# Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors

Jeremy G. Owen[a,b], Zachary Charlop-Powers[a], Alexandra G. Smith[a], Melinda A. Ternei[a], Paula Y. Calle[a], Boojala Vijay B. Reddy[a,b], Daniel Montiel[a], and Sean F. Brady[a,b,1]

[a]Laboratory of Genetically Encoded Small Molecules and [b]Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved February 24, 2015 (received for review January 18, 2015)

**In molecular evolutionary analyses, short DNA sequences are used to infer phylogenetic relationships among species. Here we apply this principle to the study of bacterial biosynthesis, enabling the targeted isolation of previously unidentified natural products directly from complex metagenomes. Our approach uses short natural product sequence tags derived from conserved biosynthetic motifs to profile biosynthetic diversity in the environment and then guide the recovery of gene clusters from metagenomic libraries. The methodology is conceptually simple, requires only a small investment in sequencing, and is not computationally demanding. To demonstrate the power of this approach to natural product discovery we conducted a computational search for epoxyketone proteasome inhibitors within 185 globally distributed soil metagenomes. This led to the identification of 99 unique epoxyketone sequence tags, falling into 6 phylogenetically distinct clades. Complete gene clusters associated with nine unique tags were recovered from four saturating soil metagenomic libraries. Using heterologous expression methodologies, seven potent epoxyketone proteasome inhibitors (clarepoxcins A–E and landepoxcins A and B) were produced from these pathways, including compounds with different warhead structures and a naturally occurring halohydrin prodrug. This study provides a template for the targeted expansion of bacterially derived natural products using the global metagenome.**

drug discovery | environmental DNA | proteasome inhibitor | nonribosomal peptide | polyketide

The advent of cost-effective high-throughput sequencing and an increasingly sophisticated understanding of bacterial secondary metabolite biosynthesis have led to two important revelations with respect to the search for new natural products: first, that the biosynthetic potential of most cultured bacteria, as judged by the number of biosynthetic gene clusters (BGCs) observed in sequenced genomes, is far greater than previously estimated (1, 2); second, that the number of bacterial species in most environments is at least 100× greater than the number of species that is readily cultured (3, 4). These observations suggest that conventional "phenotype-first" natural products isolation approaches have only examined a small fraction of earth's bacterial biosynthetic potential.

There are now a number of genomic search engines available that allow researchers to rapidly scan microbial whole genome sequences for BGCs encoding new natural products (5–7). Unfortunately, the large DNA contigs that these search strategies require as input are not readily available from complex metagenomes. In response to the need for a more robust metagenomic search strategy, our group recently developed an informatics platform called eSNaPD (8, 9) (environmental Surveyor of Natural Product Diversity) with the specific aim of facilitating sequence-guided discovery of new bacterial natural products from complex metagenomes (Fig. 1).

The eSNaPD software is designed to bioinformatically assess short DNA sequences that have been amplified from environmental metagenomes by degenerate PCR targeting conserved biosynthetic motifs (natural product sequence tags, NPSTs). NPSTs are used to predict gene content and chemical output of the BGCs present in a metagenome, in a fashion analogous to reconstructing species phylogeny using 16S rRNA sequences (8). Once NPST data are generated from environmental metagenomes or metagenomic libraries, eSNaPD searches each NPST against a curated reference database, and identifies NPSTs whose closest evolutionary relative among all previously characterized reference BGCs encodes a molecule of interest. This "closest relative" search approach is computationally inexpensive; however, the output it provides is a robust predictor of pathway gene content and chemical output (8).

Here we use the eSNaPD informatics platform, in conjunction with a refined set of metagenomic tools, to discover and characterize previously unidentified epoxyketone proteasome inhibitor (EPI) natural products from soil metagenomes. EPIs irreversibly bind and inhibit the 20S proteasome leading to a toxic accumulation of polyubiquinated proteins in the cell (10). Although none of the small number of natural EPIs identified through conventional phenotype screening has yet to complete clinical trials, they have inspired the development of synthetic EPI analogs (e.g., bortezomib, carfilzomib, oprozomib) that are rapidly becoming key therapies

## Significance

Here we use an informatics-based approach to natural product discovery that is broadly applicable to the isolation of medicinally relevant metabolites from environmental microbiomes. Combining metagenome sequencing and bioinformatics approaches with a defined set of metagenomic tools provides a template for the targeted discovery of compounds from the global metagenome. The power of this approach is demonstrated by surveying ketosynthase domain amplicon sequencing data from 185 soil microbiomes for biosynthetic gene clusters encoding epoxyketone proteasome inhibitors, leading to the isolation and characterization of seven epoxyketone natural products, including compounds with unique warhead structures. We believe this approach is applicable to any conserved biosynthetic gene and provides a higher-throughput cost-effective alternative to whole genome sequencing discovery methods.

Author contributions: J.G.O. and S.F.B. designed research; J.G.O., Z.C.-P., A.G.S., M.A.T., P.Y.C., and D.M. performed research; J.G.O., Z.C.-P., and B.V.B.R. analyzed data; and J.G.O. and S.F.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The gene cluster sequences reported in this paper have been deposited in GenBank (accession nos. KP830089–KP830097), and the natural product sequence tag data have been deposited in the BioProject database, ncbi.nlm.nih.gov/bioproject (accession no. PRJNA258222) and at esnapd2.rockefeller.edu.

[1]To whom correspondence should be addressed. Email: sbrady@rockefeller.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1501124112/-/DCSupplemental.

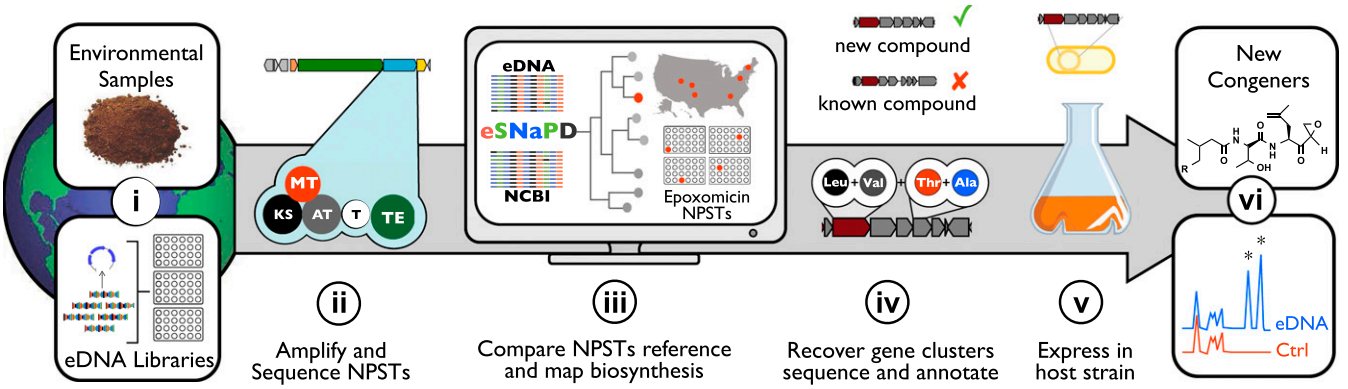


**Fig. 1.** Congener discovery using eSNaPD. (*i*) eDNA is extracted from samples collected around the globe; these can be archived as large insert libraries if desired. (*ii*) NPSTs are then generated by sequencing PCR amplicons amplified from eDNA templates with degenerate primers that target conserved biosynthetic motifs. (*iii*) Analysis of NPST data using eSNaPD identifies NPSTs that derive from biosynthetic gene clusters of interest; these are mapped to collection locations or positions within arrayed libraries using position information incorporated in the PCR primers. (*iv*) Biosynthetic gene clusters of interest are then recovered from arrayed libraries and sequenced. Bioinformatics analysis of annotated eDNA gene clusters is then used to prioritize clusters for heterologous expression studies. (*v*) Prioritized gene clusters are transferred to a laboratory-friendly host for heterologous expression. (*vi*) LCMS and/or biological activity profiles of strains harboring eDNA clusters are compared with a vector control strain to identify new metabolites for purification, structure elucidation, and bioactivity studies.

for multiple myeloma (10). We hypothesized that within the global metagenome, there were likely many undiscovered EPI BGCs that would provide a means of expanding this underexplored family of medically relevant natural products.

## Results and Discussion

**Generation and Archiving of NPST Data Sets as a Resource for Natural Product Discovery.** The starting point in our search for new EPIs was an archived collection of NPST data from hundreds of geographically diverse metagenomes and metagenomic libraries (11), which we have now compiled as an open access resource (BioProject accession no. PRJNA258222, in-house server esnapd2.rockefeller.edu). Our NPST data comprise ~$1 \times 10^6$ unique environmental sequences that were amplified from soil metagenomes using degenerate primers targeting two of the most common biosynthetic motifs: nonribosomal peptide synthetase (NRPS) adenylation (A) domains and polyketide synthase (PKS) ketosynthase (KS) domains (12). By targeting these very common biosynthetic domains, sequencing resources are focused on generating only data that are relevant to our search strategy, therefore the raw sequencing power required to generate this dataset was quite modest (~1.5 Gbps) (8, 11). We estimate that the diversity of biosynthetic pathways represented in our NPST dataset is at least 50× larger than the NRPS and PKS pathways contained in all publically available sequenced bacterial genomes, as judged by the number of equivalent domains identified by recent systematic analyses (13–15).

**Identifying EPI Biosynthetic Pathways Within 185 Globally Distributed Soil Metagenomes.** Recent cloning and sequencing of the BGCs encoding epoxomicin (epx) and eponemycin (epn) provided us with the reference sequences required to carry out a survey of our NPST data. Of particular interest as a potential NPST target was the lone KS domain that is essential for biosynthesis of the conserved epoxyketone warhead of epx and epn (16). Using this sequence as a reference, we carried out an eSNaPD survey of archived KS domain NPST data from 185 global soil metagenomes. Even a very conservative estimate of 500 different bacterial species per soil sample would suggest that within these 185 samples we have likely surveyed tens of thousands of unique bacterial genomes. This search identified 99 unique EPI-like sequences that grouped into six distinct clades (Fig. 2*A*). These unique sequences mapped to 53 of the 185 metagenomes analyzed, with as many as 7 different EPI related tags present in a single soil metagenome (Fig. 2 and *SI Appendix*, Table S12). Of the 99 unique EPI tag sequences identified, less than one-quarter (19/99) were found in more than one soil metagenome, suggesting that we have still only identified a fraction of EPI biosynthetic diversity contained within the global microbiome. Remarkably, none of the EPI hits identified in our screen is found (at ≥95% amino acid identity, BlastX) among 2,771 complete sequenced microbial genomes deposited in National Center for Biotechnology Information; in fact, the majority of our EPI hits did not have any relatives in these genomes with ≥65% amino acid identity (*SI Appendix*, Fig. S40).

**Recovery, Sequencing, and in Silico Analysis of Epoxyketone Biosynthetic Pathways from Arrayed Metagenomic Libraries.** Having identified a large number of soils that might serve as productive starting points for proteasome inhibitor discovery studies, we next sought to recover a collection of complete EPI BGCs. Three of the 53 EPI containing metagenomes we identified existed as archived large insert (~40 Kb) cosmid libraries in our laboratory, and a fourth library was constructed from a soil sample collected in Southern Arizona that showed elevated biosynthetic diversity. Each of these libraries contains ~10 million unique cosmids partially arrayed as 384 wells containing ~25,000 clones each. During library construction, barcoded KS and A domain PCR amplicons were prepared and sequenced from each of the 384 wells in all 4 libraries. The resulting NPST data were then analyzed using eSNaPD, and NPSTs related to biomedically relevant gene cluster families were mapped back to individual library wells. This information allowed us to identify 11 putative EPI BGCs in our libraries and isolate them using serial dilution PCR. Many of these BGCs spanned more than one cosmid clone. A convenient feature of the computational framework is the ability to identify overlapping clones that allow reconstruction of complete pathways by targeting multiple library wells containing the same NPST. The strategy of partially arraying libraries and generating barcoded NPSTs from each library well allows efficient storage and automated in silico screening of cloned metagenomes for diverse biomedically relevant BGCs, as well as facile recovery of entire BGCs identified in computational screens of NPST data (8, 9).

All recovered clones were sequenced, and ORFs identified using MetaGeneMark (17). Sequences were then annotated

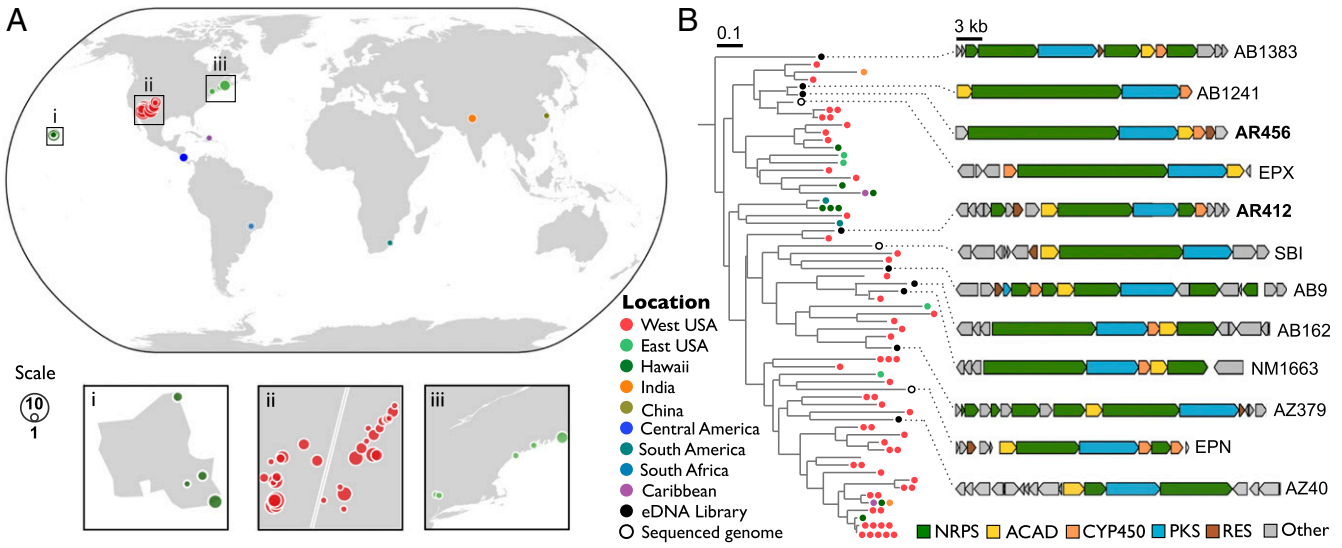


**Fig. 2.** Global survey for EPI biosynthetic gene clusters. (*A*) Results from a survey of 185 metagenomes for EPI biosynthetic gene clusters. Locations of soil metagenomes predicted to contain at least one EPI biosynthetic gene cluster are shown. Bubbles are color coded according to general geographic location around the globe (see location key). The size of each bubble indicates the number of unique EPI sequence tags identified at that location (see scale). (*Insets*) The three boxes expand three densely sampled US regions: (*i*) Hawaii, (*ii*) Arizona, and New Mexico, (*iii*) New York and New England. (*B*) Phylogeny of EPI NPSTs and representative biosynthetic pathways recovered from metagenome libraries: Annotated BGCs recovered from four metagenomic libraries are illustrated. Dashed lines indicate the original NPST sequence that led to the identification and recovery of each BGC. The phylogenetic tree presented was constructed from a pairwise alignment of a representative subset of globally distributed EPI NPST sequences. Each leaf on this tree represents a unique NPST. The colored dots indicate the number of sample sites found to contain a particular NPST. Colors correspond to geographic location around the globe (see location key). The key on the bottom right indicates predicted gene function. ACAD, Acyl-CoA dehydrogenase; CYP450, cytochrome P450; EPN, eponemycin gene cluster; EPX, epoxomicin gene cluster; RES, resistance element; SBI, orphan EPI cluster previously identified in the genome of *Streptomyces bingchenggensis* (16). Detailed annotations for each pathway are presented in *SI Appendix*, Tables S1–S9.

using AntiSMASH (6), as well as manual interrogation of individual ORFs by Blast and conserved domain searches (18). Detailed in silico analysis of the recovered BGCs revealed that the vast majority (9/11) appeared to encode an EPI, possessing biosynthetic enzymes for an *N*-acylated hydrophobic peptide backbone with C-terminal epoxyketone warhead moiety (Figs. 2*B* and 3*A* and *SI Appendix*, Tables S1–S9). The high proportion of on-target pathways we recovered shows that whereas NPST tag screening is not perfect, it does provide a remarkably reliable means for predicting BGC gene content and chemical output using a limited amount of sequence data. Whereas in silico analytical techniques are still not capable of predicting the final chemical output of a BGC, they do allow for dereplication by comparison with known biosynthetic pathways, and identification of noteworthy biosynthetic features. Such predictions are an integral step in sequence-guided natural product discovery, as they allow for prioritization of BGCs before labor-intensive heterologous expression, isolation, and structural elucidation experiments.

**Heterologous Expression of Metagenome-Derived EPI Biosynthetic Pathways.** Based on our in silico analysis we selected two biosynthetic pathways for heterologous expression studies. The first gene cluster, AR456, was selected with the hope of expanding the chemical diversity around epx, the most potent naturally occurring EPI discovered to date (10). The AR456 pathway has the same number and linear arrangement of NRPS and PKS modules as the epx cluster (Fig. 3*A* and *SI Appendix*, Table S1); however, the binding pocket of the second and fourth A domains is altered, suggesting AR456 would encode an EPI with a novel peptide backbone. The second pathway we targeted for heterologous expression, AR412, bore little resemblance to either characterized biosynthetic system (Fig. 3*A* and *SI Appendix*, Table S2) and was chosen with the hope of generating an EPI that differed significantly from either epx or epn.

Cosmids containing the complete AR412 and AR456 BGCs, as well as an empty vector control, were separately conjugated into a panel of five *Streptomyces* host strains, and small-scale heterologous expression test cultures of the resulting exconjugates were established in R5a, SMM, and ISP4 media (*SI Appendix*, section S3). Assays run on crude organic extracts derived from each culture revealed potent 20S proteasome inhibitory activity when *Streptomyces albus* J1074 transformed with either AR412 or AR456 was grown in any of the three media. Extracts from the remaining media–strain combinations for both pathways were inactive, indicating that environmental DNA (eDNA) specific metabolites were being produced only when the pathways were hosted in *S. albus* and that media choice was not important. This result adds to an accumulating body of evidence suggesting that *S. albus* J1074 is a gifted host for heterologous expression of natural product BGCs, and as such represents a good first choice for heterologous expression studies (19, 20).

**Isolation and Characterization of EPIs Encoded by Metagenome-Derived Biosynthetic Pathways.** LC (liquid chromatography)/MS analysis of active extracts confirmed the presence of clone-specific metabolites for both *S. albus:AR412* and *S. albus:AR456* (Fig. 3*B*). Bioassay guided fractionation of ethyl acetate extracts from large-scale (8L) cultures led to the isolation of five 20S inhibitory compounds (clarepoxcins A–D, **1**–**4**) from *S. albus:AR456* and two 20S inhibitory compounds (landepoxcins A and B, **6**–**7**) from *S. albus:AR412*. An additional clone-specific compound with a predicted molecular formula indicating the presence of one chlorine atom (clarepoxcin E, **5**) was purified from *S. albus:AR456* using LC/MS guided fractionation.

The structure of clarepoxcin A (**1**) was elucidated using a combination of high-resolution electrospray ionization mass spectrometry (HRESIMS) and 1D and 2D NMR data. The HRESIMS spectrum of **1** displayed a pseudomolecular ion peak at 625.4194 $[M\text{-}H]^-$, consistent with a molecular formula of $C_{32}H_{58}N_4O_8$.

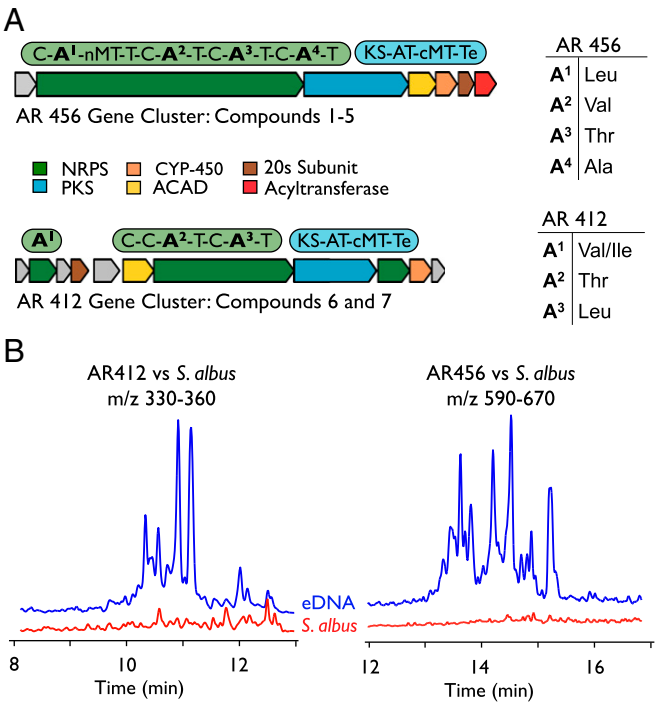


**Fig. 3.** Heterologous expression of two eDNA derived EPI biosynthetic gene clusters. (*A*) Details of the AR456 and AR412 biosynthetic gene clusters. Domain arrangement for each PKS and NRPS gene in the AR456 and AR412 biosynthetic gene clusters is shown. (*Inset*) The tables indicate the substrate specificity for each A domain, as deduced by subsequent structure elucidation. Detailed annotation of each pathway is given in *SI Appendix*, Tables S1 and S2. (*B*) Heterologous expression of selected EPI biosynthetic pathways: LC/MS chromatograms of crude extracts from *S. albus*:AR456 and *S. albus*:AR412 cultures are compared with an extract from an empty vector control culture.

The $^{1}$H NMR spectrum of **1** has a signal distribution consistent with an acylated peptide, containing signals in regions for alpha- and beta-carbon protons, as well as additional methyl and methylene signals and oxygenated methine and methylene signals. Examination of the correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY), and heteronuclear multiple bond correlation spectroscopy (HMBC) spectra of **1** established the presence of five amino acid substructures that were linked by HMBC correlations to give the peptide backbone of **1** (Fig. 4*A* and *SI Appendix*, Table S10, section S1, and Fig. S1). Each amino acid is predicted to be in the L-configuration based on a bioinformatics analysis of the individual NRPS modules in the AR456 biosynthetic pathway (*SI Appendix*, Table S1). Based on the partial structures outlined above and the molecular formula determined by HRESIMS, the *N*-acyl substituent on **1** was predicted to be a 10-carbon fatty acid containing a terminal hydroxyl group. The final structure of this fatty acid is defined by TOCSY, COSY, and HMBC correlations (Fig. 4*A*). A detailed description of structural elucidation for clarepoxcins A–E (**1–5**) is given in *SI Appendix*, Table S10, section S1, and Fig. S1.

The structure of landepoxcin A (**6**) was also elucidated using a combination of HRESIMS and 1D and 2D NMR data. The HRESIMS spectrum of **6** displayed a pseudomolecular ion peak at 339.1934 $[M-H]^{-}$, consistent with a molecular formula of $C_{17}H_{28}N_2O_5$.The $^{1}$H NMR spectrum of **6** has a signal distribution consistent with a small peptide. It contains shifts in the expected region for alpha- and beta-carbon protons, a number of methyl and methylene signals as well as oxygenated methine and methylene signals. Examination of the TOCSY, COSY, and HMBC spectra of **6** established the presence of the four substructures shown in Fig. 4*B*. These individual substructures were then linked by HMBC correlations to give the final structure of landepoxcin A (**6**). The amino acids in landepoxcin A are predicted to be in the L-configuration based on a bioinformatics analysis of the individual NRPS modules in the AR412 biosynthetic pathway (*SI Appendix*, Table S2). A complete description of the structural elucidation of landepoxcins A and B (**6** and **7**) is given in *SI Appendix*, Table S11, section S2, and Fig. S2.

With the exception of clarepoxcin E (**5**), each of the isolated compounds possesses the expected epoxyketone pharmacophore attached to a hydrophobic peptide backbone; however, the structures of the clarepoxcins and landepoxcins differ from any previously described EPI. The deduced structures of clarepoxcins A–D (**1–4**, Fig. 5*A*) correlate well with in silico NRPS–PKS predictions (Fig. 3*A* and *SI Appendix*, Table S1). They each contain a tetrapeptide backbone (*N*-methyl-L-leucine, L-valine, L-threonine, L-alanine) that is appended at the N terminus with a long-chain

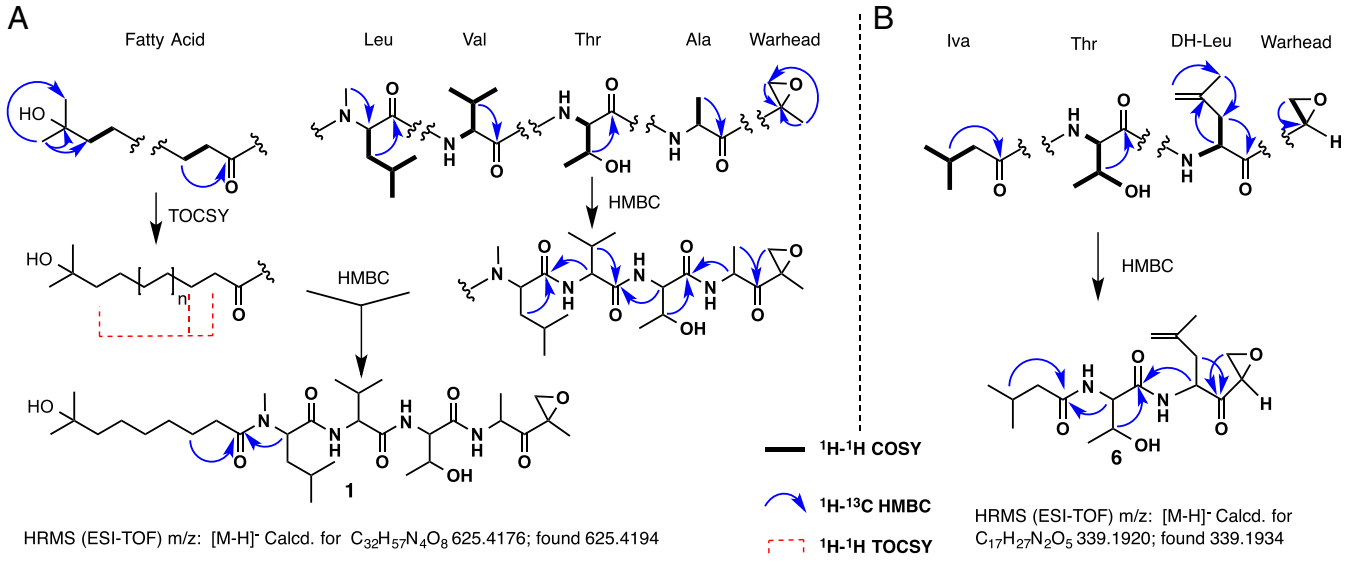


**Fig. 4.** Clarepoxcin and landepoxcin structure elucidation. (*A*) Key 2D NMR correlations used to define the structure of clarepoxcin A. (*B*) Key 2D NMR correlations used to define the structure of landepoxcin A.

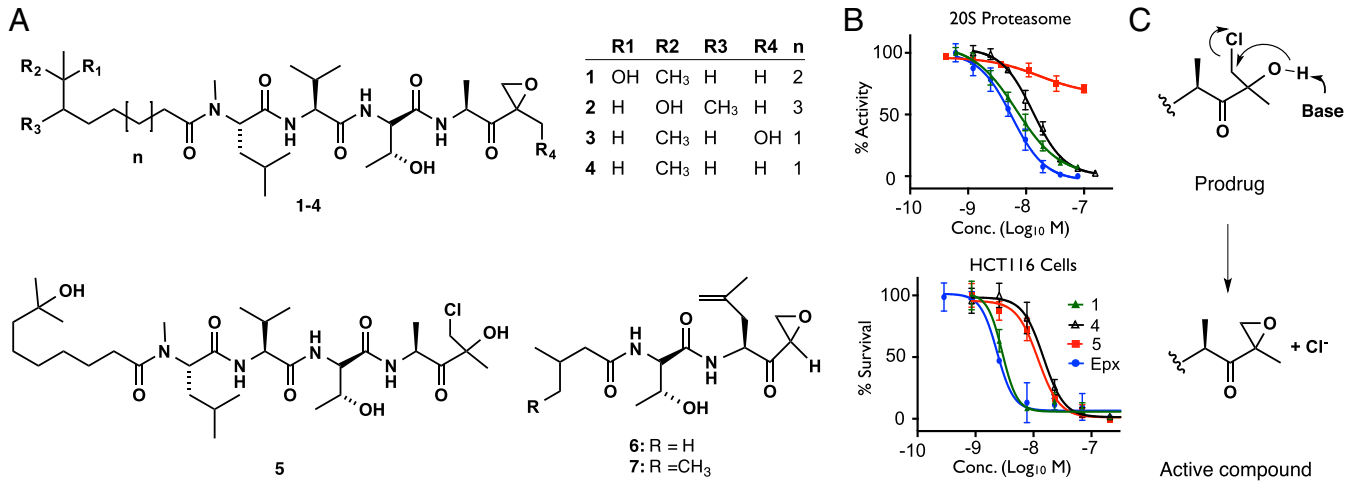


**Fig. 5.** Deduced structures and biological activities of eDNA encoded EPIs. (*A*) The elucidated structures for compounds arising from heterologous expression of the AR456 (**1**–**5**) and AR412 (**6**–**7**) biosynthetic gene clusters in *S. albus*. (*B*) Activity of eDNA encoded EPIs: $IC_{50}$ curves for selected compounds against purified human 20s proteasome (chymotrypsin-like activity) and human colon carcinoma (HCT116) cells are shown. epx, epoxomicin. Data points are an average of three independent replicates ± SD. (*C*) Proposed mechanism for activation of the halohydrin moiety **5** to give an epoxyketone warhead.

fatty acid and at the C terminus with the epoxyketone warhead. The variable long-chain fatty acid found in clarepoxcins A–D replaces the N-terminal acetate found in epx. We predict that this moiety is incorporated into the peptide by the action of an acyltransferase-type nitrilase found within the AR456 but not the epx BGC (Fig. 3*A* and *SI Appendix*, Table S1) (16, 21). The structures of landepoxcins A and B (**6** and **7**; Fig. 5*A*) are also generally congruent with in silico NRPS/PKS predictions (Fig. 3*A* and *SI Appendix*, Table S2). They each contain a dipeptide backbone (L-threonine, 4,5-dehydro-L-leucine) appended with short-chain branched acyl group at the N terminus and an epoxyketone warhead at the C terminus. Interestingly, a tertiary epoxide moiety appears in place of the expected quaternary epoxide. This structural feature was recently described in a natural product modulator of TGF-β (22), but has not previously been observed in a naturally occurring EPI. We propose that the tertiary epoxide warhead of the landepoxcins arises due to the addition of one rather than two methyl groups by the methyltransferase domain of the AR412 PKS enzyme (16). Clarepoxcin E (**5**) has an identical peptide backbone and N-terminal fatty acid to those seen in clarepoxcin A (**1**), however the epoxide moiety of **5** is replaced with a halohdyrin structure (Fig. 5*A*). In our examination of the AR456 gene cluster (*SI Appendix*, Table S1) we could not identify an obvious halogenase candidate. Detailed biosynthetically analysis will be required to determine whether the halohdyrin arises spontaneously during fermentation or through the action of enzyme encoded the AR456 gene cluster.

Compounds **1**–**4**, as well as compounds **6** and **7**, are nanomolar inhibitors of purified human 20S proteasome and possess potent cytotoxicity against cultured human cells (Table 1 and Fig. 5*B*). Compounds **1** and **2** are particularly potent, with low nanomolar $IC_{50s}$ against both purified 20S proteasome and human cells. In the case of clarepoxcins A–D we have identified, to our knowledge, the first example of naturally occurring EPIs that contain an amino acid other than leucine or dehydoleucine directly adjacent to the epoxyketone warhead. This is of particular interest given that the amino acid side chain at this position (the P1 position) is believed to be the primary determinant of inhibitor affinity for the proteasome (23). In the landepoxcins, we have identified EPIs with an unsubstituted epoxyketone warhead moiety not seen in any previously described natural or synthetic EPI. As expected, owing to the absence of an epoxyketone warhead, clarepoxcin E (**5**) lacked potent 20S inhibitory activity (Table 1 and Fig. 5*B*). Surprisingly, **5** retained cytotoxicity against HCT-116 cells; we believe that this compound is likely acting as a halohydrin prodrug that is activated either enzymatically or spontaneously in cells to yield the bioactive epoxyketone warhead (Fig. 5*C*). This supposition is supported by the recent description of a series of synthetic prodrugs synthesized by Onyx pharmaceuticals (24) that includes halohydrin structures that are believed to be activated in vivo as outlined in Fig. 5*C*.

## Conclusions

The genomic diversity captured in even a single soil metagenome library is huge, with upward of 10,000 unique species represented in some cases (3, 4). This diversity means that by constructing just a small number of metagenomic libraries, a single laboratory can access an enormous breadth of biosynthetic diversity. Harnessing this diversity for the productive discovery of biologically active small molecules, however, requires an efficient means for identifying and recovering biosynthetic pathways encoding specific target molecules of interest. The sequence-driven discovery pipeline we used here provides a highly effective means for identifying biosynthetic gene clusters encoding previously unidentified natural product congeners, and allows targeted screening of not just one but hundreds of soil metagenomes in parallel. When coupled to standardized metagenomic techniques for library construction, arraying, storage, and clone recovery it provides a simple and direct means for accessing the biosynthetic potential of the global

**Table 1. Biological activity of the clarepoxcins and landepoxcins**

| Compound | HCT- 116 cells | 20s proteasome |
|---|---|---|
| **1** | 2.9 ± 0.2 | 6.9 ± 1.8 |
| **2** | 5.4 ± 3.2 | 9.7 ± 1.7 |
| **3** | 13.7 ± 3.0 | 8.5 ± 1.4 |
| **4** | 17.4 ± 0.8 | 15.1 ± 3.7 |
| **5** | 11.0 ± 3.0 | >1,000 |
| **6** | 34 ± 5.7 | 218 ± 64 |
| **7** | 180 ± 34 | 309 ± 51 |
| epx | 2.3 ± 0.4 | 6.1 ± 1.2 |

$IC_{50}$ values (nM) for eDNA encoded compounds (**1**–**7**) and epx against human 20s proteasome (chymotrypsin-like activity) and HCT 116 cells. Values presented are the average of 3–6 independent replicates ± SD.

microbiome. In the present study we have examined 185 globally distributed soil metagenomes for EPI biosynthetic pathways, leading to the discovery of nine new EPI gene clusters and seven new EPI congeners. We anticipate that the informatics and experimental methods used to achieve this result will be widely applicable to the targeted expansion of a large number of bacterial natural product families, using both metagenomic libraries and large bacterial culture collections.

## Materials and Methods

**Identification and Analysis of Epoxyketone Markers in Soil and Library NPST Data.** The KS domains from three known EPI pathways (known EPIs) were added to existing eSNaPD reference data, and NPSTs whose closest relative among all sequenced KS domains was one of these three known EPIs were identified and mapped to soil locations and/or library wells using eSNaPD as previously described (8). Briefly, reads with any ambiguous calls and those <200 base pairs (bp) in length were removed. All remaining reads were trimmed to ≤400 bp, and then clustered at 95% identity using UCLUST (25) to generate a unique consensus (i.e., NPST). Location information for each NPST was derived from the 8-bp primer barcodes found in each read comprising the 95% identity cluster, and used to map NPSTs back to soil collection locations and/or library wells. NPSTs were then searched using BlastN against a manually curated database of KS domain sequences. NPSTs that had no Blast matches with an e value of $10^{-50}$ or less were discarded, and NPSTs that returned one of the three known EPIs as a top match were considered hits (EPI markers). The resulting set of unique EPI marker sequences was used to generate geographic and phylogenetic distribution figures. A pairwise alignment of all EPI marker sequences was generated using MUSCLE (26), and the resulting alignment file used to generate a maximum likelihood tree with FastTree (27), which was visualized using phyloseq. (28).

**Recovery of Biosynthetic Gene Clusters from eDNA Libraries.** Epoxyketone marker sequences identified within metagenomic libraries were automatically assigned to library wells by the barcode parsing functionality of the eSNaPD software package as described above. Specific primers targeting each unique EPI marker sequence were designed using BatchPrimer3 (29). To recover single clones from library wells, a serial dilution PCR strategy was used as follows: Library wells containing targets as one of ∼25,000 unique cosmids were grown overnight to confluence and diluted to a concentration of $\sim 4 \times 10^3$ CFU/mL as judged by $OD_{600}$. Then, 384 well plates were inoculated with 50 μL (200 CFU) of the resulting dilution per well, grown to confluence, and screened using real-time PCR, to identify wells containing target clones as 1 of ∼200 clones. Target positive wells were then diluted to a concentration of ∼100 CFU/mL and the process repeated to identify wells containing targets as 1 of ∼5 clones. Five clone pools were then plated on solid medium, and target clones identified by colony PCR.

**Isolation of Clarepoxcins A–E and Landepoxcins A–B.** For each recombinant strain 10 μL of a spore suspension ($\sim 2 \times 10^9$ CFU/mL) was used to establish seed cultures in 50 mL trypticase soy broth. These cultures were grown for 48 h (30 °C/200 rpm) and 5 mL of the resulting confluent culture was used to inoculate production cultures containing 1 L SMM in 2.8-L baffled flasks with 30 g Diaion HP-20 resin. After 7 d (30 °C, 200 rpm), combined resin from 8 L of culture (240 g) was collected by filtration, washed with water to remove mycelia, and air dried at room temperature. Bound material was then eluted from the washed resin with methanol ($2 \times 500$ mL) and bought to dryness by rotary evaporation. Dried extracts were suspended in 500 mL ultrapure water by sonication, and the resulting suspension was extracted twice with 1 L of ethyl acetate. Combined ethyl acetate extracts were then separated by medium-pressure chromatography (Teledyne Isco Combiflash Rf150). For extracts containing **1**–**5** (from *S. albus:*AR456 cultures), normal-phase (12 g, silica resin) chromatography was performed using a linear gradient of hexanes–ethyl acetate from 10% to 100% over 30 min and a flow rate of 30 mL $min^{-1}$. In the case of extracts containing **6** and **7** (from *S. albus:*AR412), reversed-phase chromatography (5.5 g, $C_{18}$ resin) was performed using a linear gradient from 10% to 50% acetonitrile over 30 min with a flow rate of 18 mL $min^{-1}$. To identify column fractions containing active compound, aliquots from sequential groups of five 10-mL fractions were assessed for 20S proteasome inhibitory activity. Active fractions were then analyzed by LC/MS, and like fractions were pooled. **1**–**7** were purified using two rounds of preparative HPLC ($C_{18}$, $10 \times 150$ mM, 5 μM) as follows: **1** and **5** were purified using a linear gradient from 10% to 50% acetonitrile over 60 min at a flow rate of 3 mL $min^{-1}$ and had a retention time of 51 and 53 min, respectively; **2** and **3** were purified using a linear gradient from 15% to 53% acetonitrile over 120 min, with a flow rate of 3 mL $min^{-1}$and had retention times of 108 and 83 min, respectively; **4** was purified using a linear gradient from 15% to 80% ACN over 45 min and had a retention time of 38 min; **6** was purified using isocratic elution (18% acetonitrile) with a flow rate of 7.1 mL $min^{-1}$and had a retention time of 14.5 min; **7** was purified using isocratic elution (22.5% acetonitrile) with a flow rate of 7.1 mL $min^{-1}$ and had a retention time of 15.1 min.

**ACKNOWLEDGMENTS.** This work was supported by NIH Grant GM077516. Z.C.-P. was supported by NIH Grant F32 AI1100029. S.F.B is a Howard Hughes Medical Institute Early Career Scientist.

1. Baltz RH (2008) Renaissance in antibacterial discovery from actinomycetes. *Curr Opin Pharmacol* 8(5):557–563.
2. Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat Chem Biol* 1(5):265–269.
3. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308(5721):554–557.
4. Daniel R (2005) The metagenomics of soil. *Nat Rev Microbiol* 3(6):470–478.
5. Starcevic A, et al. (2008) ClustScan: An integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res* 36(21):6882–6892.
6. Medema MH, et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web Server issue):W339–W346.
7. Boddy CN (2014) Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. *J Ind Microbiol Biotechnol* 41(2):443–450.
8. Owen JG, et al. (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci USA* 110(29):11797–11802.
9. Reddy BV, Milshteyn A, Charlop-Powers Z, Brady SF (2014) eSNaPD: A versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chem Biol* 21(8):1023–1033.
10. Kisselev AF, van der Linden WA, Overkleeft HS (2012) Proteasome inhibitors: An expanding army attacking a unique target. *Chem Biol* 19(1):99–115.
11. Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Brady SF (2014) Chemical-biogeographic survey of secondary metabolism in soil. *Proc Natl Acad Sci USA* 111(10): 3757–3762.
12. Ayuso-Sacido A, Genilloud O (2005) New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: Detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microb Ecol* 49(1):10–24.
13. Doroghazi JR, et al. (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 10(11):963–968.
14. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K (2014) Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of non-modular enzymes. *Proc Natl Acad Sci USA* 111(25):9259–9264.
15. Zaburannyi N, Rabyk M, Ostash B, Fedorenko V, Luzhetskyy A (2014) Insights into naturally minimised Streptomyces albus J1074 genome. *BMC Genomics* 15(1):97.
16. Schorn M, et al. (2014) Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem Biol* 9(1):301–309.
17. Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12):e132.
18. NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 42(Database issue):D7–D17.
19. Feng Z, Kallifidas D, Brady SF (2011) Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proc Natl Acad Sci USA* 108(31):12629–12634.
20. Yamanaka K, et al. (2014) Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci USA* 111(5):1957–1962.
21. Pace HC, Brenner C (2001) The nitrilase superfamily: Classification, structure and function. *Genome Biol* 2(1):reviews0001.
22. Tsunematsu Y, et al. (2015) Isolation, structure elucidation, and total synthesis of tryptopeptins A and B, new TGF-beta signaling modulators from *Streptomyces* sp. *Org Lett* 17(2):258–261.
23. Bogyo M, Shin S, McMaster JS, Ploegh HL (1998) Substrate binding and sequence preference of the proteasome revealed by active-site-directed affinity probes. *Chem Biol* 5(6):307–320.
24. Phiasivongsa P, Luehr G, Peng G, By K, Anik ST (2013) Prodrugs of peptide epoxy ketone protease inhibitors. US patent application 13/938,075 (July 9, 2013).
25. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
26. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
27. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641–1650.
28. McMurdie PJ, Holmes S (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8(4):e61217.
29. You FM, et al. (2008) BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253.