Published in final edited form as:
Chem Biol. 2014 August 14; 21(8): 1023–1033. doi:10.1016/j.chembiol.2014.06.007.

eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes

Boojala Vijay B. Reddy[1,2], **Aleksandr Milshteyn**[1,2], **Zachary Charlop-Powers**[2,3], and **Sean F. Brady**[2,3,*]

[2]Laboratory of Genetically Encoded Small Molecules, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

[3]Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

Summary

eSNaPD (environmental Surveyor of Natural Product Diversity) 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 dataset of characterized gene clusters. Sample mapping, molecule discovery, library mapping and new clade visualization modules facilitate the interrogation of large (meta)genomic sequence datasets 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.

Corresponding Author: Sean F. Brady, **Contact:** Laboratory of Genetically Encoded Small Molecules, The Rockefeller University, 1230 York Avenue, New York, NY 10065, **Phone**: 212-327-8280, **Fax**: 212-327-8281, sbrady@rockefeller.edu.

[1] Co-first author

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Competing interests:
The authors declare that they have no competing interests.

Authors' contributions:
Boojala Vijay B. Reddy and Aleksandr Milshteyn worked together to develop the functionality and organization of the eSNaPD web platform. Boojala Vijay B. Reddy assembled reference datasets and performed the back-end scripting for data processing, organization and presentation. Aleksandr Milshteyn developed front-end user interface, made the figures and made the major contribution to writing the manuscript.

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 only been able to access 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). While 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, bacterial DNA is captured directly from environmental samples and the small molecules are accessed through 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 sequence.

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. 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, while 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, 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 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 functional classification of novel secondary metabolite biosynthetic gene clusters based solely on the minimal amount of sequence contained within 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 within a pool of metagenomic DNA and a reference set of functionally characterized gene clusters (Figures 1 and 5).

Here we present eSNaPD (http://esnapd2.rockefeller.edu/ - environmental Surveyor of Natural Product Diversity), a web-based bioinformatics platform for automated analysis and organized aggregation of large metagenomic sequence tag datasets. eSNaPD was designed to perform four main functions in a sequence tag-based gene cluster discovery pipeline: (1) To streamline the process of pre-screening metagenomic DNA samples for library construction, in order to enable 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 rapid identification and recovery of clones containing biosynthetic gene clusters of interest. (3) To profile and catalogue global biosynthetic diversity in a systematic and meaningful way, in order to gain a better understanding of the biosynthetic richness of different ecological environments. (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; Chang et al., 2013; Chang and Brady, 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 just as easily be constructed from genomic DNA contained within culture collections or any other suitable source.

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 the number of sequence tag datasets 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 for comparative meta-analyses of secondary metabolomes.

Results

eSNaPD uses 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 5). The results of the analysis are fed into a user interface (UI) designed for easy visual exploration of large datasets 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 dataset, are predicted to be good candidates for encoding biosynthesis of novel natural product families – a function unique to eSNaPD. In addition, use of barcoded primers for arrayed metagenomic library sequencing, enables eSNaPD to pinpoint the locations of sequence tags corresponding to each family of molecules within arrayed libraries for rapid clone recovery. The displayed data can be filtered by a user-selectable e-value cutoff and all output data is 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 only populated with datasets for AD and KS domains, the data analysis pipeline permits evaluation of amplicon reads from 12 NRPS/PKS domains, namely: adenylation (AD), acyl carrier protein (AC), acyltransferase (AT), condensation (CD), dehydratase (DH), epimerization (EP), enoylreductase (ER), ketoreductase (KR), ketoacylsynthase (KS), methyltransferase (MT), peptidyl carrier protein (PC) and thioesterase (TE) domains. 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 GPS coordinates, photograph of the collection site, soil type, and, when available, physicochemical parameters of the soil sample (*e.g.* moisture content, pH, organic matter content [LOI], organic nitrogen [PMN], minerals, metals, *etc.*). As the eSNaPD database grows, this data will allow a variety of downstream, ecologically focused correlation analyses to be performed on the amplicon datasets.

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 barcoded using 8 nucleotide lead sequences, which identify each unique sample in the sequencing reaction. In the case of an arrayed library, the primer 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 position 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 barcoded arrayed libraries data 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 BLASTn to a curated reference database of similar regions from functionally characterized known molecule gene clusters (KMGCs) and all closely related sequences that have been culled from gene cluster data found in the NCBI-NT database (Figure 5). 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 < $1e^{-20}$. While 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 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 $1e^{-20}$ to $1e^{-80}$. In our experience, at low e-values ($<1e^{-40}$) this analysis has proved 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; Chang et al., 2013; Chang and Brady, 2013; Kang and Brady, 2013; Owen et al., 2013). This is an empirical observation based on our extensive experience, predominantly with AD and KS domains. However, as different domains undoubtedly evolve at different rates and the e-values are dependent on multiple parameters, including the size of the dataset and the length of the sequence, we provide the user with the option to adjust 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. Although, the utility of random shotgun sequencing data is limited relative to the PCR-targeted sequencing datasets that are designed such that the sequencing effort is focused entirely on the most bioinformatically informative natural product sequence tags.

eSNaPD Output

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, physicochemical soil parameters, and the list of molecule families identified in the sample are linked to the map marker (Figure 2). Molecule-specific output is comprised of phylogenetic dendrograms constructed for each identified molecule family on an individual dataset level as well as across all data contained in eSNaPD, using KITSCH (Felsenstein, 1993) (Figure 3C). Distributions of USRs related to each KMGC are also calculated across all samples (Figure 3D). 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 easy location of sequence tags corresponding to specific molecule families for clone recovery and provides detailed USR-specific information (Figure 4B). For all samples and libraries, sequence tags from within each dataset 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*.

1. *Map Explorer* module facilitates surveying and comparison of biosynthetic capacity across geographically distinct environments—In the *'Map Explorer'* tab, GPS coordinates of each microbiome sample are used to mark its geographical location using Google Maps API and sample specific data is linked to each marker (Figure 2). The user can opt to display only the samples corresponding to arrayed libraries by selecting 'Show libraries only' checkbox, set the e-value cutoff for displayed data, and display only samples containing hits to a specific biosynthetic system, domain or molecule family (Figure 2A). 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 2B). 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 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 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, geo-tagged data can be used to study the correlations between biosynthetic diversity and any number of parameters not captured in our analysis, ranging from average seasonal temperatures to dominant macro-flora and fauna, *etc*.

2. *Molecule Explorer* module provides visualization of phylogenetic relationships between unique sequence reads related to each KMGC and USRs distribution graph across samples—'Molecule Explorer' tab is aimed at facilitating identification of high-value targets for gene cluster recovery (Figure 3). At the heart of the *'Molecule Explorer'* module are the 'Phylogeny' and 'Distribution' panels. At the end of eSNaPD analysis, the USRs from all datasets contained in eSNaPD are pooled according to the known molecule gene cluster that they are related to and 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 dropdown 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 3C). 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 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.* 1e^{-20}), or those samples containing the most sequence tags that are very closely related to the KMGC by selecting a low e-value (*e.g.* 1e^{-80}). In addition, by clicking 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 corresponding closest related domain ID and the e-value calculated during the alignment step (Figure 3D). 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 '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 cluster organization of the reference KMGC (Figure 3B), external link to the known molecule structure in either PubChem or ChemSpider database (ChemSpider, 2014; Evan E. Bolton, 2008), and a link to a pop-up map that displays all samples containing the selected molecule family. The list of available molecules to explore can be filtered either by a biosynthetic system type or a specific sample name (Figure 3A).

3. *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 nucleotide location-specific "barcodes" 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 4). 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 10 results from a BLASTn alignment to the NCBI-NT database and locations of all wells containing sequence reads corresponding to the same USR (Figure 4B).

These data allow the user to confirm the validity of the USR assignment as being related to the selected molecule and 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).

4. *New Clade Explorer* compiles a downloadable table of USRs within each sample that form clades that do not associate with any KMGC sequence in the eSNaPD reference dataset—Only a small fraction of cleaned sequencing reads can be confidently classified as related to a KMGC (*i.e.* e-value < 1e^{-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 discovery of novel families of natural products encoded by (meta)genomic DNA (Figure 4C).

Discussion

Advances in sequencing technologies and a corresponding exponential increase in the amount of sequence data have necessitated development of bioinformatics tools that make this data useful to researchers. In the field of natural products chemistry a number of such tools have emerged, aimed at systematizing the growing body of knowledge about bacterial secondary metabolism and using its predictive power. Some of these tools, such as ClustScan (Starcevic et al., 2008), np.searcher (Li et al., 2009), 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 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 of 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 highly 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. While 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 (< 30MB or < 50,000 sequences) and designed to identify known gene clusters in these data sets. With a typical metagenomic sequence tag dataset 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 datasets of short reads (~400–500bp), which makes processing of very large datasets (up to 1GB) computationally tractable as 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, we aim 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 systematic interrogation of biosynthetic diversity they represent for promising drug discovery targets and to catalogue 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 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 biosynthetic diversity of microbiomes from across a variety of geographic and physicochemical environments, allowing for broader biosynthesis-based ecology studies. The aggregation of data on the eSNaPD platform allows for the existing data to be iteratively re-analyzed to identify additional biosynthetic pathways, as the eSNaPD reference datasets expand to include newly identified biosynthetic clusters and domains. eSNaPD data analysis and aggregation features aim to bring a degree of standardization to the bioinformatic characterization of biosynthetic diversity in the environment. While eSNaPD was developed for use with environmental microbiome data, it is equally as useful for 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 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-datasets

The eSNaPD platform has two components for sequence data housekeeping: (i) A semi-automated pipeline for updating domain sequences of characterized known molecule gene clusters (KMGC) dataset as new molecules with corresponding gene clusters are reported in literature, and (ii) An automated analysis pipeline for updating the related sequences in the natural product gene cluster domains dataset when microbial DNA sequence data substantially increases in the NCBI-NT database (nt plus other_genomic) (Garcia et al., 2011). The procedures used to create and update these data sets are described below and the most current version of all reference datasets are available upon request.

KMGC domains from natural product gene cluster (KMGC_NPD)

Experimentally characterized gene cluster were collected from open access databases and published literature. eSNaPD currently contains >450 characterized NRPS/PKS gene clusters composed of > 10,000 signature domains present in our KMGC_NPD sequence datasets. 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 dataset to include new domains that were identified in our analysis. Domains found to have > 99% sequence identity were removed from the KMGC_NPD collection if sequences were derived from domains found in gene clusters encoding identical or nearly identical molecules.

NCBI-NT_NPD sub-datasets

(Figure 5A) For each NRPS/PKS signature domain a unique dataset was created by querying NCBI-NT at $1e^{-10}$ expectation value with the appropriate KMGC_NPD datasets. Each NCBI-NT derived domain sequence was trimmed to the length of its closest relative (highest bit score) found in KMGC_NPD. KMGC_NPD sequences were added to the NCBI-NT datasets and duplicate sequences were removed, to give domains specific dataset (NCBI-NT_NPDi1). The NCBI-NT_NPDi1 was submitted to a second iteration of blast search to obtain domains specific dataset NCBI-NT_NPDi2, which was assumed to include a large fraction of the sequenced diversity for each domain.

Input data formats

eSNaPD query form accepts FASTA formatted data containing next-generation sequencing reads. The job submission form will also require a second text file containing sequencing primer sequences, which may contain an 8nt barcode that can be used to track barcoded 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

(Figure 5B) Raw amplicon sequencing data is trimmed after an ambiguous nucleotide read appears in the sequence. If the optional barcode data is provided, the FASTA formatted headers are tagged with barcode 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 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 re-clustered 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 unique sequence reads (USR) are used, along with the original member sequences present in each cluster, in downstream analyses.

eDNA reads related to KMGC

(Figure 5C) The USRs are searched against the NCBI-NT_NPDi2 sub-dataset using BLASTn sequences with identified relatives at expectation value of $1e^{-5}$ were marked as USR select one (USRS1). The USRS1 set is then searched using BLASTn against corresponding NCBI-NT_NPDi2 sub-dataset. The USRS1 sequences that have hits with expectation value of $\leq 1e^{-20}$, referred to as USRS2, and their relatives (NCBI-NT_NPDi2) are obtained for the top 50 hits. 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 among 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 top 10 hits with highest bit-scores are picked for each USRS2 as identified close relatives from the NCBI-NT_NPDi2. Within this top 10 close relatives list for each USRS2, the KMGC_NPD sequence IDs are identified and its rank value from 1 to 10 is recorded. The USRS2 sequences that hit to NCBI-NT_NPDi2 with e-value of $1e^{-20}$ or better, and have at least one KMGC_NPD ranked first, are marked as related to the corresponding KMGC and recorded as 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 un-clustered clean reads is randomized and the reads are re-ordered by length. The clustering and the process of obtaining the USRR_KMGC dataset are then repeated. These steps are iterated to produce a total of 10 USRR_KMGC datasets resulting from different clustering order. The final USRR_KMGC dataset is then composed of USRs that hit to the same KMGC_NPD on at least 6 out of 10 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 to generate the USRR_KMGC dataset. 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 files

We have found that sequence tags hitting to the KMGC_NPD database with e-values < $1e^{-45}$ correspond to gene clusters that are likely to encode congeners of a KMGC at a high frequency and those that hit to KMGC_NPD with e-values > $1e^{-45}$ have a dramatically reduced likelihood of doing so (Owen et al., 2013). Using this information we have chosen an e-value of $1e^{-30}$ as an arbitrary cutoff for considering USR as being 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 > $1e^{-30}$) 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 new clade ID, number of USRs in the clade, and the individual read ID of the centroid sequence linked to a downloadable 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 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*, (4) *New Clade Explorer*. The features incorporated in each module are described in detail in the manuscript.

Acknowledgments

This work was supported by National Institutes of Health [grant number GM077516. SFB is a Howard Hughes Medical Institute Early Career Scientist. Final UI and visual design by Bella Milshteyn, website development by Kwan Ng.

References

Banik JJ, Brady SF. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:17273–17277. [PubMed: 18987322]

Banik JJ, Brady SF. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. Current Opinion in Microbiology. 2010; 13:603–609. [PubMed: 20884282]

Bauer JD, King RW, Brady SF. Utahmycins A and B, Azaquinones Produced by an Environmental DNA Clone. Journal of Natural Products. 2010; 73:976–979. [PubMed: 20387794]

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature. 2002; 417:141–147. [PubMed: 12000953]

Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research. 2013; 41:W204–W212. [PubMed: 23737449]

Brady SF. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nature Protocols. 2007; 2:1297–1305.

Challis GL. Genome mining for novel natural product discovery. Journal of medicinal chemistry. 2008; 51:2618–2628. [PubMed: 18393407]

Chang F-Y, Brady SF. Cloning and characterization of an environmental DNA-derived gene cluster that encodes the biosynthesis of the antitumor substance BE-54017. Journal of the American Chemical Society. 2011; 133:9996–9999. [PubMed: 21542592]

Chang F-Y, Ternei MA, Calle PY, Brady SF. Discovery and synthetic refactoring of tryptophan dimer gene clusters from the environment. Journal of the American Chemical Society. 2013; 135:17906–17912. [PubMed: 24171465]

Chang FY, Brady SF. Discovery of indolotryptoline antiproliferative agents by homologyguided metagenomic screening. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:2478–2483. [PubMed: 23302687]

Charlop-Powers Z, Owen JG, Reddy BV, Ternei MA, Brady SF. Chemical-biogeographic survey of secondary metabolism in soil. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:3757–3762. [PubMed: 24550451]

ChemSpider. ChemSpider (Royal Society of Chemistry). 2014

Dewick, PM. Medicinal Natural Products: A Biosynthetic Approach. 3rd edn. Chichester, West Sussex, England; New York NY, USA: John Wiley & Sons; 2009.

Donia MS, Ruffner DE, Cao S, Schmidt EW. Accessing the hidden majority of marine natural products through metagenomics. ChemBioChem. 2011; 12:1230–1236. [PubMed: 21542088]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004; 32:1792–1797. [PubMed: 15034147]

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460–2461. [PubMed: 20709691]

Evan E, Bolton YW, Paul A. Thiessen, Stephen H. Bryant. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry. 2008:217–241.

Felsenstein, J. PHYLIP: Phylogeny Inference Package. Univ. of Washington, Seattle: 1993.

Feng Z, Kim JH, Brady SF. Fluostatins Produced by the Heterologous Expression of a TAR Reassembled Environmental DNA Derived Type II PKS Gene Cluster not been characterized from studies of cultured species . (AB) desert of southern California was used to construct a multimillion. Journal of the American Chemical Society. 2010; 132:11902–11903. [PubMed: 20690632]

Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:4601–4608. [PubMed: 18216259]

Garcia JAL, Fernández-Guerra A, Casamayor EO. A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes. Molecular Phylogenetics and Evolution. 2011; 61:650–658. [PubMed: 21864693]

Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin J-F. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011; 12:245. [PubMed: 21592414]

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nature Biotechnology. 2003; 21:526–531.

Kallifidas D, Kang HS, Brady SF. Tetarimycin A, an MRSA-Active Antibiotic Identified through Induced Expression of Environmental DNA Gene Clusters. Journal of the American Chemical Society. 2012; 134:19552–19555. [PubMed: 23157252]

Kampa A, Gagunashvili AN, Gulder TAM, Morinaka BI, Daolio C, Godejohann M, Miao VPW, Piel J, Andrésson ÓS. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:E3129–E3137. [PubMed: 23898213]

Kang H-S, Brady SF. Arimetamycin A: Improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. Angewandte Chemie International Edition. 2013; 52:11063–11067.

Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research. 2011; 39:W475–W478. [PubMed: 21470960]

Li MHT, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC bioinformatics. 2009; 10:185–185. [PubMed: 19531248]

Li X, Qin L. Metagenomics-based drug discovery and marine microbial diversity. Trends in Biotechnology. 2005; 23:539–543. [PubMed: 16154653]

MacNeil IA, Tiong CL, Minor C, August PR, Grossman TH, Loiacono Ka, Lynch Ba, Phillips T, Narula S, Sundaramoorthi R, et al. Expression and isolation of antimicrobial small molecules from soil DNA libraries. Journal of molecular microbiology and biotechnology. 2001; 3:301–308. [PubMed: 11321587]

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach Ma, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic acids research. 2011; 39:W339–W346. [PubMed: 21672958]

NCBI. New NCBI Handbook chapters: Eukaryotic and prokaryotic genome annotation pipelines. NCBI News. 2013

Owen JG, Reddy BV, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH, Brady SF. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. Proceedings of the National Academy of Sciences of the United States of America. 2013

Pop M. Genome assembly reborn: recent computational challenges. Briefings in Bioinformatics. 2009; 10:354–366. [PubMed: 19482960]

Rappé MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003; 57:369–394. [PubMed: 14527284]

Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z, Brady SF. Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. Applied and environmental microbiology. 2012; 78:3744–3752. [PubMed: 22427492]

Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. 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 research. 2008; 36:6882–6892. [PubMed: 18978015]

Torsvik V, Goksoyr J, Daae FL. High diversity in DNA of soil bacteria. Applied and environmental microbiology. 1990; 56:782–787. [PubMed: 2317046]

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS. Genome sequencing reveals complex secondary metabolome in the marine actinomycete

Salinispora tropica. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:10376–10381. [PubMed: 17563368]

Wilson MC, Gulder TA, Mahmud T, Moore BS. Shared biosynthesis of the saliniketals and rifamycins in Salinispora arenicola is controlled by the sare1259-encoded cytochrome P450. Journal of the American Chemical Society. 2010; 132:12757–12765. [PubMed: 20726561]

Winter JM, Behnken S, Hertweck C. Genomics-inspired discovery of natural products. Curr Opin Chem Biol. 2011; 15:22–31. [PubMed: 21111667]

Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS one. 2012; 7:e34064–e34064. [PubMed: 22479523]

Highlights

- Metagenomic biosynthetic diversity can be profiled, mapped, and mined using eSNaPD
- eSNaPD helps identify environments rich in unexplored natural products chemistry
- eSNaPD aids in discovery of novel chemistries and biomedically important congeners

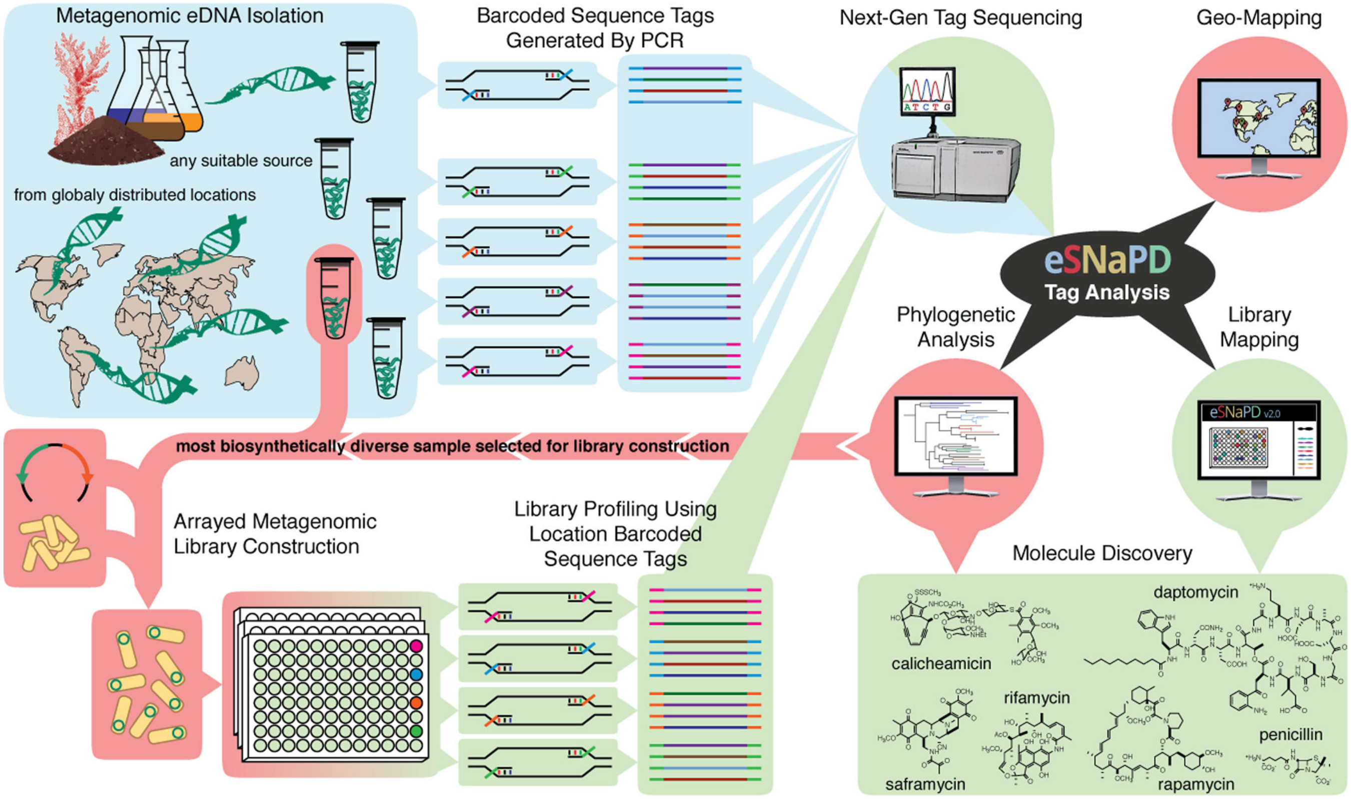
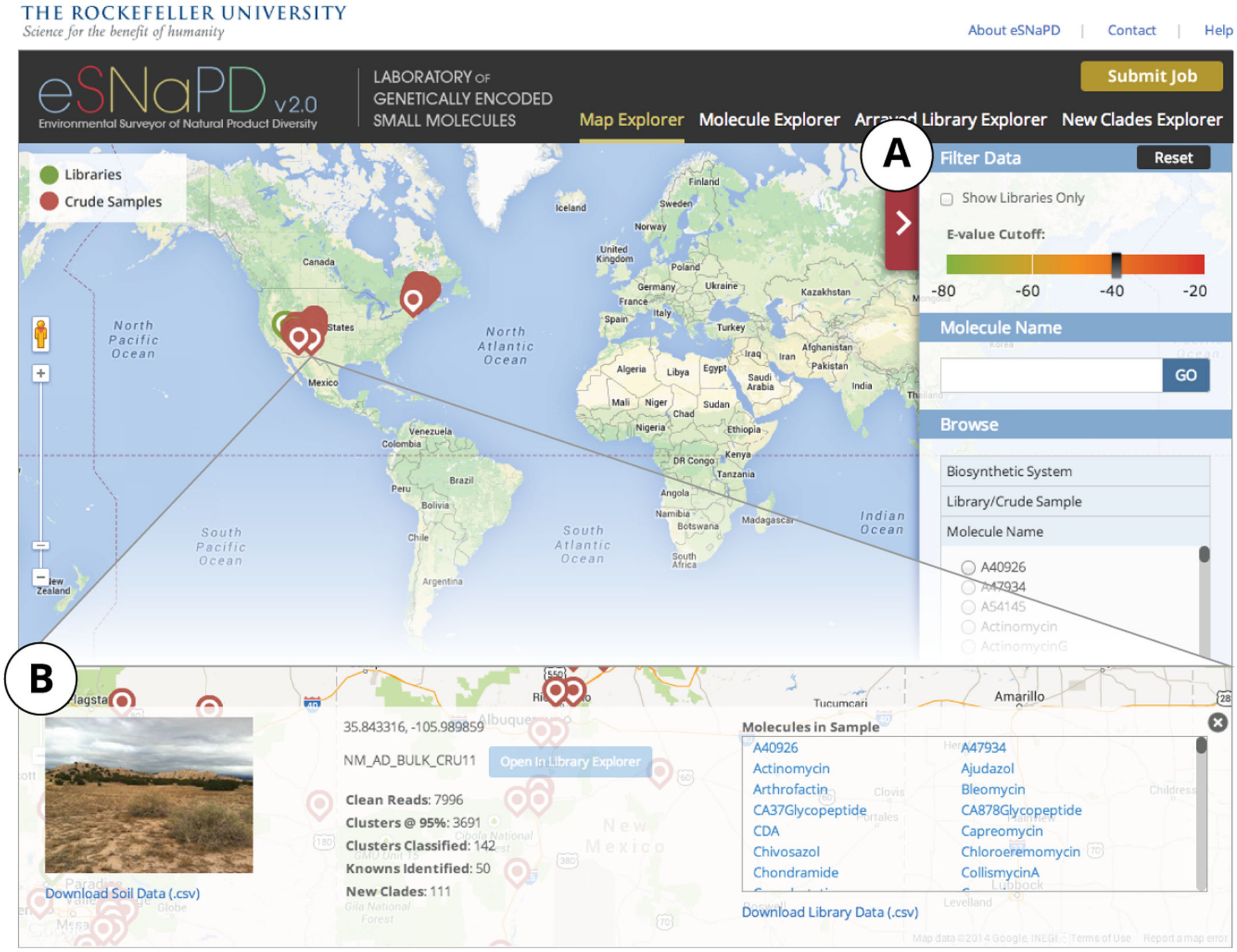


Figure 1.

An overview of eSNaPD assisted small molecule discovery pipeline. Sequence tags are generated by PCR from (meta)genomic DNA isolated from virtually any source using degenerate primers that target conserved biosynthetic domains and can be barcoded to differentiate DNA from different samples in a pooled next generation sequencing run (blue). Raw amplicon sequencing data is processed using 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). Newly generated metagenomic library is arrayed, to facilitate identification and recovery of target clones, sequenced using position-specific barcoded 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).



35.843316, -105.989859

NM_AD_BULK_CRU11

Open In Library Explorer

Clean Reads: 7996

Clusters @ 95%: 3691

Clusters Classified: 142

Knowns Identified: 50

New Clades: 111

Download Soil Data (.csv)

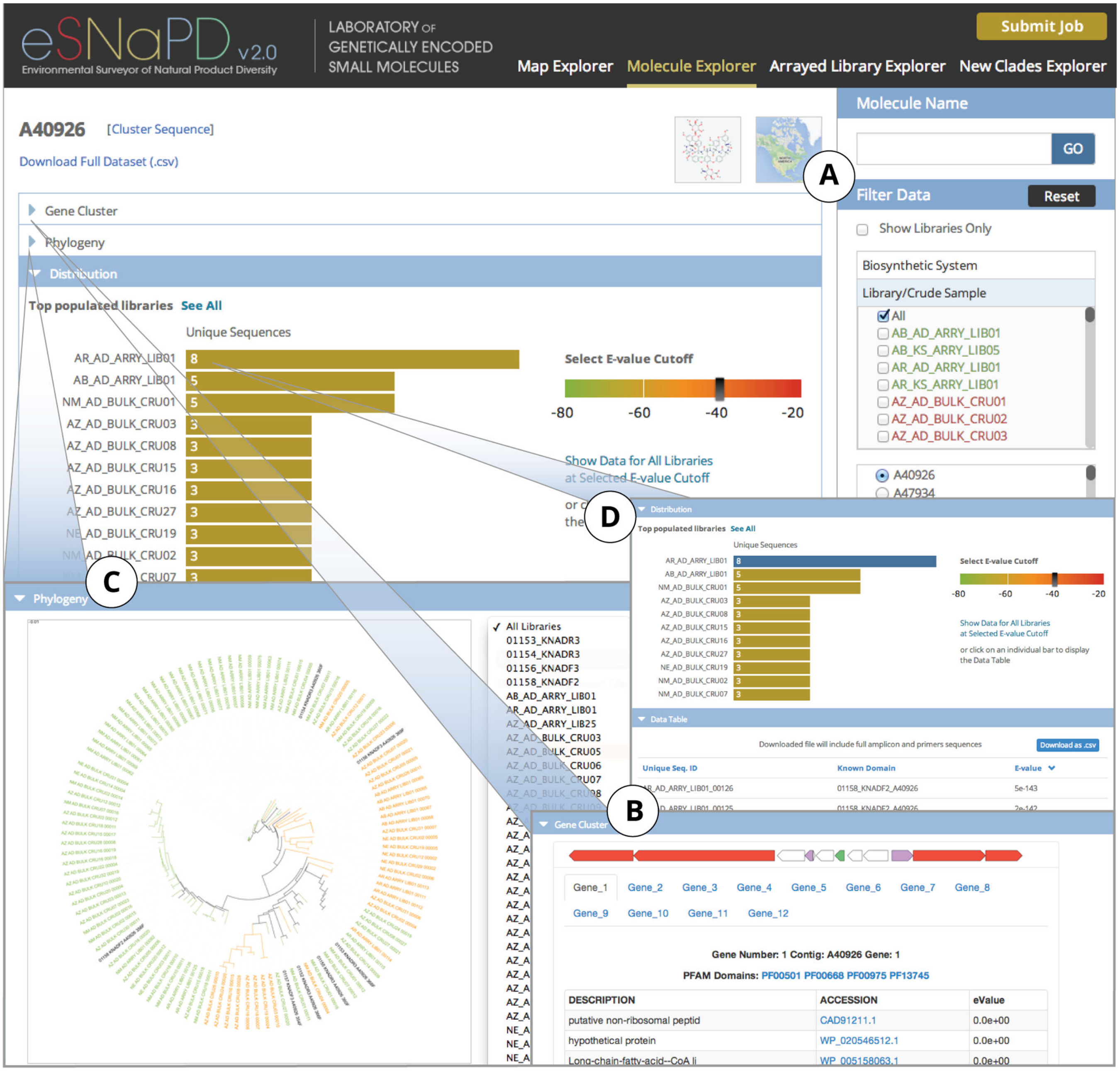
Molecules in Sample

A40926	A47934
Actinomycin	Ajudazol
Arthrofactin	Bleomycin
CA37Glycopeptide	CA878Glycopeptide
CDA	Capreomycin
Chivosazol	Chloroeremomycin
Chondramide	CollismycinA

Download Library Data (.csv)

Figure 2.

Map Explorer Module. Analyzed sequence tag data is plotted using Google Maps API to facilitate global analysis of biosynthetic diversity. The displayed data can be filtered by e-value cutoff, biosynthetic system, and specific sample or molecule family name (A). Map markers are linked to a pop-up panel containing sample specific soil data, sequence tag classification statistics and biosynthetic content (B).



Downloaded file will include full amplicon and primers sequences

Unique Seq. ID	Known Domain	E-value
AR_AD_ARRY_LIB01_00126	01158_KNADF2_A40926	5e-143
AR_AD_ARRY_LIB01_00125	01158_KNADF2_A40926	2e-142

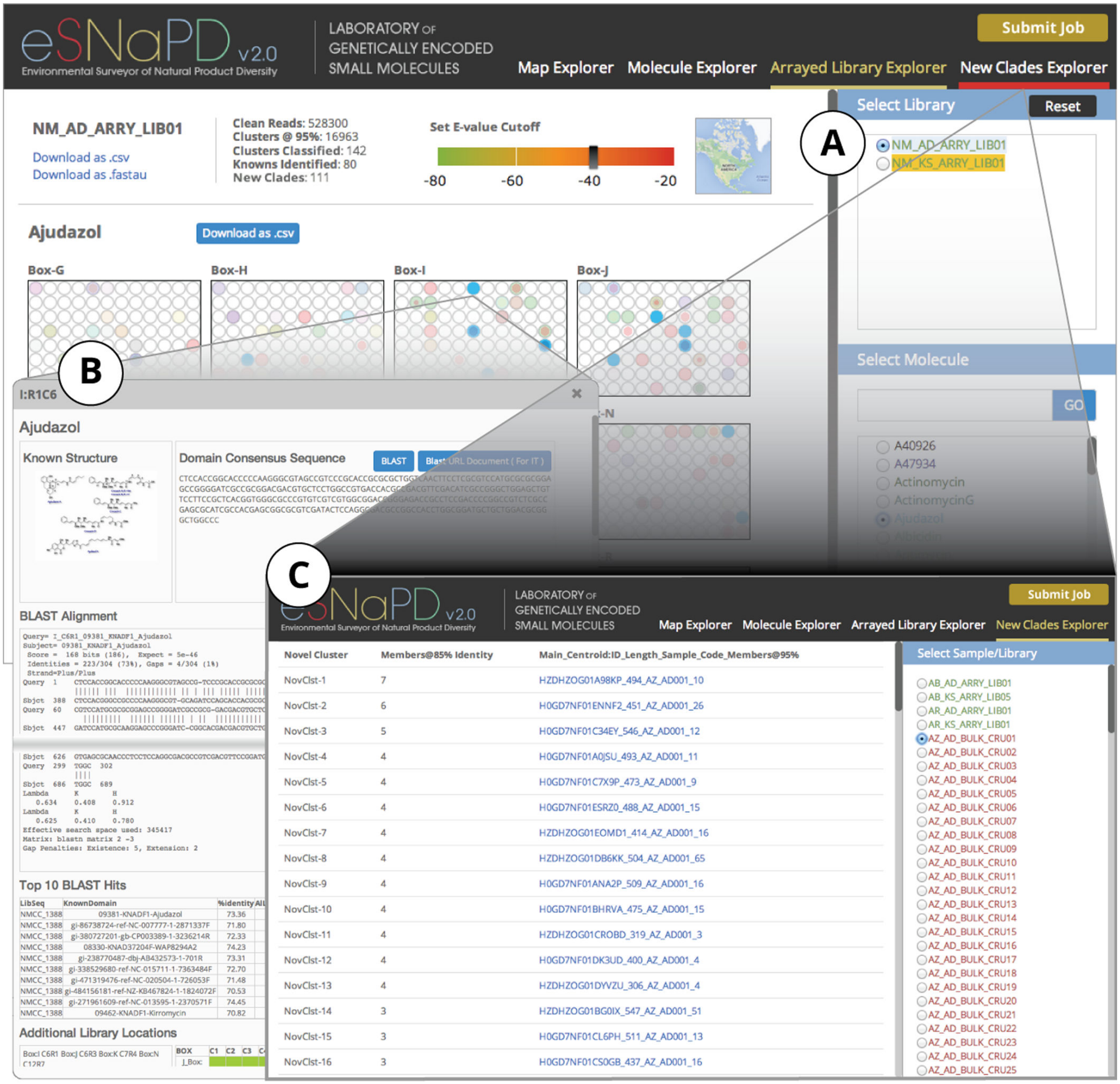
Gene Number: 1 Contig: A40926 Gene: 1

PFAM Domains: PF00501 PF00668 PF00975 PF13745

DESCRIPTION	ACCESSION	eValue
putative non-ribosomal peptid	CAD91211.1	0.0e+00
hypothetical protein	WP_020546512.1	0.0e+00
Long-chain-fatty-acid--CoA li	WP_005158063.1	0.0e+00

Figure 3.

Molecule Explorer Module. Selecting a specific molecule family from the list that can be filtered by sample or biosynthetic system (A), the user is presented with pertinent information to enable identification of samples rich in congeners of the chosen molecule. Overview of the reference gene cluster encoding the characterized molecule (B), phylogenetic tree of identified congeners for each sample or across all samples in eSNaPD database (C), and distribution of identified sequence tags across eSNaPD database (D) is provided.



NM_AD_ARRY_LIB01

Clean Reads: 528300
Clusters @ 95%: 16963
Clusters Classified: 142
Knowns Identified: 80
New Clades: 111

○ NM_AD_ARRY_LIB01 ● (selected)
○ NM_KS_ARRY_LIB01

○ A40926
○ A47934
○ Actinomycin
○ ActinomycinG
● Ajudazol
○ Albicidin

Ajudazol

I:R1C6

Ajudazol

Known Structure

Domain Consensus Sequence

```
CTCCACCGGCACCCCCAAGGGCGTAGCCGTCCCGCACCGCGCGCTGGTCAACTTCCTCGCGTCCATGCGCGCGGA
GCCGGGGATCGCCGCGGACGACGTGCTCCTGGCCGTGACCACGCCGACGTTCGACATCGCCGGGCTGGAGCTGT
TCCTTCCGCTCACGGTGGGCGCCCGTGTCGTCGTGGCGGACCGGGAGACCGCCTCCGACCCCGGCCGTCTCGCC
GAGCGCATCGCCACGAGCGGCGCGTCGATACTCCAGGCGACGCCGGCCACCTGGCGGATGCTGCTGGACGCGG
GCTGGCCC
```

BLAST Alignment

```
Query= I_C6R1_09381_KNADF1_Ajudazol
Subject= 09381_KNADF1_Ajudazol
 Score =  168 bits (186),  Expect = 5e-46
 Identities = 223/304 (73%), Gaps = 4/304 (1%)
 Strand=Plus/Plus
Query  1    CTCCACCGGCACCCCCAAGGGCGTAGCCG-TCCCGCACCGCGCGC
            |||||| |||  ||||||||||||| || | ||| ||||| ||||
Sbjct  388  CTCCACGGGCCGCCCCAAGGGCGT-GCAGATCCAGCACCACGCGC
Query  60   CGTCCATGCGCGCGGAGCCGGGGATCGCCGCG-GACGACGTGCTC
              |||||||||  |||||| |||||| | ||  |||||||||||
Sbjct  447  GATCCATGCGCAAGGAGCCCGGGATC-CGGCACGACGACGTGCTG

Sbjct  626  GTGAGCGCAACCCTCCTCCAGGCGACGCCGTCGACGTTCCGGATG
Query  299  TGGC  302
            ||||
Sbjct  686  TGGC  689
Lambda      K        H
   0.634    0.408    0.912
Lambda      K        H
   0.625    0.410    0.780
Effective search space used: 345417
Matrix: blastn matrix 2 -3
Gap Penalties: Existence: 5, Extension: 2
```

Top 10 BLAST Hits

LibSeq	KnownDomain	%identity	All
NMCC_1388	09381-KNADF1-Ajudazol	73.36	
NMCC_1388	gi-86738724-ref-NC-007777-1-2871337F	71.80	
NMCC_1388	gi-380727201-gb-CP003389-1-3236214R	72.33	
NMCC_1388	08330-KNAD37204F-WAP8294A2	74.23	
NMCC_1388	gi-238770487-dbj-AB432573-1-701R	73.31	
NMCC_1388	gi-338529680-ref-NC-015711-1-7363484F	72.70	
NMCC_1388	gi-471319476-ref-NC-020504-1-726053F	71.48	
NMCC_1388	gi-484156181-ref-NZ-KB467824-1-1824072F	70.53	
NMCC_1388	gi-271961609-ref-NC-013595-1-2370571F	74.45	
NMCC_1388	09462-KNADF1-Kirromycin	70.82	

Additional Library Locations

Box:I C6R1 Box:J C6R3 Box:K C7R4 Box:N C12R7

BOX	C1	C2	C3	C
J_Box	🟩	🟩	🟩	🟩

Novel Cluster	Members@85% Identity	Main_Centroid:ID_Length_Sample_Code_Members@95%
NovClst-1	7	HZDHZOG01A98KP_494_AZ_AD001_10
NovClst-2	6	H0GD7NF01ENNF2_451_AZ_AD001_26
NovClst-3	5	H0GD7NF01C34EY_546_AZ_AD001_12
NovClst-4	4	H0GD7NF01A0JSU_493_AZ_AD001_11
NovClst-5	4	H0GD7NF01C7X9P_473_AZ_AD001_9
NovClst-6	4	H0GD7NF01ESRZ0_488_AZ_AD001_15
NovClst-7	4	HZDHZOG01EOMD1_414_AZ_AD001_16
NovClst-8	4	HZDHZOG01DB6KK_504_AZ_AD001_65
NovClst-9	4	H0GD7NF01ANA2P_509_AZ_AD001_16
NovClst-10	4	H0GD7NF01BHRVA_475_AZ_AD001_15
NovClst-11	4	HZDHZOG01CROBD_319_AZ_AD001_3
NovClst-12	4	H0GD7NF01DK3UD_400_AZ_AD001_4
NovClst-13	4	HZDHZOG01DYVZU_306_AZ_AD001_4
NovClst-14	3	HZDHZOG01BG0IX_547_AZ_AD001_51
NovClst-15	3	H0GD7NF01CL6PH_511_AZ_AD001_13
NovClst-16	3	H0GD7NF01CS0GB_437_AZ_AD001_16

○ AB_AD_ARRY_LIB01
○ AB_KS_ARRY_LIB05
○ AR_AD_ARRY_LIB01
○ AR_KS_ARRY_LIB01
● AZ_AD_BULK_CRU01
○ AZ_AD_BULK_CRU02
○ AZ_AD_BULK_CRU03
○ AZ_AD_BULK_CRU04
○ AZ_AD_BULK_CRU05
○ AZ_AD_BULK_CRU06
○ AZ_AD_BULK_CRU07
○ AZ_AD_BULK_CRU08
○ AZ_AD_BULK_CRU09
○ AZ_AD_BULK_CRU10
○ AZ_AD_BULK_CRU11
○ AZ_AD_BULK_CRU12
○ AZ_AD_BULK_CRU13
○ AZ_AD_BULK_CRU14
○ AZ_AD_BULK_CRU15
○ AZ_AD_BULK_CRU16
○ AZ_AD_BULK_CRU17
○ AZ_AD_BULK_CRU18
○ AZ_AD_BULK_CRU19
○ AZ_AD_BULK_CRU20
○ AZ_AD_BULK_CRU21
○ AZ_AD_BULK_CRU22
○ AZ_AD_BULK_CRU23
○ AZ_AD_BULK_CRU24
○ AZ_AD_BULK_CRU25

Figure 4.

Arrayed Library Explorer and New Clade Explorer Modules. For arrayed libraries sequenced with location-tagged barcoded primers, the sequence tag classification data is mapped onto a graphical representation of the library array. When a molecule family is selected from the menu (A), 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 menu. Selecting a specific location in the array provides an overlay with detailed information about the specific USR in the selected array position that facilitates validation of the sequence tag classification (B). 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.

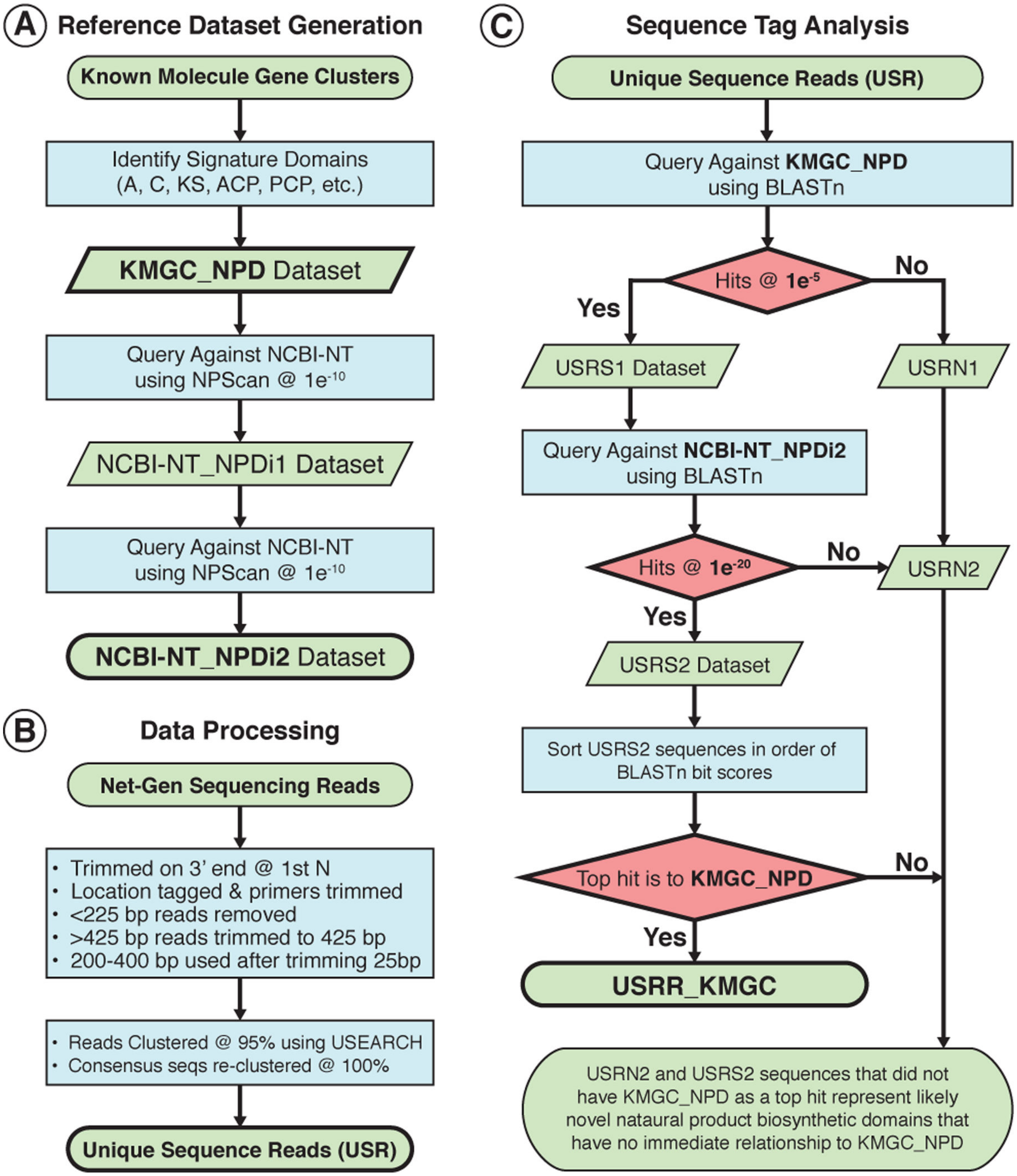

A Reference Dataset Generation
Known Molecule Gene Clusters
Identify Signature Domains (A, C, KS, ACP, PCP, etc.)
KMGC_NPD Dataset
Query Against NCBI-NT using NPScan @ 1e-10
NCBI-NT_NPDi1 Dataset
Query Against NCBI-NT using NPScan @ 1e-10
NCBI-NT_NPDi2 Dataset
B Data Processing
Net-Gen Sequencing Reads
• Trimmed on 3' end @ 1st N
• Location tagged & primers trimmed
• <225 bp reads removed
• >425 bp reads trimmed to 425 bp
• 200-400 bp used after trimming 25bp
• Reads Clustered @ 95% using USEARCH
• Consensus seqs re-clustered @ 100%
Unique Sequence Reads (USR)
C Sequence Tag Analysis
Unique Sequence Reads (USR)
Query Against KMGC_NPD using BLASTn
Hits @ 1e-5
Yes
No
USRS1 Dataset
USRN1
Query Against NCBI-NT_NPDi2 using BLASTn
Hits @ 1e-20
No
USRN2
Yes
USRS2 Dataset
Sort USRS2 sequences in order of BLASTn bit scores
Top hit is to KMGC_NPD
No
Yes
USRR_KMGC
USRN2 and USRS2 sequences that did not have KMGC_NPD as a top hit represent likely novel nataural product biosynthetic domains that have no immediate relationship to KMGC_NPD

Figure 5.

Overview of the eSNaPD data analysis pipeline. Steps and datasets used for generating reference natural product biosynthetic domains datasets (NCBI-NT_NPDi2) (A), cleaning and clustering of sequence tag next-generation sequencing reads to yield Unique Sequence Reads (USRs) (B), and classification of USR relationships to known molecule gene clusters (C).