature and rapid evolution of BGCs[25] may hamper MAG binning where computational similarity would need to link core genomic regions and varied BGC content. We rationalized that our long-read sequencing data would allow us to explore biosynthetic systems found in complex microbiomes without these limitations. To gain a broad overview of the biosynthetic content captured within our complete and near-complete metagenomic genome dataset, we bioinformatically identified and summarized BGC content at the taxonomic family level (Fig. 2a). On average, there were 1.3 ± 0.7 BGCs per Mbp across the families, with four families (UBA5704, Pseudonocardiaceae, Streptomycetaceae and UBA11063) exceeding 3 BGCs per Mbp. UBA stands for 'uncultured bacteria or archaea' and was coined to phylogenetically characterize MAGs that provided the first example of a bacterial or archaeal lineage[26]. UBA5704 is a family within the Acidobacteriota phylum under the order Multivoradales, while UBA11063 is a family within the Pseudomonadota phylum under the order Burkholderiales. UBA5704 contains five genera and UBA11063 only contains a single genus (*Aquella*), for which a cultured representative was isolated and sequenced in 2019 (ref. 27). In Fig. 2c, we highlight the two most BGC-rich assemblies within the taxa with >3 BGCs per Mbp, as well as one of the five UBA11063 (*Aquella*) assemblies. Even with their relatively small genome sizes, our data indicated that *Aquella* species within our dataset contained a disproportionally large biosynthetic repertoire. Notably, the clusters were overwhelmingly nonribosomal peptide synthase (NRPS)-type BGCs (Fig. 2d). Even though the GTDB contained 37 *Aquella* entries (one isolate and 36 MAGs), analysis of the total 76.4-Mbp sequence yielded only 85 total BGCs (1.1 BGC per Mbp). This may indicate that, because of difficulties in assembling NRPS BGCs using short reads, this genus may have been previously overlooked as a rich source of natural products. Overall, our large contiguous assemblies offered an avenue for natural product discovery by pointing toward BGC-rich

https://doi.org/10.1038/s41587-025-02810-w

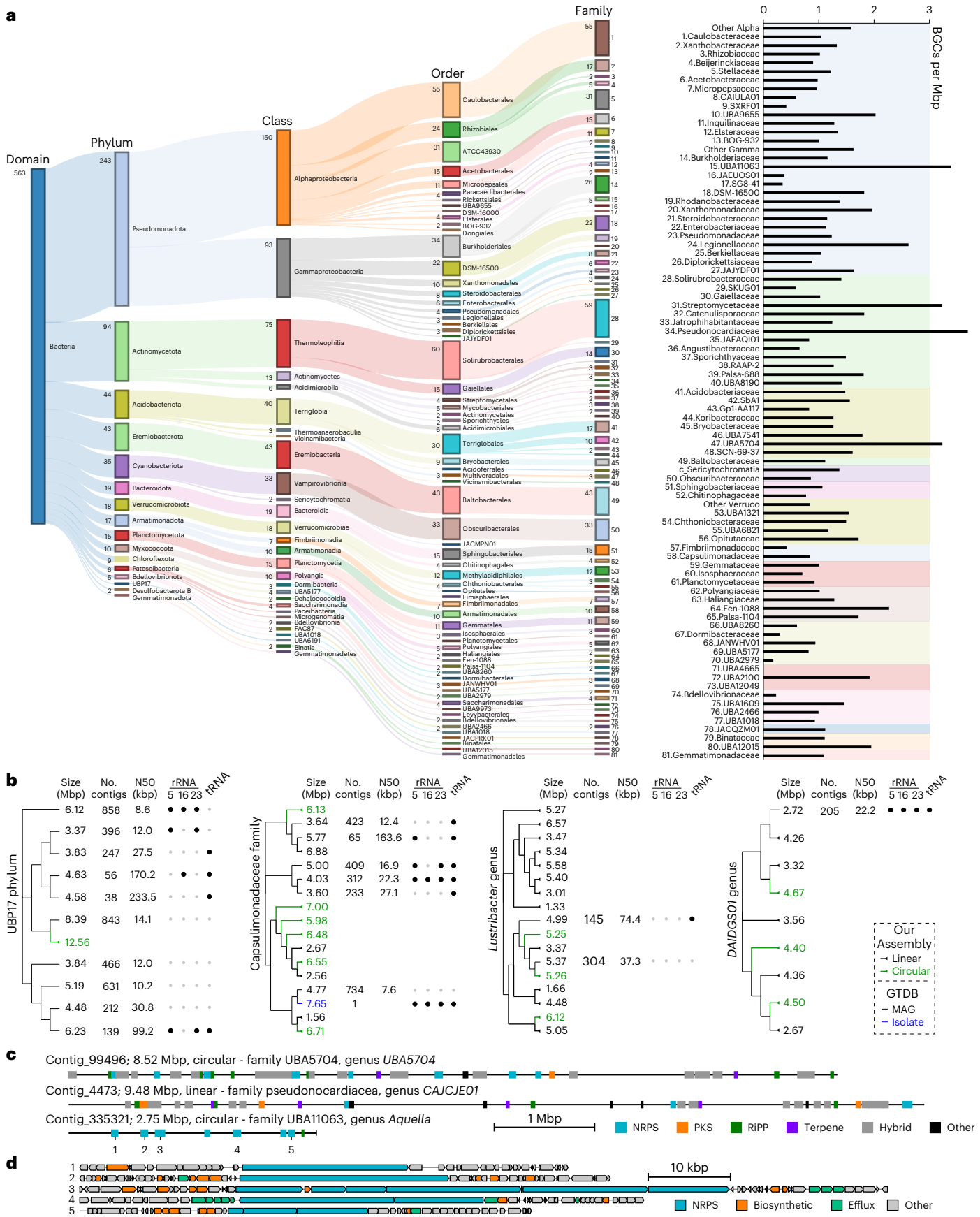

Fig. 2 | Bioinformatic exploration of the complete and near-complete metagenomic genomes. a, Phylogeny of high-quality contiguous assemblies from Fig. 1d (left) with BGCs per Mbp (right). Counts are indicated if >1. **b**, Representative assemblies expanding underrepresented taxa. **c**, Example BGC-rich assemblies. **d**, NRPS BGCs numbered in **c**.

https://doi.org/10.1038/s41587-025-02810-w

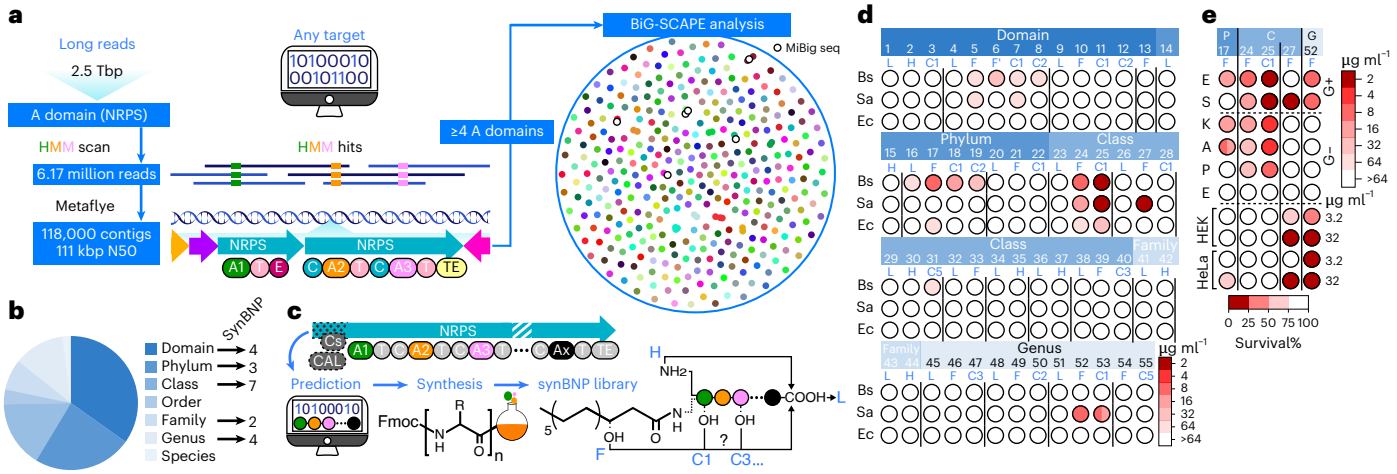

Fig. 3 | Metagenome-guided synthesis of bioactive small molecules. **a**, Left, general outline of targeted metagenomic NRPS assembly through HMM identification of A-domain-containing reads. Right, BiG-SCAPE analysis of 328 assemblies containing ≥4 A domains. Colored circles are metagenomic BGCs, with white circles denoting clusters matched from the MiBig database. **b**, Phylogeny breakdown of assemblies from **a**. **c**, Outline of the synBNP approach. **d**, Antibiotic activity screening of 55 synBNPs synthesized based on 20 metagenomic NRPS BGCs against *B. subtilis* (Bs), *S. aureus* (Sa) and *E. coli* (Ec). Lines separate predictions generated by single BGCs. **e**, Expanded activity screening of synBNPs with MICs ≤8 μg ml^{-1} from **b** against the ESKAPE pathogens and activity against HEK and HeLa human cell lines. Results in **d**,**e** are representative of two and three biologically independent replicates, respectively.

taxa that have remained hidden or overlooked because of a lack of quality assemblies. To access this previously hidden biosynthetic diversity, we used a total synthesis approach of bioinformatically predicted nonribosomal peptides.

Sequence-only discovery of bioactive molecules from microbial dark matter

NRPS BGCs encode large modular proteins where each module functions to incorporate a monomer building block into a growing peptide-like structure[28]. The synBNP approach exploits this modular organization to predict the NRPS product by decoding which monomer is selected by each NRPS module to predict the assembled peptide; then, the predicted structure is accessed by total chemical synthesis.

Targeted assembly of NRPS BGCs

A small number of bacterial species dominate soil biomass[29] and obscure the rare microbiome that contains the majority of genetic diversity within a metagenome. To explore the complete set of genetic diversity captured in our whole 2.5-Tbp dataset, we devised a partitioned assembly approach where a hidden Markov model (HMM) search was used to identify reads containing specific sequences of interest for subassembly. While we targeted NRPS adenylation (A) domains, this approach could be used to target any sequence of interest. The HMM search identified 6.17 million reads containing an A-domain, which assembled into 118,000 contigs with an N50 of 111 kbp (Fig. 3a). We then narrowed our search to contigs that contained four or more A domains to focus on large NRPS BGCs, which are generally problematic for short-read assemblies to fully capture. This yielded 338 contigs. BiG-SCAPE analysis of the captured NRPS BGCs on these contigs revealed 366 clusters, of which only six could be found in the MiBig database of characterized molecules (Fig. 3a). Notably, only 20.5% of the clusters from our targeted assembly could be mapped to the global assembly dataset, indicating that our partitioned assembly approach did indeed access a larger proportion of the rare microbiome.

A possible limitation of the targeted assembly approach was that removing genomic context by restricting assembly to reads only containing NRPS genes may have led to erroneous or chimeric assemblies. While it is not possible to know the extent to which this problem may exist, we believe that it is unlikely to be pervasive because of two key points. First, previously characterized NRPS BGCs identified in the BiG-SCAPE analysis were correctly assembled. Second, as a control, we cultured and sequenced a collection of microorganisms from the same forest soil sample used in this study, which yielded multiple NRPS BGCs that matched those found in our targeted assembly dataset (Supplementary Table 7, Supplementary Fig. 3 and Supplementary Note 2).

NRPS BGC selection, prediction and synthesis

We selected 20 NRPS BGCs with diverse predicted phylogenic origins for synBNP structure prediction and synthesis (Fig. 3b and Supplementary Tables 8–10). While the peptide sequence generated by an NRPS BGC can be predicted through the A-domain signatures present within each module, the final product can be released as either a linear or cyclic peptide, which remains difficult to predict bioinformatically. Therefore, for each of the 20 predictions, we synthesized a range of possible products (Fig. 3c; linear (L), cyclized through the terminal amine (H), cyclized through the 3-OH fatty acid (F) and cyclized through a nucleophilic amino acid side chain (C#)) to generate a library of 55 synBNPs (Supplementary Table 11). Myristic or hydroxymyristic acid was used when lipidation was predicted to be present as it is the most commonly seen lipid in characterized lipopeptides.

Bioactivity screening

All synBNPs were assayed for antibiotic activity against Gram-positive model organisms *Bacillus subtilis* and *Staphylococcus aureus*, as well as the Gram-negative model organism *Escherichia coli* (Fig. 3d and Supplementary Table 11). Five synBNPs, based on predictions from four NRPS clusters, had notable activity (minimum inhibitory concentration (MIC) ≤ 8 μg ml^{-1}) against at least one of these bacteria and were, therefore, screened for wider activity against the ESKAPE pathogens (*Enterococcus faecium*, *S. aureus*, *Klebsiella pneumoniae*, *Acetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter cloacae*) and for human cell toxicity (Fig. 3e). We then selected the two most potent synBNPs for further characterization. SynBNP 25, inspired by the BGC found on contig 112034, was a potent broad-spectrum antibiotic and lacked cytotoxicity. This compound was named erutacidin after the Latin erutus for 'dug out'. SynBNP 27, inspired by the BGC found on contig 97355, was selected for its potent activity against *S. aureus* (MIC = 2 μg ml^{-1}). This compound was named trigintamicin after the

https://doi.org/10.1038/s41587-025-02810-w

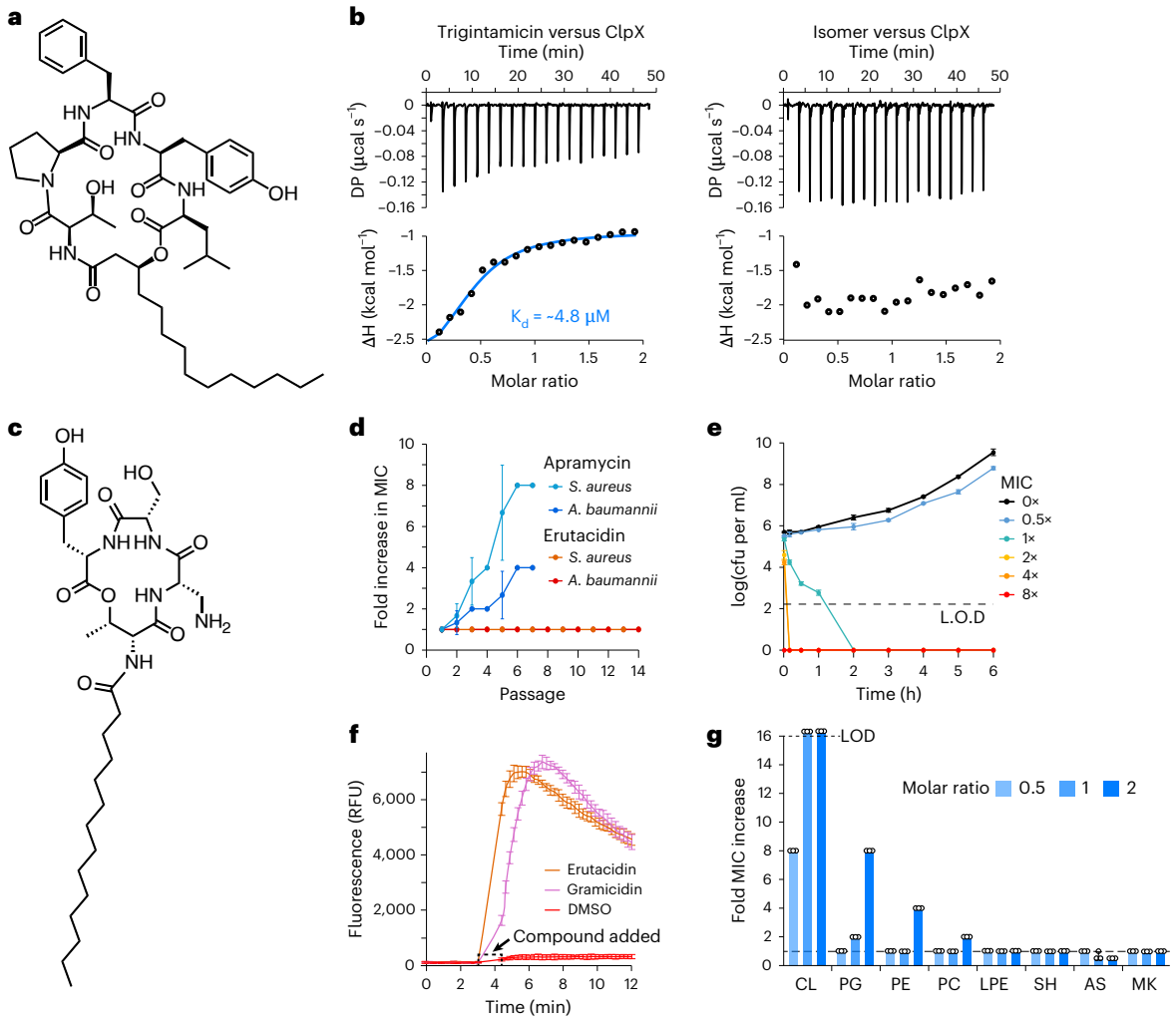

Fig. 4 | MOA studies for trigintamicin and erutacidin. a, Trigintamicin structure. **b**, ITC (top) and isotherm curves (bottom) of *S. aureus* ClpX against trigintamicin and its biologically inactive *R*-isomer. Data are representative of at least two independent determinations. **c**, Erutacidin structure. **d**, Resistance changes of *S. aureus* and *A. baumannii* during serial passaging in the presence of erutacidin or apramycin. **e**, *S. aureus* kill curve elicited by erutacidin at varied concentrations. **f**, $DiSC_3(5)$ depolarization assay of *S. aureus* in the presence of antibiotics at 2× MIC or DMSO. **g**, Fold change in erutacidin MIC against *S. aureus* caused by exogenous supplementation of various lipids. LOD, limit of detection; CL, cardiolipin; LPE, lysophosphatidylethanolamine; SH, sphingomyelin; AS, *N*-acetylsphingosine; MK, menaquinone-4. Data in **d**–**g** are presented as the mean values ± s.d. ($n = 3$ biologically independent replicates).

Latin word for '30', the temperature at which activity was detected. In our initial screen, racemic 3-OH myristic acid was used for the synthesis of fatty acid cyclized synBNPs that included trigintamicin. Upon resynthesis and more thorough purification, the two trigintamicin isomers were resolved, with only the *S*-isomer showing antibiotic activity against *S. aureus* (Supplementary Table 12).

Mode of action (MOA) determination

To determine the MOAs of erutacidin and trigintamicin, we first attempted to raise resistant mutants. Direct plating of *S. aureus* on trigintamicin (Fig. 4a) at 8× its MIC generated resistant mutants. Genomic sequencing of resistant clones revealed mutations at two genetic loci, *ytrA* and *clpX* (Supplementary Table 13). Mutations in *ytrA* have been previously linked to nonspecific antibiotic resistance[30,31]; hence, we focused on *clpX*. ClpX is an ATPase that unfolds and translocates proteins into the ClpP serine protease for degradation[32]. In *S. aureus*, ClpX is essential at 30 °C but dispensable at 37 °C (ref. 33). Our antibiotic activity assays were conducted at 30 °C and we found that shifting the assay conditions to 37 °C abolished trigintamicin activity, further suggesting ClpX as the target (Supplementary Table 12). We then performed isothermal titration calorimetry (ITC) to confirm the interaction between trigintamicin and ClpX. These studies found that trigintamicin bound ClpX, whereas its biologically inactive *R*-isomer did not (Fig. 4b). No thermal changes were observed in the absence of ClpX (Supplementary Fig. 4). Collectively, our MOA studies indicated that trigintamicin was a ClpX-targeting antibiotic. The targeting of ClpX by trigintamicin explains the narrow antibacterial spectrum, as *clpX* is not essential in most bacteria. It also explains human cell cytotoxicity, as mitochondrial ClpXP is essential and an emerging target for cancer therapy[34].

Neither direct plating nor prolonged passage of *S. aureus* or *A. baumannii* in the presence of erutacidin (Fig. 4c) generated resistant mutants (Fig. 4d). Time-dependent killing curve analysis of *S. aureus* treated with erutacidin showed a rapid bactericidal effect (<10 min at ≥2× MIC) (Fig. 4e). This suggested disruption of membrane integrity as a potential MOA. To test erutacidin for membrane depolarization activity, we used a $DiSC_3(5)$-based fluorescence assay. Treatment of *S. aureus* with erutacidin increased fluorescence similarly to the known depolarizer gramicidin, confirming membrane depolarization (Fig. 4f). Antibiotics that depolarize the bacterial membrane frequently interact with specific membrane components. For example, the antibiotics telomycin and daptomcyin interact with cardiolipin

(also known as diphosphatidylglycerol) and phosphatidylglycerol (PG), respectively[35,36]. In each case, the activity of the antibiotic can be suppressed by the addition of its binding partner to the growth medium (Supplementary Fig. 5)[36,37]. To further investigate erutacidin's MOA, we tested whether its activity could be suppressed by supplementation with various membrane lipids. Erutacidin's antibacterial activity was strongly inhibited by the addition of cardiolipin, reducing its activity eightfold at just a 0.5:1 molar ratio of cardiolipin to erutacidin (Fig. 4g). Supplementation of PG also caused a notable inhibition of activity, with an eightfold decrease in MIC at a 2:1 molar ratio of PG to erutacidin. Minor protection was provided by the supplementation of membrane lipids related to cardiolipin and PG, including phosphatidylethanolamine (PE) and phosphatidylcholine (PC) but no effect was observed for any other lipids tested or menaquinone (Fig. 4g). These data suggest that erutacidin interacts strongly and preferentially with cardiolipin but may also interact with related lipid structures. This may explain the lack of observed in vitro resistance for erutacidin, as acquiring mutations to overcome multiple inhibited molecular targets is very difficult. Cardiolipin binding is a rare MOA that is not represented among clinically used antibiotics. As such, we expected that multidrug-resistant (MDR) isolates would not show cross-resistance to erutacidin. Indeed, erutacidin maintained potent activity against all MDR *S. aureus* and MDR *A. baumannii* clinical isolates we tested (Supplementary Table 14).

Discussion

By achieving a substantial increase in read length, we show that it is possible to resolve complete contiguous metagenomic genomes from a complex soil sample. This not only advances our understanding of the composition of environmental bacterial communities but also provides a deeper understanding of encoded biosynthetic potential. Notably, the number of NRPS BGCs uncovered in the global assembly (0.53 Tbp) compared to the targeted A-domain assembly accessing the full dataset increased linearly, suggesting that even 2.5 Tbp was not sufficient sequencing depth to saturate the diversity found in the soil sample. The distribution of bacterial taxa in soil is uneven, with a small number dominating the biomass[29], and it is unclear what sequencing depth would be required to fully uncover very rare microbial dark matter. However, similar to trends observed with short-read sequencing[38], continued improvements in efficiency and accuracy should allow deep long-read sequencing to increasingly become a routine tool in metagenomics. As with the rapid growth of short-read data[38], this will necessitate the development of improved bioinformatic tools and computation infrastructure to assemble and explore terabase-scale long-read datasets. Nanopore specific assemblers are already becoming increasing powerful, with a very recent preprint introducing nanoMDBG as a fast and less computationally intensive assembly algorithm[39]. Bioinformatic algorithms for predicting the chemical output of BGCs are also improving, particularly with the incorporation of artificial intelligence and machine learning methods[40,41], and we expect that the number and diversity of BGCs that can be decoded using a synBNP approach will continue to increase. Even in the small number of BGCs examined in this study through the synBNP approach, we could identify a broad-spectrum antibiotic with no detected resistance and with activity against very-difficult-to-treat MDR pathogens including the Gram-negative *A. baumannii*. By combining improved long-read access to metagenome sequence space with synBNP discovery, we established a pipeline to leverage the vast genetic diversity of microbial dark matter for the systematic detection, synthesis and screening of previously inaccessible bioactive molecules. This represents a scalable strategy to identify molecules inspired by BGCs from within the near-limitless uncultured bacterial majority.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-025-02810-w.

References

1. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
2. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
3. Meziti, A. et al. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* **87**, e02593-20 (2021).
4. Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
5. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
6. Daniel, R. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).
7. Gaulke, C. A. et al. Evaluation of the effects of library preparation procedure and sample characteristics on the accuracy of metagenomic profiles. *mSystems* **6**, e0044021 (2021).
8. Meleshko, D. et al. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352–1362 (2019).
9. Meslier, V. et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. *Sci. Data* **9**, 694 (2022).
10. Orellana, L. H., Kruger, K., Sidhu, C. & Amann, R. Comparing genomes recovered from time-series metagenomes using long- and short-read sequencing technologies. *Microbiome* **11**, 105 (2023).
11. Singleton, C. M. et al. Connecting structure to function with the recovery of over 1,000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021).
12. Eisenhofer, R. et al. A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. *Microbiol. Spectr.* **12**, e0359023 (2024).
13. Bickhart, D. M. et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* **40**, 711–719 (2022).
14. Zhao, W. et al. Oxford Nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing. *Front. Microbiol.* **14**, 1179966 (2023).
15. Sereika, M. et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
16. Verma, S. K., Singh, H. & Sharma, P. C. An improved method suitable for isolation of high-quality metagenomic DNA from diverse soils. *3 Biotech* **7**, 171 (2017).
17. Chu, J., Vila-Farres, X. & Brady, S. F. Bioactive synthetic–bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome. *J. Am. Chem. Soc.* **141**, 15737–15741 (2019).
18. Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* **2**, 1297–1305 (2007).
19. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

20. Riley, R. et al. Terabase-scale coassembly of a tropical soil microbiome. *Microbiol. Spectr.* **11**, e0020023 (2023).
21. Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440–444 (2018).
22. Van Goethem, M. W. et al. Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. *Commun. Biol.* **4**, 1302 (2021).
23. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
24. Kazarina, A., Wiechman, H., Sarkar, S., Richie, T. & Lee, S. T. M. Recovery of 679 metagenome-assembled genomes from different soil depths along a precipitation gradient. *Sci. Data* **12**, 521 (2025).
25. Jensen, P. R. Natural products and the gene cluster revolution. *Trends Microbiol.* **24**, 968–977 (2016).
26. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
27. Chan, K. G. et al. *Aquella oligotrophica* gen. nov. sp. nov.: a new member of the family Neisseriaceae isolated from laboratory tap water. *Microbiologyopen* **8**, e00793 (2019).
28. Sussmuth, R. D. & Mainz, A. Nonribosomal peptide synthesis—principles and prospects. *Angew. Chem. Int. Ed. Engl.* **56**, 3770–3821 (2017).
29. Delgado-Baquerizo, M. et al. A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
30. Johnston, P. R., Dobson, A. J. & Rolff, J. Genomic signatures of experimental adaptation to antimicrobial peptides in *Staphylococcus aureus*. *G3 (Bethesda)* **6**, 1535–1539 (2016).
31. Kawada-Matsuo, M., Le, M. N. & Komatsuzawa, H.Antibacterial peptides resistance in *Staphylococcus aureus*: various mechanisms and the association with pathogenicity. *Genes (Basel)* **12**, 1527 (2021).
32. Baker, T. A. & Sauer, R. T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta* **1823**, 15–28 (2012).
33. Stahlhut, S. G. et al. The ClpXP protease is dispensable for degradation of unfolded proteins in *Staphylococcus aureus*. *Sci. Rep.* **7**, 11739 (2017).
34. Nouri, K., Feng, Y. & Schimmer, A. D. Mitochondrial ClpP serine protease—biological function and emerging target for cancer therapy. *Cell Death Dis.* **11**, 841 (2020).
35. Taylor, S. D. & Palmer, M. The action mechanism of daptomycin. *Bioorg. Med. Chem.* **24**, 6253–6268 (2016).
36. Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.* **12**, 233–239 (2016).
37. Kleijn, L. H. J. et al. A high-resolution crystal structure that reveals molecular details of target recognition by the calcium-dependent lipopeptide antibiotic laspartomycin C. *Angew. Chem. Int. Ed. Engl.* **56**, 16546–16549 (2017).
38. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
39. Benoit, G. et al. High-quality metagenome assembly from nanopore reads with nanoMDBG. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.22.649928 (2025).
40. Huang, J. et al. DeepAden: an explainable machine learning for substrate specificity prediction in nonribosomal peptide synthetases. Preprint at *bioRxiv* https://doi.org/10.1101/2025.05.21.655435 (2025).
41. Mongia, M. et al. AdenPredictor: accurate prediction of the adenylation domain specificity of nonribosomal peptide biosynthetic gene clusters in microbial genomes. *Bioinformatics* **39**, i40–i46 (2023).

Methods

Bacterial strains and growth conditions

All strains, unless otherwise specified, were grown in Luria–Bertani (LB) at 30 °C with 200-rpm shaking. Antibiotic activity test strains and ESKAPE pathogens were: *B. subtilis* 168, *S. aureus* SH1000, *E. coli* ATCC 25922, *E. faecium* COM15, *K. pneumoniae* ATCC 10031, *A. baumannii* ATCC 17978, *P. aeruginosa* PAO1 and *E. cloacae* ATCC 13047.

Sample collection

In late May 2023, we collected samples from Rockefeller University's 1,200-acre Center for Field Research in Ethology located in Dutchess County, New York. In order of collection, these comprised soil from a healthy forest (forest 1), sediment from a muddy bog, soil from the edge of a lake (forest 2), soil from a grove of dead trees (forest 3), decaying material from a dead tree and soil from a dry forest area (forest 4). Soil from a maximum depth of 10 cm was collected from an area of 50 cm^2 into a gallon-sized Ziploc bag. A single collected bag of forest soil 1 was used for testing and scaling. After collection, bags were stored at 4 °C.

Nycodenz isolation of soil microorganisms

Gradient centrifugation for separation of cells from the soil matrix and testing washes for removal of copurified contaminants were motivated by previous approaches in the field[42–44]. Soil (62.5 g) was added to 200 ml of prechilled 10 mM sodium pyrophosphate (Sigma) and blended (Hamilton Beach 10 speed) for 1 min using the LO/batter setting. The container was then placed on ice for 5 min. The 1-min blending and 5-min ice break were repeated two more times. Blended soil was then filtered through a cheesecloth (grade 90, bleached) to remove large insoluble particles and the filtrate was gently layered onto 8 ml of 36% w/v nycodenz (Accurate Chemical and Scientific Corporation) to the top of 50-ml conicals (Sarstedt). The tubes were centrifuged at 4 °C, 5,000*g* for 1.5 h in a swing bucket rotor yielding a fuzzy beige cell layer at the interface of the top (aqueous) and bottom (nycodenz) layer. The cell layer was gently aspirated and combined in a fresh 50-ml conical. The cell suspension was mixed 1:1 with sterile water and the mixture centrifuged at 4 °C, 5,000*g* for 1 h. The supernatant was discarded and the cell pellet was washed twice with 40 ml of JSM buffer (0.9% w/v NaCl and 1% w/v instant nonfat dry milk (Nestle)), followed by two washes with 40 ml of TE lysis buffer (25 mM Tris, 25 mM EDTA and 50 mM glucose). Lastly, the cell pellet was resuspended in a small volume (1–5 ml) of lysis buffer, the optical density at 600 nm (OD_{600}) was checked and the bacterial suspension diluted as necessary for the desired DNA extraction protocol.

DNA extractions

Direct soil extraction. DNA extraction directly from soil was performed as previously described[18]. Briefly, 2.4 ml of CTAB lysis buffer (100 mM Tris-HCl, 100 mM Na EDTA, 1.5 M NaCl, 1% w/v CTAB and 2% w/v SDS, pH 8.0) was added to 2 g of soil in a 5-ml tube and the slurry was incubated at 70 °C for 2 h. For the nycodenz-isolated bacteria, 3 ml of a 30-OD_{600} bacterial suspension was pelleted, then resuspended in 4 ml of CTAB lysis buffer and incubated at 70 °C for 2 h. After incubation, the suspensions were centrifuged at 20,000*g* for 10 min and the supernatant was moved to new tubes. Supernatants were combined with 0.7 volumes of isopropanol and incubated at room temperature for 10 min. Samples were centrifuged at 20,000*g* for 30 min and the resulting DNA pellets was washed twice with 1 ml of 70% ethanol. The pellets were then air-dried for 1–2 h at room temperature and resuspended in 100–200 µl of 5 mM Tris pH 8 overnight at 4 °C.

Gel plug. This protocol is an adaptation of gel plug lysis described by Walter et al.[45] Nycodenz-isolated bacteria were resuspended in lysis buffer to 10 OD_{600}. The bacterial suspension was mixed 1:1 with 1.5% UltraPure low-melting-point agarose (Thermo Fisher Scientific) prepared in lysis buffer and cooled to 50 °C. The mixture was used to cast gel plugs in CHEF Mapper XA system 50-well plug molds (BioRad, 1703713). Gel plugs were then suspended in a solution of 5 mg ml^{-1} lysozyme and 1 mg ml^{-1} achromopeptidase prepared in lysis buffer and gently rotated at 37 °C overnight using a Roto-therm plus (Benchmark Scientific). The buffer was decanted and the plugs were topped up with TESP buffer (50 mM Tris, 100 mM EDTA, 1% w/v SDS and 1 mg ml^{-1} proteinase K) and incubated overnight with gentle rotation at 50 °C. The gel plug suspension was decanted and the plugs were topped up with fresh TESP and once again incubated overnight with gentle rotation at 50 °C. TESP was then decanted and the gel plugs were rinsed several times with milli-Q water before being suspended in T10E50 (10 mM Tris and 50 mM EDTA, pH 8) with 1 mM PMSF and incubated at room temperature for 2 h. The liquid was decanted and the plugs washed four times with T10E50 with each wash lasting 1 h on ice. Gel plugs were then stored suspended in T10E50 in the fridge.

Direct lysis. Cell lysis was performed similarly to previously established approaches[46,47]. Nycodenz-isolated bacteria were resuspended in lysis buffer to 10 OD_{600} and 500 µl of the suspension was pelleted by centrifugation for 1 min at 20,000*g*. The cell pellet was resuspended in 800 µl of lysis buffer containing 5 mg ml^{-1} lysozyme and 1 mg ml^{-1} achromopeptidase and the suspension was incubated at 37 °C for 2 h. After incubation, 1 mg of proteinase K powder was added, followed by 200 µl of 5% w/v SDS. The sample was very gently inverted to mix and incubated for 3 h at 55 °C. The viscous solution was then layered onto a sucrose gradient for size selection.

Monarch HMW DNA extraction kit for tissue. About 1 ml of 10-OD_{600} nycodenz-isolated bacteria was centrifuged at 20,000*g* for 1 min and the cell pellet was resuspended in 300 µl of TE lysis buffer containing 5 mg ml^{-1} lysozyme and 1 mg ml^{-1} achromopeptidase. The sample was incubated for 2 h at 37 °C and then carried forward using the New England Biolabs protocol for HMW DNA extraction from bacteria using the Monarch HMW DNA extraction kit for tissue (New England Biolabs, T3060). Briefly, 20 µl of proteinase K was added and the sample was incubated at 56 °C with varied shaking speeds (1,900, 1,700, 1,400 or 650 rpm) using a Fisherbrand Heat/Cool Thermal Mixer II (Fisher Scientific). After 30 min, 10 µl of RNase A was added and the sample was incubated a further 10 min with shaking at the same speed. Then, 300 µl of protein separation solution was added and the sample was mixed by inversion for 1 min. The sample was centrifuged for 10 min at 16,000*g* and the upper phase was then transferred to a Monarch 2-ml tube using a wide-bore pipette tip. Two DNA capture beads were added, followed by 550 µl of isopropanol. The sample was mixed by gentle rotation using the Roto-therm Plus on the lowest speed setting for 5 min. The liquid was then discarded and the beads were washed twice with 500 µl of genomic DNA wash buffer. Finally, DNA was eluted by addition of 100 µl of elution buffer II and incubation at 56 °C for 10 min with 300-rpm agitation. The DNA solution was left to resuspend at 4 °C overnight.

DNA size-selection methods

Electroelution. The electroelution protocol was performed as previously described[18] with minor modifications. Briefly, a 0.9% agarose gel was run in 0.5× TBE (VWR) at 6 V cm^{-1} for 2 h to size-select crude DNA extracts. The compression band containing HMW DNA was then excised, the gel slice was placed in dialysis tubing (Spectra/Por; 12,000–14,000-kDa molecular weight cutoff (MWCO)) and topped up with 0.5× TBE. DNA was then electroeluted in 0.5× TBE at 6 V cm^{-1} for 2 h. The gel slice was removed from the dialysis bag and the remaining liquid was concentrated to dryness against 25% PEG 8000 with 5 mM Tris pH 8. DNA was then rehydrated in the dialysis bag against 5 mM Tris pH 8 for isolation.

Sucrose gradient. Linear 10–40% sucrose gradients, with 10 mM Tris pH 8, were prepared using the freeze–thaw method[48]. Gradients were prepared by liquid nitrogen flash-freezing at the end of the day and immediately placed to thaw at 4 °C until use 22–24 h later. Two sizes of polypropylene tubes were used, 25 × 89 mm and 14 × 95 mm (Beckman), with 8-ml and 2.9-ml sucrose layers, respectively. Larger tubes were centrifuged using an SW28 rotor, while the smaller tubes were centrifuged using an SW40 Ti rotor. In both cases, centrifugation was at 4 °C, 40,000*g* for 16 h. After centrifugation, fractions were gently pipetted from the top of the tubes using wide-bore tips. Pulse-field gel electrophoresis was used to analyze the fractions. For shorter-fragment purification (that is, DNA from hot CTAB and Monarch 1,900 rpm), analysis was conducted using 0.9% agarose with 1× TAE at 6 V cm^{-1} for 4 h with a 5-s switch time. For ultralarge DNA, analysis was conducted using 0.9% agarose with 1× TAE at 6 V cm^{-1} for 16 h with a linear 3–7-s switch time. Ultralarge DNA from direct lysis was isolated using large tubes from fractions at 14 ml to 32 ml. Hot CTAB DNA was purified using small tubes from fractions at 2.5 ml to 4.5 ml. DNA from Monarch 1,900 rpm was purified using large tubes from fractions at 6 ml to 8.5 ml, aiming to isolate DNA ~40 kb in size.

Short fragment eliminator (SFE) kit. DNA size selection with the SFE kit (Oxford Nanopore EXP-SFE001) was performed according to the manufacturer's instructions.

A-domain survey

DNA from gel-plug-extracted metagenomic DNA was isolated using the NucleoSpin gel and PCR cleanup kit (Macherey-Nagel) according to the manufacturer's instructions. A-domain amplicon sequencing was performed as previously described[49]. Read counts were 1,424,301 for forest soil 1, 923,413 for forest soil 2, 551,873 for forest soil 3, 1,965,137 for forest soil 4, 717,416 for the muddy bog and 467,043 for the decaying tree. A-domain amplicons were trimmed to remove primer sequences using vsearch (version 2.28.1)[50] fastx_filter and then assembled into operational taxonomic units (OTUs) clustered at 95% nucleotide similarity using vsearch cluster_size. A-domain OTUs for each soil were then compared across all samples at 95% similarity. Parallel processing was achieved using GNU parallel (version 20230922)[51].

Nanopore ultralong DNA sequencing

Ultralong DNA sequencing kit V14 (SQK-ULK114) preparation was performed according to the manufacturer's instructions. Three DNA inputs were attempted, one following Oxford Nanopore directions on the Monarch HMW DNA extraction kit for tissue extraction, the manufacturer-recommended protocol for Monarch HMW DNA extraction kit for tissue and sucrose gradient purification. An additional attempt using the Monarch extracted DNA input was tried with a slightly altered protocol; instead of heat-killing the fragmentation mix, proteinase K was added to 1 mg ml^{-1} for 1 h at room temperature, followed by the addition of 3 mM PMSF for 1 h at room temperature to inhibit the proteinase K before continuing the protocol.

Nanopore ligation kit sequencing

Oxford Nanopore's ligation sequencing kit V14 (SQK-LSK114) was used to prepare sequencing samples, with DNA extraction and size selection as indicated, according to the manufacturer's instructions and minor modifications. Input DNA was increased to 3 µg, blunting and end-repair mix reaction incubation times were extended to 15 min at 20 °C and 10 min at 65 °C and adaptor ligation incubation time was increased to 30 min. Library solution was used for all samples and the DNA concentrations loaded are indicated in Supplementary Table 2. On the basis of our optimization experiments, the 1,700-rpm Monarch isolation and SFE size-selection protocol applied to 62.5 g of the forest soil would be sufficient to generate >2 Tbp of sequence.

Bioinformatic analysis

Sequencing. PromethION R10.4.1 flow cells were used to generate all of the sequencing data. Nanopore sequencing was conducted using a P2 solo with MinKnow 23.04.6 (ultra kits) or 23.11.4 (Ligation kits) and a P24 with MinKnow 23.07.12 (scaled ligation). All raw data were duplex-basecalled with standalone Dorado 0.4.1 + 6c4c636 for subsequent analysis.

Global assembly. The longest, highest-quality reads (>20 kbp, Q20+) were extracted from the 2.5-Tbp sequence dataset using Chopper (version 0.6.0)[52]. The resulting 528.5-Gbp dataset was assembled using Flye (version 2.9.3-b1797) in metagenomic mode (metaFlye) with parameters '--nano-hq --i 0 --threads 64 --meta' (ref. 19). The assembly was run on Rockefeller University's large-memory high-performance computing node (64 cores, 3 TB of RAM) and took just under 8 days to complete with RAM usage peaking at 845 GB. The assembly was polished once (the emerging best practice[53]) with Medaka (version 1.11.3; https://github.com/nanoporetech/medaka) and annotated by Bakta (version 1.10.4)[54]. Assembled contigs > 1 Mbp were extracted and their taxonomy assigned using the GTDB toolkit (GTDB-tk version 2.4.0)[55] classify workflow with GTDB release 09-RS220. The presence of rRNA was detected within the Bakta annotation using a GNU AWK script (version 5.3.0), while tRNAscan-SE (version 2.0.9) was used to quantify total unique tRNAs. A set of 563 contigs > 1 Mbp were identified containing all rRNAs and at least 18 unique tRNAs to establish the complete or near-complete metagenomic genome dataset. The dataset was then frame-shift-corrected using proovframe[56] and completeness was estimated by CheckM (version 1.2.3)[57]. The taxonomic classification of these contigs was plotted as a Sankey diagram using the networkD3 R package 0.4 (Fig. 2a). AntiSMASH (version 6.1.1)[58] results from the output JSON files were used to generate the appended BGC summaries. Subtrees expanding previously underexplored clades were generated by extracting the alignment information from the GTDB-tk classify workflow for desired sequences and plotting with FastTree (version 2.1.11)[59] using default settings. Additional information was then manually added.

Targeted NRPS assembly. HMMER (version 3.1b2)[60] was used to scan the total sequencing dataset for AMP-binding domains using an available HMM (AMP-binding, Pfam PF00501.23). Reads containing AMP-binding domains were assembled (metaFlye), polished (Medaka) and annotated (Bakta), as described above. The AMP-binding domain search using HMMER was repeated on the assemblies to extract the subset containing at least four AMP-binding domains. Similar to the complete and near-complete genome assembly dataset, frameshifts were minimized using proovframe and then the subset was analyzed by antiSMASH (version 6.1.1). For analysis of NRPS content, the antiSMASH results were analyzed by BiG-SCAPE (version 1.1.5) to generate Fig. 3a. The phylogeny of NRPS containing BGCs was determined using the mmseqs2 (ref. 61) toolkit's easy-taxonomy workflow and the National Center for Biotechnology Information nonredundant protein database.

16S rRNA analysis. Reads containing 16S rRNA sequences were extracted from the 2.5 Tbp of sequencing data using barrnap 0.9 (https://github.com/tseemann/barrnap). The 16S rRNA read subset was then assembled (metaFlye), polished (Medaka) and annotated (Bakta) as described above. Annotated 16S rRNA gene sequences were then extracted and, using vsearch, dereplicated and clustered at 99%, the recommended threshold for clustering full-length 16S sequences[62]. Lastly, to ensure full-length sequences, only OTUs between 1,400 and 1,700 bp were retained. The same extraction, dereplication and clustering process was used to identify 16S rRNA genes within the global assembly dataset. Phylogeny was assigned using mmseqs2 and the SILVA database (release 138.2)[63].

Forest soil 1 cultured isolates

Culturing. Forest soil 1 microorganisms were isolated using the nycodenz method with the last two washes replaced with Dulbecco's PBS (DPBS; Gibco). The purified suspension was diluted to 0.05 OD_{600} in DPBS and 200 µl was spread on individual agar plates (150 mm × 15 mm). Ten plates each were used with three different media: R2A agar (Fisher Scientific), which was previously used to isolate the *Aquella* species[27], and two versions of dilute nutrient broth (Criterion). In one case, nutrient broth was diluted 200-fold compared to the manufacturer's recommendation; in the second case, dilute nutrient broth was with supplemented 1:1 (v/v) with soil extract[64]. In all cases, nystatin (50 µg ml^{-1}) was added to limit fungal growth. Plates were incubated at 30 °C with R2A colonies picked after 3 days and nutrient broth colonies picked after 10 days. Attempting to maximize isolate diversity, colonies were picked to maximize different morphologies. Liquid R2A or unmodified nutrient broth were used to culture isolates for sequencing (3 ml, 30 °C, 200-rpm shaking).

Genomic DNA extraction and sequencing. Bacterial cells were pelleted from turbid cultures (1–1.5 ml) by centrifugation for 1 min at 20,000*g*. The supernatant was removed and the cell pellets were caried forward for genomic DNA isolation using the PureLink microbiome DNA purification kit (Thermo Fisher Scientific) as per manufacturer's instructions for soil samples with elution volume lowered to 50 µl. Oxford Nanopore's native barcoding kit 24 V14 (SQK-NBD114.24) was used as per the manufacturer's instructions to prepare sequencing samples. From R2A medium culturing, 22 genomes were individually barcoded and sequenced. From nutrient medium culturing, 20 genomes were individually barcoded and 12 were barcoded as pools of three. Each medium set was sequenced using a PromethION flow cell.

Bioinformatics. Isolate genomes were assembled using Flye with Medaka polishing and Bakta annotation. Genome assemblies were dereplicated by extracting 16S rRNA genes, clustering sequences at 99% identity and then selecting a single assembly per cluster. This yielded 30 genomes (Supplementary Table 7). BGCs were identified in each dereplicated cultured genome using antiSMASH. Large NRPS BGCs (that is, ≥4 A domains) were compared to those from the targeted metagenomic assembly using mmseqs2. Matches are shown in Supplementary Fig. 3.

SynBNP predictions

SynBNP prediction was carried out according to our standard workflow[65]. Briefly, 'bioinformatically tractable clusters' were identified and their products were predicted as follows. BGCs were scanned for NRPS domains and modules. The order in which these domains are likely to function was predicted on the basis of canonical NRPS module organization combined with the principle of collinearity. The following BGCs were then removed: (1) BGCs lacking predicted initiation or termination domains; (2) BGCs with modules that do not follow canonical NRPS domain organization (that is, condensation domain, A-domain and thiolation domain); (3) BGCs rich in tailoring genes; and (4) BGCs having fewer than four A domains, as these make poor synBNP targets. Structure predictions were then generated on the basis of A-domain specificity and derived structural modifications predicted on the basis of an analysis of additional domains present in the BGC (for example, epimerization and *N*-methylation). These data were combined on the basis of the initial domain order analysis to generate the linear structure encoded by a BGC. Specificity codes were compared to both the antiSMASH output and an in-house specificity code table collated from previously characterized natural products. The highest A-domain matches were solely considered for predictions. In rare cases where multiple amino acids share the same code, one was selected from the consensus of predictions generated by NRPSPredictor2 within antiSMASH. If multiple predictions were tied after the consensus stage, one was selected at random. The set of bioinformatically tractable clusters considered and their corresponding predictions are given in Supplementary Table 8. This represented roughly 19% of detected NRPS BGCs. On average, A domains matched at 90.1% ± 13.9%. The average A-domain match per peptide was 89.9% ± 8.0%. From the subset of predictions that contained at most one <70% A-domain specificity code (~83% of the predictions passed this cutoff), we randomly selected 20for synthesis. These covered a range of phylogenetic classifications (Fig. 3b and Supplementary Table 9). Contig_44957 had two mismatched codes but was retained as the codes were closely related and predicted to specify the same amino acid. While we used a cutoff of NRPs containing at most a single A-domain that matched <70%, this can be loosened or more restrictive if desired. A breakdown of predictions that passed different cutoff levels is provided in Supplementary Table 10.

SynBNP synthesis

Reagents and solvents used for synthesis were obtained from commercial sources and used without further purification. Chromatography solvents were of high-performance liquid chromatography (HPLC) grade or higher. Peptides were purified using a CombiFlash EZ Prep purification system with ultraviolet detection and equipped with a RediSep gold HP C18 column (30 g), RediSep gold HP C8 column or Phenomenex Luna 5 µm C18 prepHPLC column using a 5–95% acetonitrile–water gradient supplemented with 0.1% formic acid. High-resolution mass spectrometry (HRMS) data were acquired on a SCIEX ExionLC ultra-HPLC (UPLC) instrument coupled to an X500R quadrupole time-of-flight mass spectrometer with a Phenomenex Kinetex PS C18 100-Å column (2.1 mm × 50 mm, 2.6 µm) and analyzed using SCIEXOS software. ^{1}H and ^{13}C nuclear magnetic resonance (NMR) spectra were obtained on a Bruker Avance DMX 600-MHz spectrometer equipped with cryogenic probe (Rockefeller University) and all spectra were analyzed using MestRe Nova software.

Linear peptides were synthesized using standard Fmoc-based solid-phase peptide synthesis (SPPS) methods. 2-Chlorotritylchloride resin (0.1 mmol) was swollen in dichloromethane (DCM) at room temperature for 30 min. After draining, the resin was washed with DCM (three washes, 5 ml each). The C-terminal Fmoc-protected amino acid (four equivalents, 0.4 mmol) was dissolved in 5 ml of DCM and treated with *N*,*N*-diisopropylethylamine (DIPEA; five equivalents, 0.5 mmol). This mixture was added to the swollen resin and agitated for 1 h. The solution was drained and the resin was washed with DCM (three washes, 5 ml each). The resin was then capped by treatment with a 17:2:1 mixture of DCM, methanol and DIPEA. The resin was agitated for 30 min, drained and washed with DCM (three washes, 5 ml each) and dimethylformamide (DMF; three washes, 5 ml each). For peptide couplings, the Fmoc group was removed by treating the resin with 20% piperidine in 5 ml of DMF for 7 min. The solution was drained, the deprotection reaction was repeated and the resin was washed with DMF (five washes, 5 ml each). The next Fmoc amino acid to be coupled (three equivalents, 0.3 mmol) and hexafluorophosphate azabenzotriazole tetramethyl uronium (HATU; three equivalents, 0.3 mmol) was dissolved in 4 ml of DMF and treated with DIPEA (three equivalents, 0.3 mmol). This solution was added to the resin and agitated for 45 min. The solution was drained and the resin was washed with DMF (three washes, 5 ml each). This deprotection and coupling process was repeated for each subsequent amino acid. Finally, the N-terminal Fmoc amino acid was deprotected with 20% piperidine in DMF (two cycles of 5 ml for 7 min) and reacted with myristic acid (three equivalents, 0.3 mmol), HATU (three equivalents, 0.3 mmol) and DIPEA (three equivalents, 0.3 mmol) in DMF (4 ml). The reaction was agitated for 45 min, at which point the resin was washed with DMF (three washes, 5 ml each) and DCM (three washes, 5 ml each). The peptide was cleaved from resin and the side-chain-protecting groups were removed by treating the resin with a trifluoroacetic acid (TFA) cleavage cocktail containing 95% TFA, 2.5% water and 2.5% triisopropylsilane for 2 h. The solution was evaporated

under air flow and the resulting residue was purified on a CombiFlash EZ Prep system (30-g HP C18 column, water–acetonitrile with 0.1% v/v formic acid 5%–95%).

Peptides cyclized through the side chain of an amino acid were synthesized starting from the penultimate amino acid following the same Fmoc SPPS loading, Fmoc deprotection and coupling procedures described above. Side-chain serine and threonine residues at which cyclization was predicted to occur were added to the peptide with the side chain unprotected. After the N-terminal amino acid was deprotected, myristic acid was coupled to the terminal amine as described above (for peptides not predicted to have a lipid tail, the Boc-protected N-terminal amino acid was coupled). Next, the C-terminal amino acid was esterified onto the free serine or threonine side chain by adding a solution of Fmoc amino acid (15 equivalents, 1.5 mmol), *N*,*N*′-diisopropylcarbodiimide (DIC; 15 equivalents, 1.5 mmol) and 4-dimethylaminopyridine (DMAP; 0.5 equivalents, 0.05 mmol) in 5 ml of DMF to the resin and agitating for 16 h. The solution was drained and the resin was washed with DMF (three washes, 5 ml each) and DCM (three washes, 5 ml each), followed by deprotection of the Fmoc group using 20% piperidine in DMF (two cycles of 5 ml for 7 min). The linear precursor peptide was gently cleaved from resin by treating with a solution of 20% hexafluoroisopropanol in 6 ml of DCM for 45 min. The solution was collected and the process repeated once more. After removal of solvent under air flow, the resulting peptide was dissolved in DMF (50 ml, 0.002 M) and treated with (7-azabenzotriazol-1-yloxy) tripyrrolidinophosphonium hexafluorophosphate (seven equivalents, 0.7 mmol) and DIPEA (30 equivalents, 3 mmol). After stirring for 2 h, the solution was diluted with ethyl acetate and washed with brine (three washes, 100 ml each). The organic layer was dried (Na_2SO_4), filtered and concentrated in vacuo. The resulting cyclic peptide was globally deprotected by treatment with TFA cleavage cocktail for 2 h. After evaporation under air flow, the crude cyclic peptide was purified on a CombiFlash EZ Prep system as described above.

Peptides cyclized through the fatty acid were synthesized starting from the penultimate amino acid residue following the Fmoc SPPS method described above. After synthesis of the linear precursor was completed, the N-terminal Fmoc group was removed using 20% piperidine in DMF (two cycles of 5 ml for 7 min) and reacted with (±)-3-hydroxymyristic acid (three equivalents, 0.3 mmol), HATU (three equivalents, 0.3 mmol) and DIPEA (three equivalents, 0.3 mmol) in DMF (4 ml). The reaction was agitated for 45 min, at which point the resin was washed with DMF (three washes, 5 ml each) and DCM (three washes, 5 ml each). Next, the C-terminal amino acid was esterified onto the hydroxyl group of the fatty acid by adding a solution of Fmoc amino acid (15 equivalents, 1.5 mmol), DIC (15 equivalents, 1.5 mmol) and DMAP (0.5 equivalents, 0.05 mmol) in 5 ml of DMF to the resin and agitating for 16 h. The solution was drained and the resin was washed with DMF (three washes, 5 ml each) and DCM (three washes, 5 ml each), followed by deprotection of the Fmoc group using 20% piperidine in DMF (two cycles of 5 ml for 7 min). The linear precursor was cleaved from resin, cyclized, deprotected and purified as described above.

Successful peptide synthesis was confirmed by UPLC, HRMS and/or NMR (Supplementary Table 15 and Supplementary Figs. 6–9). All compounds that displayed antibacterial activity were confirmed to have >95% purity.

MIC assay

To avoid solubility issues, each synBNP concentration was prepared individually in DMSO before addition to LB. A series of eight concentrations were tested, with the last one being a DMSO only control. First, the synBNP was diluted to 3.2 mg ml^{-1} in DMSO and then serially diluted twofold six times in DMSO. Then, 4 µl of each concentration was added to 100 µl of LB dispensed across a 96-well plate (Costar, 3370). Test bacteria were grown overnight shaking at 200 rpm at 30 °C and the culture was diluted to an OD_{600} of 0.005. Next, 100 µl of the dilution was added to the antibiotic dilution series, yielding a test range of 64 to 1 µg ml^{-1} along with a control (0 µg ml^{-1}). The plate was then incubated statically at 30 °C for 18 h, followed by visual inspection. The MIC was determined as the lowest concentration within the dilution series that did not exhibit turbidity. MICs were determined by two biologically independent replicates for the initial synBNP screen and three independent replicates for all other determinations.

Human cell toxicity assay

The inhibitory activity of synBNPs was assessed using a 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay[66]. HEK239 and HeLa cells were cultured in DMEM supplemented with 10% FBS, 10 µg ml^{-1} penicillin–streptomycin, 2 µM glutamine and nonessential amino acids. Cells were passaged and seeded in 96-well plates at the exponential growth phase, inoculated at ~1×10^5 cells per ml and maintaining a cell density below 1×10^6 cells per ml. The plates were incubated at 37 °C with 5% CO_2. All assays were performed using 5,000 cells per well, with 50 µl of a homogeneous cell suspension at 1×10^5 cells per ml seeded in each well. For screening, synBNPs were diluted to 6.4 µg ml^{-1} in DMSO to prepare a 200× stock solution. For the high-concentration assays (32 µg ml^{-1}), 1.5 µl of synBNP stocks were used to prepare a master solution by dilution in 150 µl of DMEM without phenol red. Two 50-µl aliquots of the master solution were added to cell-seeded plates, obtaining duplicate 100-µl wells with synBNPs at 32 µg ml^{-1} (0.5% DMSO). For the low concentrations (3.2 µg ml^{-1}), 15 µl of the master solution was diluted to 150 µl with DMEM and 50 µl was added to cell-seeded plates to obtain duplicate wells of 100 µl at 3.2 µg ml^{-1} (0.05% DMSO). DMEM solutions containing 0.5% and 0.05% DMSO were used as vehicle controls to define 100% cell viability. Camptothecin at 40 µg ml^{-1} was used as a positive control to define 0% cell viability. After treatment, plates were incubated for 48 h followed by addition of 10 µl of 5 mg ml^{-1} MTT solution in DPBS and the plate was incubated for an additional 4 h (SK-N-SH was incubated for 5–6 h). After incubation, 90 µl of solubilization solution (40% DMF, 16% SDS and 2% acetic acid) was added to each well and formazan crystals were allowed to dissolve overnight. The absorbance of each well was measured at 570 nm using a TECAN Infinite M NANO+. Absorbance values were normalized to the positive and negative controls and percentage survival values were calculated.

Kill curve assay

S. aureus SH1000 was streaked onto a LB agar plate and incubated overnight at 30 °C. Three independent colonies were then inoculated into 3 ml of LB and grown overnight at 30 °C with shaking at 200 rpm. The cultures were diluted 1:100 and grown at 30 °C and 200 rpm until they reached an OD_{600} between 0.4 and 0.6. The growing cultures were then diluted into a preprepared series of 3 ml of LB containing the indicated antibiotic concentrations to ~6×10^6 colony-forming units (cfu) per ml using the expectation that 1 OD_{600} of *S. aureus* contained 5×10^8 cfu per ml. When a series was inoculated, a 100-µl sample was taken, representing 1 min of treatment. The cultures were then placed shaking at 200 rpm and 30 °C and samples were taken at 10 min, 30 min, 1 h and then every subsequent hour for a total of 6 h. For cfu quantification, 100-µl samples were placed immediately into the top well of a 96-well plate and serially diluted tenfold (20 µl into 180 µl of LB) down to 10^{-7}. Then, 6 µl of each dilution from the series was spotted onto LB agar, the spots were air-dried and the plate was incubated overnight at 30 °C. Resulting colonies were enumerated to calculate the cfu per ml.

Mutant isolation

Direct plating. *S. aureus* SH1000 was streaked onto an LB agar plate and incubated overnight at 30 °C. A single colony was inoculated into 3 ml of LB and grown to confluence overnight at 30 °C. The culture was then diluted to roughly 1×10^7 cells per ml in LB containing antibiotic at 8× MIC. Five 96-well plates were then seeded with 100 µl of the

dilution and grown stationary overnight at 30 °C. Mutants appeared as small colonies at the bottom of the well and were picked into fresh 8× MIC medium to ensure resistance. The confirmation culture was then streaked onto LB agar and incubated overnight at 30 °C; two colonies per mutant were inoculated for an overnight culture in LB at 30 °C. The MIC of each mutant was checked to confirm resistance. For erutamicin, no resistant colonies appeared and the direct plating was repeated at 4× and 2× MIC concentrations, which also did not produce resistant mutants.

Passaging. *S. aureus* SH1000 and *A. baumannii* ATCC 17978 were streaked onto LB agar and incubated overnight at 30 °C. Three independent colonies for each of the bacteria were then inoculated into 3 ml of LB and grown overnight at 30 °C with shaking at 200 rpm. The cultures were used to prepare a MIC assay as described above with the indicated antibiotics. The following day, 1 µl was taken from the well containing the highest concentration of antibiotic exhibiting turbid growth and mixed into 1 ml of fresh LB. This dilution was used to seed a new MIC assay. This was repeated for each subsequent day and the change in MIC was plotted.

Mutant sequencing and identification

Genomic DNA from resistant mutants was extracted for sequencing using the PureLink microbiome DNA purification kit (Thermo Fisher Scientific) per the manufacturer's instructions. Isolated DNA was then Illumina-sequenced using standard Nextera preparation and the MiSeq reagent kit v3 with 2× 300-bp reads. Reads for both parent and mutants were mapped to a reference genome using minimap2 (version 2.28-r1209)[67]. Changes detected in the parent strain were eliminated from consideration and mutations were uncovered using snippy (https://github.com/tseemann/snippy).

ClpX expression vector

The *clpX* gene was amplified from *S. aureus* SH1000 genomic DNA by PCR using the primers ClpXF (TATTACTCGAGATGTTTAAATTCAATGAAGATGAAG) and ClpXR (TAATAGGATCCACATCAATGATTAAGCTGATG). The pET19b (Sigma) vector backbone was amplified by 19BF (TAATAGGATCCGAAAGGAAGCTGAGTTGG) and 19BR (TATTACTCGAGCATATGCTTGTCGTCGTCGTC). The two PCR products were then digested with XhoI and BamHI and ligated together; the ligation reaction was transformed into *E. coli* EPI300 for colony screening on LB agar plates containing 100 µg ml^{-1} carbapenem. The successful construction of the ClpX expression vector pHisClpX (which added a 10× His tag to the N terminus of ClpX) was confirmed by Sanger sequencing performed by Genewiz using the universal T7 primer.

ClpX expression and purification

The pHisClpX vector was transformed into *E. coli* Rosetta 2(DE3) (Sigma) and a single colony was inoculated into 5 ml of LB containing carbapenem to grow overnight at 37 °C and 200-rpm shaking. The overnight culture was then added to 1 L of LB + carbapenem prewarmed to 37 °C and grown at 37 °C and 200-rpm shaking until an OD_{600} of 0.2–0.4. Once the desired OD range was achieved, the culture was transferred to 16 °C and 200-rpm shaking for 30 min, after which IPTG was added to 1 mM to induce protein expression, which continued overnight at 16 °C and 200 rpm. The culture was pelleted by centrifugation at 4,000*g* for 45 min and the cell pellet was resuspended in 10 ml of protein lysis buffer (50 mM Tris pH 7.5, 100 mM NaCl, 5 mM imidazole, 0.1 mM EDTA and 1 mM β-mercaptoethanol). The suspension was sonicated using a Fisherbrand sonicator for a total of 7.5 min of sonication time, with 45 s on and 45 s off at 45% amplitude. The sonicated suspension was then centrifuged at 4,000*g* for 30 min. The supernatant was then centrifuged at 23,000*g* for an additional 30 min. The His-tagged ClpX was isolated from the cleared supernatant using Ni-NTA agarose (Quiagen) according to the manufacturer's instructions. Briefly, a column containing 2 ml of Ni-NTA was washed with 20 ml of protein lysis buffer, followed by passage of the cleared lysate. Bound ClpX was then washed with 20 ml of increasing concentrations of imidazole in protein lysis buffer (10 mM, 20 mM and 30 mM), followed by elution with protein lysis buffer containing 250 mM imidazole, while collecting 0.5-ml fractions. The protein content in the eluted fractions was checked by absorbance at 280 nm, with any fractions at an OD > 0.6 pooled together. About 1 ml of the pooled fractions were then dialyzed (MWCO: 12,000–14,000 kDa; Spectra/Por) against 1 L of protein storage buffer (50 mM Tris pH 7.5, 200 mM KCl, 25 mM $MgCl_2$, 1 mM β-mercaptoethanol and 10% glycerol) for 3 h at 4 °C, followed by an additional dialysis against 3 L of protein storage buffer overnight. The protein concentration was then determined to be 5.83 mg ml^{-1} (117 µM) by spectroscopy at 280 nm with the calculation of 1 OD corresponding to 3.69 mg ml^{-1} of His-ClpX. A sample (50 ml) of the 3-L storage buffer after dialysis was saved for resuspending trigintamicin and its isomer for ITC.

ITC

ITC measurements were carried out at 25 °C using a Malvern ITC PEAQ system. The ligands (trigintamicin and its biologically inactive isoform) were dissolved in storage buffer taken post dialysis of ClpX. The concentration of ClpX was adjusted to ~30 µM and the ligand concentrations were adjusted to ~300 µM using the postdialysis storage buffer batch. Then, 300 µl of ClpX solution was placed in the sample cell and titrated by ligand solution through 19 successive injections of 2 µl with a 2.5-min interval between injections and a reference power of 5 µcal s^{-1}. Titration data were analyzed using the PEAQ-ITC analysis software (version 1.41). The isotherm curves were best fitted to a single-site binding model.

Depolarization assay

An overnight culture of *S. aureus* SH1000 was pelleted, washed twice in equal volume of DPBS (Thermo Fisher Scientific) and diluted to 0.35 OD_{600} in DPBS. Then, 100 µl of the dilution was added to 300 µl of DPBS, followed by the addition of 50 µl of 20 µM $DISC_3(5)$ dye in DPBS. The mixture was incubated at room temperature in the dark for 15 min, followed by the addition of 50 µl of 2 M KCl and another 15-min incubation. The samples were then transferred into a 384-well flat, clear-bottom black microtiter plate with 30 µl per well. The initial fluorescence intensity of each well was recorded using a TECAN Infinite M NANO$^+$ (excitation, 620 nm; emission, 675 nm) at 10-s intervals until the baseline stabilized. After signal stabilization, 30 µl of test compounds diluted in DPBS were added to the indicated concentrations and readings resumed. Gramicidin at 16 µg ml^{-1} (2× MIC) was used as a positive control for depolarization.

Lipid antagonism assay

Test lipids were dissolved in the manufacturer's recommended solvent to 0.5–1 mg ml^{-1}. An aliquot containing the appropriate amount of lipid for the desired molar ratio to a target antibiotic was transferred to a 1.5-ml Eppendorf and dried under vacuum for 1 h. The dried pellet was then resuspended by pipetting with 6 µl of the test antibiotic dissolved in ethanol at 1.28 µg ml^{-1}. The suspension was then topped up with 114 µl of LB and the sample was mixed by pipetting. Then, 100 µl was transferred to a 96-well assay plate and 50 µl was transferred down a column to generate a 1:2 dilution series. An overnight culture of *S. aureus* SH1000 was diluted 1:1,000 into LB and 50 µl was added to the dilution series to generate a final antibiotic concentration range of 0.25 to 32 µg ml^{-1}. The plate was then incubated overnight at 30 °C for 18 h, after which growth was visually inspected. For experiments with telomycin and daptomycin, all LB was supplemented with Ca^{2+} to 50 mg L^{-1} (183 mg L^{-1} $CaCl_2$).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 563 high-quality metagenomic assemblies, the 338 large NRPS assemblies, the genomes of the 30 cultured isolates and the extracted 16S rRNA datasets were deposited to Zenodo (https://doi.org/10.5281/zenodo.16782139)[68]. Publicly available GTDB release 09-RS220 was used for taxonomic assignment and comparison. Publicly available AMP-binding protein (Pfam PF00501.23) was used for HMM searches. Publicly available SILVA database release 138.2 was used for taxonomic assignment of 16S rRNA sequences. Source data are provided with this paper.

Code availability

Nanopore sequencing was performed using a P2 solo with MinKnow 23.04.6 (ultra kits) or 23.11.4 (ligation kits) and a P24 with MinKnow 23.07.12 (scaled ligation). All raw data were duplex-basecalled with standalone Dorado 0.4.1+6c4c636. Publicly available vsearch 2.28.1, GNU parallel 20230922, Chopper 0.6.0, Flye 2.9.3-b1797, Medaka 1.11.3, Bakta 1.10.4, GTDB-tk 2.4.0, GNU AWK script 5.3.0, tRNAscan-SE 2.0.9, networkD3 R package 0.4, antiSMASH 6.1.1, FastTree 2.1.11, HMMER 3.1b2, proovframe 0.9.7, CheckM 1.2.3, Barrnap 0.9, BiG-SCAPE 1.1.5, minimap2 2.28-r1209 and Snippy 4.6.0 were used for sequence assembly and analysis. PEAQ-ITC analysis software 1.41 was used for ITC analysis.

References

42. Lindahl, V. & Bakken, L. R. Evaluation of methods for extraction of bacteria from soil. *FEMS Microbiol. Ecol.* **16**, 135–142 (1995).
43. Cheng, F. et al. Soil pretreatment and fast cell lysis for direct polymerase chain reaction from forest soils for terminal restriction fragment length polymorphism analysis of fungal communities. *Bras. J. Microbiol.* **47**, 817–827 (2016).
44. Amalfitano, S. & Fazi, S. Recovery and quantification of bacterial cells associated with streambed sediments. *J. Microbiol. Methods* **75**, 237–243 (2008).
45. Walter, J., Mangold, M. & Tannock, G. W. Construction, analysis, and β-glucanase screening of a bacterial artificial chromosome library from the large-bowel microbiota of mice. *Appl. Environ. Microbiol.* **71**, 2347–2354 (2005).
46. Morita, H. et al. An improved DNA isolation method for metagenomic analysis of the microbial flora of the human intestine. *Microbes Environ.* **22**, 214–222 (2007).
47. Barsotti, O. et al. Rapid isolation of DNA from Actinomyces. *Ann. Inst. Pasteur Microbiol.* **138**, 529–536 (1987).
48. Luthe, D. S. A simple technique for the preparation and storage of sucrose gradients. *Anal. Biochem.* **135**, 230–232 (1983).
49. Libis, V. et al. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).
50. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
51. Tange, O. GNU Parallel 2018. *Zenodo* https://doi.org/10.5281/zenodo.1146014 (2018).
52. De Coster, W. & Rademakers, R.NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
53. Luan, T. et al. Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. *BMC Genomics* **25**, 679 (2024).
54. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021).
55. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
56. Hackl, T. et al. proovframe: frameshift-correction for long-read (meta)genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.23.457338 (2021).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
59. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
60. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
61. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
62. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
63. Yilmaz, P. et al. The SILVA and 'All-Species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
64. George, I. F., Hartmann, M., Liles, M. R. & Agathos, S. N. Recovery of as-yet-uncultured soil acidobacteria on dilute solid media. *Appl. Environ. Microbiol.* **77**, 8184–8188 (2011).
65. Wang, Z., Koirala, B., Hernandez, Y., Zimmerman, M. & Brady, S. F. Bioinformatic prospecting and synthesis of a bifunctional lipopeptide antibiotic that evades resistance. *Science* **376**, 991–996 (2022).
66. Heo, D. S. et al. Evaluation of tetrazolium-based semiautomatic colorimetric assay for measurement of human antitumor cytotoxicity. *Cancer Res.* **50**, 3681–3690 (1990).
67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Burian, J. et al. Datasets for 'Bioactive molecules unearthed by terabase-scale long-read sequencing of a soil metagenome'. *Zenodo* https://doi.org/10.5281/zenodo.16782139 (2025).

Acknowledgements

We thank J.-N. Audet at the Rockefeller University Field Research Center for coordinating sample collection. We are grateful to C. Mason at Weill Cornell Medicine for access to a P24 sequencer and his technician K. Ryon for coordinating use. We also thank the Rockefeller University High-Throughput Screening and Spectroscopy Resource Center for assistance with ITC experiments. This work was supported by National Institutes of Health grant R35GM122559 (S.F.B.).

Author contributions

S.F.B. conceptualized and supervised the study. J.B. conceptualized, designed, performed and analyzed the experiments. C.P. assisted J.B. in the sample collection and DNA preparation. J.B. and Y.H. performed the bioinformatic analysis. R.B., L.J. and A.B. synthesized the peptides. L.J. performed the ITC experiment. A.M.-A. performed the human cell toxicity assays. M.A.T. performed the MiSeq analysis. S.F.B. and J.B. prepared the manuscript. All authors were involved in reviewing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02810-w.

Correspondence and requests for materials should be addressed to Sean F. Brady.

Peer review information *Nature Biotechnology* thanks Rayan Chikhi, Kim Lewis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Sean F Brady

Last updated by author(s): Mar 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Nanopore sequencing was done using a P2 solo with MinKnow 23.04.6 (ultra kits) or 23.11.4 (ligation kits), and a P24 with MinKnow 23.07.12 (scaled ligation). All raw data was duplex basecalled with standalone Dorado 0.4.1+6c4c636
Data analysis	Publicly available vsearch 2.28.1, GNU parallel 20230922, Chopper v0.6.0, Flye v2.9.3-b1797, Medaka v1.11.3, Bakta v1.10.4, GTDB-tk v2.4.0, GNU AWK script 5.3.0, tRNAscan-SE v2.0.9, networkD3 R package 0.4, antiSMASH v6.1.1, FastTree v2.1.11, HMMER v3.1b2, proovframe-v0.9.7, CheckM v1.2.3, Barrnap 0.9, BigSCAPE v1.1.5, minimap2 2.28-r1209 and Snippy 4.6.0 were used for sequence assembly and analysis. PEAQ-ITC analysis software 1.41 was used for ITC analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The 563 high-quality metagenomic assemblies, the 338 large NRPS assemblies, the genomes of the 30 cultured isolates, and the extracted 16S rRNA datasets have been submitted to GenBank under Bioproject PRJNA1226572. Publicly available Genome Taxonomy Database release 09-RS220 was used for taxonomic assignment and comparison. Publicly available AMP-binding protein PFAM PF00501.23 was used for Hidden Markov Model searches. Publicly available SILVA database release 138.2 was used for taxonomic assignment of 16S rRNA sequences. All other data is available in the main text or the supplementary materials.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size, biologically independent repeats were performed to ensure result accuracy and reproducibility.
Data exclusions	No data was excluded
Replication	Compound characterization results were confirmed through three biologically independent experiments, ITC experiments were multiple independent measurement experiments, synBNP MIC screening was performed in duplicate with all other MIC experiments performed in biologically independent triplicate
Randomization	Randomization was not required as each experimental replicate was grown under identical conditions.
Blinding	Blinding was not relevant as each experimental replicate was grown under identical conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293 (ATCC, CRL-1573), origin of HeLa is unknown
Authentication	cell lines were not authenticated
Mycoplasma contamination	MycoAlert detection kit (Lonza Bioscience)
Commonly misidentified lines (See ICLAC register)	n/a

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*