Chemistry & Biology
Resource

eSNaPD: A Versatile, Web-Based Bioinformatics Platform for Surveying and Mining Natural Product Biosynthetic Diversity from Metagenomes

Boojala Vijay B. Reddy,[1,3] Aleksandr Milshteyn,[1,3] Zachary Charlop-Powers,[1,2] and Sean F. Brady[1,2,*]
[1]Laboratory of Genetically Encoded Small Molecules
[2]Howard Hughes Medical Institute
Rockefeller University, 1230 York Avenue, New York, NY 10065, USA
[3]Co-first author
*Correspondence: sbrady@rockefeller.edu
http://dx.doi.org/10.1016/j.chembiol.2014.06.007

SUMMARY

Environmental Surveyor of Natural Product Diversity (eSNaPD) is a web-based bioinformatics and data aggregation platform that aids in the discovery of gene clusters encoding both novel natural products and new congeners of medicinally relevant natural products using (meta)genomic sequence data. Using PCR-generated sequence tags, the eSNaPD data-analysis pipeline profiles biosynthetic diversity hidden within (meta)genomes by comparing sequence tags to a reference data set of characterized gene clusters. Sample mapping, molecule discovery, library mapping, and new clade visualization modules facilitate the interrogation of large (meta)genomic sequence data sets for diverse downstream analyses, including, but not limited to, the identification of environments rich in untapped biosynthetic diversity, targeted molecule discovery efforts, and chemical ecology studies. eSNaPD is designed to generate a global atlas of biosynthetic diversity that can facilitate a systematic, sequence-based interrogation of nature's biosynthetic potential.

INTRODUCTION

Many important therapeutic agents currently in use have been isolated from cultured bacteria. It is clear, however, that traditional, culture-dependent methods for small-molecule drug discovery have been able to access only a small fraction of bacterial biosynthetic diversity found in the environment (Rappé and Giovannoni, 2003; Torsvik et al., 1990). Furthermore, advances in sequencing technologies and a corresponding increase in the number of newly sequenced genomes have shown that even extensively studied, cultured bacteria contain an abundance of previously undetected, cryptic biosynthetic gene clusters (Bentley et al., 2002; Ikeda et al., 2003; NCBI, 2013). These findings have led to a renaissance in genome mining as a means of natural product drug discovery (Challis, 2008; Winter et al., 2011). Although much of the focus has been placed on accessing molecules from newly identified biosynthetic gene clusters found in bacteria housed in culture collections, culture-independent, or metagenomic, methods provide an alternative approach that can unlock access to a vast pool of biologically active small molecules encoded by environmental bacteria by bypassing the initial culturing step (Banik and Brady, 2010; Brady, 2007; Kampa et al., 2013; Li and Qin, 2005; MacNeil et al., 2001). When using metagenomic methods, researchers capture bacterial DNA directly from environmental samples and access the small molecules through the expression of biosynthetic pathways in easily cultured, model-heterologous hosts. Small-molecule discovery efforts from metagenomes, however, face a key challenge in the identification of biosynthetic gene clusters of interest from among a much larger pool of undesired DNA sequences.

Although shotgun-sequencing approaches have been useful for guiding the identification of biosynthetic targets in individual genomes (Bentley et al., 2002) and small, endosymbiont metagenomes (Donia et al., 2011; Kampa et al., 2013), their application to more complex metagenomes is very limited. The repetitive use of highly conserved biosynthetic domains and the fact that a typical soil metagenome may contain 10^4–10^5 unique bacterial species make it impractical to assemble large numbers of complete biosynthetic gene clusters from metagenomic sequence data (Pop, 2009; Rappé and Giovannoni, 2003; Torsvik et al., 1990). Fortunately, although the high degree of conservation seen in natural product biosynthetic genes makes the accurate assembly of metagenomic gene clusters difficult, it also allows for a substantial amount of information pertaining to the biosynthetic pathways present in a metagenome to be gleaned through the use of a PCR-based sequence-tag approach that targets these conserved genes (Udwary et al., 2007). Much of the biosynthetic diversity arises from a relatively small number of biosynthetic classes (e.g., nonribosomal peptide synthase [NRPS], polyketide synthase [PKS], isoprene, sugar, shikimic acid, alkaloid, and ribosomal peptide) related through the common use of highly conserved domains (Dewick, 2009). With the exception of rare cases of convergent evolution, gene clusters encoding structurally related metabolites are predicted to share a common ancestry and therefore exhibit high sequence identity among these conserved domains (Fischbach et al., 2008; Wilson et al., 2010). We have exploited this correlation to develop a general method for the functional classification of novel secondary metabolite biosynthetic gene clusters based

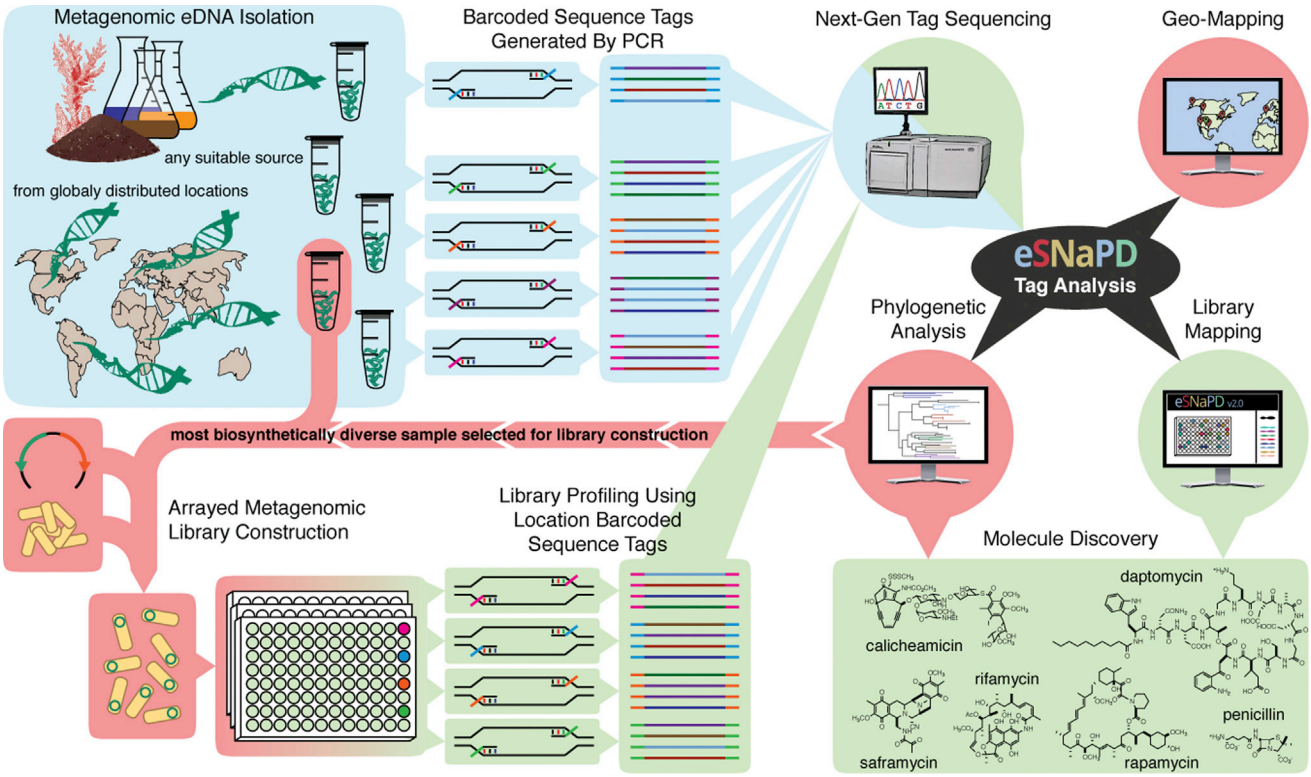

Figure 1. An Overview of the eSNaPD-Assisted Small-Molecule-Discovery Pipeline

Sequence tags are generated using PCR from (meta)genomic DNA isolated from virtually any source using degenerate primers that target conserved biosynthetic domains and can be bar-coded to differentiate DNA from different samples in a pooled next-generation sequencing run (blue). Raw amplicon sequencing data are processed using the eSNaPD bioinformatics platform, which automatically cleans the input data; classifies biosynthetic gene clusters present in the samples by performing a phylogenetic comparison to a reference set of characterized, known molecule gene clusters; and visualizes the results in a number of ways that aid in identifying the most desirable samples for library generation (black and red). The newly generated metagenomic library is arrayed to facilitate the identification and recovery of target clones, sequenced using position-specific bar-coded primers, and eSNaPD analysis is performed to generate a detailed biosynthetic profile of the library that facilitates the identification of high-value target clones for recovery and heterologous expression (green).

solely on the minimal amount of sequence contained in a PCR amplicon (Owen et al., 2013; Reddy et al., 2012). In this approach, conserved biosynthetic domains are PCR amplified from (meta)genomic DNA and the individual next-generation sequencing reads derived from these amplicons (termed natural product sequence tags or, simply, sequence tags) are used to establish the relationship between gene clusters present in a pool of metagenomic DNA and a reference set of functionally characterized gene clusters (Figure 1).

Here we present Environmental Surveyor of Natural Product Diversity (eSNaPD; http://esnapd2.rockefeller.edu/), a web-based bioinformatics platform for the automated analysis and organized aggregation of large metagenomic sequence-tag data sets. eSNaPD was designed to perform four main functions in a sequence-tag-based gene-cluster-discovery pipeline: (1) to streamline the process of prescreening metagenomic DNA samples for library construction in order to enable the prioritization of those samples containing the most overall biosynthetic diversity or the largest population of sequences related to a family of molecules of particular biomedical interest; (2) to enable high-throughput profiling of arrayed metagenomic libraries and facilitate the rapid identification and recovery of clones containing biosynthetic gene clusters of interest; (3) to profile and catalog global biosynthetic diversity in a systematic and meaningful way in order to gain a better understanding of the biosynthetic richness of different ecological environments; and (4) to identify and characterize novel secondary metabolite biosynthetic systems, not related to any others characterized to date, as a source for new classes of bacterially encoded bioactive small molecules. We have extensively validated the robustness and general utility of our bioinformatics approach in a number of studies, where it was used to identify metagenomic gene clusters that were found to encode novel biologically active secondary metabolites (Banik and Brady, 2008; Bauer et al., 2010; Chang and Brady, 2011, 2013; Chang et al., 2013; Feng et al., 2010; Kallifidas et al., 2012; Kang and Brady, 2013; Owen et al., 2013). In addition, we have used our biosynthetic gene-cluster-classification pipeline to survey diverse environmental metagenomes (Charlop-Powers et al., 2014; Reddy et al., 2012). Although our work has been geared primarily toward environmental metagenomic DNA (eDNA), libraries can be constructed just as easily from genomic DNA contained in culture collections or any other suitable source.

The aggregation of data on the eSNaPD server will allow the unclassified sequence tags to be revisited as the reference databases are updated with newly characterized gene clusters. As

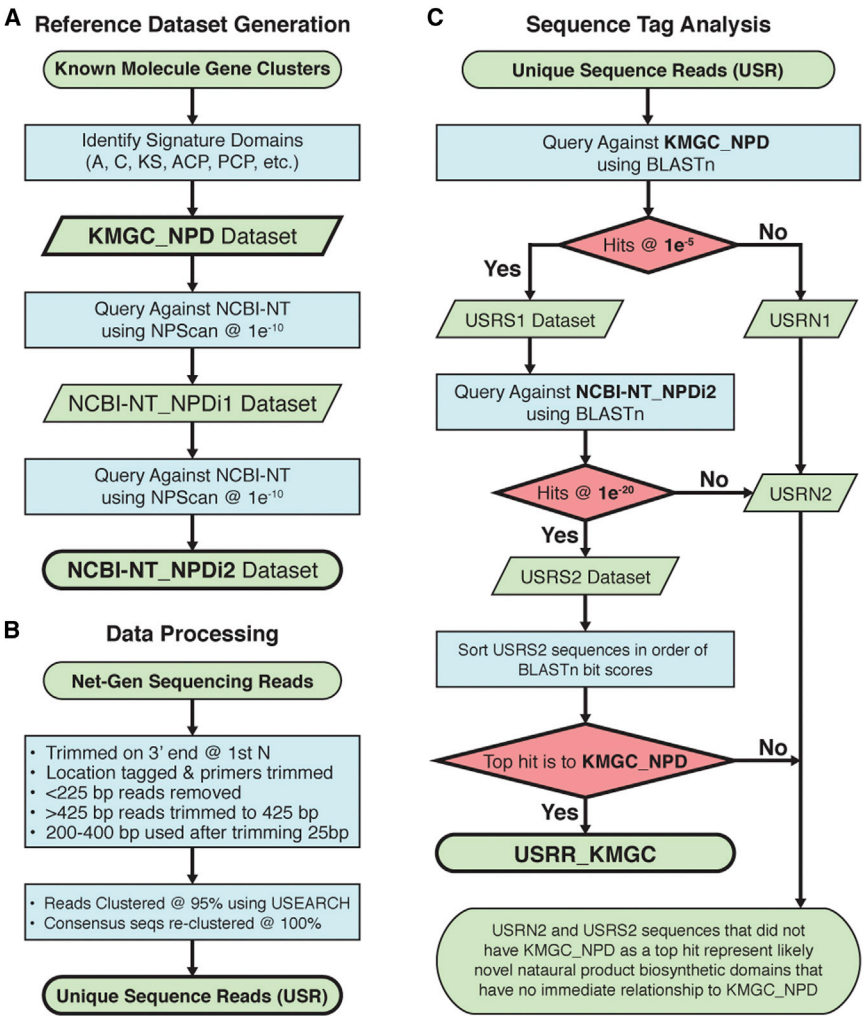

Figure 2. Overview of the eSNaPD Data-Analysis Pipeline

(A) Steps and data sets used for generating the reference natural product biosynthetic-domains data sets (NCBI-NT_NPDi2).

(B) Steps and data sets used for cleaning and clustering the sequence-tag next-generation sequencing reads to yield unique sequence reads (USRs).

(C) Steps and data sets used for classifying USR relationships to known molecule gene clusters.

the number of sequence-tag data sets of diverse origins in eSNaPD grows, this web-based platform will represent a robust mechanism for cataloging global biosynthetic diversity, identifying novel gene clusters that can be fed into natural-product-discovery pipelines, and comparing meta-analyses of secondary metabolomes.

RESULTS

eSNaPD uses a next-generation sequencing data file containing sequence-tag reads as input and carries out data processing and clustering of identical sequences, followed by a search against a curated reference database of functionally characterized natural product gene clusters (referred to here as known-molecule gene clusters [KMGCs]) and phylogenetic analysis (Figure 2). The results of the analysis are fed into a user interface (UI) designed for the easy visual exploration of large data sets of natural-product-associated amplicons. The data visualization modules of eSNaPD make it possible to (1) map and assess overall biosynthetic richness across microbiomes contained in geographically distinct samples; (2) systematically identify gene clusters that are likely to encode congeners of biomedically relevant families of natural products; and (3) identify biosynthetic systems that, through their associations with conserved biosynthetic domains that are not closely related to any sequences in the reference data set, are predicted to be good candidates for encoding the biosynthesis of novel natural product families—a function unique to eSNaPD. In addition, the use of bar-coded primers for arrayed metagenomic library sequencing enables eSNaPD to pinpoint the locations of sequence tags corresponding to each family of molecules in the arrayed libraries for rapid clone recovery. The displayed data can be filtered using a user-selectable e-value cutoff, and all output data are available in click-and-see architecture with downloadable CSV- and FASTA-formatted files to facilitate subsequent user-specific analyses.

At present, even though eSNaPD is populated only with data sets for the adenylation (AD) and ketoacylsynthase (KS) domains, the data-analysis pipeline permits the evaluation of amplicon reads from 12 NRPS/PKS domains: AD, acyl carrier protein (AC), acyltransferase (AT), condensation (CD), dehydratase (DH), epimerization (EP), enoylreductase (ER), ketoreductase (KR), KS, methyltransferase (MT), peptidyl carrier protein (PC), and thioesterase (TE). It is also designed for easy expansion to include any conserved biosynthetic domain in the analysis pipeline. In addition to sequence data, eSNaPD stores and maps sample details, such as global positioning system (GPS) coordinates, photographs of the collection site, soil type, and, when available, physicochemical parameters of the soil sample (e.g., moisture content, pH, organic matter content [loss on ignition, LOI], organic nitrogen content [potentially mineralizable nitrogen, PMN], minerals, and metals). As the eSNaPD database grows, these data will allow a variety of downstream, ecologically focused correlation analyses to be performed on the amplicon data sets.

Job Submission and Data Processing

eSNaPD accepts FASTA-formatted sequence-reads files from most sequencing technologies. A second, space-delimited text file, containing primer sequences used to generate the reads,

is required for processing the raw reads. For arrayed libraries or large sets of environmental samples, this file will contain primer sequences that are bar-coded using 8-nt lead sequences that identify each unique sample in the sequencing reaction. In the case of an arrayed library, the primer identification (ID) is appended with a serial number corresponding to the specific location in the library array, which is used to map the classified tags back to their positions in the library.

In the analysis pipeline, next-generation sequencing data from PCR-amplified signature domains are cleaned and tagged with location information, if needed for bar-coded arrayed libraries or when analyzing multiple environment samples simultaneously. Cleaned reads are clustered at 95% sequence identity and a consensus sequence is generated from each cluster with each such consensus sequence representing a unique sequence read (USR). USRs are then compared by the Nucleotide Basic Local Alignment Search Tool (BLASTn) to a curated reference database of similar regions from functionally characterized KMGCs and all closely related sequences that have been culled from the gene cluster data found in the NCBI-NT database (Figure 2). The eSNaPD reference database currently contains ~450 unique, functionally characterized gene clusters collected from publically available databases. It is designed to be easily updatable with additional functionally characterized gene clusters as they become available. The eSNaPD BLASTn analysis identifies USRs that are more closely related to a functionally characterized gene cluster than to any other sequence with an expectation value (e-value) of $< 1 \times 10^{-20}$. Although a high degree of sequence similarity between a sequence tag and the corresponding conserved biosynthetic domain from a characterized gene cluster is often indicative of the two gene clusters encoding molecules in the same structural family, the final structure of a natural product is determined by the set of tailoring enzymes present in the gene cluster. Thus, when a sequence tag clades with, but is not identical to, a reference sequence, it is likely to be indicative of a gene cluster that encodes a novel congener of the metabolite encoded by the reference sequence. The user can then define the e-value cutoff for most displayed data in the range from 1×10^{-20} to 1×10^{-80}. In our experience, at low e-values ($<1 \times 10^{-40}$) this analysis has proven to be a robust indicator that the corresponding USRs represent a biosynthetic cluster that belongs to the same family of molecules as the matching KMGC (Banik and Brady, 2008; Bauer et al., 2010; Chang and Brady, 2011, 2013; Chang et al., 2013; Kang and Brady, 2013; Owen et al., 2013). This is an empirical observation based on our extensive experience, predominantly with the AD and KS domains. However, because different domains undoubtedly evolve at different rates and the e-values are dependent on multiple parameters, including the size of the data set and the length of the sequence, we provide the user with the option to adjust the expectation values for displayed output as deemed appropriate, based on user's experience. Our eSNaPD bioinformatics pipeline can also be used to process shotgun-sequencing data. However, the utility of random shotgun-sequencing data is limited relative to the PCR-targeted sequencing data sets, which are designed such that the sequencing effort is focused entirely on the most bioinformatically informative natural product sequence tags.

eSNaPD Output

The eSNaPD analysis pipeline automatically generates several types of output. These include sample-specific, molecule-specific, and arrayed-library-specific data. For the sample-specific output, the GPS coordinates of the sample collection site are marked on a map and a photograph of the collection site, the physicochemical soil parameters, and the list of molecule families identified in the sample are linked to the map marker (Figure 3). Molecule-specific output is made up of phylogenetic dendrograms constructed for each identified molecule family on an individual data-set level as well as across all data contained in eSNaPD, using KITSCH (Felsenstein, 1993) (Figure 4C). The distributions of USRs related to each KMGC are also calculated across all samples (Figure 4D). Arrayed libraries are treated as any other metagenomic sample as far as sequence-tag classification goes, but the resulting output is also mapped onto a clickable graphical map of the library array, which enables the easy location of sequence tags corresponding to specific molecule families for clone recovery and provides detailed USR-specific information (Figure 5B). For all samples and libraries, sequence tags from within each data set that form clades lacking a close phylogenetic relationship to any KMGC are also identified and compiled. All data can be navigated via four display modules: (1) Map Explorer, (2) Molecule Explorer, (3) Arrayed Library Explorer, and (4) New Clades Explorer.

Map Explorer Module Facilitates the Surveying and Comparison of Biosynthetic Capacity across Geographically Distinct Environments

In the Map Explorer tab, the GPS coordinates of each microbiome sample are used to mark its geographical location using Google Maps application programming interface (API), and sample-specific data are linked to each marker (Figure 3). The user can opt to display only the samples corresponding to arrayed libraries by selecting "Show libraries only" checkbox, setting the e-value cutoff for displayed data, and displaying only samples containing hits to a specific biosynthetic system, domain, or molecule family (Figure 3A). Clicking on a map marker displays a brief overview of the sample information that can be used to quickly estimate the biosynthetic richness of the sample (Figure 3B). This information includes the summary of sequencing, clustering, and classification statistics, as well as a photograph of the collection site (if available). A list of all molecule families identified in the selected sample appears on the right of this page. From this panel, the user can download CSV-formatted files containing physicochemical soil data (if available), detailed sequencing statistics, and a list of all the classified USRs with corresponding KMGC IDs, e-values, and primer sequences used.

Mapping the data generated by eSNaPD will help guide future environmental sampling for drug-discovery efforts by identifying the most biosynthetically rich geographical areas overall or specific microbiomes rich in a particularly biomedically interesting class of molecules. Using stored physicochemical soil data, one can examine metagenomic data for correlations between soil properties and the biosynthetic diversity it contains (Charlop-Powers et al., 2014). In addition, geotagged data can be used to study the correlations between biosynthetic diversity and any number of parameters not captured in our analysis,

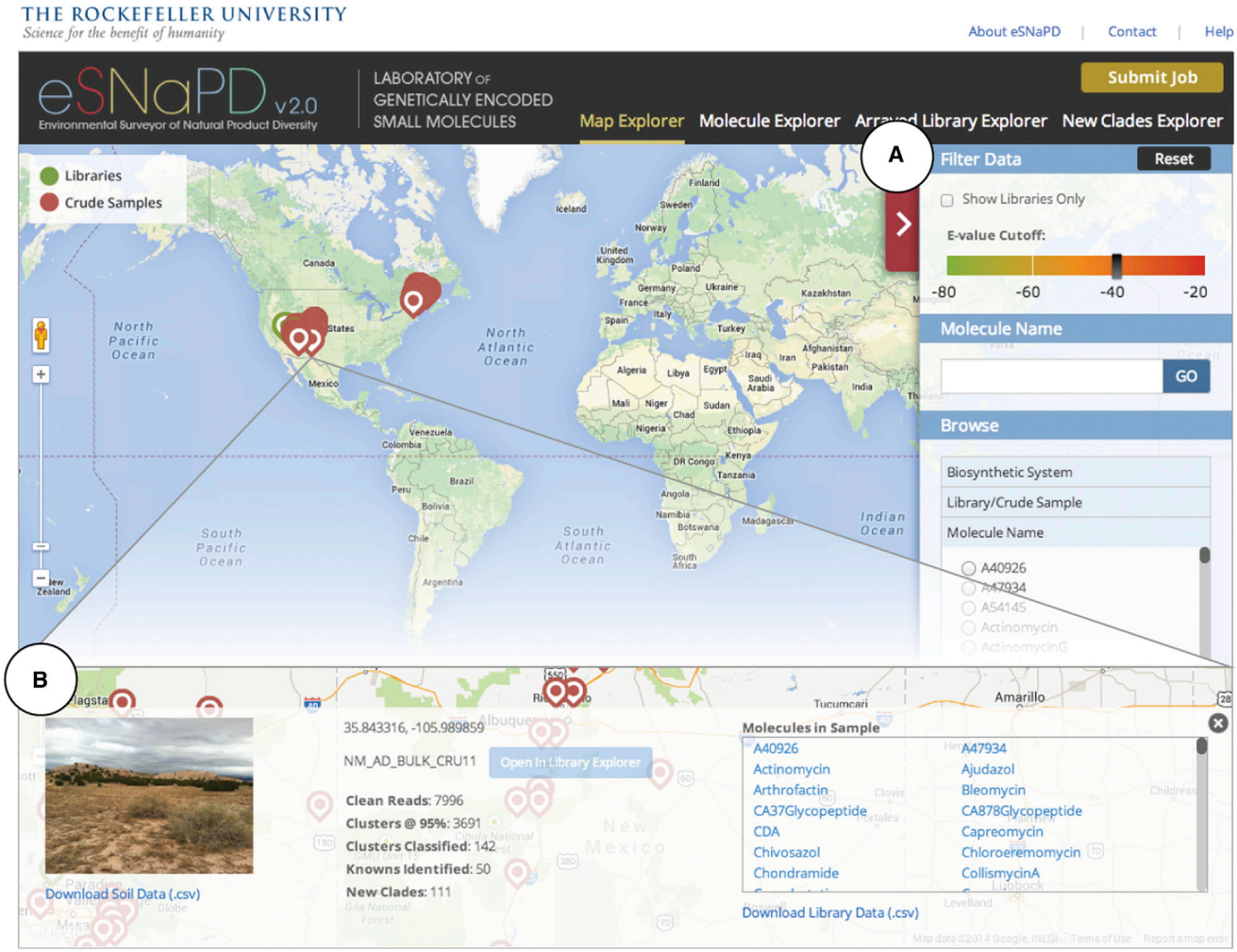

Figure 3. Map Explorer Module

(A) The analyzed sequence-tag data are plotted using Google Maps API to facilitate the global analysis of biosynthetic diversity. The displayed data can be filtered by the e-value cutoff, biosynthetic system, and specific sample or molecule family name.

(B) Map markers are linked to a pop-up panel containing sample-specific soil data, sequence-tag classification statistics, and biosynthetic content.

ranging from average seasonal temperatures to dominant macroflora and -fauna.

Molecule Explorer Module Provides the Visualization of Phylogenetic Relationships between USRs Related to Each KMGC and the USR Distribution Graph across Samples

The Molecule Explorer tab facilitates the identification of high-value targets for gene-cluster recovery (Figure 4). At the heart of the Molecule Explorer module are the Phylogeny and Distribution panels. At the end of eSNaPD analysis, the USRs from all the data sets contained in eSNaPD are pooled according to the known molecule gene cluster that they are related to and then aligned using MUSCLE (Edgar, 2004) as a whole set as well as subsets by individual known domains and individual samples. The resulting distance matrix files from MUSCLE alignments are then used to plot dendrograms via batch job submission to the iTOL server (Letunic and Bork, 2011). The dendrograms are accessible via a drop-down menu in the Phylogeny panel and are color coded by e-value range to reflect the phylogenetic distance relationship to a known characterized domain (Figure 4C). This feature provides a qualitative and quantitative visual overview of the data, allowing the user to rapidly identify samples that are either rich or scarce in sequence tags that are related to the KMGCs of interest. The selected set or subset of sequences used to construct the dendrograms is also available for download as a FASTA file to enable user-specific analysis.

The Distribution panel provides a graph quantifying the number of USRs related to the selected molecule family. A key feature of the Distribution panel is the user-selectable e-value cutoff for displayed data, which automatically updates the distribution graph to reflect the selection. This allows the user to readily identify the samples containing the most overall hits to the selected molecule by choosing a high e-value (e.g., 1×10^{-20}) or the samples containing the most sequence tags that are very closely related to the KMGC by selecting a low e-value

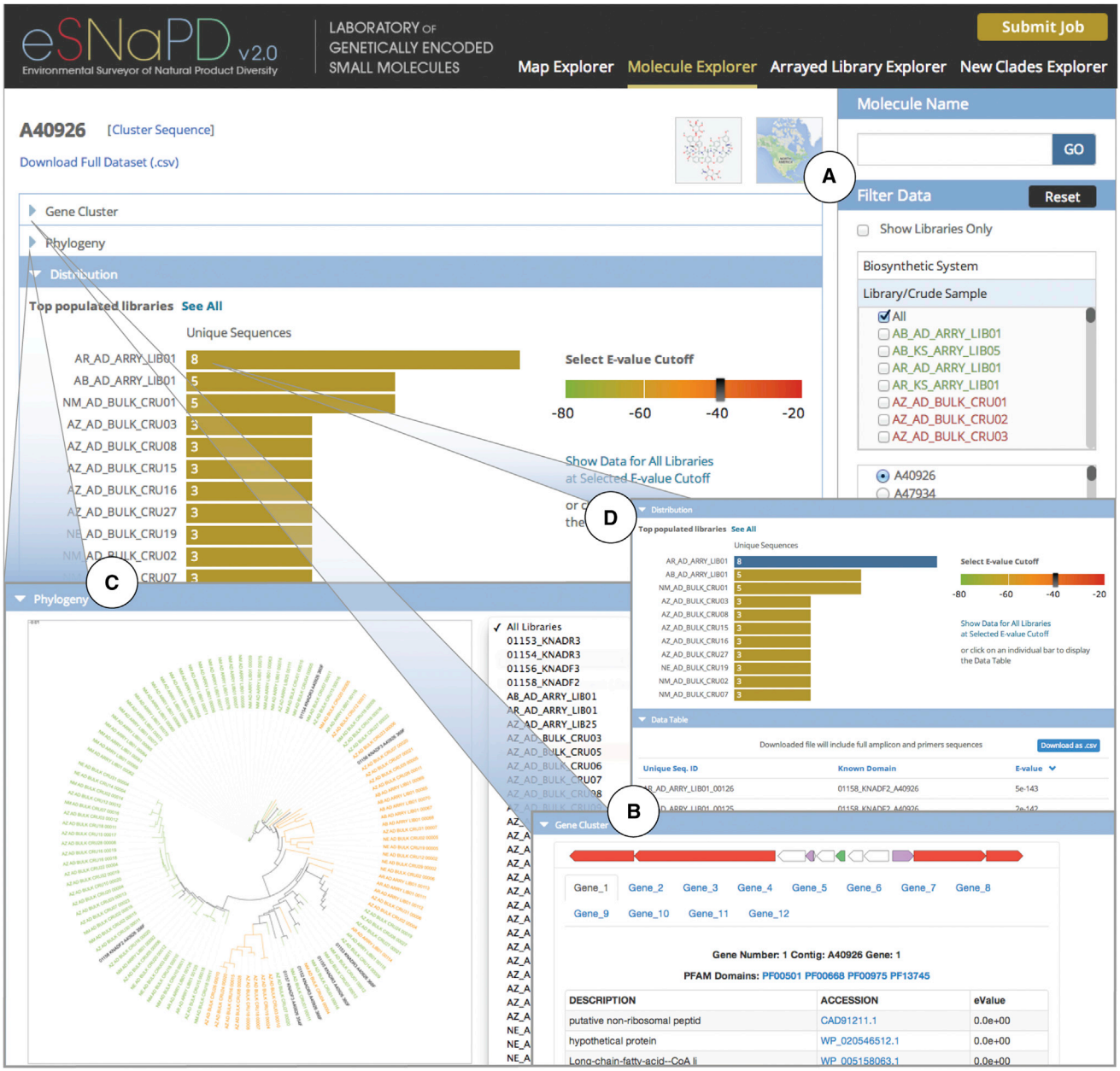

Figure 4. Molecule Explorer Module

(A) Selecting a specific molecule family from the list that can be filtered by sample or biosynthetic system, the user is presented with pertinent information to enable the identification of samples rich in congeners of the chosen molecule.

(B–D) Overview of the reference gene cluster encoding the characterized molecule (B), phylogenetic tree of identified congeners for each sample or across all samples in the eSNaPD database (C), and distribution of identified sequence tags across the eSNaPD database (D).

(e.g., 1×10^{-80}). In addition, when the user clicks on a bar corresponding to a desired sample, the Data Table section is populated with a sortable list of all USRs in the selected sample, along with the corresponding closest-related domain ID and the e-value that was calculated during the alignment step (Figure 4D). This feature enables the user to obtain a downloadable CSV-formatted file containing the subset of the data for a selected molecule family in a selected sample, or across all samples if the "Show Data for All Libraries" link is selected, based on chosen set e-value cutoff. In addition to the sequence IDs, domain IDs, and e-values displayed in the Data Table section, the downloaded file also contains sequences for all USRs.

Additional features of the Molecule Explorer tab include an overview of the cluster organization of the reference KMGC (Figure 4B), an external link to the known molecule structure in either the PubChem or ChemSpider database (Evan et al., 2008), and a link to a pop-up map that displays all samples containing the selected molecule family. The list of available molecules to

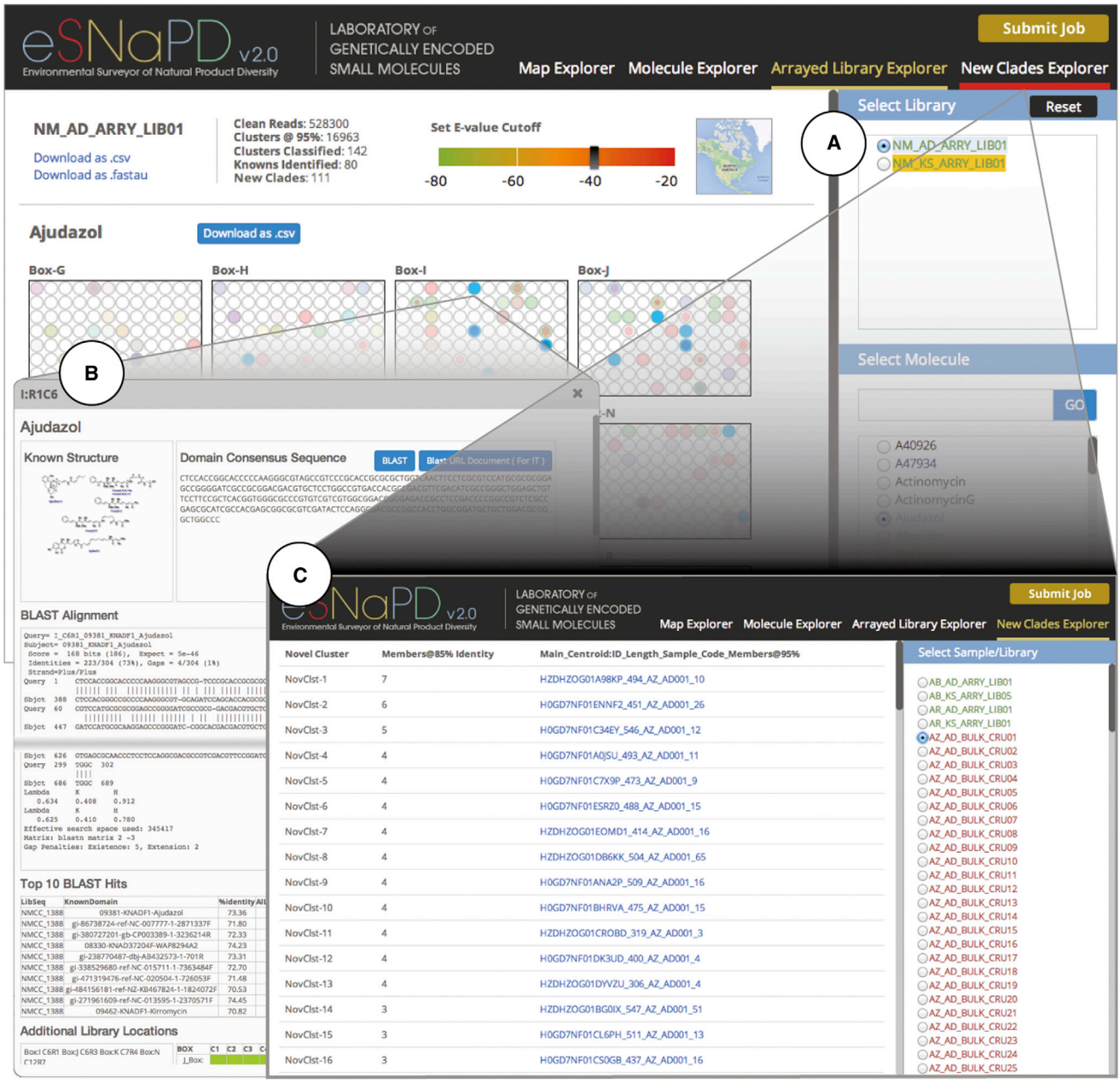

Figure 5. Arrayed Library Explorer and New Clade Explorer Modules

(A) For arrayed libraries sequenced with location-tagged bar-coded primers, the sequence-tag classification data are mapped onto a graphical representation of the library array. When a molecule family is selected from the Select Molecule menu, the locations of all identified USRs for the selected molecule are highlighted on the array map and all other libraries containing the selected molecule are highlighted in the Select Library menu.

(B) Selecting a specific location in the array provides an overlay with detailed information about the specific USR in the selected array position that facilitates the validation of the sequence-tag classification.

(C) The New Clades Explorer module produces a list of all novel clades identified in the selected sample with the unique next-generation sequencing read IDs for the centroid sequence in the clade linked to a downloadable file containing all sequences included in the clade.

explore can be filtered using either a biosynthetic system type or a specific sample name (Figure 4A).

Arrayed Library Explorer Simplifies Molecule Discovery from Arrayed Metagenomic Libraries

The arraying and positional sequencing of large (meta)genomic libraries, or even large culture collections, can greatly facilitate the identification and recovery of specific gene clusters of interest from these genomic pools. eSNaPD is designed to accept sequence tags amplified using primers that contain 8-nt location-specific bar-codes that allow each sequence read obtained from the arrayed library or culture collection to be mapped to its physical position in the original array. The Arrayed Library

Explorer tab provides an interactive visual display that delivers access to detailed information about individual sequence reads related to each KMGC, which helps to more easily prioritize gene clusters for downstream analysis and to quickly identify all other array locations that contain sequence reads related to the KMGC of interest (Figure 5). The information provided for each sequence read includes a PubMed- or ChemSpider-linked image of the structure encoded by KMGC, the consensus USR that the read belongs to, the top ten results from a BLASTn alignment to the NCBI-NT database, and the locations of all wells containing sequence reads corresponding to the same USR (Figure 5B).

These data allow the user to confirm the validity of the USR assignment as being related to the selected molecule and to identify all locations in the arrayed library containing the same sequence tag. Our original eSNaPD analysis tool was confined to a beta version of the Library Explorer module (Owen et al., 2013).

New Clade Explorer Compiles a Downloadable Table of USRs in Each Sample that Forms Clades that Do Not Associate with Any KMGC Sequence in the eSNaPD Reference Data Set

Only a small fraction of cleaned sequencing reads can be confidently classified as related to a KMGC (i.e., e-value $< 1 \times 10^{-40}$). A portion of the remaining sequences, considered unclassified or unreliably classified, will cluster at 95% and will be recognized as USRs. A further portion of these will form clades that are distinct from clades that are associated with KMGCs. Such clades contain USRs that have a high likelihood of being associated with gene clusters that may encode natural products that are fundamentally different from those encoded by KMGCs. The New Clade Explorer provides an interactive and downloadable table containing unclassified USRs from each sample in the eSNaPD database to assist in the discovery of novel families of natural products encoded by (meta)genomic DNA (Figure 5C).

DISCUSSION

Advances in sequencing technologies and a corresponding exponential increase in the amount of sequence data have necessitated the development of bioinformatics tools that make these data useful to researchers. In the field of natural products chemistry, a number of such tools have emerged with the aims of systematizing the growing body of knowledge about bacterial secondary metabolism and of using its predictive power. Some of these tools, such as ClustScan (Starcevic et al., 2008), np.searcher (Li et al., 2009), and antiSMASH (Blin et al., 2013; Medema et al., 2011), are designed primarily to scan large genomic assemblies, identify biosynthetic clusters, and attempt to predict the structure of the molecule they encode based on the predicted substrate specificities of the conserved biosynthetic domains they contain. Another recently developed tool, NaPDoS (Ziemert et al., 2012), identifies candidate KS and condensation (C) domains in user sequences and uses a simpler, phylogenomic approach to infer the structural family of the product metabolite. At the core of all these tools are carefully curated databases of individual biosynthetic domain sequences as the secondary metabolism biosynthetic systems are highly modular and, ultimately, empirical evidence prescribes that sufficiently high sequence conservation predicts conserved function (Wilson et al., 2010).

These tools are extremely useful for mining the rapidly growing number of newly sequenced microbial genomes, but they are of little use for interrogating large metagenomic data sets. Although NaPDoS is intended to accept a variety of input data ranging from whole genomes to PCR amplicons, it is computationally limited to relatively small data sets (<30 MB or <50,000 sequences) and designed to identify known gene clusters in these data sets. With a typical metagenomic sequence-tag data set approaching 500 MB and exceeding 1,000,000 reads, there are currently no bioinformatics tools suitable for the analysis of this type of data for novel gene clusters other than eSNaPD. eSNaPD was specifically designed to process large, targeted sequence-tag data sets of short reads (~400–500 bp), which makes processing of very large data sets (up to 1 GB) computationally tractable because only short sequences need to be aligned. Without knowing the full biosynthetic cluster sequence with all of the tailoring enzymes, it is not possible for eSNaPD to predict the exact structure of the molecule encoded by a gene clusters associated with sequence tag. However, we find that, as a rule, phylogenomics-based analysis, as described here and proven in numerous studies conducted in our lab, provides accurate predictions with respect to the structural family and potential novelty of the small molecule encoded by biosynthetic cluster corresponding to a sequence tag.

SIGNIFICANCE

By providing an open-access, web-based user interface, our aim is to facilitate the analysis of metagenomic data by the greater natural products community and to start building a permanent, systematic database of the microbial biosynthetic diversity across the globe. Each natural microbiome contains thousands of unique sequence tags, and new bioinformatics tools are needed to facilitate the systematic interrogation of the biosynthetic diversity they represent for promising drug-discovery targets and to catalog and geographically map the global biosynthetic diversity in the environment. The eSNaPD platform performs an automated analysis of metagenomic sequence-tag data and parses the results into a user-friendly UI that is designed to facilitate the rapid identification of high-value targets for library construction or biosynthetic-pathway recovery and characterization. In addition to its usefulness for target identification for drug-discovery pipelines, eSNaPD provides an open-access survey of the biosynthetic diversity of microbiomes from a variety of geographical and physicochemical environments, allowing for broader biosynthesis-based ecology studies. The aggregation of data on the eSNaPD platform allows the existing data to be iteratively reanalyzed to identify additional biosynthetic pathways as the eSNaPD reference data sets expand to include newly identified biosynthetic clusters and domains. The aim of the eSNaPD data-analysis and aggregation features is to bring a degree of standardization to the bioinformatic characterization of biosynthetic diversity in the environment. Although eSNaPD was developed for use with environmental microbiome data, it is equally as useful for the

analysis of large culture collections, which are now known to contain many previously unseen, silent pathways (Bentley et al., 2002; Ikeda et al., 2003; NCBI, 2013). To accelerate the initial population of the eSNaPD database with sequences from diverse environments, we have launched a citizen science effort (http://www.drugsfromdirt.org/) with the initial goal of obtaining an evenly distributed set of approximately 500 soils from across the United States.

EXPERIMENTAL PROCEDURES

Preparation of Required Domain Sub-Data-Sets

The eSNaPD platform has two components for sequence data housekeeping: (1) a semiautomated pipeline for updating the domain sequences of the characterized, known-molecule-gene-clusters (KMGCs) data set as new molecules with corresponding gene clusters are reported in literature and (2) an automated analysis pipeline for updating the related sequences in the natural product gene-cluster-domains data set when microbial DNA sequence data substantially increase in the NCBI-NT database (databases nt and other_genomic) (Garcia et al., 2011). The procedures used to create and update these data sets are described next, and the most current version of all the reference data sets are available upon request.

KMGC Domains from Natural Product Gene Clusters in KMGC_NPD

Experimentally characterized gene clusters were collected from open access databases and the published literature. eSNaPD currently contains >450 characterized NRPS/PKS gene clusters composed of >10,000 signature domains present in our KMGC_NPD sequence data sets. The domains were collected from these gene clusters and then trimmed to similar lengths, based on BLAST comparisons to our curated collection of reference domain sequences (KMGC_NPD). Our reference domain collection was iteratively expanded from the ClustScan data set to include domains that were identified in our analysis. Domains found to have >99% sequence identity were removed from the KMGC_NPD collection if the sequences were derived from domains found in gene clusters encoding identical or nearly identical molecules.

NCBI-NT_NPD Sub-Data-Sets

For each NRPS/PKS signature domain, a unique data set was created by querying NCBI-NT at e-value 1×10^{-10} with the appropriate KMGC_NPD data sets (Figure 2A). Each NCBI-NT-derived domain sequence was trimmed to the length of its closest relative (highest bit score) found in KMGC_NPD. The KMGC_NPD sequences were added to the NCBI-NT data sets and duplicate sequences were removed to give a domains-specific data set (NCBI-NT_NPDi1). The NCBI-NT_NPDi1 was submitted to a second iteration of BLAST search to obtain the domains-specific data set NCBI-NT_NPDi2, which was assumed to include a large fraction of the sequenced diversity for each domain.

Input Data Formats

The eSNaPD query form accepts FASTA-formatted data containing next-generation sequencing reads. The job-submission form will also require a second text file containing the sequencing primer sequences, which may contain an 8 nt bar code that can be used to track bar-coded amplicons in the sequencing-reads file. Specific examples of each input file type are provided on the Submit Job page in eSNaPD.

eDNA Sequence Data Cleaning and Clustering

Raw amplicon sequencing data are trimmed after an ambiguous nucleotide read appears in the sequence (Figure 2B). If the optional bar-code data are provided, the FASTA-formatted headers are tagged with bar-code information. Subsequently, primers are trimmed, and the reads shorter than 225 bp are removed. The reads longer than 425 bp are trimmed to 425 bp. The resulting 225–425 bp long reads are uniformly trimmed by 25 bp on the 3′-end to reduce error at the end of short sequences (Gilles et al., 2011) and are sorted in descending order by length. The clean sequence reads for each sample are clustered at 95% identity using USEARCH (Edgar, 2010). The consensus seed sequences, resulting from the clustering, are reclustered at 100% identity to merge identical consensus reads that arise from distinct cluster groups. Representative sequences from each cluster (i.e., consensus seeds), which we have termed USRs, are used, along with the original member sequences present in each cluster, in downstream analyses.

eDNA Reads Related to KMGC

The USRs are searched against the NCBI-NT_NPDi2 sub-data-set using BLASTn, and sequences with identified relatives at e-value 1×10^{-5} are marked as USR select one (USRS1) (Figure 2C). The USRS1 set is then searched using BLASTn against the corresponding NCBI-NT_NPDi2 sub-data-set. The USRS1 sequences that have hits with e-value $\leq 1 \times 10^{-20}$, referred to as USRS2, and their relatives (NCBI-NT_NPDi2) are obtained for the top 50 hits. The BLAST hits to each USRS2 are combined, and redundant hits are removed by keeping the first hit in the alphanumerical order of the accession ID. If a sequence of an added KMGC domain is in the redundant set of hits, that KMGC domain ID with its bl6-formatted BLASTn scores is preserved. The bl6 statistics file is sorted in descending order of the bit scores, and the top ten hits with the highest bit scores are picked for each USRS2 as identified close relatives from the NCBI-NT_NPDi2. In this top-ten-close-relatives list for each USRS2, the KMGC_NPD sequence IDs are identified and their rank values from 1 to 10 are recorded. Each USRS2 sequence that has a hit to NCBI-NT_NPDi2 with an e-value of 1×10^{-20} or better and that has at least one KMGC_NPD ranked first is marked as related to the corresponding KMGC and recorded as a unique read from eDNA library related to KMGC (USRR_KMGC). To remove bias from the clean-reads clustering process, at this point the order of the unclustered clean reads is randomized and the reads are reordered by length. The clustering and the process of obtaining the USRR_KMGC data set are then repeated. These steps are iterated to produce a total of ten USRR_KMGC data sets resulting from different clustering orders. The final USRR_KMGC data set is then composed of USRs that hit to the same KMGC_NPD on at least six out of ten iterations. Arrayed eDNA libraries, which contain sequence-location information in their FASTA header, are processed by combining all cleaned reads and following the same steps used to generate the USRR_KMGC data set. Once the classification of the USRs is complete, the sequence-location information is used to map the data back to the library array.

New Clade CSV File

We have found that sequence tags hitting to the KMGC_NPD database with e-values $< 1 \times 10^{-45}$ correspond to gene clusters that are likely to encode congeners of a KMGC at a high frequency and that those that hit to KMGC_NPD with e-values $> 1 \times 10^{-45}$ have a dramatically reduced likelihood of doing so (Owen et al., 2013). Using this information, we have chosen an e-value of 1×10^{-30} as an arbitrary cutoff for considering a USR to be related to KMGC. USRs that are composed of at least two individual sequence reads that cluster at 95% yet do not have an immediate relationship to a KMGC_NPD (e-value $> 1 \times 10^{-30}$) are clustered at 85% identity to generate new clades that are displayed in the New Clade Explorer module. The table that appears in this tab contains the new clade ID, the number of USRs in the clade, and the individual read ID of the centroid sequence linked to a downloadable CSV file containing all member sequences of the new clade with their corresponding individual read IDs.

Output Data and Visualization

A web-based UI was designed for the visualization of eSNaPD output across all samples processed by eSNaPD. This interface contains four modes of navigation: (1) Map Explorer, (2) Molecule Explorer, (3) Arrayed Library Explorer, and (4) New Clade Explorer. The features incorporated in each module have been described in detail.

AUTHOR CONTRIBUTIONS

B.V.B.R. and A.M. worked together to develop the functionality and organization of the eSNaPD web platform. B.V.B.R. assembled reference data sets and performed the back-end scripting for data processing, organization, and

presentation. A.M. developed the front-end user interface, made the figures, and made the major contribution to writing the manuscript.

ACKNOWLEDGMENTS

This work was supported by the NIH (grant number GM077516). S.F.B. is a Howard Hughes Medical Institute Early Career Scientist. Bella Milshteyn provided the final UI and visual design; Kwan Ng provided the website development.

Received: April 14, 2014
Revised: May 29, 2014
Accepted: June 10, 2014
Published: July 24, 2014

REFERENCES

Banik, J.J., and Brady, S.F. (2008). Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proc. Natl. Acad. Sci. USA *105*, 17273–17277.

Banik, J.J., and Brady, S.F. (2010). Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. Curr. Opin. Microbiol. *13*, 603–609.

Bauer, J.D., King, R.W., and Brady, S.F. (2010). Utahmycins a and B, azaquinones produced by an environmental DNA clone. J. Nat. Prod. *73*, 976–979.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature *417*, 141–147.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. *41*, W204–W212.

Brady, S.F. (2007). Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nat. Protoc. *2*, 1297–1305.

Challis, G.L. (2008). Genome mining for novel natural product discovery. J. Med. Chem. *51*, 2618–2628.

Chang, F.-Y., and Brady, S.F. (2011). Cloning and characterization of an environmental DNA-derived gene cluster that encodes the biosynthesis of the antitumor substance BE-54017. J. Am. Chem. Soc. *133*, 9996–9999.

Chang, F.Y., and Brady, S.F. (2013). Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. Proc. Natl. Acad. Sci. USA *110*, 2478–2483.

Chang, F.-Y., Ternei, M.A., Calle, P.Y., and Brady, S.F. (2013). Discovery and synthetic refactoring of tryptophan dimer gene clusters from the environment. J. Am. Chem. Soc. *135*, 17906–17912.

Charlop-Powers, Z., Owen, J.G., Reddy, B.V., Ternei, M.A., and Brady, S.F. (2014). Chemical-biogeographic survey of secondary metabolism in soil. Proc. Natl. Acad. Sci. USA *111*, 3757–3762.

Dewick, P.M. (2009). Medicinal Natural Products: A Biosynthetic Approach, Third Edition. (Chichester, UK: John Wiley & Sons).

Donia, M.S., Ruffner, D.E., Cao, S., and Schmidt, E.W. (2011). Accessing the hidden majority of marine natural products through metagenomics. ChemBioChem *12*, 1230–1236.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Evan, E., Bolton, Y.W., Thiessen, P.A., and Bryant, S.H. (2008). PubChem: integrated platform of small molecules and biological activities. In Annual Reports in Computational Chemistry, *Vol. 4*, R.A. Wheeler and D.C. Spellmeyer, eds. (Oxford: Elsevier), pp. 217–241.

Felsenstein, J. (1993). PHYLIP: phylogeny inference package. (Seattle: University of Washington).

Feng, Z., Kim, J.H., and Brady, S.F. (2010). Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. J. Am. Chem. Soc. *132*, 11902–11903.

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. Proc. Natl. Acad. Sci. USA *105*, 4601–4608.

Garcia, J.A.L., Fernández-Guerra, A., and Casamayor, E.O. (2011). A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes. Mol. Phylogenet. Evol. *61*, 650–658.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics *12*, 245.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat. Biotechnol. *21*, 526–531.

Kallifidas, D., Kang, H.S., and Brady, S.F. (2012). Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. J. Am. Chem. Soc. *134*, 19552–19555.

Kampa, A., Gagunashvili, A.N., Gulder, T.A.M., Morinaka, B.I., Daolio, C., Godejohann, M., Miao, V.P.W., Piel, J., and Andrésson, Ó. (2013). Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. Proc. Natl. Acad. Sci. USA *110*, E3129–E3137.

Kang, H.-S., and Brady, S.F. (2013). Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. Angew. Chem. Int. Ed. Engl. *52*, 11063–11067.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. *39*, W475–W478.

Li, X., and Qin, L. (2005). Metagenomics-based drug discovery and marine microbial diversity. Trends Biotechnol. *23*, 539–543.

Li, M.H.T., Ung, P.M.U., Zajkowski, J., Garneau-Tsodikova, S., and Sherman, D.H. (2009). Automated genome mining for natural products. BMC Bioinformatics *10*, 185.

MacNeil, I.A., Tiong, C.L., Minor, C., August, P.R., Grossman, T.H., Loiacono, K.A., Lynch, B.A., Phillips, T., Narula, S., Sundaramoorthi, R., et al. (2001). Expression and isolation of antimicrobial small molecules from soil DNA libraries. J. Mol. Microbiol. Biotechnol. *3*, 301–308.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. *39*, W339–W346.

NCBI (2013). New NCBI Handbook chapters: eukaryotic and prokaryotic genome annotation pipelines. NCBI News, http://www.ncbi.nlm.nih.gov/news/12-17-2013-new-handbook-chapters-genome-annotation-pipelines/.

Owen, J.G., Reddy, B.V., Ternei, M.A., Charlop-Powers, Z., Calle, P.Y., Kim, J.H., and Brady, S.F. (2013). Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. Proc. Natl. Acad. Sci. USA *110*, 11797–11802.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. Brief. Bioinform. *10*, 354–366.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394.

Reddy, B.V., Kallifidas, D., Kim, J.H., Charlop-Powers, Z., Feng, Z., and Brady, S.F. (2012). Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. Appl. Environ. Microbiol. *78*, 3744–3752.

Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J., and Hranueli, D. (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*, 6882–6892.

Torsvik, V., Goksøyr, J., and Daae, F.L. (1990). High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. *56*, 782–787.

Udwary, D.W., Zeigler, L., Asolkar, R.N., Singan, V., Lapidus, A., Fenical, W., Jensen, P.R., and Moore, B.S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. Proc. Natl. Acad. Sci. USA *104*, 10376–10381.

Wilson, M.C., Gulder, T.A., Mahmud, T., and Moore, B.S. (2010). Shared biosynthesis of the saliniketals and rifamycins in Salinispora arenicola is controlled by the sare1259-encoded cytochrome P450. J. Am. Chem. Soc. *132*, 12757–12765.

Winter, J.M., Behnken, S., and Hertweck, C. (2011). Genomics-inspired discovery of natural products. Curr. Opin. Chem. Biol. *15*, 22–31.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and Jensen, P.R. (2012). The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One *7*, e34064, http://dx.doi.org/10.1371/journal.pone.0034064.